Al-Otaibi, R., Jin, N., Wilcox, T., & Flach, P. (2016). Feature Construction and Calibration for Clustering Daily Load Curves from Smart-Meter Data. *IEEE Transactions on Industrial Informatics*, *12*(2), 645-654. https://doi.org/10.1109/TII.2016.2528819

Peer reviewed version

Link to published version (if available):
10.1109/TII.2016.2528819

Link to publication record on the Bristol Research Portal
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via IEEE at http://dx.doi.org/10.1109/TII.2016.2528819. Please refer to any applicable terms of use of the publisher.

University of Bristol – Bristol Research Portal

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/red/research-policy/pure/user-guides/brp-terms/

Feature Construction and Calibration for Clustering Daily Load Curves from Smart Meter Data

Reem Al-Otaibi, Nanlin Jin, *Member, IEEE,* Tom Wilcox, and Peter Flach

***Abstract*—This paper proposes and compares feature construction and calibration methods for clustering daily electricity load curves. Such load curves describe electricity demand over a period of time. A rich body of the literature has studied clustering of load curves, usually using temporal features. This limits the potential to discover new knowledge which may not be best represented as models consisting of all time points on load curves.**

This paper presents three new methods to construct features: conditional filters on time-resolution based features, calibration and normalization, and using profile errors. These new features extend the potential of clustering load curves. Moreover, smart metering is now generating high-resolution time series, and so the dimensionality reduction offered by these features is welcome.

The clustering results using the proposed new features are compared with clusterings obtained from temporal features as well as clusterings with Fourier features, using household electricity consumption time series as test data. The experimental results suggest that the proposed feature construction methods offer new means for gaining insight in energy consumption patterns.

***Index Terms*—Feature construction; feature transformation; clustering; meter data analytics**

I. Introduction

THE smart grid and smart metering play an essential role in future energy management [1]. This study demonstrates how greater time-resolution household electricity meter readings are analyzed to extract typical daily usage patterns.

In the UK, Elexon profiles are the industry standard, used to represent presumably typical consumption load curves [2]. For domestic consumers, there are two profile classes: customers choosing tariff “economy 7” and the rest. For non-domestic users, on the other hand, there are six profile classes. The usefulness of these Elexon “profiles” for domestic customers is unsatisfactory. It has been reported that the use of the profiles has made about 9×10^{12} watt-hours electricity losses yearly in the UK [3]. To design better profiles is an open challenge. This paper aims to contribute to this challenge by proposing methods to segment and extract households’ typical daily load curves from their actual consumption. One experiment also uses data about households’ Gas connection.

Advanced data mining methods have not been fully adopted in practice yet, mainly due to the limited quality of data available. At present, most UK domestic consumers still have their electricity meter data read quarterly or half-yearly. This sparseness limits the potential to accurately separate load curves at daily, weekly, or even monthly level.

In the UK, the introduction of smart metering has started to generate half-hourly electricity usage data. Such data enables meter data analytics at a much finer resolution, so as to gain a better understanding of energy usage.

This work generates clusters of load curves. To choose and design appropriate features for clustering is vital. This paper proposes three new types of features for clustering and applies them on real smart meter data. The representative load curves from the resulting clusters provide insights for refining the existing profiles, and might even be used as the basis of new profiles. The clustering results with these new features are assessed and compared with the clustering results using two other methods of feature construction, which have been reported in the literature.

The paper is organized as follows: Section II discusses related work and Section III introduces the basic concept of feature construction. The data set is then described in Section IV, followed by experimental work to determine the appropriate number of clusters in Section V. After that, three clustering experiments, each using one of the newly constructed features, are reported in Sections VI, VII and IX. Control experiments are conducted in Sections VIII and X to compare clustering results. Section XI conducts a comparative study to evaluate the newly constructed features and reports the main findings. Finally, Section XII concludes the paper.

II. Related Work

Load profiling often includes three stages [4], [5]: firstly to group consumption behaviors using clustering methods; secondly to generate typical load patterns (load curves) for each resulting group using statistical criteria such as mean or median; finally, to associate customers’ characteristics, such as locations and incomes, with the typical load patterns, using classification methods. This work focuses on the first stage.

A rich body of the literature is available for clustering load curves. The electricity consumption or load data used for

Manuscript received November 25, 2014. Revised July 20, 2015 and November 6, 2015. Accepted for publication November 17, 2015. This work was supported in part by the U.K. Technology Strategy Board (Project 100923, TP 3981-33147). Paper no. TII-15-1176. Copyright 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

R. Al-Otaibi is a PhD student at the Department of Computer Science, University of Bristol, Bristol BS8 1UB, U.K. She is working at the Faculty of Computing, King Abdul-Aziz University, Saudi Arabia. (e-mail: ra12404@bristol.ac.uk)

N. Jin was with the Department of Computer Science, University of Bristol, Bristol BS8 1UB, U.K. She is now with the Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne NE1 8ST, U.K. (e-mail: nanlin.jin@northumbria.ac.uk)

T. Wilcox was with the Centre for Sustainable Energy, Bristol BS3 4AQ, U.K. He is now with the Mobile Robotics Research Group, University of Oxford, Oxford, U.K. (e-mail: tomw@robots.ox.ac.uk)

P. Flach is with the Department of Computer Science, University of Bristol, Bristol BS8 1UB, U.K. (e-mail: Peter.Flach@bristol.ac.uk).

clustering load curves form time series. The meter resolution (sampling rate) determines the number of time points sampled within a time period. The time period of interest is user-defined. The common examples are daily curve and weekly curve.

Given a data set with observations (data records) in d number of dimensions, the input data of load curves is conventionally set up in a matrix. It has a number of rows, each representing a customer. And it has a number of columns, each representing the consumption at a time point. In the literature, a column is also called a dimension, feature, or variable. The consumptions at all time points in a data set are called *default features* in this paper. They can be used directly for clustering. The data in the default features are typically aggregated or normalized values [4]. For example, data might be collected at a 15-minute sampling rate and the features of daily load curves for clustering are the corresponding 96 time points [6]; or the data might be sampled at hourly intervals and the features of daily load curves for clustering are the 24 time points [7].

It is NP-hard to find optimal clusterings even for two clusters [8], [9]. Therefore, dimensionality reduction methods have been extensively studied in the literature to reduce the number of dimensions. The known benefits include (a) to simplify the outputs models for easier interpretation by users [10], (b) to save computational resources and reduce time, and (c) to reduce over-fitting [11].

The methods of dimensionality reduction can be grouped in two categories: (1) feature selection, which selects a subset of features to replace the full set of all dimensions in the data set; and (2) feature construction, which creates new features by applying operations or functions on the default features. This is the focus of this paper.

Expert knowledge has often been applied to construct a set of application-dependent new features. Feature construction has been used in meter data analytics, where the four major ways to create new features in the literature are:

1) On the basis of default features, feature construction can be applied to reduce the time resolution [12]. For example, a created feature for the morning consumption combines the consumption from 7am to 12noon.
2) Previous work has designed a set of shape-related features to model the specific aspects of "signature" of the load patterns, for example, dimensionless ratios [13], load factor [14], and variability [15].
3) New features can be generated in the frequency domain, such as the harmonics-based coefficients, the coefficients derived from the wavelet transform, surveyed in [4], the Fourier series coefficients [16], and the fast Fourier transform (FFT) algorithm [17].
4) New features can also be constructed by Principal Component Analysis, Curvilinear Component Analysis, and Canonical Variate Analysis, surveyed in [4].

This paper will propose and demonstrate new methods of feature construction to generate processed data as inputs for clustering of daily curves. Limited research on this has been reported, although extensive research has been published on clustering methods.

This work studies how to construct new features that will improve clustering performance tested with two popular clustering methods, rather than finding features that only impact the performance of highly specialized clustering algorithms. For this reason the experiments are carried out with straightforward and widely used clustering methods such as K-means and K-medoids. The underlying hypothesis is that constructed features that enable improved performance with these baseline clustering algorithms are also likely to benefit more sophisticated algorithms.

III. Feature Construction

Machine learning models are only as good as the features they use, and this is particularly true for unsupervised learning methods that do not have access to labeled training data. Well-conceived new features can capture information which are unavailable from the default features in a data set [18]. Raw features often need to be transformed or combined with other features in order to be useful.

For example, many distance-based methods are sensitive to the scale of the feature, and careful normalization is therefore important. So, for example, instead of reporting the average daily energy consumption of a particular household in kWh, it is worth reporting that this household is 1.3 standard deviations below the mean.

Another common transformation is discretization into a relatively small number of bins, as in a histogram. This paper reports the design of new features and the results with a new discretization method that can be related to a binary signal. For example, given two demographic groups A and B, it is worth investigating a particular energy consumption range in terms of the percentage of group A households that fall in that range (out of all group A and group B households). Following [18], we call this feature calibration, as the process is akin to building a univariate binary classifier that outputs calibrated class probabilities.

In this paper, three new types of features are designed:

- New consumption-based and time-based features on the basis of prior knowledge of the aggregated daily consumption in the data set (Section VI). In addition, feature construction applies operations on existing features to create new features and feature transformation: scaling and normalization.
- Calibrated features incorporating additional information on households' gas connection (Section VII). The new features ideally contain additionally useful information to discriminate outcomes.
- The use of profile error as features (Section IX) incorporates domain-specific and problem-specific knowledge. In addition, it significantly reduces dimensionality.

Two comparative studies are undertaken: one uses the default features (Section X); and the other uses Fourier transform feature vector which converts data from time domain to frequency domain (Section VIII).

This paper does not intend to practise "feature subset selection" which only selects a subset of features. Instead, this paper focuses on the construction of new features. The

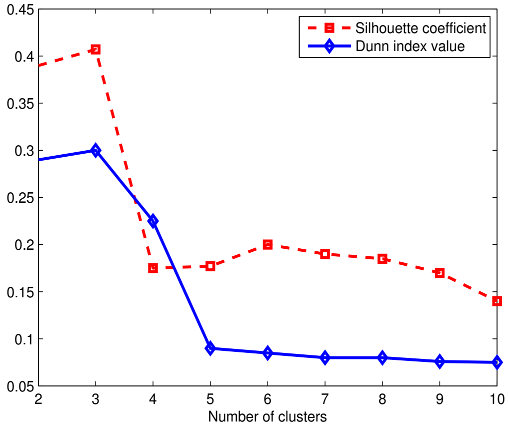


Fig. 1. Silhouette coefficients and Dunn index values for $K = 2$ to $K = 10$.

number of new features is less than the default ones: the number of default features in a daily load curve at half-hourly sampling rate is 48, which is manageable, but when clustering a weekly or monthly load curve, the dimensionality becomes a concern. For example, a weekly load curve has 336 default features at half-hourly sampling rate. So it becomes necessary to use a smaller number of features in order to reduce the computational burden especially when the data set is large.

IV. Data Set

SSE Energy Supply Ltd, UK has collected electricity usage data of 5000 households, at a temporal sampling rate of one reading per 30-minute, from April 2009 to Oct 2010.

Many energy suppliers offer time-of-use tariffs. One popular differential tariff, called Economy 7, charges at a higher price from 7am to 12 midnight, and charges a lower price at the rest of time. This tariff economically discourages consumption during the peak time. The consumption data in the SSE data set was collected when a flat tariff was applied to any time, so these readings reflected the actual demands in absence of any impact from economic considerations to change consumption behaviors.

The consumption data is averaged to yield an aggregate daily load curve over 48 time points. The set of default features is thus the average daily energy consumption at these 48 time points. Normalized consumption data are also generated. A common practice of generating typical load patterns is to define the reference power in kWh and then to compute normalized representative load patterns (RLP) [4]. Here Relative Average consumption (RAC) of a time point is defined as the normalized average consumption at this time point relative to the average daily total consumption. Some of the experiments in this paper use the normalized consumption to capture the shape rather than the magnitude consumption.

A load curve can thus be represented as a vector C^T with T default features: $C^T = \{c_t, t = 1, \ldots, T\}$.

V. Determining the Number of Clusters

Before applying clustering methods to segment customers' load curves into groups, the appropriate number of clusters needs to be set. Many clustering algorithms, including K-means, K-medoids and fuzzy c-means, require a parameter which specifies the number of clusters to detect, here denoted by K. The appropriate values of this parameter are determined by data sets, prior knowledge, users' preferences and the properties of clustering algorithms of choice. While increasing the number of clusters tends to increase cluster compactness, an overly large number of clusters is practically useless and lacks representativeness. Therefore, an appropriate number of clusters balances these two considerations.

This paper uses data mining methods to determine K, with a practical constraint. Thus K is determined mainly by the nature of the data set, reflecting its characteristics. And the practical consideration was advised by industrial experts that the appropriate number of clusters should not be more than 10. From an industry point of view, the resulting number of clusters may be used for planning tariffs; or providing evidence for marketing. The operational cost for serving 10 types of different tariffs or marketing strategies will be within an affordable cost limit. However, the methodology is generic; users who would like to have a larger number of clusters can still use the same algorithms/methods to be presented. Two methods are used to search for the appropriate number(s) of clusters within the range $K \in [2, 10]$, as explained below.

The Silhouette coefficient combines a measure of how close samples within the same cluster are to each other with a measure of how well-separated one cluster is from other clusters [19]. The Silhouette value s of a sample load curve, which is assigned to a cluster is $s = (b-a)/\max\{b,a\}$, where b is the smallest average distance between this sample and the samples in another cluster; and a is the average distance of this sample to the other samples in its cluster. Silhouette values range from 1 to -1, with a value close to 1 indicating that this sample is much closer to samples from its own cluster than to samples from other clusters, and a value of -1 indicating that this sample might have been assigned to a wrong cluster. The Silhouette coefficient is then the average Silhouette value over all sample curves and can be used to quantitatively compare clustering results.

As an alternative to the Silhouette coefficient, we also use the Dunn index to determine the number of clusters. The Dunn index is defined as the ratio between the minimal within-cluster distance and the maximal between-cluster distance [20]. In our experiments, the within cluster distance has been chosen as the distance between the farthest two points inside one cluster. The among-cluster distance has been chosen as the farthest two data points, one data point from each cluster. Higher index values indicate better clustering.

Fig.1 shows the Silhouette coefficients and Dunn indices for $K = 2$ to $K = 10$, where K-means has converged within 100 iterations. As can be seen, both metrics indicate $K = 3$ as the most suitable number of clusters, with $K = 2$ the second-best choice. We hence set the number of clusters to 3 in our experiments. The next five sections will report the experiments using five different sets of features for clustering by means of K-means and K-medoids clustering. We used Matlab's K-means clustering implementation that applies K-means++ seeding by default.

VI. Constructed Features on Load Shapes

This section presents consumption-based and time-based new features. Furthermore, the Silhouette coefficient is employed to indicate the quality of the resulting clusters.

A. Feature Construction

The average daily usage in the data set demonstrates one morning peak and one evening peak, as seen in Fig.2. Capturing the consumption values at these two peaks characterizes households' patterns. The default features can be abstracted into a simpler model which uses a smaller number of features to reflect these two characteristic peaks. Based on this concept, six new features are designed to incorporate this observation.

These six new features model a household's consumption at three time points which vary from one household to another. This model is called "V-shape" shown in Fig. 2. The operations used to construct the new features are a conditional filter, for example, "before 2 p.m", and the maximum and the minimum functions. These new features include both the consumption-based features and the time-based features, replacing the 48 default features. The difference of the default features from the newly created time-based features is that given a default feature, its time is fixed, and its corresponding consumption for a household is known; but given a new time-based feature, its value on time is uncertain before finding the satisfying consumption.

If there are more than one consumption values satisfying the same condition, for example, being the maximum consumption before 2 p.m, the corresponding time of the latest one will be chosen as the value of its respective time.

To add complexity, the "M-Shape" is designed to include two more time points, shown in Fig. 2.

The new features are:

- $Cmin1A$ and $Tmin1A$ are the minimum consumption from midnight to 2 p.m. and its corresponding time: $Cmin1A_{Tmin1A} = \min\{c_t, t = 1, \ldots, 27\}$
- $Cmax1$ and $Tmax1$ are the maximum consumption before 2 p.m. and its corresponding time: $Tmax1 = \arg\max_{t\in\{1,\ldots,27\}} c_t; Cmax1 = \max_{t\in\{1,\ldots,27\}} c_t = c_{t=Tmax1}$
- $Cmax2$ and $Tmax2$ are the maximum consumption after 2 p.m. and its corresponding time: $Tmax2 = \arg\max_{t\in\{28,\ldots,48\}} c_t; Cmax2 = c_{t=Tmax2}$
- $Cmin1$ and $Tmin1$ are the minimum consumption between $Tmax1$ and $Tmax2$ and its corresponding time: $Tmin1 = \arg\min_{t\in\{Tmax1+1,\ldots,Tmax2\}} c_t; Cmin1 = c_{t=Tmin1}$
- $Cmin1B$ and $Tmin1B$ are the minimum consumption from 2 p.m to midnight and its corresponding time: $Cmin1B_{Tmin1B} = \min\{c_t, t = 28, \ldots, 48\}$

B. Feature Transformation

The distributions of a consumption related new feature or a time-based new feature vary greatly, so they have to be normalized for clustering to improve the results. Two approaches are applied: statistical normalization and scaling. In the former one, the values to a feature, x, are normalized to $x' = (x - \mu)/\sigma$, where μ and σ are the mean and standard deviation of x of all households in the data set. The purpose of this normalization is to convert data from any normal distribution into the standard normal distribution with mean zero and variance 1.

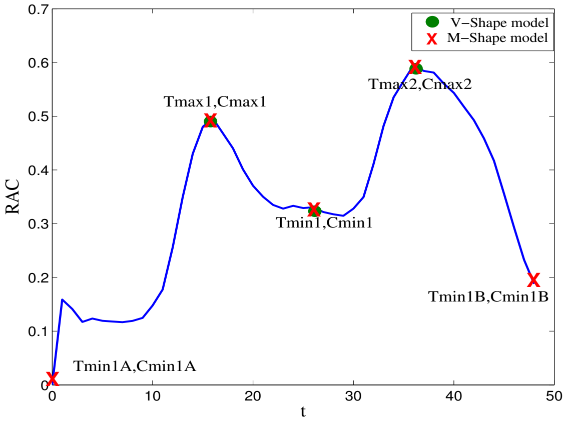


Fig. 2. V-Shape model with 6 features and M-Shape model with 10 features, where x-axis is time "t" of the 48 time points; and y-axis is the related daily consumption.

For the scaling method, the value of x is divided by its maximum and scaled into a value between zero and one. The transformed features are notated as x', for example $T'max1_i, C'max1_j$. The motivation of the use of scaling is because of the differences of the constructed features' values (both consumption and time). And it is known that the Euclidean distance is sensitive to these differences [18].

C. Distance Measures

The distance measure to be used for clustering is the total Euclidean distance over the points making up the V-shape or M-shape in (T, C) space. For the V-shape model this gives

$$dV(i,j) = \sqrt{(T'max1_i - T'max1_j)^2 + (C'max1_i - C'max1_j)^2} + \sqrt{(T'max2_i - T'max2_j)^2 + (C'max2_i - C'max2_j)^2} + \sqrt{(T'min1_i - T'min1_j)^2 + (C'min1_i - C'min1_j)^2} \quad (1)$$

A similar formula is used for comparing two M-shapes using 5 2-D points. Then the Silhouette values are calculated over the resulting clusters to measure how well separated they are, as shown in Figures 3 and 4. Each horizontal line represents the Silhouette value of one household, in decreasing order within each cluster.

It is observed that scaling constructed features to [0,1] may not be the better choice compared with the normalization.

More negative values of the Silhouette coefficient on the scaled features, as seen on Fig. 3 (right) and Fig. 4 (right), reveal the fact that there is less homogeneity within these clusters than the normalized features, as seen on Fig. 3 (left) and Fig. 4 (left).

V and M models are simple but informative, which in fact has considered the variability of the maximal and the minimal consumptions within a time period of interest, which probably offers richer information than the variability of consumption at given fixed time points. In addition, V and M models are representative and easy to operate.

VII. Calibrated Features

We now present the first of the three new feature types we propose for clustering daily load curves. Using calibration techniques from supervised machine learning, the values of the five new features created in the previous section will be calibrated before applying the clustering algorithm. The consumption-based features ($Cmax1$, $Cmax2$, $Cmin1$, $Cmin1A$, and $Cmin1B$) in the M-Shape model are calibrated using a Boolean variable, "main gas flag". The variable "main gas flag" is used in the industry to indicate whether a household is connected to the main gas network. If 'Yes' (the positive class), it is assumed that the household uses gas in addition to electricity.

A. Feature Calibration

Classifier scores can be calibrated in various ways in order to take empirical probabilities observed in the data into account. Numerical features can be seen as univariate scoring models and hence are amenable to such calibration methods. In our work, the consumption-based features are transformed using isotonic feature calibration [18]. The purpose here is to discretise each consumption feature into a smaller range, meaning many consumption values will be mapped to the same calibrated value. If the calibrated value is 0.7, for instance, it means that 70% of the households with consumption values falling in this range are connected to the main gas network.

Specifically, the algorithm is as follows:

- Sort the households descending on a consumption-based feature.
- Create the ROC curve, which depicts the trade off between the true positive rate and the false positive rate [21]. Here we use the "main gas flag" to construct the ROC curve (if the "main gas flag" is positive, move up, otherwise, move right).
- Construct the convex hull of the ROC curve. This ensures that the proportion of positives is monotonically non-increasing along the curve.
- Obtain the calibrated feature by computing the proportion of positives in each segment of the ROC convex hull as shown below:

$$v = \frac{g+1}{g+1+p(e-g+1)} \tag{2}$$

where: g is the total number of households with a positive main gas flag in the segment; e is the total number of households in a segment regardless of the class sign; and p is the prior positive class probability of the main gas flag.

An example of the ROC Convex Hull for $Cmax1$ is shown in Fig. 5. On the left is an example ROC curve of $Cmax1$ for 20 households while on the right is ROC curve of $Cmax1$ for all households in the data set. The solid line is the ROC curve and the dashed line is the convex hull, each segment of which corresponds to a discrete calibrated feature value. "B" on Fig. 5 (left) refers to the second segment. It has 4 households in total ($e = 4$). Two of them have a positive main gas flag ($g = 2$). These segments of the convex hull represent a discretization of the values of $Cmax1$.

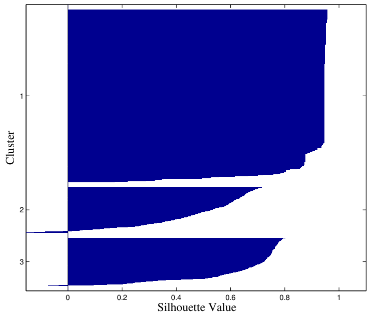


Fig. 6. Silhouette output using calibrated features, average silhouette=0.7888, where x-axis is Silhouette value; and y-axis is the clusters.

The ROC curve and its convex hull are used in this section to calibrate the features before clustering. One of the advantages of this approach is that ROC curve ignores the magnitude of the features and only takes their rank order into account (i.e., the lowest value gets rank 1, the next value gets rank 2, etc.). This rank order requires consideration of all points at once, but does not depend on the order of presentation of the points.

B. Clustering

K-means clustering method is applied to the calibrated features with $K = 3$, and the Silhouettes are shown in Fig. 6. As can be seen in this figure, fewer households returning negative Silhouette values compared to the clustering results in Fig. 3 and Fig. 4, showing that fewer households are assigned to the wrong clusters. Average silhouette using the calibrated features is 0.7888. In addition, as Fig. 6 shows, the top cluster which is also the largest one, has the smallest average error. The finding itself may suggest that a large number of households share very similar consumption patterns.

VIII. Fourier features

Fourier analysis transforms temporal data into the frequency domain, providing a robust method for extracting the major frequency components of a time series. It has been used to forecast daily patterns of electricity consumption [17]. Frequency components representing the major patterns in the temporal data are collected in a Fourier feature vector. We then use Euclidean distance on these Fourier feature vectors to cluster daily load curves by means of K-medoids clustering.

Fourier Transforms provide desirable properties: they are stable mechanisms that produce the same value given similar inputs; they are robust to missing data which can be a concern; they are phase/translation-invariant, meaning time series with similar frequency patterns but different start and end points will be represented by similar transforms.

By applying a Fast Fourier Transform to daily load curves, a corresponding series of coefficients for component frequencies

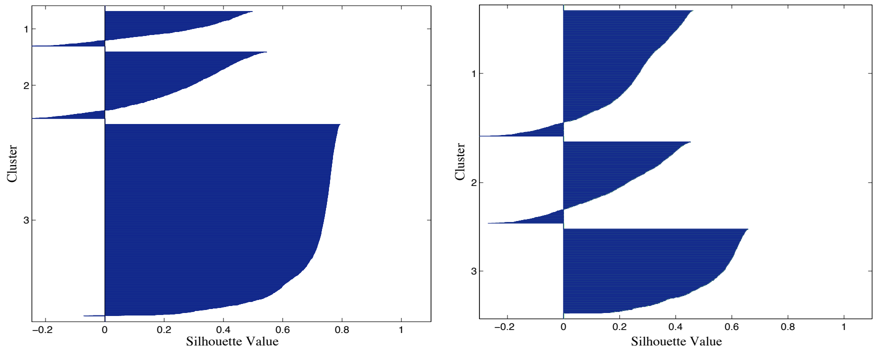


Fig. 3. (left) Silhouette output using V-Shape model (Normalized), average silhouette=0.5383. (right) Silhouette output using V-Shape model (Scaled), average silhouette=0.3081. X-axis is Silhouette value; and y-axis is the clusters.

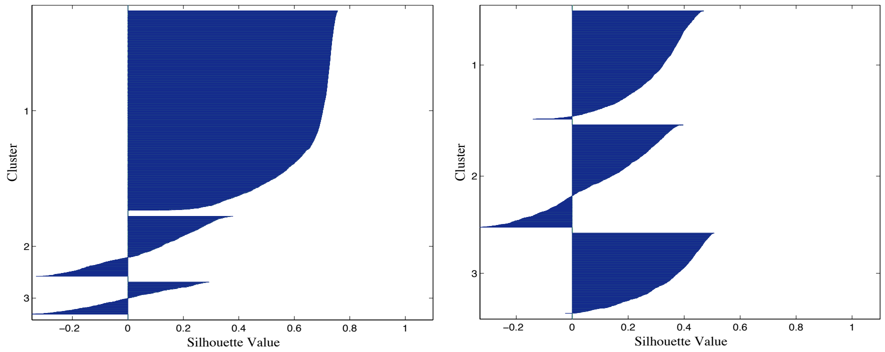


Fig. 4. (left) Silhouette output using M-Shape model (Normalized), average silhouette=0.4599. (right) Silhouette output using M-Shape model (Scaled), average silhouette=0.2420. X-axis is Silhouette value; and y-axis is the clusters.

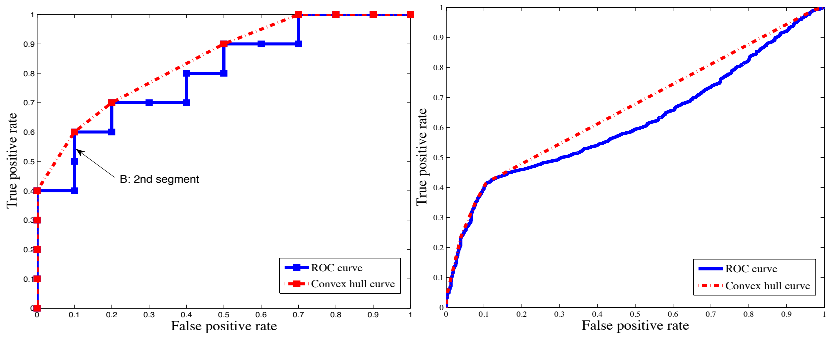


Fig. 5. (left) an example of ROC curve and Convex Hull of *Cmax*1 for 20 households, as an demonstrative example, where x-axis is the false positive rate; and y-axis is the true positive rate. "B" refers to the second segment. (right) ROC curve and Convex Hull of *Cmax*1 for all households in the data set, where x-axis is the false positive rate; and y-axis is the true positive rate.

is produced. This will reduce the dimensionality of the data for each household, whilst preserving the most significant periodic patterns present within each time series in a manner which enables comparison of household consumptions by those key features. From here two approaches were considered for building a feature vector to best represent the data.

A. Highest-Variance Frequency Component Coefficients

One approach is to evaluate the variance of the amplitude of each frequency across the data set and then rank frequencies in decreasing order by variance. This would identify the best frequencies to use in differentiating between households.

The 25 most variable frequencies are selected to create a feature vector for each household containing the corresponding amplitudes for those frequencies for each time series. Issues with this approach arose from the fact that the length of time series varied significantly across the data set. In addition, the training data included some extreme (possibly anomalous) time series values, which may have exaggerated the variance of amplitudes. They may have caused some frequency components to be incorrectly promoted higher in the ranking process.

More specifically, data curation consisted of an initial phase applied to the raw data to remove data points above and below manually set thresholds from the time series. We then continued to remove those time series from the training data set which contained too few points to satisfy a minimum coverage/overlap when compared over a specified time period (e.g. 2 years) with all other time series included in the training data set. This was in an effort to ensure that all time series were being compared over the same time period. Finally, for the purposes of constructing the Fourier features, we removed those sample time series that contained frequency amplitudes which were beyond 5 standard deviations of the average, as the threshold, which accounted for approximately less than 5% of the entire training data set. For example, a sample time series would be removed from the training data if it contained an amplitude of 13 for a frequency with an average amplitude of 2 and standard deviation of 2 across the entire data set. This resulted in the removal of those most extreme samples, containing atypical patterns, which would skew the overall distribution of the training data set when evaluating these amplitude-based features for identifying general cluster.

The similarity between two feature vectors is evaluated using the Manhattan distance between each coefficient pair. Issues with missing data resulting in missing frequency component values in the transform may affect the reliability of this distance metric.

B. *Amplitude-ranked Frequency Component Coefficients*

An alternative approach is to create a feature vector to represent the important frequency components of a time series. First, the complex amplitudes of each frequency component of a time series are measured, and then they are ranked by the absolute value of those amplitudes. Finally the feature vector is created by selecting the top 25 frequency-amplitude pairs in this ranking. The similarity of two features are calculated, taking the Manhattan distance between the amplitude and frequency values.

Highest-Variance Frequency Component Coefficients and Amplitude-ranked Frequency Component Coefficients yield similar results in evaluation on a subset of samples, but the first one has not been tested on the whole data set due to the aforementioned considerations and the limits on computational resources. In Table I, FFTFeatureVector(A) refers to Amplitude-ranked Frequency Component Coefficients.

IX. Subgroup Discovery Using Profile Error

Subgroup discovery is a data mining method to uncover unusual patterns associated with selected features [22]. It has been used in analyzing smart meter data [14]. The induced rules can be used to divide the data into two exclusive groups: one satisfying the rule and the rest. The following experiment will apply subgroup discovery to partition samples on their profile errors. Profile errors are widely used in industry to evaluate the accuracy of profiles, using the difference between the profile estimate and the actual consumption. One of such error measures is mean absolute percentage error (MAPE) [23].

In this experiment, one cluster is generated by grouping 12% of the samples with the highest MAPE values. The rest of the data is given to a subgroup discovery algorithm to separate into two more groups with different MAPE distributions. The resulting three clusters show a cluster of households with small MAPE, and a cluster of households with medium MAPE and a cluster with very high MAPE.

This experiment has one target feature, namely MAPE and 13 socio-demographic pattern features. The definitions of the socio-demographic pattern features can be found in [14] which used a similar data set. This approach is usually used for rule generation. In this special case, it is also used for sample segmentation.

X. K-means Clustering using Default Features

Finally, as a baseline, the 48 default features are used as features for clustering. Households' RAC values at 48 time points are used for clustering. K-means clustering with squared Euclidean distance has been applied with $K = 3$, as seen in Fig. 1.

XI. Evaluation and Comparison

This section evaluates the constructed new features for clustering. Generally speaking, clustering can be evaluated by two approaches: external and internal. External approach compares the resulting clusters with externally supplied class labels. Class labels are not used during clustering, but used to assess the resulting clusters. The SSE data set has no class labels on electricity consumption, therefore this approach is impractical. The internal approach requires no knowledge of external class labels. Two measurement criteria have been widely used for evaluating clustering results, namely compactness and separation. They are combined in a single score by means of a modified version of the Clustering Dispersion Indicator [24].

Notations

W	within-cluster distance
A	among-cluster distance
K	number of clusters, in this case, $K = 3$
k	cluster numbered k and $k \leq K$
u_k	centroids of cluster k
m	medoid
n_k	the number of samples in cluster k
N	total samples
S_k	the set of samples in cluster k
y	a household daily load curve
T	the number of time values in each time series

A. *Evaluation Metrics*

First a distance measure is introduced to evaluate the similarity of two load curves. A variety of distance measures are considered, include dynamic time warp analysis [25]. However, limited by computational resources, a simple yet sufficient measure is chosen on the basis of the City-Block or the Manhattan distance: a mean of a ratio of the absolute differences between the normalized consumption values at the same time t for two samples i and j, $i \neq j$:

$$p_t(i,j) = \frac{|c_{i(t)} - c_{j(t)}|}{c_{i(t)} + c_{j(t)}} \tag{3}$$

The symmetric mean of a ratio of consumption for two samples i and j, is their distance metric:

$$f(i,j) = \frac{\sum_{t=1}^{T} p_t(i,j)}{T} \tag{4}$$

Among clusters distance, A, is the average distance between all the cluster centroids u_k:

$$A = \frac{\sum_{k=1}^{K-1}\sum_{l=k+1}^{K} f(u_k,u_l)}{K(K-1)/2} \tag{5}$$

Within cluster distance, W, is the mean distance between households' load curves and their corresponding centroid values:

$$W = \frac{\sum_{k=1}^{K}\sum_{\forall x\in S_k}^{n_k} f(y,u_k)}{K} \tag{6}$$

To evaluate clusters produced by using different features, a score formula is designed. It favors clustering whose distance among every cluster's centroid, as the separation measure, is large, and the distance between every member to its cluster's centroid, as the compactness measure, is small, in principle:

$$Score = \frac{A}{W+\varepsilon} \tag{7}$$

The parameter ε handles the trivial cases where clusters consist of a single sample, therefore resulting in the within-cluster distance being equal to zero which would result in an undefined score. Therefore, we include an additional epsilon with an arbitrary, small value to handle this case without allowing infinite scores or significantly affecting the results when comparing clustering methods.

B. Experimental Results and Discussion

Table I compares the quality of the resulting clustering of the seven different sets of features, on the basis of nearly 5000 households data. Table II explains these seven sets of features. This shows that constructed features, namely calibrated features on M-shape, and subgroup discovery on profile error are among the best performing. The use of the default features for clustering yields a reasonably good score. The 2^{nd} column is on the measure of mean of within cluster distance: the smaller values the better; the 3^{rd} column is the measure of mean of between cluster distance: the larger the better; the 4^{th} column gives the trade-off measure to balance the two aforementioned measures. The experimental evidences have shown that the constructed features have achieved competitive clustering results, as shown in Table I. It is clear that the two constructed features, namely Calibration (on M-shape) and subgroup discovery return the best results. Calibration is 7.87% better than the default one; and the subgroup discovery is 6.06% better than the default features. V shape-S is almost identical to the default ones. Importantly, using the constructed features can reduce the computational load significantly.

Fig.7 shows representative load curves of the resulting clusters each using a different set of features. For each plot, the three households whose load curves are the medoids of their respective clusters are chosen as the representatives of the resulting clusters. The average consumptions at 48 time points on a day in unit kWh of such representative households are plotted.

The three medoid curves in Fig.7 (i) are generated using the default features. The medoid "i-b" has a high morning peak, and much less of an afternoon peak. In contrast, the medoid "i-a" has an unnoticeable morning peak, and a long lasting afternoon peak. The medoid "i-c" is the regular one similar to the overall average of the data set. The remaining three plots, using the constructed features, display some of the discovered consumption behaviors which are distinct to those demonstrated in (i). In plot (ii) the medoid "ii-b" consumes a larger and stable volume over a day; the medoid "iii-b" in plot (iii) consumes a larger volume only at day time, in particular in the morning and late afternoon peak time; the medoid "iv-c" on (iv) consumes the highest in kWh but only during two peak time periods. The flat load curves, namely "iii-a" on (iii) and "iv-a" and "iv-b" on (iv) are only found by V shape features. We see that clustering using the constructed features successfully separates such distinctive patterns from the rest.

The data set is for household consumption, but the concepts of feature construction and calibration are generic, easily applied to industrial and commercial consumption as well. The time series chosen here are daily load curves. However, the approach is applicable to weekly, monthly and even yearly time series, although the models of these time series are more complicated (e.g., V-shape and M-shape features will need to be adapted to longer patterns).

As Table II indicates, the 6 ways to create new features return much smaller numbers of features to be used for clustering. To replace the default features with the constructed features, the number of dimensions is reduced from 48 to 25, 14,10, 6 or 5. Then, computational resource and time are saved by using a set of new features.

XII. Conclusion

Meter data analytics is one of the most important parts of smart grids. To analyze the recently available fine-grained data delivered by smart metering systems will help achieve the full potential of smart grids. This paper focuses on clustering daily load curve and proposes three new types of features that are generated by applying conditional filters on meter-resolution based features integrated with shape signatures, calibration and normalization, and profile errors.

Given the shape signatures, such as peak and off-peak consumption time widely used in industry, conditional filters have been used to create new consumption-based features and time-based features. They form alternatives to the 48 default time-based features for clustering. The second new method of feature construction utilizes two feature transformation techniques, namely statistical normalization and scaling. They have been further integrated to the first feature construction method to improve the performance. The third newly proposed feature construction method modifies the consumption-based features developed by the first method, using ROC convex hull and calibration.

The first advantage of the proposed techniques of feature construction is related to computational complexity. The constructed features produce smaller numbers of dimensions. This

TABLE I
EVALUATION ON CLUSTERING OUTCOMES ORDERED BY SCORE.

Feature(s)	Mean of Within Cluster Distance (W) Eq.6	Mean of Between Cluster Distance (A) Eq.5	Score Eq.7
Calibration	0.3181	0.1485	0.4661
SD MAPE	0.3208	0.1437	0.4480
Default	0.3217	0.1249	0.3874
V shape-S	0.3234	0.1242	0.3834
V shape-N	0.3234	0.1168	0.3606
M shape-N	0.3256	0.1135	0.3480
FFTFeatureVector(A)	0.3329	0.0380	0.1143

TABLE II
SUMMARY OF CONSTRUCTED FEATURES

Method of Feature(s)	Description	Definition	Size of the Feature Set	Detail
Calibration	Calibrated features by Gas Connection	Section VII	5	Calibrated $\{Cmax1, Cmax2, Cmin1, Cmin1A, Cmin1B\}$
SD MAPE	Subgroup discovery using profile error	Section IX	14	MAPE and 13 socio-demographic features
Default	Default features	Section X	48	$\{c_1, c_2, \ldots, c_{48}\}$
V shape-S	V shape-scaled	Section VI	6	Scaled $\{C'max1, T'max1, C'max2, T'max2, C'min1, T'min1\}$
V shape-N	V shape-normalized	Section VI	6	Normalized $\{C'max1, T'max1, C'max2, T'max2, C'min1, T'min1\}$
M shape-N	M shape-normalized	Section VI	10	Normalized $\{C'max1, T'max1, C'max2, T'max2, \ldots, C'min1B, T'min1B\}$
FFTFeatureVector(A)	Fourier Transform Feature Vector	Section VIII-B	25	25 most variable frequencies

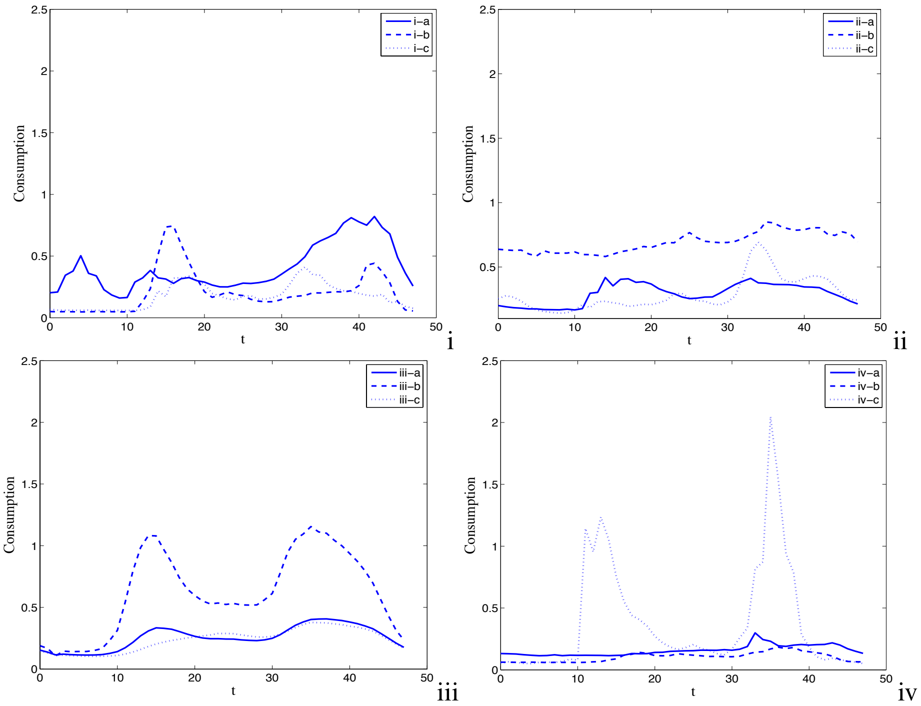


Fig. 7. The average consumptions at daily 48 time points in kWh of the representative households, using (i): default features; (ii):M shape-S; (iii): V shape-S; (iv): V shape-N. X-axis is time "t" of the 48 time points; and y-axis is the related daily consumption.

will consequently reduce computational demand. Secondly, the clustering performance of the constructed features are compared, measured by compactness and separation. Our experiments showed that two sets of the constructed features outperform the use of default features. Thirdly, another advantage of adopting newly constructed features is to improve comprehensibility. As shown, the models using the new features are informative, comprehensive and understandable in describing the electricity usage of daily periodicities and trends.

Hence, this study offers approaches and experiences on consumption pattern recognition, potentially useful to utility companies for tariff design and recommendation; consumption estimation; and demand response management.

REFERENCES

[1] V. Gungor, D. Sahin, T. Kocak, S. Ergut, C. Buccella, C. Cecati, and G. Hancke, "A survey on smart grid potential applications and communication requirements," *Industrial Informatics, IEEE Transactions on*, vol. 9, no. 1, pp. 28–42, 2013.

[2] K. Spencer, "Load profiles and their use in electricity settlement," *Elexon*, 2013.

[3] J. Andrews, "Review of gsp group correction scaling weights," *Elexon*, no. 150/04, 2013.

[4] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern grouping," *Energy*, vol. 42, no. 1, pp. 68 – 80, 2012.

[5] F. McLoughlin, A. Duffy, and M. Conlon, "A clustering approach to domestic electricity load profile characterisation using smart metering data," *Applied Energy*, vol. 141, pp. 190 – 199, 2015.

[6] W. Labeeuw and G. Deconinck, "Residential electrical load model based on mixture model clustering and markov models," *Industrial Informatics, IEEE Transactions on*, vol. 9, no. 3, pp. 1561–1569, 2013.

[7] A. M. Ferreira, C. A. Cavalcante, C. H. Fontes, and J. E. Marambio, "A new method for pattern recognition in load profiles to support decision-making in the management of the electric sector," *International Journal of Electrical Power and Energy Systems*, vol. 53, no. 0, pp. 824 – 831, 2013.

[8] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "Np-hardness of euclidean sum-of-squares clustering," *Machine Learning*, vol. 75, no. 2, pp. 245–248, 2009.

[9] S. Dasgupta, C. La Jolla, and Y. Freund, "Random projection trees for vector quantization," *Information Theory, IEEE Transactions on*, vol. 55, no. 7, pp. 3229 – 3242, 2009.

[10] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. Springer, August 2013.

[11] M. L. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. F. Wright, J. F. Wilson, F. Agakov, P. Navarro, and C. S. Haley, "Application of high-dimensional feature selection: evaluation for genomic prediction in man," *Sci. Rep.*, vol. 5, no. 10312, 2015.

[12] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, "Customer characterization options for improving the tariff offer," *Power Systems, IEEE Transactions on*, vol. 18, no. 1, pp. 381–387, Feb 2003.

[13] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern grouping," *Energy*, vol. 42, no. 1, pp. 68 – 80, 2012, 8th World Energy System Conference, {WESC} 2010.

[14] N. Jin, P. Flach, T. Wilcox, R. Sellman, J. Thumim, and A. Knobbe, "Subgroup discovery in smart electricity meter data," *Industrial Informatics, IEEE Transactions on*, vol. 10, no. 2, pp. 1327–1336, May 2014.

[15] I. Dent, T. Craig, U. Aickelin, and T. Rodden, "Variability of behaviour in electricity load profile clustering; who does things at the same time each day?" in *Advances in Data Mining. Applications and Theoretical Aspects*, ser. Lecture Notes in Computer Science, P. Perner, Ed. Springer International Publishing, 2014, vol. 8557, pp. 70–84.

[16] S. Verdu, M. Garcia, C. Senabre, A. Marin, and F. Franco, "Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps," *Power Systems, IEEE Transactions on*, vol. 21, no. 4, pp. 1672–1682, Nov 2006.

[17] M. Manera and A. Marzullo, "Modelling the load curve of aggregate electricity consumption using principal components," *Environ. Model. Softw.*, vol. 20, no. 11, pp. 1389–1400, 2005.

[18] P. Flach, *Machine Learning: The art and science of algorithms that make sense of data*. Cambridge University Press, September 2012.

[19] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, no. 0, pp. 53 – 65, 1987.

[20] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973.

[21] P. Flach, "ROC analysis," in *Encyclopedia of Machine Learning*, 2010, pp. 869–875. [Online]. Available: http://dx.doi.org/10.1007/978-0-387-30164-8_733

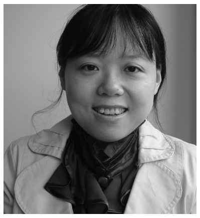
[22] M. Meeng and A. Knobbe, "Flexible enrichment with cortana – software demo." in *Proceedings of BeneLearn*, 2011, pp. 117–119.

[23] C. Borges, Y. Penya, and I. Fernandez, "Evaluating combined load forecasting in large power systems and smart grids," *Industrial Informatics, IEEE Transactions on*, vol. 9, no. 3, pp. 1570–1577, Aug 2013.

[24] I. Panapakidis, G. Christoforidis, and G. Papagiannis, "Modifications of the clustering validity indicators for the assessment of the load profiling procedure," in *Power Engineering, Energy and Electrical Drives (POWERENG), 2013 Fourth International Conference on*, May 2013, pp. 1253–1258.

[25] F. Petitjean, A. Ketterlin, and P. Ganarski, "A global averaging method for dynamic time warping, with applications to clustering," *Pattern Recognition*, vol. 44, no. 3, pp. 678 – 693, 2011.

Reem Al-Otaibi received the M.Sc. degree in computer science from King Abdulaziz University, Jeddah, Saudi Arabia in 2009. Since 2012, she is a PhD student at Intelligent System Laboratory, University of Bristol, Bristol, U.K. Her research interests are machine learning, data mining and multi-label classification. She is also working as a Lecturer in Faculty of Computing and Information Technology, King Abdulaziz University, Saudi Arabia where she got her scholarship.



Nanlin Jin received the Ph.D. degree in computer science from the University of Essex, Essex, U.K. Since 2013, she has been a Lecturer at the Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne, U.K. Her research interests include computational intelligence, heuristic optimization, data mining, and multi-disciplinary research. She was a recipient of the IEEE Computational Intelligence Society (CIS) Student Travel Grant.

Tom Wilcox received the M.Eng. degree in computer science from the University of Bristol, Bristol, U.K., in 2008. In 2011, he joined the Centre for Sustainable Energy, Bristol, U.K., as a Research Software Engineer. He is now working as a Research Engineer in the Mobile Robotics Research Group, University of Oxford, Oxford, U.K. His research interests include designing and developing web, geographic information system (GIS), and database software for modelling tools, computer games, and robotics.

Peter Flach received the Ph.D. degree in computer science from Tilburg University, Tilburg, The Netherlands, in 1995. He is a Professor of Artificial Intelligence with the University of Bristol, Bristol, U.K. His research interests include the evaluation and improvement of machine learning models using receiver operating characteristic (ROC) analysis, and learning from highly structured data. He is the author of Machine Learning: The Art and Science of Algorithms for Making Sense of Data (Cambridge University Press, 2012). Dr. Flach is the Editor-in-Chief of Machine Learning.