Predicting Consumer Load Profiles Using Commercial and Open Data

Dauwe Vercamer, Bram Steurtewagen, Dirk Van den Poel, *Senior Member, IEEE,* and Frank Vermeulen

***Abstract*—Automated Metering Infrastructure (AMI) has gradually become commonplace within the utilities industry and has brought with it numerous improvements in all related fields. Specifically in tariff setting and demand response models, classification of smart meter readings into load profiles helps in finding the right segments to target. This paper addresses the issue of assigning new customers, for whom no AMI readings are available, to one of these load profiles. This post-clustering phase has received little attention in the past. Our framework combines commercial, government and open data with the internal company data to accurately predict the load profile of a new customer using high performing classification models. The daily load profiles are generated using Spectral Clustering and are used as the dependent variable in our model. The framework was tested on over 6000 customers from GDF SUEZ in Belgium and six relevant load profiles were identified. The results show that the combination of internal data with commercial and cartographic data achieves the highest accuracy. Using external data alone, the model was still able to adequately place customers into their relevant load profile.**

***Index Terms*—Advanced Metering Infrastructure, Load Profile, Spectral Clustering, Classification, Open Data**

I. Introduction

WITH the rise of the Advanced Metering Infrastructure (AMI) over the past few years, vastly more energy consumption data is being collected today, and at a higher resolution than before. This facilitates innovative technologies and smart analytics to gain deeper insight into both the micro- and macro-level power consumption patterns of consumers. In addition, they also effectively improve control over the whole energy demand/supply system. These insights facilitate the realization of management techniques such as demand response, dynamic payment programs and flexible consumption hours [1], [2]. Programs such as targeted tariff schemes are getting more and more attention as energy retailers in competitive markets seek to better balance their loads and to maximize their total profitability [3]. The customers also benefit from them as they can gain insights into their energy consumption and lower it, or shift it to less expensive times in order to reduce the total cost of their energy bill. To execute these programs effectively, the electricity retailers have to predict which customers will participate in a certain demand-response measure and to allocate each customer to one of the schemes.

Prediction and classification of energy customer behavior is therefore an important field in domain-driven data mining. For the consumption side, besides forecasts, one is interested in customer load profiles that provide insights about their daily, weekly, or monthly behavior. While previously these profiles were made by the producers, more and more they are based on real data coming from AMI that is installed on-site. As behaviors change dramatically over time, decision makers are in need of learning algorithms that enable them to act upon these changes. Unsupervised methods are means for this kind of data analysis. Clustering algorithms detect groups of customers showing highly similar behavior, without any prior knowledge about these groups. For instance, in demand-side management and tariff setting, these clusters are used for specific strategies in each customer group. Current literature indicates that spectral clustering is very well suited to this end as it has a stronger focus on the shape of the loads and detects similar patterns with a small shift [4]. For a full overview of methods that cluster these load patterns, we refer the reader to the surveys of Chicco [5] and Zhou [6].

However, once a load profile has been generated, assigning new and existing customers to one of these profiles remains a difficult task, not in the least because of the lack of an ubiquitous AMI. This classification step relates to the post-clustering phase [5]. Given the importance of custom tariff offerings, this is still highly relevant. To do this, consumers are often associated with a predefined customer load profile. Currently, a customer is assigned to one of these profiles based on commercial codes or his type of application [5], [7]. Within commercial codes however, often there can still be big differences in consumption patterns [5]. Nowadays, much more information is available that can help in this assignment. To our knowledge, little research has been done to assess whether adding information from government bodies and other commercial players can enhance this classification. Preliminary research illustrated the potential of adding such variables, but it did not provide conclusive results yet [8].

In response to this gap, this paper enriches the internal company data a producer may possess with commercially available data to predict load profiles for the business-to-business market. For this we will first generate load profiles with a spectral clustering algorithm [4]. After this step we will try to predict these clusters using internal company information combined with data from government bodies and commercial companies. These additional data types include publicly available municipal data (demographics, crime rates, ...), publicly available cartographic data (lot size, building size, usable area, heatmap) and commercially available company data (turnover, number of employees, ...). To verify the quality of our model, we will compare two popular classification algo-

D. Vercamer, B. Steurtewagen and D. Van den Poel are with the Department of Marketing, Universiteit Gent, Gent, Belgium (email: Dauwe.Vercamer@UGent.be)

F. Vermeulen is with the Center of Expertise in Economic Modelling and Studies (CEEME) at Electrabel - GDF SUEZ, B-1000 Brussels, Belgium.

rithms, namely Adaptive Stochastic Boosting [9] and Random Forests [10]. We use data from 6000+ smart meters throughout Flanders in Belgium from non-residential customers over a period of two years. The aim of this paper is to determine whether we can succesfully assign a customer, for whom no smart meter data is available, to its relevant load profile so that the energy producer can act upon this information.

The rest of this paper is organized as follows. Section II dives deeper into the clustering algorithms used for creating load profiles and methodologies to classify them. Section III details our solution approach and Section IV provides the results. Lastly, in Section V we conclude our research and give indications for future research.

II. Literature Review

The idea of using demand response models and load-based tariff setting is becoming increasingly important with a more widespread AMI. One paper used clusters of AMI time series to develop optimized pricing schemes [3]. Another one also focused on creating better tariffs using load profile information [11]. Other papers that have focused on these types of applications can be found in [12]–[14]. Crucial in the development of these models is the generation of the load profiles.

The load classification cycle consists of four phases: (i) data gathering, (ii) pre-clustering, (iii) clustering and (iv) post-clustering [5]. The first phase is focused on gathering and cleaning the data. In the second phase, the data is processed so as to prepare it for clustering, resulting in the generation of the input data set for the clustering algorithm. Then the actual clustering is performed and the centroids are formulated to represent them as Typical Daily Profiles (TDP). Finally, in the post-clustering phase, meaning is given to the generated load profiles by identifying the relevant customer attributes that define them. This is an important step in tariff setting as it allows new customers to be easily assigned to their relevant profile.

The clustering phase of this cycle has gained a significant amount of attention in recent years and different methodologies have arisen to tackle this problem. A recent survey identified five different clustering methods for load profiling: partitioning based methods, hierarchical methods, density-based methods, grid-based methods and model-based methods [6]. The most commonly used clustering techniques in this context are k-means [15], fuzzy c-means [16] and Self-Organizing Maps [17], [18]. [5] also adds the Follow The Leader [11] and Probabilistic Neural Networks [19] to the methodologies. Recent studies have shown the merits of using other techniques as well. In [4], [20], spectral clustering using weighted kernel principal component analysis was used and Tsekouras [21] employed a two-stage approach.

With regard to assigning attributes to the load profiles, the paper by Figueiredo [7] provides the first insights. In this paper, commercial indices, combined with load shape indices are used to predict the representative load classes. Here it became clear that the Load Factor index and the Night Impact index were most crucial in determining the correct load profiles. However, this paper did not include any information that could be collected freely outside of the organization. [14] also stressed the post-clustering phase but again neglected to include external information, be it freely or commercially available, for the attribution to the load profiles.

III. Algorithmic Framework

Our proposed framework addresses the need of utility companies for a clear segmentation of customers, for whom no individual-level consumption patterns are readily available, by combining both supervised and unsupervised techniques. Figure 1 details our procedure. First, we need to define the different segments the company wishes to focus their incentive programs on. We do this by looking for customers with similar consumption behavior. The rationale is that consumers with similar behavior will likely be more receptive to the same kinds of incentives. To detect these different segements, we aggregate similar load patterns into clusters of similar consumption behavior. Once we have identified these clusters, the next phase is to assign new customers to these pre-existing clusters. This is not a trivial task as we typically do not have consumption information on these new customers. Therefore, we first identify the relationship between the load profile and the company's internal database without using smart meter data. This is done by predicting the probability of belonging to a certain cluster by means of classification algorithms. As typically the internal database of a company alone does not provide adequate predictions, we extend this internal database with government data, open data and commercially available data of the customer. The result of the classification is a model that relates the internal and external data to certain load profiles. Afterwards, when a new consumer applies to the provider, we can use these relationships to calculate the probability for this consumer to belong to a specific load profile. The following sections will dive deeper into every single pillar and discuss every step in the process more thoroughly.

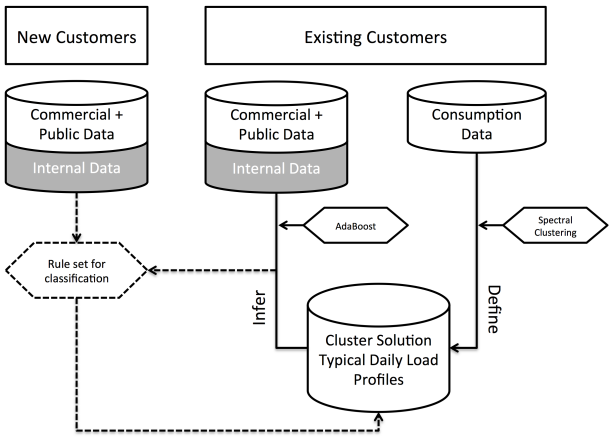


Fig. 1. An overview of the proposed methodology

A. Load Profiling

The initial pillar is a comprehensive load profiling step. This step aims to provide accurate clusters of consumers with

similar consumption patterns. Formally, using the notation of [11] we have a set $L = \{l^m, m = 1, ..., M\}$ of load diagrams for M customers that we wish to partition into K consumer clusters. For each of these clusters, we will then determine the representative load profile $T = \{t^k, k = 1, ..., K\}$. We follow the procedure introduced by Zhou [6] to determine our representative profiles. This implies we first correct our data and delete any possible outliers. Next, we aggregate our full dataset to an average daily 15 minute pattern (96 measurements per customer). This new and shorter time series is then rescaled so we confine ourselves to analyzing the shape of the load curve and not the absolute value of the load over the period. This approach of using *Typical Daily Profiles* is a common method within the industry and in literature to reduce data dimensionality, problem complexity and absolute load impact. This is achieved by first taking the mean values for any given quarter of an hour for every customer followed by rescaling every time series objects in the aim to only focus on consumption patterns and not absolute load. Time series consisiting only of 0-measurements are then removed. Afterwards, these daily patterns per customer are used as the basis for a clustering algorithm in order to attain the load profiles from the dataset. In [6] several clustering methodologies for load classification are proposed. With regard to choosing a clustering algorithm, literature suggests that Self Organizing Maps [22], k-means based approaches [23] or a combination of both [7] perform well. However, recent research [20], [24] suggests that a Spectral Clustering Algorithm (SCA) using weighted kernel principal component analysis (WKPCA) performs better for this type of application. To this end, we will utilize a spectral clustering approach in our methodology as well and we will compare it with a classical k-means clustering. However, unlike their methodology, we use the observed daily load profiles rather than model-based estimates. For a summary and formal definition of the SCA with WKPCA, we refer the reader to [4], [20].

This method offers two main advantages: (i) it is mainly based on the similarity matrix of the dataset where it needs the pairwise similarities of the time series, as a result it ignores the high-dimensionality problem; (ii) it can be used to cluster time series with arbitrary length with the condition that the similarity measures between them are well defined. The spectral clustering algorithm provides a powerful unsupervised tool to identify similar patterns across time-series data. Spectral clustering makes use of the eigenvalues of the similarity matrix of the time series data in order to achieve a dimension reduction (In essence, it is equivalent to running a linear principal component analysis in a high dimensional kernel space). Clustering is performed on the modified data by comparing pairwise similarities within this high-dimensional space. The spectral clustering algorithm has its roots in graph partitioning theory and it operates without making specific assumptions on the form of the clusters. This makes it specifically robust and also leads to good results, as the clustering itself happens in a high-dimensional space (like in the context of Support Vector Machines) and the clusters become linearly separable. It is argued that spectral clustering is, under certain conditions, equivalent to kernel-based k-means, which is a k-means clustering on a kernel-based PCA analysis [20].

procedure LOAD PROFILING
  Make n time series objects:
    Remove outliers and aggregate to daily patterns
    Check data quality
    Filter time series to remove missing values
  Normalize Time Series:
  Select Time Series (EANS) to be clustered
  Perform Spectral Clustering:
  for $ClusterSize$ = 2 to 10 **do**
    Execute Spectral Clustering on selected Time Series
    Calculate cluster membership
    $c_t \leftarrow$ total number of clusters
    for k = 1 to K **do**
      $k_i^{count} \leftarrow$ count of cluster members
    end for
    Calculate Davis-Bouldin index
    Calculate Dunn index
  end for
  $ClusterSize_{opt} \leftarrow$ best avg rank of Dunn and D-B
  Determine Representative Load Profile T
  Export optimal clusters
end procedure

Fig. 2. Load Profiling Methodology

One drawback of SCA is that the optimal number of clusters still has to be determined a priori. To analyze which number is optimal for our dataset, we use two internal validity constructs, the Davies-Bouldin-index [25] and the Dunn-index [26]. These two criteria aim to maximize internal cluster consistency and minimize overlap. We vary K from 2 to 10 for our algorithm, as most studies indicate that there are between 4-10 clusters [17], [20]. Algorithm 1 outlines this procedure. After we run the procedure, we have an optimal amount of clusters with their respective Daily Load Profiles. Figure 2 summarizes our approach.

B. Data Enrichment

In our second component we aim to enrich our internal customer data. This is a necessary step as our internal data set only contains contractual information on our customers as well as information on their credit-worthiness, their total consumption over the year and the commercial (NACE) code. Survey data was not available. Therefore it was necessary to extend our basetable with more elaborate information on the individuals themselves as well as information on their surroundings. To this end, we mine publicly or commercially available data on the customers, given their exact location, registration number, company name, etc. We collect this data in order to get a descriptive and complete profile of the end-users. To link our customers to open data, we follow five steps.

Locate:
a possible data source is identified

TABLE I
INTERNAL DATA TABLE

Variable	Source	Region	Completeness
Operational Segmentation	internal	Belgium	all
Volume Segmentation	internal	Belgium	all
AccountClass	internal	Belgium	all
Name	internal	Belgium	all
NACECode	internal	Belgium	all
Creditworthiness	internal	Belgium	all
PayMethod	internal	Belgium	all
EnergyManager	internal	Belgium	all
hasSmartTermostat	internal	Belgium	all
hasSmartEnergyBox	internal	Belgium	all
EAN	internal	Belgium	all
Address	internal	Belgium	all
PostalCode	internal	Belgium	all
internalDivision	internal	Belgium	all
hasPhotoVoltaic	internal	Belgium	all
GridOperator	internal	Belgium	all
ContractType	internal	Belgium	all
MeterType	internal	Belgium	all
GasConsumption	internal	Belgium	all

TABLE II
COMMERCIAL DATA TABLE

Variable	Source	Region	Completeness
Revenues	Graydon	Belgium	5290
Nbr of Employees	Graydon	Belgium	5290
Revenue / Nbr of Employees	Graydon	Belgium	5290

Evaluate:
: the available variables are evaluated on their ability to describe the entity to be researched (a customer) or the environment in which the entity operates

Export:
: the variables are exported or mined according to best practice techniques

Clean:
: the data is cleaned of missing values and it is checked for completeness

Link:
: the linking variable is identified and the newly acquired data is integrated into the full database

The result of this process is an extended database that includes, besides the internal company data, information on their municipality, commercially available company information and information their individual lot and building sizes. The commercial data is linked with the internal data by means of their VAT number. The information on the municipality is linked by the postal code and the cartographic data is linked through the exact address. An overview of these different data sources and their completeness level can be found in Tables I, II, III and IV.

TABLE III
MUNICIPAL DATA TABLE

Variable	Source	Region	Completeness
Net total taxable income	BEstat	Belgium	5653
Average net taxable income	BEstat	Belgium	5653
Median net taxable income	BEstat	Belgium	5653
Income Asymmetry	BEstat	Belgium	5653
Percentage Foreigners	BEstat	Belgium	5653
Birth Rate	BEstat	Flanders	3939
Death Rate	BEstat	Flanders	3939
Residents in 5-year age-groups	BEstat	Flanders	5653
Crime Rate	FedPol	Flanders	3836
Cultural Events	UiT-DB	Belgium	5653

TABLE IV
CARTOGRAPHIC DATA TABLE

Variable	Source	Region	Completeness
Lot Size	CADgis	Belgium	3070
# Buildings	CADgis	Belgium	3068
Building Size	CADgis	Belgium	3070
Useable Area	CADgis	Belgium	3070

C. Load Profile Prediction

The third and final pillar is combining the newly mined data with the results from the 'load profiling' as performed in step one. The main goal of this module is to ensure that we can identify a set of rules with which we can properly attribute a new customer, or a customer for which his Smart Meter-load profile is not readily available, to his specific segment. Using classification algorithms we can identify the relation between a customer's attributes and his load profile. Bearing in mind that our dependent variable is not binary in nature (unless we have a two-cluster solution), we will always have to either use a capable model or Round Robin Classification [27] also often called one-vs-all classification, which can be easily described as turning each level (n) of your dependent variable into a binary classifier and running the classification algorithm n times. The accuracy of these models is evaluated by looking at their individual accuracies. This makes it easy to compare different classification algorithms and also helps us in iteratively adding/removing variables from the extended database in order to maximize predictive performance. We focused on two popular and well-performing algorithms, as verified in numerous studies, to assess the performance of our model [9], [10], [28], [29]. Both algorithms use an ensemble of models to come up with their predictions.

1) Random Forests: Random Forests (RF) is an extension of decision tree algorithms [10]. Rather than estimating a single decision tree using all data to predict the class memberships, RF creates a number of different decision trees that are combined into one single estimate by using majority voting. The different trees are created by taking random samples on both the observations and the predictors. The result is that RF is very robust to outliers, overfitting and noise in the data. Furthermore, it is a computationally efficient methodology as

all trees can be estimated in parallel and as the sampling reduces the dimensions of the problem.

2) Stochastic Boosting: The Stochastic Boosting (SB) algorithm has the reputation of the best off-the-shelf classifier in the world and is an efficient and simple approach for building classification models [9]. Similar to RF, it creates multiple models for a given dataset. However, unlike RF, it does not create the models in parallel. Instead, it uses a sequential approach to create the classifiers. After every iteration the model checks which observations it was able to predict accurately. Those that were difficult to predict, get a higher weight (i.e., they are 'boosted') in the next iteration. This pushes the model to learn different types of rules, also for those observations that are not so easy to predict accurately. Boosting algorithms require some tuning and can fail to perform well if there is insufficient data due to sensitivity to noise. As a meta-learner algorithm, Boosting is capable to employ any simple learning algorithm to create multiple models.

D. Performance Evaluation

In order to assess the performance of our models, we opt to use the Area Under the receiver operator Curve (AUC) measure [30]. The rationale for this is that it is a more objective measure to evaluate the performance of a predictive model than accuracy (Percentage Correctly Classified, PCC). The reason is that the accuracy of a model crucially depends on the cut-off value that is applied to the posterior probabilities [31]. The AUC measure overcomes this by looking at all possible cut-off values. Theoretically it is defined in Equation 1 with TP as the True Positives, FN as the False Negatives, FP as the False Positives, TN as the True Negatives, P as the Positives and N as the Negatives. It is the area under the Receiver Operator Curve (ROC). This curve plots for every cut-off value between 0 and 1 the True Positive Rate (TP / P) and the False Positive Rate (FP / N). The area under this curve then approximates how well the model performs across all cut-off values. As both the True Positive Rate and the False Positive Rate are between 0 and 1, the maximum value for the AUC is 1. A random model should have an AUC value of 0.5 so it's value is only meaningful between 0.5 and 1. A higher score implies a higher True Positive Rate and therefore a better model.

$$AUC = \int_0^1 \frac{TP}{TP+FN} d\frac{FP}{FP+TN} = \int_0^1 \frac{TP}{P} d\frac{FP}{N} \quad (1)$$

To ensure we are not biased by our chosen training set, we perform five times two-fold crossvalidation (5x2 cv) in order to assess our final performance [32]. We assign the data randomly to either the training or validation set and perform the algorithm. At every replication we use the same split between training and validation. The reported AUCs are the averages across these models.

IV. Computational Results on Belgian Consumption Data

A. Implementation Details

All of our results were obtained using R v3.1.0 on a 3.2 GHz CPU running Windows with 32 GB of RAM. Our dataset contains 6975 customers of an energy provider in Belgium. For every customer we were given two years of smart meter measurements in 15-minute intervals. This leaves us with 70040 measurements for every customer. After our pre-processing stage, we ended up with 3068 customers for whom all relevant information was available. For our spectral clustering, we employed the *kernlab* package by [33]. We used the standard parameters with the exception of the number of random starts. This was experimentally changed to 1000 to generate the best results. We used Euclidean distances for our similarity measure. For our classification algorithms, we used the *ada* R package by [34] and the *RandomForest* package from [35]. In both cases we employed the standard parameters as suggested by the authors.

B. Cluster Quality

Table V reports the Dunn and Davis-Bouldin indices across the different cluster sizes. These indices are the average results after running the clustering 50 times. We can see that both indices find an optimum of 2 clusters for our SCA as we seek a minimum index for Davies-Bouldin and a maximum for Dunn.

However, in the two-cluster solution, the second cluster contains less than 0.1 % of the total observations. This cluster represents companies that are open all day, which heavily skews the results. We obtained similar results when looking at the three-cluster solution. In this case, one group contained over 80 % of the observations. Looking at our indices, It is noteworthy that for six clusters the indices reach values close to the optimum, indicating that a selection of six clusters could also be viable. When looking at the representative load profiles for the six cluster solutions, they are also easily interpretable. Therefore, we ultimately settled for the six cluster solutions. Theoretically, we could further break down the biggest cluster in this solution into smaller sub-clusters as we did find some meaningful ones. However, we chose to follow the results of our indices to stop at our optimal number. Table VII shows the number of customers per cluster and their respective percentage over the population. The robustness of this solution was also checked by making 50 bootstraps of the original data. In 47 cases the same results were found. In the other 3, the results were only slightly different. We employed the same methodology to compare our results with k-means. There we found an optimal cluster size of two using our indices as can be seen in Table VI. Again, these two clusters provided little to no meaning. Furthermore, it proved very difficult to detect the amount of clusters that represented meaningful load profiles using this approach.

Looking at Figure 3 you can see the TDP for all clusters using the SCA. We provide both the representative profile and the rolling box plot around our profile. This enables us to assess the spread around our centroid.. We did not plot all time series as this would give graphs that are too difficult to interpret. There are clear distinctions between each of the profiles that provide clearly separated segments. Load Profile 1 clearly contains companies that are less active during the day and start to consume energy very early in the morning. The

TABLE V
PERFORMANCE FOR SPECTRAL CLUSTERING

	2	3	4	5	6	7	8	9	10
Dunn	1.32	0.91	0.38	0.29	0.65	0.21	0.31	0.35	0.36
Davies-Bouldin	1.45	2.01	3.61	4.52	1.88	3.10	2.54	2.49	2.51

TABLE VI
PERFORMANCE FOR K-MEANS CLUSTERING

	2	3	4	5	6	7	8	9	10
Dunn	1.55	1.20	1.18	0.72	0.68	0.61	0.58	0.63	0.55
Davies-Bouldin	1.10	1.19	1.24	1.39	1.48	1.61	1.64	1.59	1.60

TABLE VII
NUMBER OF CUSTOMERS PER CLUSTER

	C 1	C 2	C 3	C 4	C 5	C 6
Amount	83	428	1828	81	460	188
%	2.7%	13.9%	59.6%	2.6%	15.0%	6.1%

TABLE VIII
CROSSVALIDATED AUCS WITH SCA

Used Data	SB	RF
Internal	0.754	0.733
Internal + Commercial	0.767	0.765
Internal + Mun	0.747	0.696
Internal + Location	0.762	0.743
Internal + Commercial + Mun	0.754	0.732
Internal + Commercial + Location	**0.778**	**0.763**
Internal + Location + Mun	0.748	0.728
Internal + Commercial + Mun + Location	0.765	0.740
Commercial + Location + Mun	0.687	0.649
Commercial + Location	0.699	0.639

TABLE IX
CROSSVALIDATED AUCS WITH K-MEANS

Used Data	SB	RF
Internal	0.730	0.693
Internal + Commercial	0.734	0.720
Internal + Mun	0.705	0.663
Internal + Location	0.733	0.704
Internal + Commercial + Mun	0.720	0.682
Internal + Commercial + Location	**0.737**	**0.730**
Internal + Location + Mun	0.748	0.728
Internal + Commercial + Mun + Location	0.712	0.693
Commercial + Location + Mun	0.641	0.610
Commercial + Location	0.640	0.595

third profile is a company mainly active during typical business hours. When compared to the fifth profile there is a dip during lunch times and a quicker descent in the evening. Overall these clusters represent interesting segments for incentive schemes.

C. Basic Model

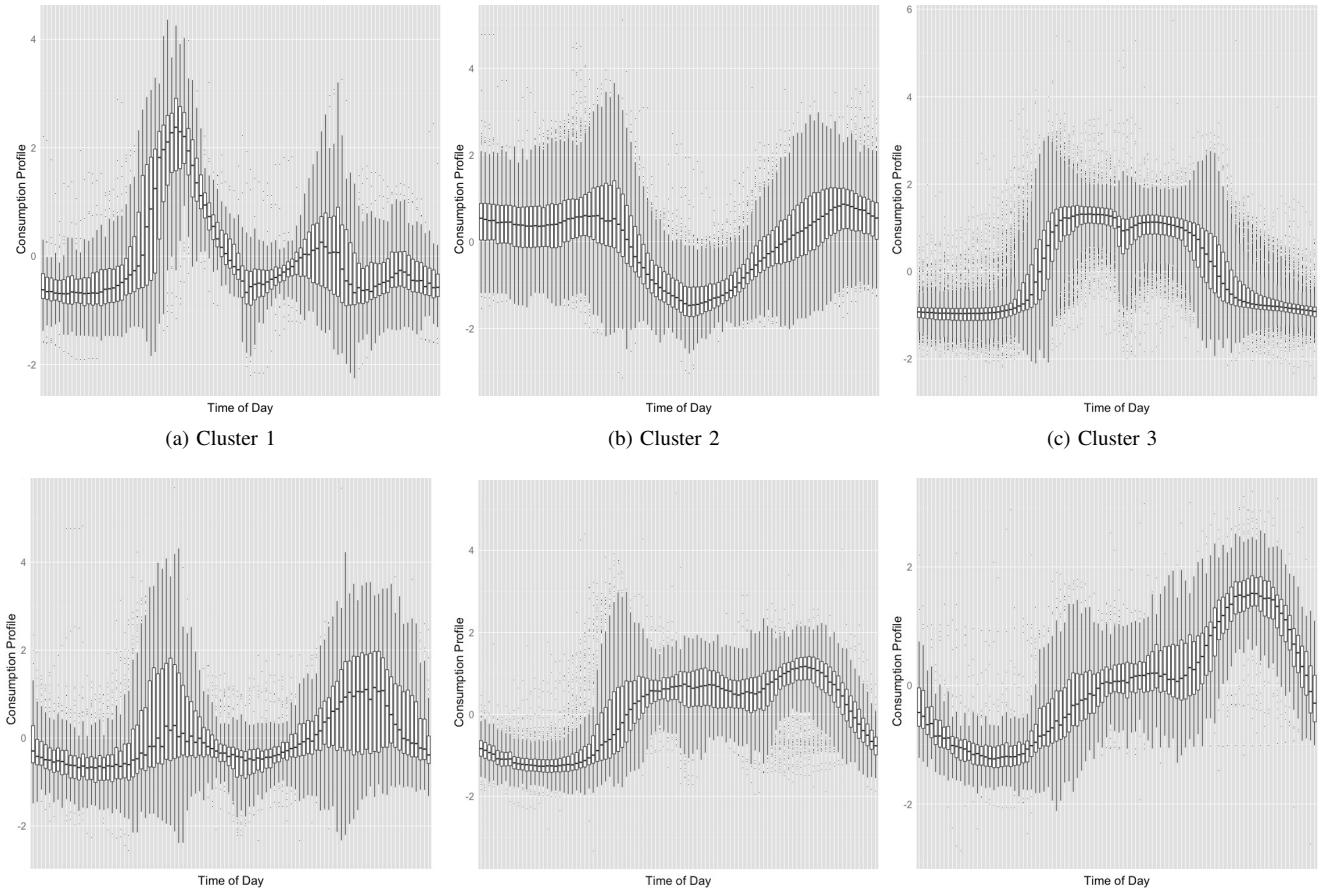
Table VIII reports the weighted crossvalidated AUCs for all models. The weights are calculated based on the size of the cluster relative to the total amount of observations. In our benchmark model we only use the internal company data to classify the customers to their respective load profiles. We can see that our benchmark model already performs quite well. This implies that the contractual and basic customer data alone is already able to predict with relative good accuracy what kind of consumption pattern a customer will have. If we look at the variable importances of this model in Figure 4a, it is clear that the commercial codes have the biggest influence on the model performance. This is in line with the results of Figueiredo [7] that also identified commercial codes as a big factor. It is noteworthy that credit risk is another important factor for identifying the consumption behavior as well as the area in which you are operating. This area is identified through the different electricity grid variables.

D. Added Value of Commercial Data

Looking at the addition of the commercial data, we can see that this has a positive effect on the predictive performance. This commercial data represents the number of employees and the total turnover of the company. These give an indication of the size of the company and it makes sense that this can help in predicting the consumption behavior. The addition of this data to the internal dataset is a logical extension and should be considered as extended internal information. In practice, it should be no problem to gather this data from the company itself, thereby eliminating the search procedure.

E. Added Value of Municipal Data

The municipal data reduces the predictive power rather than improving it. We see this negative effect every time we use it in combination with other data sources and across the different algorithms. Only in combination with the cartographic variables, the effect is not strongly negative. Looking at the variable importances of the open data model including municipal variables in Figure 4b, it is obvious they are dominating the other variables. As they have a negative impact on total performance, this shows that it is not wise to include them in the final model. The reason for this negative impact on



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

(d) Cluster 4 (e) Cluster 5 (f) Cluster 6

Fig. 3. Typical Load Profiles for all clusters

#	Predictor
1	NACE = 43320
2	NACE = 45113
3	Payer Risk Grade Bad
4	NACE = 25620
5	NACE = 56101
6	Operational Segmentation
7	NACE = 47716
8	PrivateGridOperator
9	IndustrialGridOperator
10	Volume Segmentation

(a) Internal Data

#	Predictor
1	TaxableNetIncome
2	Deaths
3	% residents in 25-29
4	% residents in 0-4
5	% residents in 80-84
6	% residents in 15-19
7	% residents in 10-14
8	% residents in 70-74
9	% residents in 60-64
10	% residents in 55-59

(b) Open data

#	Predictor
1	LotSize
2	Turnover
3	Ratio
4	usableArea
5	NrOfEmployees
6	buildingSize

(c) Open data (No Municipal)

#	Predictor
1	usableArea
2	Turnover
3	NACE = 56101
4	Ratio
5	LotSize
6	buildingSize
7	Grid Operator
8	numberOfHardLines
9	NACE = 68321
10	Volume Segmentation

(d) Final Model

Fig. 4. Top 10 Predictors for the different models

the total model performance can likely be found in the lack of differentiating power. As these data can be the same for any kind of company in the same area, they cannot easily discriminate between them.

F. Added Value of Cartographic Data

The value of the cartographic data is very good. All models including these data, except when combined with municipal data, perform better than those without. This seems to make sense as the area that a company uses as well as their number of buildings and the building size heavily impact

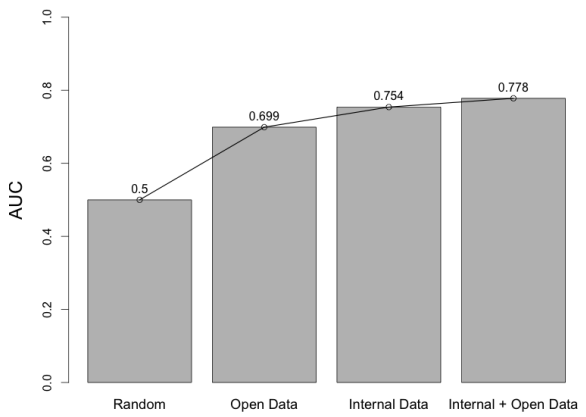


Fig. 5. Predictive Power of the different models

energy consumption. We can see this again when looking to the variable importances. Every time the cartographic data is included, they are among the top predictors, except when they are dominated by the municipal variables.

G. Model Value without Internal Data

We also tested whether it could be possible to predict the consumption patterns when no information with regard to the consumers is available from the company itself. The results indicate that when only relying on open data, the predictive power of the model is reduced. However, when removing the municipal data, we can see that the SB algorithm is able to produce an AUC of 0.7. This implies that the open data model already performs better in determining the correct consumption pattern than a random model. Even when no contractual or commercial code information is available, energy producers can already assign customers to pre-made load profiles. The combination of turnover and lot size are the best predictors in this case as seen in Figure 4c.

H. Final Model

The best performing model is the one that includes all internal data as well as commercial and cartographic data. This model outperforms all others and is the best model for assigning new customers to a load profile. With an AUC of 0.778, it is highly effective. When running this model, the most important explanatory variables are the commercial codes in combination with the usable area, lot size, building size and turnover of the company. This can be observed in Figure 4d. Finally, Figure 5 compares the performance of the different models and highlights the strength of the chosen model.

We also replaced our spectral clusters by the k-means clusters and ran our AUC analysis again. As can be seen from Table IX the results are comparable to those with SCA, although the SCA clusters generate a higher AUC on all models. This indicates applying k-means is possible, but it comes at a cost. The advantage of k-means is that it is more scalable. However, it deteriorates the interpretability of the profiles and it is more difficult to derive meaningful links with other variables.

V. Conclusion

We have shown that it is not necessary to have smart meter data for all your customers in order to classify them into a specific load profile. Using internal company data alone, our model is able to accurately predict the load profiles. This is interesting in the design of incentive schemes for different types of customers. As smart meter data is not available for all customers, this enables a company to propose adequate schemes. Furthermore, extending internal company data with open data on the consumers increases the predictive performance of these consumption patterns. Especially commercial data about these companies as well as cartographic data are effective in improving these models. Our analysis also shows that only having this information, without any other internal data, is enough to provide a basic model that classifies customers into their respective patterns. Having the ability to classify customers without smart meters into their likely load profiles is not only interesting from a customer relationship management point of view, it is also meaningful for the daily operations of electricity producers. Being able to predict the consumption behavior of the different customers will enable the producers to better balance the electricity grid.

Toward future research, it could be interesting to assess the value of weekly or monthly profiles rather than using the typically used TDP for generating our load profile clusters. Furthermore, it would be meaningful to assess how the top predictors change when looking at seasonal TDPs. Especially in corporate environments, these profiles could highlight extra information. However, processing the increased amount of measurements is currently difficult as it requires very high computation times and memory consumption. Finding efficient algorithms and structures for doing this should be beneficial. A first paper with these types of profiles was undertaken by Mutanen [14]. However, no link to any external data was included here.

Acknowledgment

We would like to thank Electrabel - GDF SUEZ for its financial and technical support. We also would like to thank in particular Mr. Pierre Garbit and Dr. Marcelo Espinoza for their valuable insights and input as well as Dushyant Khosla and Evangelos Lemonis - members of the student team from the Master of Science in Marketing Analysis - who collaborated on this project.

References

[1] M. Negnevitsky and K. Wong, "Demand-Side Management Evaluation Tool," *IEEE Transactions on Power Systems*, vol. 30, no. 1, pp. 212–222, Jan. 2015.

[2] A. Albert and R. Rajagopal, "Smart Meter Driven Segmentation: What Your Consumption Says About You," *Ieee Transactions on Power Systems*, vol. 28, no. 4, pp. 4019–4030, Nov. 2013.

[3] N. Mahmoudi-Kohan, M. P. Moghaddam, and M. K. Sheikh-El-Eslami, "An annual framework for clustering-based pricing for an electricity retailer," *Electric Power Systems Research*, vol. 80, no. 9, pp. 1042–1048, Sep. 2010.

[4] C. Alzate and J. Suykens, "Multiway Spectral Clustering with Out-of-Sample Extensions through Weighted Kernel PCA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 335–347, Feb. 2010.

[5] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern grouping," *Energy*, vol. 42, no. 1, pp. 68–80, Jun. 2012.
[6] K.-l. Zhou, S.-l. Yang, and C. Shen, "A review of electric load classification in smart grid environment," *Renewable & Sustainable Energy Reviews*, vol. 24, pp. 103–110, Aug. 2013.
[7] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, "An electric energy consumer characterization framework based on data mining techniques," *Ieee Transactions on Power Systems*, vol. 20, no. 2, pp. 596–602, May 2005.
[8] J. Saarenpaa, M. Kolehmainen, M. Mononen, and H. Niska, "A data mining approach for producing small area statistics-based load profiles for distribution network planning," in *Industrial Technology (ICIT), 2015 IEEE International Conference on*, March 2015, pp. 1236–1240.
[9] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, Feb. 2002.
[10] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
[11] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. M. Toader, "Customer characterization options for improving the tariff offer," *Ieee Transactions on Power Systems*, vol. 18, no. 1, pp. 381–387, Feb. 2003.
[12] L. A. Greening, "Demand response resources: Who is responsible for implementation in a deregulated market?" *Energy*, vol. 35, no. 4, pp. 1518–1525, Apr. 2010.
[13] J. Torriti, M. G. Hassan, and M. Leach, "Demand response experience in Europe: Policies, programmes and implementation," *Energy*, vol. 35, no. 4, pp. 1575–1583, Apr. 2010.
[14] A. Mutanen, M. Ruska, S. Repo, and P. Jarventausta, "Customer Classification and Load Profiling Method for Distribution Systems," *IEEE Transactions on Power Delivery*, vol. 26, no. 3, pp. 1755–1763, Jul. 2011.
[15] G. Chicco, R. Napoli, and F. Piglione, "Comparisons among clustering techniques for electricity customer classification," *Ieee Transactions on Power Systems*, vol. 21, no. 2, pp. 933–940, May 2006.
[16] X. Li, X. JIANG, J. QIAN, H. CHEN, J. SONG, and L. HUANG, "A Classifying and Synthesizing Method of Power Consumer Industry Based on the Daily Load Profile [J]," *Automation of Electric Power Systems*, vol. 10, p. 012, 2010.
[17] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader, "Load pattern-based classification of electricity customers," *IEEE Transactions on Power Systems*, vol. 19, no. 2, pp. 1232–1239, May 2004.
[18] S. Verdu, M. Garcia, F. Franco, N. Encinas, A. Marin, A. Molina, and E. Lazaro, "Characterization and identification of electrical customers through the use of self-organizing maps and daily load parameters," in *Power Systems Conference and Exposition, 2004. IEEE PES*, Oct. 2004, pp. 899–906 vol.2.
[19] D. Gerbec, S. Gasperic, I. Smon, and F. Gubina, "Allocation of the load profiles to consumers using probabilistic neural networks," *IEEE Transactions on Power Systems*, vol. 20, no. 2, pp. 548–555, May 2005.
[20] C. Alzate, M. Espinoza, B. De Moor, and J. A. K. Suykens, "Identifying Customer Profiles in Power Load Time Series Using Spectral Clustering," in *Artificial Neural Networks - Icann 2009, Pt Ii*, C. Alippi, M. Polycarppou, C. Panayiotou, and G. Ellinas, Eds. Berlin: Springer-Verlag Berlin, 2009, vol. 5769, pp. 315–324.
[21] G. Tsekouras, N. Hatziargyriou, and E. Dialynas, "Two-Stage Pattern Recognition of Load Curves for Classification of Electricity Customers," *IEEE Transactions on Power Systems*, vol. 22, no. 3, pp. 1120–1128, Aug. 2007.
[22] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, Sep. 1990.
[23] J. MacQueen and others, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1. California, USA, 1967, pp. 281–297.
[24] M. Espinoza, C. Joye, R. Belmans, and B. De Moor, "Short-term load forecasting, profile identification, and customer segmentation: A methodology based on periodic time series," *Ieee Transactions on Power Systems*, vol. 20, no. 3, pp. 1622–1630, Aug. 2005.
[25] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.
[26] 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, Jan. 1973.
[27] J. Frnkranz, "Round Robin Classification," *J. Mach. Learn. Res.*, vol. 2, pp. 721–747, Mar. 2002.
[28] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning*, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 161–168.
[29] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," *Expert Systems with Applications*, vol. 36, no. 3, pp. 4626–4636, 2009.
[30] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve." *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982.
[31] D. Thorleuchter and D. Van den Poel, "Predicting e-commerce company success by mining the text of its publicly-accessible website," *Expert Systems with Applications*, vol. 39, no. 17, pp. 13 026–13 034, Dec. 2012.
[32] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998.
[33] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "kernlab - An S4 Package for Kernel Methods in R," *Journal of Statistical Software*, vol. 11, no. 9, pp. 1–20, 2004.
[34] M. Culp, K. Johnson, and G. Michailides, "ada: An R Package for Stochastic Boosting," *Journal of Statistical Software*, vol. 17, no. 2, pp. 1–27, 2006.
[35] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

Dauwe Vercamer was born in 1987. He received his M.Sc. in Commercial Engineering from Ghent University in 2010. After a brief time as a supply chain analyst at Accenture, he is currently working toward his PhD at the department of marketing at Ghent University. His focus is mainly on methodologies and applications in predictive and prescriptive analytics.

Bram Steurtewagen received his M.Sc. degree in Commercial Engineering (2013) and his M.Sc. degree in Marketing Analytics (2014) from Ghent University in Belgium. Since then, he has been pursuing a PhD in Marketing Analytics at the Faculty of Economics and Business Adminstration of Ghent University.

Dirk Van den Poel (SM'10) was born in 1969. He obtained his M.Sc. and PhD at KULeuven in Belgium. He is now Full Professor of Business Analytics / Big Data at Ghent University (Belgium). He teaches courses such as: Statistical Computing, Big Data, Analytical CRM, Marketing Information Systems/Database Marketing. He co-founded the advanced M.Sc. in Marketing Analysis, the first (predictive) analytics master program in the world as well as the M.Sc. in Statistical Data Analysis. His major research interests are in the field of analytical CRM: customer acquisition, churn, upsell/cross-sell, and win-back modeling. His methodological interests include ensemble classification and big data analytics. He has been the supervisor of 12+ PhDs and has co-authored over 70 international peer-reviewed publications.

Frank Vermeulen obtained an electrical engineering degree from Ghent University in 1989 and a M.Sc. in Control and Information Technology from UMIST Manchester in 1991. He obtained an MBA from Leuven University in 1999. He started his career at the Flemish Institute of Technology working on the design and construction of hydrogen fuelled and hybrid vehicle drive trains. He moved to Elia, the Belgian Power Transmission System Operator to take up several functions in transmission grid planning, environmental impact assessments and economic studies for grid tariff design. in 2008 he became responsible for the Economic Studies department at GDF-SUEZ dealing with the economic analysis of wholesale and downstream gas and power business activities in Central Western Europe. His domains of interest are power system design, energy economics and business development.