energies

MDPI

Article

Definition of Residential Power Load Profiles Clusters Using Machine Learning and Spatial Analysis

Mario Flor *, Sergio Herraiz and Ivan Contreras

Institut d'Informatica i Applicacions, Universitat de Girona, 17003 Girona, Spain; sergio.herraiz@udg.edu (S.H.); ivan.contreras@udg.edu (I.C.)
* Correspondence: mario.flor@udg.edu

Abstract: This study presents a novel approach for discovering actionable knowledge and exploring data-based models from data recorded by household smart meters. The proposed framework is supported by a machine learning architecture based on the application of data mining methods and spatial analysis to extract temporal and spatial restricted clusters of characteristic monthly electricity load profiles. In addition, it uses these clusters to perform short-term load forecasting (1 week) using recurrent neural networks. The approach analyses a database with measurements of 1000 smart meters gathered during 4 years in Guayaquil, Ecuador. Results of the proposed methodology led us to obtain a precise and efficient stratification of typical consumption patterns and to extract neighbour information to improve the performance of residential energy consumption forecasting.

Keywords: energy consumption clustering; spatial analysis; machine learning; recurrent neural network; smart meter; load profiles forecasting

check for updates

Citation: Flor, M.; Herraiz, S.; Contreras, I. Definition of Residential Power Load Profiles Clusters Using Machine Learning and Spatial Analysis. *Energies* **2021**, *14*, 6565. https://doi.org/10.3390/en14206565

Academic Editor: Dimitrios Katsaprakakis

Received: 8 September 2021
Accepted: 6 October 2021
Published: 12 October 2021

1. Introduction

Smart meters (SMs) provide a granularity and precision that make it possible to overcome classical methods such as the definition of a single typical consumption curve for residential or commercial clients. In addition, this rich source of data for energy consumption analysis shows that conventional methods are unable to cope with such volume or speed. This situation, together with the increasing availability of more powerful computers, has promoted the use of intelligent techniques to study patterns in time series data. The application of machine learning (ML) to smart grid data enables designers to address problems in a more tailored way and seek more accurate results.

One of the most topical research areas taking advantage of the aforementioned premise is the application of clustering techniques to determine energy consumption efficiency. Clustering can be used on a daily electricity demand time series to group identical profiles to reveal the most typical load profiles [1,2]. An example is the work proposed by Lavin et al. in [3] that applied the k-means heuristic partitioning method to compare the structural similarity of daily time series with a focus on finding clients with similar energy consumption profiles. Clustering can extract consumption patterns at different time periods (e.g., monthly [4], seasonal [5], or annual). [6,7]. This reveals important information about households consumption habits and their relationship with time variables.

Energy consumption habits can be affected by elements, such as geographical barriers (rivers, hills), political boundaries (district divisions), commercial areas (free trade zones or industrial parks and harbours), soil characteristics or orientation of buildings among others. Geographic information systems (GIS) collect these spatial elements and their relationships to perform a spatial-temporal analysis that may prove extremely useful for spatial clustering. Spatial-temporal analyses are also used to forecast energy demand; for instance, authors in [8] provided a framework that facilitates the exploitation of low-dimensional structures to govern the interactions between the surrounding residential SM users. In [9], a k-nearest Vector Autoregressive framework with exogenous input for models with spatial-temporal variation of electricity consumption in individual household load forecasting was proposed. Another example is the work presented in [10], where authors studied how an electrical load is distributed in a city using area-independent agents and the relationships among neighbouring areas. In [11], the authors applied a temporal and spatial analysis to study the evolution patterns of alternative energies and improve the planning and construction of energy systems by a cellular automata model. In [12], load forecasting is studied using spatial regression to determine the probability of rural regions becoming urban areas as part of urban sprawl by spatially relating installed load and socioeconomic variables distributed in the study area.

Several studies have shown that forecasting independent residential loads is more challenging than forecasting commercial or aggregated loads [13,14]. The main reason of this larger complexity is an increment in the variability of the load profile. This is normally produced by the use of household appliances that generate significant fluctuations in the consumption patterns. These fluctuations are often unpredictable due to the dynamic nature of the behaviour of the household residents. Furthermore, there are studies that show that the hybridisation of clustering with forecasting techniques improves prediction performance [15,16]. Recent works on training artificial neural networks (ANNs) with residential SM data for energy consumption predictions can be found in the literature. ANN methodologies vary from classic implementations [17] to more complex approaches, such as the recurrent neural network (RNN) with long short-term memory (LSTM) architecture [18] or restricted Boltzmann machines [19,20].

This study analyses data from residential SMs by applying ML techniques to extract knowledge and build prediction models. First, we applied soft dynamic time warping [21] clustering methodology to find sets of users with similar monthly load profiles. Then, we used the clusters in a second process where a spatial-temporal analysis was applied to find sets of users adjusted to the particular reality of the geographical zones. Finally, users in the same geographical zone with similar energy consumption patterns were used in conjunction with a ML method to forecast future energy consumption. Hence, this study yielded two important contributions. On the one hand, it applied a temporal and a constrained spatial analysis to a large dataset of residential SMs, proving that a spatial neighbourhood is a significant source of information that can improve decision support and predictability. On the other hand, it supported our experiments with a novel hybrid ML framework that applied a constrained spatial clustering technique together with LSTM architecture. Although previous studies that have applied methodologies of time series clustering and spatial analysis, to our knowledge, there is no single study that incorporates time series clustering using soft dynamic time warping with a spatial-temporal analysis to define load profiles in specific geographic areas and then uses those profiles to feed a LSTM-based neural network to forecast energy consumption estimates.

The identification of a geographic area with characteristic behaviours can provide an accurate and updateable feedback for advanced technical analysis of utilities and regulators, including planning and decision making for activities such as a deeper understanding of electricity demand [22,23], analysing criteria for establishing dynamic tariffs [24], setting up tariff adjustment mechanisms for energy efficiency [25] and optimising energy demands [6,26]. Profiling and forecasting of energy consumption are also necessary for optimising local energy systems and energy communities [27,28]. In this regard, the E-LAND Horizon 2020 project is developing a set of software tools, including an energy forecaster, to support energy management in communities [29].

2. Materials and Methods

The general methodology proposed in this study is divided into four main steps:

- Data collection, pre-processing and time series generation,
- Clustering time series generation,
- Spatial-temporal analysis, and
- Applications for energy consumption forecasting.

First, we apply a series of pre-processing procedures to obtain a clean, complete and consistent dataset to generate a time series representing a user's monthly load profile. Secondly, we deploy a classification methodology to cluster the load profile time series according to characteristic monthly behaviour. We then perform a spatial-temporal analysis to extract spatial clusters restricted in nearby areas. Finally, to demonstrate the benefits of obtaining information on consumer behaviour in spatial areas, we model the extracted knowledge using a RNN to forecast the week-ahead hourly energy consumption.

To improve the readability and follow up of this study, Figure 1 shows the phases at each step.

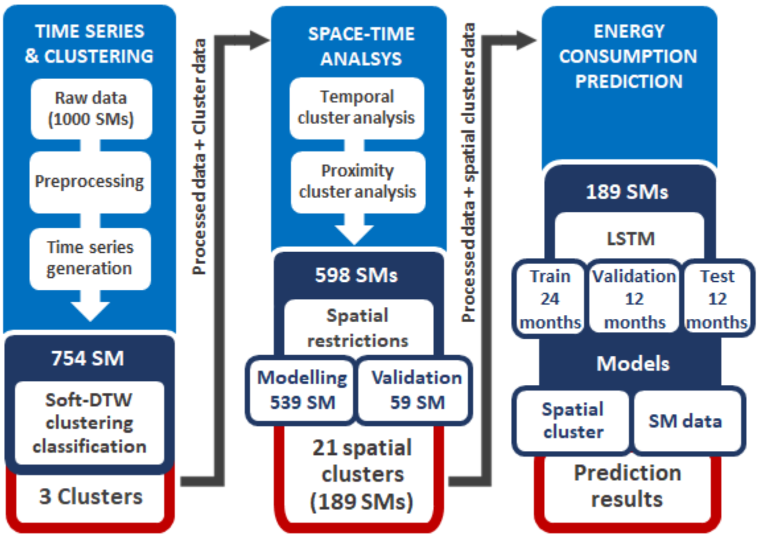

Figure 1. Summary of the methodology and filtered dataset obtained at each stage.

2.1. Data Pre-Processing and Time Series Generation

Data were gathered from a sample of 1000 residential SMs—one for each house—over 4 years in the city of Guayaquil, Ecuador. Measurements are taken every 15 min, which generated a database with approximately 130 million records. The collected measurements include the following variables:

- Geographic position of the SMs,
- Timestamp with the date and time of measurement,
- Customer code to identify the client (anonymised data), and
- Active power (kW).

Although our database initially had 1000 SMs, the percentage of available SMs with respect to the total number of meters in the city of Guayaquil is significantly low. The city has a very high concentration of electricity consumers with the number reaching more than 700,000 in an urbanised area of approximately 190 km^2. Therefore, the sample is quite dispersed, mainly due to the existence of old electromechanical meters that had not yet been replaced. Therefore, our initial data set was strictly analysed through different processes to obtain a final dataset focused on validating the objectives of this study.

A series of data pre-processing procedures was applied to obtain a consistent dataset. First, an exploratory analysis pointed out missing active power values and outliers on the gathered SM information. Missing data (fails in the SMs or in the data hub transmitters) was found either in short periods of less than or equal to one hour or in longer periods (weeks), usually because some properties were holiday rental houses and the SMs aere turned off when they are uninhabited. Thus, a missing value between two valid measurements is imputed by linear interpolation and, in cases where up to 96 measurements (1 day) are missing, these values are imputed using the corresponding measurements of the previous week. After processing the missing data, an outlier detection methodology was applied to detect unusual values and replace them using the earlier imputation mechanism. Then, we extracted time series to comply with the purposes of this study. First, time series were simplified by resampling to hourly measurements with the average active power value in this period. Secondly, since long periods of no consumption could represent a change of tenant or owner, to ensure that we were always analysing the same customer, only time series with the consecutive and complete measurements of at least 10 months of the year were selected.

To obtain comparable energy consumption behaviour profiles, a normalisation procedure was applied to time series in the range of 0 to 1. Furthermore, characteristic time series representing typical weekly behaviour for each month of the year (e.g., a time series representing a characteristic week of January) were generated by averaging the value of active energy in each hour of a day (24 measurements) for each particular day of the week (Monday–Sunday). After all the aforementioned pre-processing procedures, summarised in Figure 2, we hada total of 754 SMs which representeds a database of approximately 100 million records.

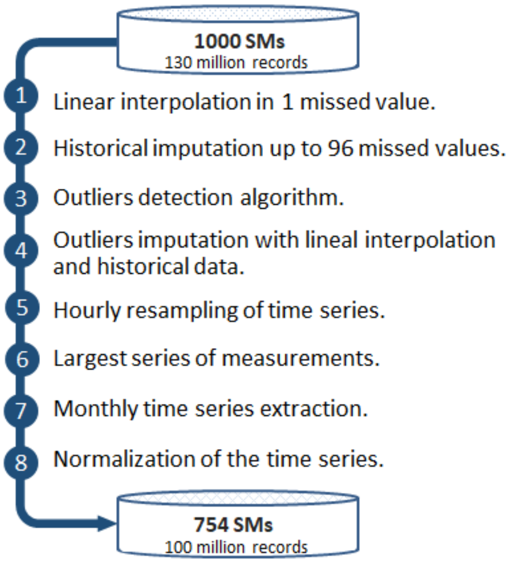
1000 SMs
130 million records
1 Linear interpolation in 1 missed value.
2 Historical imputation up to 96 missed values.
3 Outliers detection algorithm.
4 Outliers imputation with lineal interpolation and historical data.
5 Hourly resampling of time series.
6 Largest series of measurements.
7 Monthly time series extraction.
8 Normalization of the time series.
754 SMs
100 million records

Figure 2. Flow diagram showing raw data moving to prepared data.

2.2. *Time Series Clustering*

Analysing and extracting information from the SM time series data was complex. A set of cluster analysis techniques were used to evaluate their performance and understand the macroscopic structure and relations among the analysed time series. Four different clustering methodologies were evaluated. First, a classic k-means based on Euclidean distance had been implemented [3]. Euclidean distance is the most commonly used metric for measuring similarities between profiles in clusters [30]. However, it can generate considerable errors when calculating the distance between time series [31]. To deal with this possible issue, specific time series metrics were also implemented. Dynamic time warping (DTW) allowed the finding of the minimum distance between profiles by shifting on the time axis, which made it possible to group profiles with similar shapes regardless of their temporality. In addition, a gradient-based version of DTW was implemented. The metric is known as soft-DTW [21] and is a differentiable loss function that depends on a hyperparameter that controls the smoothness of the resulting metric. On the other hand, the k-shape clustering methodology [32] was also implemented. K-shape is based on the computation of cluster centroids and cross-correlation measurements with the cluster time series to preserve the shapes of the time series sequences.

This stage includes a clustering validation method; namely, the Silhouette coefficient [33], which provides a representation of how well each piece of data lies within its cluster. The values of the index are in the range of -1 and 1. Values close to 1 indicate that the instances are well classified, values close to -1 indicate a misclassification and values close to 0 indicate that the instances fall between two natural clusters.

Finally, the last step of this stage was to classify (type 1 to n) each SM according to its memberships in a characteristic cluster.

2.3. Spatial-Temporal Analysis

At this stage, SMs are georeferenced into a geographic information system (GIS) to perform a spatial-temporal analysis that aims to determine if similarities in energy consumption patterns are related to their specific location or to the spatial-temporal proximity of other SMs. The hybrid analysis combining clustering and spatial-temporal analysis is one of the highlighted contributions of this study, which models our problem geographically and then allow us to explore, interpret and detect important patterns hidden in the dataset.

The temporal analysis is concerned with the variability of the SM membership to the generated clusters and the spatial analysis defines areas that belong to the same type of behaviour using spatially restricted clustering by applying the minimum spanning tree [34]. Previously, a proximity analysis was performed to discard isolated SMs.

Thus, this analysis evaluated the hypothesis that the energy consumption behaviour of users within the same geographic area is better represented by also taking into consideration their nearest neighbours. The definition of a geographical area in this study involves SMs located consecutively without any one belonging to another behaviour type or having a significant spatial element between them, such as a river or non-residential zone.

To validate the generated spatial clusters, we had randomly selected and separated 10% of the SMs. The validation process consisted of evaluating whether the SMs in the validation set were located in areas with same type of behaviour as defined by spatial clustering.

2.4. Forecasting of Energy Consumption

Here, we validate the contribution of this study using the spatial clusters generated by the spatial and temporal analysis outcomes as a component of a recurrent neural network (RNN). Specifically a long short-term memory (LSTM) [35] configuration was implemented. Although RNNs exhibit a superior ability to model sequences [36], they suffer from so-called gradient fading during the backpropagation process explained in [37,38], so they are unable to learn long-term dependencies; that is, the relationship between entities that are separated by several steps. Hochreiter demonstrates in his article [37] that when neural networks have multiple steps, the error gradient will decrease exponentially with each step in the backpropagation process, so the training of a basic RNN with a long-term dependency becomes very slow and does not fit properly. To solve this problem, Hochreiter and Schmidhuber designed a special type of recurrent neural networks called long short-term memory (LSTM) networks [35]. LSTMs, like RNNs, have a chain-like structure, but instead of having a single activation function in their memory cell, LSTM networks have three structures called gates (forgetting, input and output), through which information can be removed or added to a cell state, which is like a conveyor belt that runs directly through each memory cell with interactions in each of them that do not affect it exponentially.

The LSTM network was applied by two different approaches to compare the resulting predictions. In the first approach, a coded LSTM network used only measurements from the same SM in a univariate analysis. In the second approach, a multivariate analysis was implemented in the LSTM network that included measurements from 5 SMs within the same spatial cluster, including the SM to be predicted. Considering that we had information for 4 years, the measurements for the years 2014–2015 (50% of the data) were defined to execute the training. To adjustment the hyperparameters, the data from 2016 (25%) was used as a validation set, and the data for 2017 (25%) was initially reserved as a test set. As more LSTM hidden layers were added, the network was able to infer more complex behaviour in our time series and increase the accuracy of the prediction, so two hidden layers that assumed hourly and daily behaviour were used in our model. Finally, an output layer with 168 neurons was included. The results of this output layer matched the predicted values for 24 hours a day for a week. In addition, a neuron dropout layer was alternated between each LSTM layer to speed up learning and avoid over fitting. This consisted of updating only a percentage of the neuron weights in the iterations, while the rest remained constant. In our case, a 20% dropout was applied to each layer. Additionally, the setup included an Adam optimiser and a root mean square error (RMSE) loss function. Neural network training was established with a maximum of 100 epochs with early stopping if the RMSE did not improve in 5 epochs (see Figure 3).

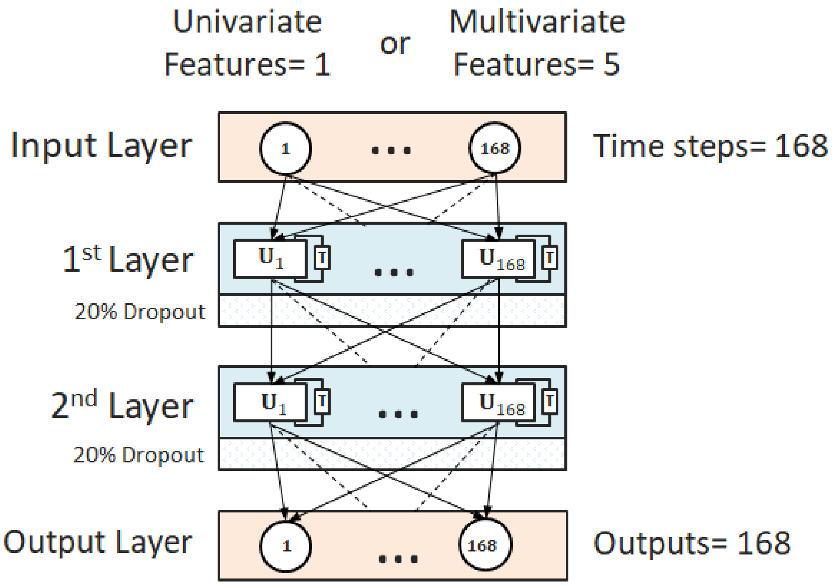

Figure 3. The proposed deep LSTM network with 2 LSTM layers with dropout layers. For clarity, the temporal recurrent structure is not shown.

The performance of the forecasting model was measured using root mean square error (RMSE) and symmetric mean absolute percentage error (sMAPE) [39]. The metric sMAPE is defined as:

$$sMAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|y_t - \hat{y}_t|}{(|y_t| + |\hat{y}_t|)/2} \tag{1}$$

where y_t is the actual value and $\hat{y}_t$ the forecast value.

3. Results

Results were analysed using Python software while TSLearn library was used for time series cluster generation [40]. ESRI® ArcGIS Pro software, a Geographic Information System (GIS), was used for spatially restricted cluster generation.

3.1. Time Series Clustering

First, we present the results of the silhouette validation method for the combination of each of the clustering methods (k-means, k-shape, DTW and soft-DTW) with a different number of target clusters ($k = 3, 4, 5$) and methodology hyperparameters (soft-DTW with $\gamma = 0.5, 1, 2$). The results are in Table 1, which presents a ranked classification of the configurations. The cluster with $k = 3$ was the best performing configuration among all the methodologies, achieving a better differentiation between the clusters. Since soft-DTW ranked in the first positions, it was therefore, the clustering method that provided the best results, so the results of soft-DTWs with hyperparameters $k = 3$ and $\gamma = 1$ are used in the remainder sections.

Table 1. Silhouette coefficient results for the clustering methodologies analysed. High value indicates that the object is well matched to its own cluster.

Cluster Method	k	Silhouette Index
Soft-DTW ($\gamma = 1$)	3	0.5775
Soft-DTW ($\gamma = 2$)	3	0.5297
Soft-DTW ($\gamma = 2$)	4	0.5280
DTW	3	0.5234
K-Shape	3	0.5213
Soft-DTW ($\gamma = 1$)	4	0.5187
Euclidean	3	0.5132
K-Shape	5	0.5119
Soft-DTW ($\gamma = 0.5$)	3	0.5083
K-Shape	4	0.4907
Soft-DTW ($\gamma = 2$)	5	0.4479
DTW	4	0.4235
Soft-DTW ($\gamma = 1$)	5	0.3714
Soft-DTW ($\gamma = 0.5$)	5	0.3675
Soft-DTW ($\gamma = 0.5$)	4	0.3617
Euclidean	4	0.3341
Euclidean	5	0.3023
DTW	5	0.2755

Figure 4 shows the different profiles obtained from the application of soft-DTW with $k = 3$ and γ = 1.

Subsequently, a temporal analysis of the extracted clusters was performed to explore the variability of cluster memberships during the 12 months of the year. A summary of the results are presented in Table 2. It can be seen that 594 SMs had 12 months of complete information and a constant cluster membership over the whole year. Similarly, there were another 82 SMs with more than 10 months of complete information and a constant cluster membership. The remaining SMs changed their cluster membership up to 2 months.

To perform the spatial analysis, only SMs that contained at least 10 months of complete information and had a constant cluster membership were used, resulting in 676 SMs (89.66% of SMs analysed in this section); that is, the 78 SMs with variable behaviour during the analysis period were not considered in the analysis.

Table 2. Result of the analysis of the space-time cubes. First column indicates the number of months with complete measurements per year. Behaviour type indicates if the SM stayed constant in its behaviour categorisation (membership of the cluster).

Months with Complete Information per Year (2014–2017)	Behaviour Type	SMs	SMs (%)
12 months	constant	594	78.78
10–11 months	constant	82	10.88
10–12 months	variable	78	10.34
Total		754	100.00

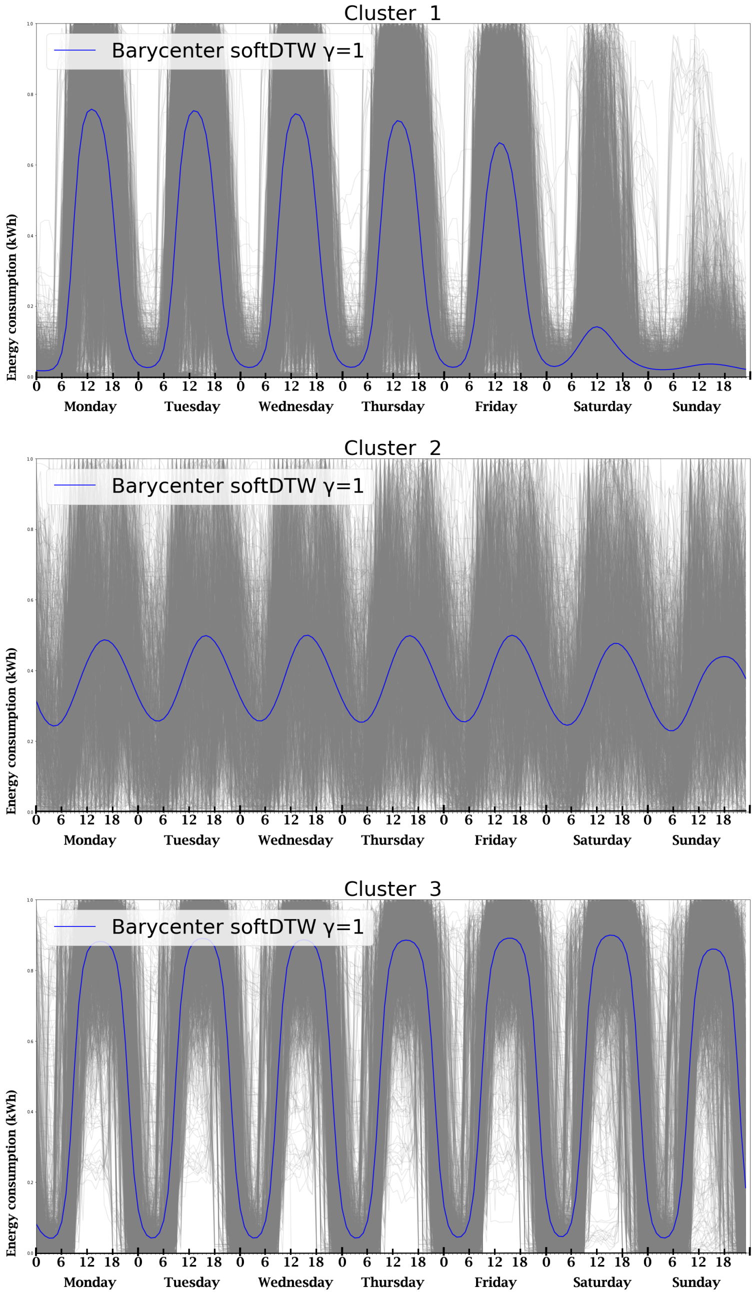

Figure 4. The figure shows the generated time series clusters, which model a typical week of consumption in each month for each SM. The time series for each SM are represented in grey, and blue represents the barycenter of each cluster obtained by soft-DTW with $k = 3$ and $\gamma = 1$.

3.2. Spatial and Temporal Analysis

Following the insights provided by the previous temporal analysis, a proximity analysis was performed using a GIS over the resulted sub-sample. First, the average circle radius in which a SM had at least 4 neighbours was calculated. The number of 4 neighbours was defined considering at least the 4 sides of a two-dimensional space. The result of the analysis resulted in 980 m, which was rounded up to 1 km. Next, the spatial proximity analysis was performed, resulting in 20 SMs located more than 1 km from any other SM or had less than 4 neighbours 1 km around. Table 3 presents the proximity analysis results of clusters of this sub-sample. Results show that the majority of isolated SMs belonged to Cluster 2, mainly country houses at least 1 km away from other dwellings. The 20 SMs located in isolated houses were not considered in the next spatial analysis.

Table 3. Spatial analysis to identify SMs with at least 5 neighbours within 1 km.

Cluster	SMs	>1 km	<1 km
Type 1	370	2	368
Type 2	74	16	58
Type 3	232	2	230
Total	676	20	656

Table 4 presents a series of metrics to evaluate the consistency of the elements in each of the clusters. The table presents RMSE and sMAPE between the load profile of electrical consumers and the barycenter of each cluster. The mean RMSE of the 58 SMs in Cluster 2 is greater than twice the RMSE of Cluster 1 and Cluster 3. Similarly, the sMAPE is 17.47 and 22.43% greater than the sMAPE of Cluster 1 and Cluster 3, respectively. These values indicate that the members of Cluster 2 had a high variability and represented SMs that were not identified in Cluster 1 or Cluster 3. For this reason, and since our objective was to focus on sets of SMs that were similar and closely located, the SMs in Cluster 2 were excluded from the analysis, resulting in a final dataset of 598 SMs.

Table 4. RMSE and sMAPE metrics between the time series of each SM and the barycenter of the cluster in which they were classified.

Cluster	RMSE	sMAPE	SMs	SMs
Type 1	0.5541	45.78	368	56.10%
Type 2	1.5594	63.25	58	8.84%
Type 3	0.6078	40.82	230	35.06%
		Total:	656	100.00%

The final database was subsequently set to extract spatially constrained clusters and validate them with a sample. Therefore, 59 SMs were randomly selected to later validate the spatial clusters. The remaining 539 SMs, georeferenced in a GIS together with a layer of rivers and a layer of non-residential areas in the city, were geoprocessed to find the spatially constrained clusters using a minimum spanning tree [34]. In this way, the SMs that belonged to the same type of behaviour and were spatially contiguous were selected, i.e., they were not separated by a river, and there were no other commercial SMsor SMs belonging to other clusters between them. The results of the analysis are presented in the Table 5, which points to 21 different spatial clusters, i.e., 21 zones where consumption patterns were more similar among them. In the case of the spatial sub-clusters belonging to Cluster 1, the results defined 14 zones having a total of 143 SMs. On the other hand, for the case of the spatial sub-clusters belonging to Cluster 3, the results showed 7 zones with a total of 46 SMs. This fact indicated that the behaviour of Cluster 1 was more common, frequent and geographically stable than Cluster 2. In addition, Figure 5 graphically presents a cluster map of the georeferenced SMs in the city of Guayaquil. The blue points represent the SMs belonging to Cluster 1 and the red are those belonging to Cluster 3. The polygons were generated to visualise the spatial cluster zones to which the SMs belong, likewise in blue for the zones comprising SMs belonging to Cluster 1 and red for SM clients belonging to Cluster 3.

Table 5. Number of spatial clusters found for each type of behaviour and number of SMs that were grouped into spatial clusters.

Behaviour Type	SMs	Assigned SMs in Spatial Cluster	Number of Spatial Clusters
Type 1	332	143	14
Type 3	207	46	7
Total	539	189	21

Table 6 shows the results of the validation process. Results were satisfactory since all the metric values were close to 90%.

Table 6. Recall and precision for classification of sample SMs for validation.

Metric	Total (%)
Specificity	90.32
Precision	86.96
Recall	95.24
F-Score	90.91

3.3. Forecasting of Energy Consumption

In this subsection the results of a RNN incorporating spatial clustering information are shown. To compare the performance of the inclusion of this new information, the prediction scenario was applied to two different RNNs, one with and one without the added information of the neighbours. The training of both RNNs was performed with 100 epochs. In the first scenario, a univariate analysis was done, applying a RNN with a LSTM architecture to forecast the energy consumption at each location for the following week, using only the hourly active power measured by the SMs. In the second scenario, the same RNN architecture was applied for a multivariate analysis, adding the active power measurements of the SMs belonging to the same spatial cluster. The results of both forecasting models are shown in Table 7. It can be seen that using data from the closest neighbours improved the results by 2.46%.

Table 7. Comparison of the results of energy consumption forecasting with a RNN that uses the information of the SMs belonging to each spatial subgroup (5 variables) and a RNN that only uses the historical information of the SM itself (1 variable).

Type	Univariate	Multivariate	Increment
	$\overline{sMAPE}$ (%)	$\overline{sMAPE}$ (%)	$\overline{sMAPE}$ (%)
Type 1	19.72	16.84	2.88
Type 3	14.70	12.65	2.05
Total	17.21	14.75	2.46

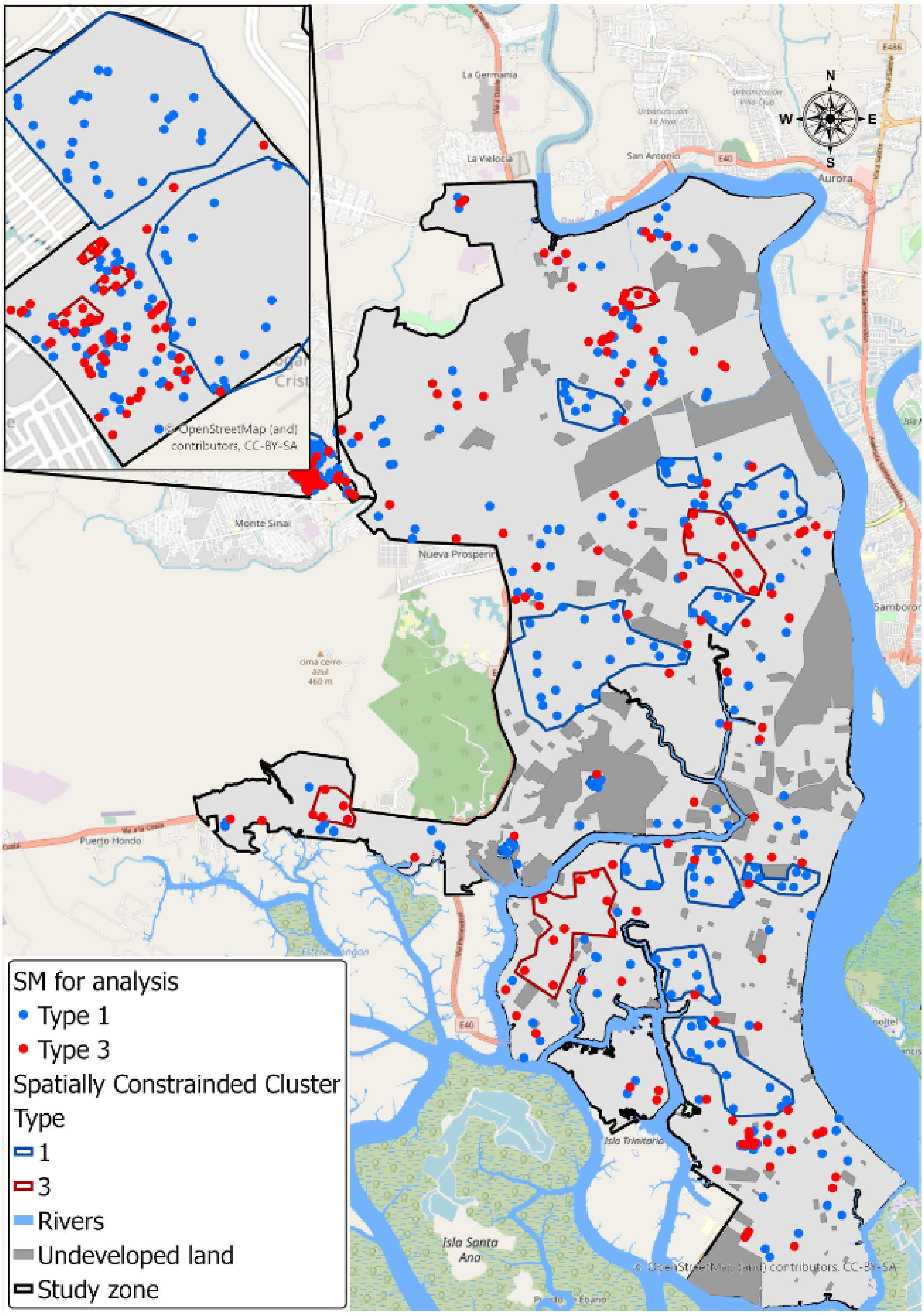

Figure 5. Spatial cluster location map after performing spatially constrained clustering.

4. Discussion

The temporal clustering of the initial set of measurements gathered by SMs in Guayaquil over 4 years, once preprocessed, resulted in 3 different clusters. Further analysis of the clusters revealed that only 2 were significant since the third contained a low percentage of meters that had a distinct behaviour. Next, for the meters classified in each of the clusters, a typical monthly consumption profile was generated. After performing an analysis of these profiles, it was observed that, on one hand, consumer behaviour did not change significantly over time due to the slight climatic variability in the area, where the temperature is usually between 21 and 30 °C [41]. On the other hand, months in which users had variant behaviour were those with long holidays periods, such as December (Christmas) or February (Carnival).

To illustrate the valuable information found in the spatial clusters, load profiles without normalising the measurements were graphed. Figure 6 shows the average hourly load profiles from Monday to Sunday for the types of behaviour found. Load profiles within each cluster were stratified according to monthly energy consumption to avoid smoothing them out when averaging users with higher consumption but with the same behaviour: (i) less than 130, (ii) 130–500, (iii) 500–1000 and (iv) more than 1000 kWh/month.

The first two groups (i, ii) are those that received the largest economic subsidies from the government, whereas the other two had benefits—group (iv) did not receive any subsidy. Type 1 customers demanded a greater amount of energy on Mondays, while type 3 customers demanded it on Saturdays, regardless of their monthly consumption. Then, for type 1 clients, it would be better to plan maintenance activities on weekends, while for type 3 clients,they should be avoided on Saturdays. Furthermore, if maintenance were planned on working days, they would have a lesser impact from 5:00 p.m. for type 1 clients and after 7:00 p.m. for type 3 clients. For both types, mornings until 10:00 a.m. would be critical periods for carrying out maintenance activities due to the rapid increase in energy demand in that period. This precise, updated and geographically zoned consumption information will substantially improve maintenance planning and optimise resources.

Furthermore, the electrical utility reported that there are users who do not respect the use of the assigned electrical energy since uses other than residential (e.g., by bars, mini-markets, cybercafes or micro-business workshops) were detected during the verification of the quality of the measurements. This fact highlighted the valuable information provided by this investigation since the results can be an input for detecting such illicit usage and verify it in the field. For example, the upper graph in Figure 6 shows a lower consumption during the weekend, which may not correspond to a usual residential behaviour.

The spatial analysis aimeds to define the geographical zones where all meters had the same behaviour; that is, they were classified in the same temporal cluster and were exploited for forecasting purposes. To predict the energy consumption of one consumer, not only data gathered by its SM was used, but also data gathered by other SMs that belonged to the same spatial sub-cluster. The results showed that the accuracy of forecasting improved by 2.46% on average when information about neighbouring SMs was included.

A more precise knowledge of energy consumption patterns of clients is valuable for both technical and commercial management. From a commercial perspective, the methodology allows the accurate estimate of energy being supplied to different zones of the city during blackouts and the ability to prioritise zones in consumption awareness campaigns. From a technical point of view, it allows better planning of maintenance activities and a more accurate estimation of future demand factors, which are useful for network planning and fro reducing investment in networks or power plants.

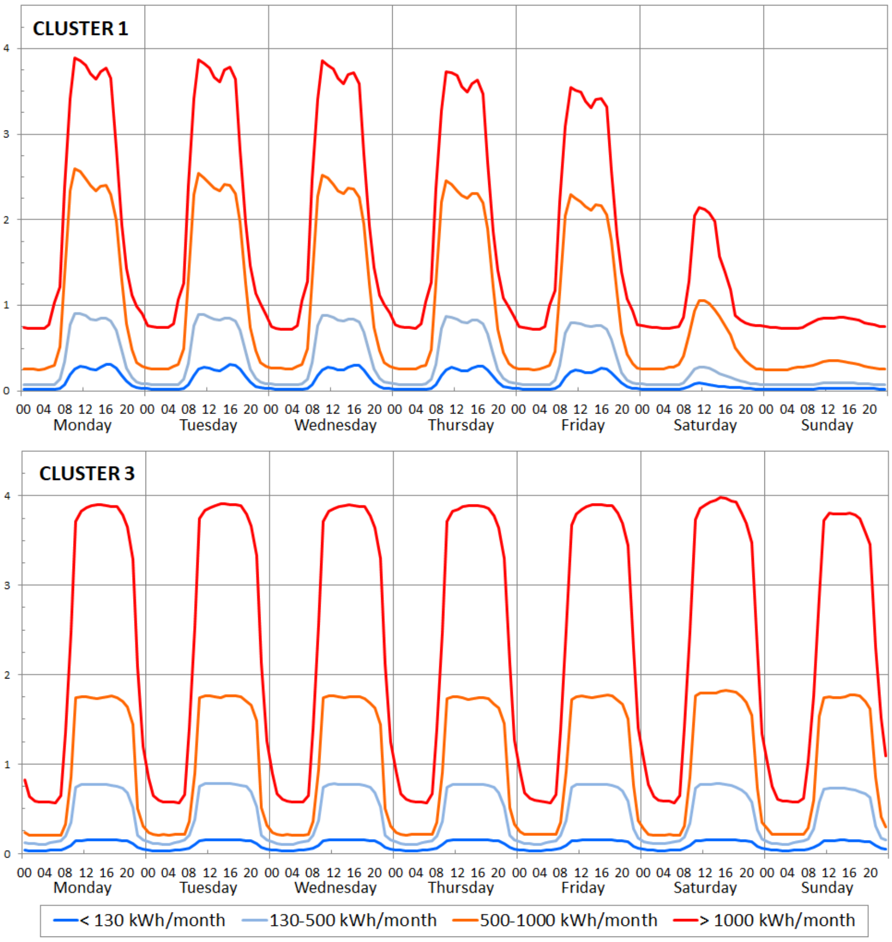

Figure 6. Average hourly load profiles are graphed from Monday to Sunday for type 1 and type 3 behaviours divided according to the energy consumed: less than 130, 130–500, 500–1000 and more than 1000 kWh/month

5. Conclusions

In this study we presented a methodology to demonstrate that energy consumption patterns in nearby areas are related and to extract models that use this information as an advantage. The use of ML tools helps define and discover new consumption behaviour profiles of residential users and determine geographic zones where behaviour is more marked and stable, thereby allowing us to improve the forecasting of energy consumption for the members of each sub-cluster.

Author Contributions: Conceptualization, M.F. and S.H.; methodology, M.F. and I.C.; software, M.F.; validation, M.F., S.H. and I.C.; formal analysis, M.F.; investigation, M.F., S.H. and I.C.; resources, S.H.; data curation, M.F.; writing—original draft preparation, M.F., S.H. and I.C.; writing—review and editing, M.F., S.H. and I.C.; visualization, M.F. and I.C.; supervision, S.H. and I.C.; project administration, S.H.; funding acquisition, S.H. All authors have read and agreed to the published version of the manuscript.

Funding: The University of Girona and SENESCYT-Ecuador awarded the author with a pre-doctoral grant of Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación, (SENESCYT)—Ecuador. This work has been partially funded by the grant PID2020-117171RA-I00 funded by MCIN/AEI/ 10.13039/501100011033, the Government of Catalonia under 2017SGR1551 and the E-LAND project which received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824388. This document reflects only the author's views and the commission is not responsible for any use that may be made of the information contained there.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación, (SENESCYT)—Ecuador.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AMI	Advanced metering infrastructure
ANN	Artificial neural networks
DTW	Dynamic time warping
GIS	Geographic Information System
LSTM	Long short-term memory
RMSE	Root mean squared error
RNN	Recurrent neural network
SM	Smart meter
sMAPE	Symmetric mean absolute percentage error

References

1. Hsiao, Y.H. Household electricity demand forecast based on context information and user daily schedule analysis from meter data. *IEEE Trans. Ind. Inform.* **2014**, *11*, 33–43. [CrossRef]
2. Zhou, K.; Yang, C.; Shen, J. Discovering residential electricity consumption patterns through smart-meter data mining: A case study from China. *Util. Policy* **2017**, *44*, 73–84. [CrossRef]
3. Lavin, A.; Klabjan, D. Clustering time-series energy data from smart meters. *Energy Effic.* **2015**, *8*, 681–689. [CrossRef]
4. Gouveia, J.P.; Seixas, J. Unraveling electricity consumption profiles in households through clusters: Combining smart meters and door-to-door surveys. *Energy Build.* **2016**, *116*, 666–676. [CrossRef]
5. Viegas, J.L.; Vieira, S.M.; Melício, R.; Mendes, V.; Sousa, J.M. Classification of new electricity customers based on surveys and smart metering data. *Energy* **2016**, *107*, 804–817. [CrossRef]
6. Kwac, J.; Flora, J.; Rajagopal, R. Household energy consumption segmentation using hourly data. *IEEE Trans. Smart Grid* **2014**, *5*, 420–430. [CrossRef]
7. Abreu, J.M.; Pereira, F.C.; Ferrão, P. Using pattern recognition to identify habitual behavior in residential electricity consumption. *Energy Build.* **2012**, *49*, 479–487. [CrossRef]
8. Tascikaraoglu, A.; Sanandaji, B.M. Short-term residential electric load forecasting: A compressive spatio-temporal approach. *Energy Build.* **2016**, *111*, 380–392. [CrossRef]
9. Xu, J.; Yue, M.; Katramatos, D.; Yoo, S. Spatial-temporal load forecasting using AMI data. In Proceedings of the 2016 IEEE International Conference on Smart Grid Communications (SmartGridComm), Sydney, NSW, Australia, 6–9 November 2016; pp. 612–618.
10. Melo, J.D.; Carreno, E.M.; Padilha-Feltrin, A. Multi-agent simulation of urban social dynamics for spatial load forecasting. *IEEE Trans. Power Syst.* **2012**, *27*, 1870–1878. [CrossRef]
11. Zhang, L.; Feng, J.; Jian, X. Model of energy alternative in spatial load forecasting. In Proceedings of the 2016 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC), Xi'an, China, 25–28 October 2016; pp. 2106–2110.
12. Melo, J.; Padilha-Feltrin, A.; Carreno, E. Spatial pattern recognition of urban sprawl using a geographically weighted regression for spatial electric load forecasting. In Proceedings of the 2015 18th International Conference on Intelligent System Application to Power Systems (ISAP), Porto, Portugal, 11–16 September 2015; pp. 1–5.
13. Wijaya, T.K.; Vasirani, M.; Humeau, S.; Aberer, K. Cluster-based aggregate forecasting for residential electricity demand using smart meter data. In Proceedings of the 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, USA, 29 October–1 November 2015; pp. 879–887.
14. Sevlian, R.; Rajagopal, R. Short term electricity load forecasting on varying levels of aggregation. *arXiv* **2014**, arXiv:1404.0058.
15. Shahzadeh, A.; Khosravi, A.; Nahavandi, S. Improving load forecast accuracy by clustering consumers using smart meter data. In Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 12–17 July 2015; pp. 1–7.
16. Ilic, D.; Karnouskos, S.; Da Silva, P.G. Improving load forecast in prosumer clusters by varying energy storage size. In Proceedings of the IEEE Grenoble PowerTech, Grenoble, France, 16–20 June 2013.

17. Biswas, M.R.; Robinson, M.D.; Fumo, N. Prediction of residential building energy consumption: A neural network approach. *Energy* **2016**, *117*, 84–92. [CrossRef]
18. Marino, D.L.; Amarasinghe, K.; Manic, M. Building energy load forecasting using deep neural networks. In Proceedings of the IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society, Florence, Italy, 23–26 October 2016; pp. 7046–7051.
19. Mocanu, E.; Nguyen, P.H.; Gibescu, M.; Kling, W.L. Deep learning for estimating building energy consumption. *Sustain. Energy Grids Netw.* **2016**, *6*, 91–99. [CrossRef]
20. Ryu, S.; Noh, J.; Kim, H. Deep neural network based demand side short term load forecasting. *Energies* **2017**, *10*, 3. [CrossRef]
21. Cuturi, M.; Blondel, M. Soft-DTW: A differentiable loss function for time-series. *arXiv* **2017**, arXiv:1703.01541.
22. Cano, E.L.; Groissböck, M.; Moguerza, J.M.; Stadler, M. A strategic optimization model for energy systems planning. *Energy Build.* **2014**, *81*, 416–423. [CrossRef]
23. Esther, B.P.; Kumar, K.S. A survey on residential demand side management architecture, approaches, optimization models and methods. *Renew. Sustain. Energy Rev.* **2016**, *59*, 342–351. [CrossRef]
24. Mahmoudi-Kohan, N.; Moghaddam, M.P.; Sheikh-El-Eslami, M. An annual framework for clustering-based pricing for an electricity retailer. *Electr. Power Syst. Res.* **2010**, *80*, 1042–1048. [CrossRef]
25. Rhodes, J.D.; Cole, W.J.; Upshaw, C.R.; Edgar, T.F.; Webber, M.E. Clustering analysis of residential electricity demand profiles. *Appl. Energy* **2014**, *135*, 461–471. [CrossRef]
26. Beaudin, M.; Zareipour, H. Home energy management systems: A review of modelling and complexity. *Renew. Sustain. Energy Rev.* **2015**, *45*, 318–335. [CrossRef]
27. Subramanian, A.S.R.; Gundersen, T.; Adams, T.A. Modeling and simulation of energy systems: A review. *Processes* **2018**, *6*, 238. [CrossRef]
28. Lilla, S.; Orozco, C.; Borgethhi, A.; Napolitano, F.; Tossani, F. Day-ahead scheduling of a local energy community: An alternating direction method of multipliers approach. *IEEE Trans. Power Syst.* **2020**, *35*, 1132–1142. [CrossRef]
29. E-LAND Horizon H2020. Available online: https://elandh2020.eu (accessed on 1 October 2021).
30. Chicco, G. Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy* **2012**, *42*, 68–80. [CrossRef]
31. Hino, H.; Shen, H.; Murata, N.; Wakao, S.; Hayashi, Y. A versatile clustering method for electricity consumption pattern analysis in households. *IEEE Trans. Smart Grid* **2013**, *4*, 1048–1057. [CrossRef]
32. Paparrizos, J.; Gravano, L. k-shape: Efficient and accurate clustering of time series. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, VIC, Australia, 31 May–4 June 2015; pp. 1855–1870.
33. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
34. Assunção, R.M.; Neves, M.C.; Câmara, G.; da Costa Freitas, C. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *Int. J. Geogr. Inf. Sci.* **2006**, *20*, 797–811. [CrossRef]
35. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
36. Medsker, L.R.; Jain, L. Recurrent neural networks. *Des. Appl.* **2001**, *5*, 64–67.
37. Hochreiter, S.; Bengio, Y.; Frasconi, P.; Schmidhuber, J. Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies. In *A Field Guide to Dynamical Recurrent Networks*; Kolen, J.F., Kremer, S.C., Eds.; IEEE Press: New York, NY, USA, 2001; pp. 237–244.
38. Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **1994**, *5*, 157–166. [CrossRef]
39. Chen, Z.; Yang, Y. Assessing Forecast Accuracy Measures. 2004. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.69.1016&rep=rep1&type=pdf (accessed on 1 October 2021).
40. Tavenard, R.; Faouzi, J.; Vandewiele, G.; Divo, F.; Androz, G.; Holtz, C.; Payne, M.; Yurchak, R.; Rußwurm, M.; Kolar, K.; et al. Tslearn, A Machine Learning Toolkit for Time Series Data. *J. Mach. Learn. Res.* **2020**, *21*, 1–6.
41. Average Weather in Guayaquil. 2018. Available online: weatherspark.com (accessed on 1 July 2021).