Improving energy self-sufficiency of a renovated residential neighborhood with heat pumps by analyzing smart meter data

Citation for published version (APA):
Walker, S., Bergkamp, V., Yang, D., van Goch, T. A. J., Katic, K., & Zeiler, W. (2021). Improving energy self-sufficiency of a renovated residential neighborhood with heat pumps by analyzing smart meter data. Energy, 229, [120711]. https://doi.org/10.1016/j.energy.2021.120711

Document license: CC BY

DOI: 10.1016/j.energy.2021.120711

Document status and date:
Published: 15/08/2021

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.

Download date: 08. Jun. 2022
Improving energy self-sufficiency of a renovated residential neighborhood with heat pumps by analyzing smart meter data

Shalika Walker a,*, Vince Bergkamp a, Dujuan Yang a, T.A.J. van Goch b, Katarina Katic a, Wim Zeiler a

a Eindhoven University of Technology, PO Box 513, 5600 MB, Eindhoven, the Netherlands
b BAM Bouw en Techniek, Postbus 109, 3980 CC, Bunnik, the Netherlands

ARTICLE INFO
Article history:
Received 13 July 2020
Received in revised form 12 April 2021
Accepted 19 April 2021
Available online 22 April 2021

Keywords:
K-means clustering
Neighborhood level
Electrical storage system
Random-forest
Multi-objective optimization

ABSTRACT
In the energy renovation process, usually, buildings are upgraded to become energy-neutral annually with installed photovoltaic systems and heat pumps. However, the energy self-sufficiency of these buildings is surprisingly low. Therefore, the rapid deployment of heat pump based heating systems creates a shift of natural-gas consumption from the previously consumed building side (boilers) towards the electricity production side (power-plants). Fortunately, the development of information and communication technology enables access to consumption/generation data of building-related energy systems. Thus, there is an opportunity to strategically use this data and improve energy self-sufficiency and accommodate heat pump based heating systems. In this study, the improvement of self-sufficiency is discussed using a renovated neighborhood. The presented method incorporates a smart-grid application with a data-driven clustering, prediction, and an energy management strategy. First, clustering of similar demand-profiled dwellings with the k-means algorithm, and demand-prediction using the random-forest technique was performed. Afterwards, electric energy storage was introduced and multi-objective optimization reducing annualized costs and carbon emissions have been performed. For the carbon-dioxide optimal case, when aimed at the entire neighborhood, an annual self-sufficiency increment of more than 25% can be achieved, while four months out of the twelve being 100% energy self-sufficient.

© 2021 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction
Ever since the Paris agreement to pursue efforts to limit the temperature increase to 1.5 °C above pre-industrial levels was accepted [1], governments have started stimulating energy efficiency and sustainability measures by means of subsidies and regulations. One of the most important focus fields of these regulations is the buildings since they account for 40% of the global energy demand [2] and 30% of global CO2 emissions [2]. This has led to a growth in utilizing private renewable energy sources [3], primarily photovoltaic (PV) panels, and has changed the centralized energy system towards a more decentralized infrastructure [4].

Along with this transformation, most of the European countries seek out measures to eliminate the natural-gas utilization of the buildings' heating sector. All-electric heating systems [5], for instance, heat pumps (HPs) [6] gain a great deal of attention as a substitute. Nevertheless, with the current composition of the electricity production (central coal and natural-gas power plants) [7], a shift of fossil energy consumption from the demand side (boilers) to the central production side (power-plants) [5] can be observed when operating all-electric heating systems [8]. Some can argue, this problem can be eliminated with local renewable energy sources. Yet, most of the renewable energy sources, especially wind and solar energy, do not have a high energy density and are irregularly available [9]. Mainly for PV systems of residential buildings, the solar intensity is at its peak during daytime whereas electricity demand is generally low. Therefore, methods are needed to improve energy self-sufficiency, especially for neighborhoods with residential buildings.

1.1. Motivation
Even though employing new energy systems or smart control
strategies is a possibility to improve energy self-sufficiency with new buildings, applying the same to the existing buildings remains a huge challenge. In the existing buildings’ energy renovation process, the first step is to efficiently insulate them so that the heating demand can be reduced by a considerable amount. Then, the rest of the demand can be provided by a suitable energy system. In this process most of the time, buildings are upgraded to become energy neutral annually, yet energy self-sufficiency has been identified to be surprisingly low [10]. Therefore, most of the produced energy from PV panels is fed-back to the grid without using locally. At the moment, in some countries, energy taxes are only applied to the net electricity consumption of prosumers. However, this net metering system will gradually disappear. Meaning, the electricity that is directly used from the national grid will be costly and feed-in electricity may not bring any advantage to the building owners. This study is motivated by the fact that the level of energy self-sufficiency is not adequate as elements associated with the heating demand electrify especially with HPs.

On a positive note, in parallel with the installation of PV and HP systems, the replacement of traditional analog meters with smart meters is also being done which has the possibility to generate and accumulate enormous amounts of data [13]. Smart meter data has enabled to cluster dwellings into the energy demand patterns of individuals and neighborhoods [12]. However, this data is often considered privacy sensitive. Therefore, not many studies have been able to use such detailed energy consumption data of large samples of households. Making use of smart-meter data with available control strategies to improve energy self-sufficiency of existing renovated buildings, and accommodate HP based heating systems is the main objective of this work.

1.2. Literature survey

As of many other sectors, the built environment has gained momentum over the past decade in taking advantage of machine learning algorithms for enduring smart-grid applications. Out of these applications, the highest focus is drawn to multi-step (Ex: day-ahead) [13] or single-step (Ex: Next hour) [14] energy demand prediction. These studies pointed out it is challenging to determine one specific best machine learning model for demand prediction. It is stated that the choice of the model is determined by the nature of the data [15]. Therefore, it is essential to analyze the available data and application, to determine which model suits best in the given situation [16]. Nonetheless, most of these studies agree that when the resolution of the data points is high (hourly, 30-min, 15-min, etc.) the prediction accuracy reduces with commonly used machine learning algorithms, therefore advanced deep-learning methods or unsupervised clustering methods (depending on the application) are useful in these situations.

To fully grasp the information available from smart meters, evaluating data at the collected high resolution is important. When the granularity of the data set is high, rather than demand prediction, researchers have focused on identifying representative typical demand patterns. Clustering algorithms have served in this area under discussion. Shin et al. [17] used a data set of 22 houses with a sampling interval of 1/15 s for 29–122 days with the target of finding typical daily patterns. In their study, the k-means method was used as the clustering algorithm. In another study, Yilmaz et al. [18] compared two different approaches using the k-means method to identify the best possible representative electricity demand profiles. In this study, 15-min data of 690 households for a complete year has been exploited. When the demand prediction is a necessity, Schectenin [19] has proposed to cluster the demand patterns first and then perform the prediction in order to improve the prediction accuracy.

1.3. Contribution

This research shows one of the many possible applications that analyzing smart meter data makes possible. Using demand prediction and smart control strategies, this study evaluates how to efficiently micro-manage the smart grid, expose energy-saving possibilities, and improve energy self-sufficiency of renovated neighborhoods.

1.4. Organization of the paper

First, in section 2, an approach is suggested to perceive information from 15-min granularity smart meter data with a clustering method. Then, this knowledge is used innovatively for demand prediction. In section 3, a currently used control algorithm is utilized to improve the energy self-sufficiency of a renovated neighborhood along with a battery storage system. Section 4 provides the concluding remarks of the study.

2. Methodology

The case study dwellings are located in the center of the Netherlands, have no natural gas connection, and are provided with a PV-system, HP, and high levels of insulation. These renovated dwellings are meant to be ‘net-zero-energy’ when considering a full year (i.e. yearly consumption = yearly production). The living space of the dwellings is about 85–100 m² with an installed HP capacity of 1.2 kW and a PV installation capacity of 5–8 kWp. Two years (2016 and 2017) of 15-min interval meter data from these 70 renovated dwellings were used in this study.

These households use a constant tariff structure for their electricity consumption. The collected meter data from the dwellings are namely; smart meter (SM) consumed, smart meter produced, and heat pump consumed. Smart meter produced indicates electricity produced by the PV panels. The total grid consumption is defined as in equation (1). The total grid consumption can carry a value of plus or minus depending on the PV production as illustrated in Fig. 1.

\[
\text{Total}_{\text{Grid}} = \text{SM}_\text{Consume} - \text{SM}_\text{Produce} + \text{HP}_\text{Consume} \ [\text{kWh}]
\]  

The cleaned smart meter data of the 70 houses (see Annex.A) was used to follow the steps described in the next sections, namely; dynamic clustering, demand prediction, and self-sufficiency improvement of the neighborhood as demonstrated in Fig. 2. The underlying theoretical principles of the used methods and models will be clarified consequently.

2.1. Step 1: Dynamic clustering

The goal of this clustering step is to exploit and identify similar electricity demand patterned dwellings as illustrated in Fig. 3. By doing a dynamic clustering process (change every day) it is possible to recognize how similar the dwellings behave each day. As the clustering technique, the k-means method [20] is employed since it is one of the most widely used clustering algorithms with proven performances (see Annex.B).

In order to identify the suitable number of clusters ‘k’ to segment the data, two validity indices have been checked, namely: Davies-Bouldin validity Index (DBI) [21] and Silhouette Coefficient (SC) [21]. Davies-Bouldin index evaluates inter-cluster differences. As shown by equation (2), the minimum score that can be obtained for DBI is zero. Smaller the values obtained for DBI, the better the clustering [11]. In the equation, the average intra-cluster distance is given by \(diam(C_i)\) and the distance between two cluster centroids is
given by \(d(C_i, C_j)\).

\[
DBI = \frac{1}{K} \sum_{i=1}^{k} \max_{j \neq i} \frac{\text{diam}(C_i) + \text{diam}(C_j)}{d(C_i, C_j)}
\]  

(2)

The Silhouette Coefficient is a normalized value that can be used to determine the degree of separation between the clusters [18]. The coefficient is bounded by the values \(-1\) and \(1\). A higher value indicates better clustering, while negative values represent misclustering [11]. In equation (3), \(C_x\) shows the average intra-cluster distance and \(C_0\) shows the average minimum distance to another cluster. In this study, these two indices were calculated for each day of the analyzed period. Then, the average value per month was compared and a suitable number for \(k\) is selected.

\[
SC = \frac{C_x - C_0}{\max(C_x, C_0)}
\]

(3)

Fourteen days prior to the evaluation-day (= \(Day: x+1\)) electricity consumption data is used for the clustering process (see Fig. 3). Thus, \(Day: x\) to \(Day: x-14\). So, the 70 houses are grouped into \(k = M\) number of clusters according to their past 14-day demand.
behavior. Then, on **Day: x+1** (which is the evaluation day) the energy demand of dwellings are aggregated virtually according to the clustering obtained (illustrated in Fig. 4).

At the end of each day (24:00), the clustering process will be repeated. So, every new day a different group and a different number of houses are available in each cluster according to the demand behavior of their past 14 days. The 14-day mark was selected with a holistic approach. The 14-day mark provided a balanced behavior for the desired process. This step allows understanding in each day how many and which dwellings group together. The dwelling numbers in each cluster are known.

### 2.2. Step 2: Demand prediction

Usually, the demand prediction step is needed to reduce real-time monitoring expenses, initial costs of hardware components, and long-term maintenance costs. For the demand prediction, the *Random Forest* regression algorithm (see Annex.C) is used due to its performance superiority. With aggregated level predictions it is not possible to identify information about individual buildings, but the prediction accuracy is high. With individual building prediction, no information is lost but prediction accuracy is low. Clustering is used so that the information about individual buildings is not completely lost and prediction accuracy can be improved. Fig. 5 illustrates the demand prediction procedure.

In this study, after the clustering is performed at the end of each day, demand prediction is performed for each cluster for the coming day-ahead (**Day: x+1**). Since the number of dwellings inside the clusters is changing each day, model training for each cluster should also be performed at the end of each day (re-trained). For the model training and prediction, 14-day prior data is utilized (1344 data points per dwelling). In demand prediction also 14-day mark is chosen after performing several trial and error exercises with smaller and larger data sample sizes. It was identified for this case; larger data samples do not significantly influence the prediction accuracy. Since the prediction variable TotalGrid includes plus and minus values, normalization of the data set was performed before prediction. Normalization brings all the values into the range [0,1]. Outdoor temperature, irradiance, and previous day demand of the dwellings are chosen as the input features for the random forest model (see Fig. 5). To understand the added value of clustered predictions, a similar process was conducted to predict the day-ahead individual demand of the houses. Then, the results of individual predictions are compared with clustered predictions.

The performance quality of the trained models was evaluated using the coefficient of determination ($R^2$) [22]. $R^2$ is a dimensionless indicator that shows how much variability in the dependent variable is accounted for by the independent variables. This value varies between 0 and 1. The closer the value to 1, the better the predictions. For this study, the pseudo-$R^2$ value above 0.8 is considered sufficient for the trained models. $R^2$ is a better evaluation metric for trained models than the predictions [23]. The coefficient of the variance of root mean square error (CVRMSE) [22] was used to determine the prediction accuracy. ‘ASHRAE guideline for measurement of energy demand’ [24] states for hourly demand predictions, CVRMSE value below 0.3 ($\sim$30%) is sufficiently close to physical reality, and it is a good indicator of predictions. This same standard is considered in this study for 15-min demand predictions. In equations (4) and (5), $\bar{x}_i$ and $x_i$ symbolize the predicted and observed values of data point $i$, while $\bar{x}$ shows the mean of all the observed data points. $N$ represents the total number of data points.

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (\bar{x}_i - x_i)^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \quad (4)$$

$$CVRMSE = \frac{RMSE}{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x}_i)^2}{N}} \frac{1}{\bar{x}} \quad (5)$$

### 2.3. Step 3: Self-sufficiency improvement of the neighborhood

After knowing the demand predictions of the neighborhood, the next step is to improve the self-sufficiency. This is done by using
appropriately energy sharing measures between dwellings. If the dwellings operate under a separate micro-grid, at a certain time interval 'T', it is possible to calculate the direct-shared energy potential which is the allocation of excess produced electricity from PV panels directly between the dwellings. However, by comparing the demand patterns of individual houses, the exploitation and identification of direct energy sharing potential is cumbersome. In this case, clustered demand patterns significantly help in identifying such parameters. The notion of direct sharing is illustrated in Fig. 6.

To improve the self-sufficiency (equation (6)), next, battery storage is introduced, so that, the excess energy produced by PV panels can be stored and used at a later time period as illustrated in Fig. 7. The battery storage systems can be installed at a neighborhood scale or local level. Due to the development of technology and the gradual price reduction [25], it is foreseeable that the vast scale or local level. Due to the development of technology and the gradual price reduction [25], it is foreseeable that the vast utilization of battery systems at the neighborhood [26] level will soon become a preference.

\[
\text{Self-sufficiency} = \frac{\text{Usable energy from renewable sources}}{\text{Actual energy demand}}
\] (6)

2.3.1. Model formulation for the operation of neighborhood battery storage system

To calculate the energy sharing potential through a battery storage system, a Matlab-based optimization problem is implemented using linear programming [27]. Modeling toolbox YALMIP is used as the interface to the solver Mosek. The model formulation is based on the ‘energy hub’ concept [28] and is defined in 2.3.1. sections A-C. In the formulations, a scalar variable is represented by an italic letter while a bold letter stands for a vector. Scalar and vector annotation mainly differentiate the design value of devices (one value) and optimal control of devices (per sample time). An independent decision variable is presented by a superscript \( d \).

2.3.1.1. Battery storage. The storage system modeled in this work is battery storage and a simplified linear model is used for the optimization problem as described by the references [28]. Equations (7)–(10) describes the equality and inequality constraints. In equation (7), \( E_t \) is the storage level at each time interval \( t \) in kWh, \( \eta_{\text{Charge/Discharge}} \) is the charging and discharging efficiencies in kWh and \( \eta_{\text{Charge/Discharge}} \) is the charging and discharging efficiencies. \( \text{Decay} \) is the rate at which stored energy declines.

\[
E_t = E_{t-1} - \eta_{\text{Discharge}} \frac{1}{\eta_{\text{Discharge}}} Bat_{\text{Discharge}}^d t + \frac{\eta_{\text{Charge}}}{\eta_{\text{Discharge}}} Bat_{\text{Charge}}^d t
\] (7)

\[
0 \leq E_t \leq C_{\text{Bat}}^d
\] (8)

\[
X_{\text{Bat}}^d, C_{\text{min}}^d \leq E_{t,\text{Bat}}^d \leq X_{\text{Bat}}^d, C_{\text{max}}^d
\] (9)

\[
0 \leq \frac{Bat_{\text{Charge/Discharge}}^d - C_{\text{Bat}, b}^d}{\eta_{\text{Charge/Discharge}}}
\] (10)

Both charging and discharging efficiencies are chosen to be 90% and the decay is set to 0.1% per hour. These values were chosen based on the work of [28]. \( C_{\text{Bat}}^d \) is an integer and \( X_{\text{Bat}}^d \) is a binary variable that acquires a value of ‘1’ when the device is included in the optimization. \( \eta_{\text{Charge/Discharge}} \) set the limit on the charging and discharging rates based on a C-rate [29] of 0.5C, and depth of discharge (DOD) of 80%.

2.3.1.2. Electricity grid and energy balance. The model includes an external grid for electricity. The equality constraint for the energy balance is presented by equation (11). Electricity consumed from the external electricity grid is represented by \( \text{Grid}_{\text{Energy}}^d \) while electricity fed to the grid is represented by \( \text{Grid}_{\text{In}}^d \). Note: \( \text{Total}_{\text{Grid}}^d \) is described in equation (1).

\[
\text{Grid}_{\text{Energy}}^d t + Bat_{\text{Discharge}}^d t - Bat_{\text{Charge}}^d t - \text{Grid}_{\text{In}}^d t = \text{Total}_{\text{Grid}}^d t
\] (11)

2.3.1.3. Objective function. The objective of this problem is to schedule the battery storage optimally so that the electricity feedback to the grid (\( \text{Grid}_{\text{In}}^d \)) can be minimized. Multi-objective optimization is performed with the weighted sum method [30]. Equation (12) shows the objective function (U) which minimizes both system costs and carbon emissions associated with grid electricity consumption. For utilizing PV-produced electricity no additional carbon emission is considered. Equations (14)–(18) demonstrates the notations shown by the objective function (12). The parameter \( \alpha \) in equation (12) is used as the weighting factor for multi-objective optimization. When \( \alpha \) set to zero, the objective represents the minimization of \( CO_2 \) emissions only. In this case, the battery storage is utilized at its maximum. On the other hand, when \( \alpha \) equals one objective represents cost optimization.

\[
U = \min \sum_{f=1}^{2} \text{w}_f E_f = (\alpha) C_{\text{Total}} + (1 - \alpha) CO_2_{\text{Total}}
\] (12)
The total costs ($C_{\text{Total}}$) comprises of the investment cost of the battery storage ($C_{\text{Investment}}$), fixed operation and maintenance costs ($C_{\text{OMF}}$), and variable costs ($C_V$). ACF in equations (14) and (15) is the annualized cost factor calculated based on the discount rate ($r$) and an estimated lifetime of the battery storage system. Other relevant parameters associated with the objective function are presented in Table 1.

In order to obtain the Pareto optimal solutions, $\alpha$ in equation (12) is increased in 0.05 intervals starting from 0 (CO2 optimal) to 1 (cost-optimal). For the investment cost of the battery storage and carbon factor for grid electricity, two scenarios were exploited; case-1 with 584 €/kWh [28] and 0.5 kgCO2-eq/kWh while case-2 with 260 €/kWh [28] and 0.3 kgCO2-eq/kWh. Case-1 shows a conventional circumstance and case-2 shows a futuristic sustainable situation with reduced prices for battery storage and reduced CO2 emissions for grid electricity (Netherlands mix).

### 3. Results and Discussion

Following the steps formulated in Section 2, the simulations were conducted, and the results are discussed accordingly.

#### 3.1. Step 1: Clustering

In this step, importance is given to the number of clusters that should be selected. The two validity indices, DBI and SC, have been calculated for each clustering day and averaged per month as shown in Fig. 8. For DBI, it is possible to observe, when $k = 2, 3$ the distribution of the index over the months is dispersed while the rest of the ‘$k$’ values provide closer behavior over the months. From the rest of the ‘$k$’ values, the lowest point occurs in July for $k = 4$ clusters. The lower the DBI, the better the cluster performance. Even though higher numbers for ‘$k$’ show acceptable performances, it is also detected that the distribution of the number of houses among the clusters is not satisfactory.

Similarly, for SC also the clusters show scattered behavior over the months when $k = 2, 3$ and show closer performances for the other $k$ values. Out of the other $k$ values, many months show the best performance with the highest SC value occurring when $k = 4$. Considering all these aspects described and the distribution of the number of houses among the clusters, $k = 4$ is chosen as the finest value for this study.

Fig. 9 shows the distribution of five houses among the clusters for the month of January in the years 2016 and 2017. The illustration gives a clear indication of the diversity of the demand profiles of the dwellings over different days.

#### 3.2. Step 2: Demand prediction

This section discusses the day-ahead demand prediction for individual dwellings and clustered dwellings. For the trained models, pseudo-$R^2$ value of all houses remained at the desired margin ($>0.8$). Fig. 10 demonstrates the $R^2$ variation of daily trained models of a randomly selected house. Fig. 11 demonstrates the CVRMSE variation for the day-ahead normalized individual demand predictions of all the 70 houses. The worst prediction day, best prediction day and the mean of predictions are disclosed in the figure. From the evaluated total number of 358 days in each of the years 2016 and 2017, 54 and 56 houses were able to maintain the CVRMSE margin ($\leq 0.3$) for over 300 days respectively. Worst performing house fulfilled the CVRMSE requirement 161 days for the year 2016 and 148 days for 2017.

Fig. 12 illustrates the daily $R^2$ variation of the clustered trained models. When the number of houses in a cluster increases, the trained models perform better than the days where the number of houses is less in number. Fig. 13 shows the CVRMSE error calculation for the clustered normalized predictions. For all clusters, in both the years 2016 and 2017, over 345 days managed to maintain the required CVRMSE margin. This shows a prediction improvement of the dwellings in the neighborhood while adequately preserving the insights.

When accumulated and compared the actual peak energy consumption (max: $Total_{\text{peak}}$) and maximum energy fed back to the grid (min: $Total_{\text{feed}}$) of the whole neighborhood with individual predictions and clustered predictions, clustered demand showed closer values than the individual predictions (see Table 2).

Other than that, when it is requested to predict each dwellings’ energy demand, the energy management system experiences a complex time-consuming problem. In comparison, clustering the demand patterns before prediction significantly reduced the required computational time.

#### 3.3. Self-sufficiency improvement

After the predictions are known, next, a battery storage system
Fig. 8. Cluster performance evaluation over the months.

Fig. 9. Distribution of houses among the clusters for January in 2016 and 2017.

Fig. 10. $R^2$ variation of the daily trained models of a randomly selected house (year 2016).
is introduced at the neighborhood level. If the battery storage is introduced at a larger level, one can argue the purpose of cluster-level demand predictions within the neighborhood. If the demand prediction is performed for the whole neighborhood, identification of characteristics of individual buildings is not a possibility. However, with clustering, the characteristics can still be identified for similarly performing buildings. Therefore, clustered sub-level predictions are more useful than a single aggregated prediction for the whole neighborhood.

Before introducing battery storage, as mentioned in Section 2.3, if the neighborhood operates under a separate micro-grid, the direct energy sharing possibility between the dwellings can be calculated as 20 MWh for 2016 and 28 MWh for the year 2017. Table 3 shows a summary of the results for direct sharing and the self-sufficiency increment of the entire neighborhood. Grasping this type of information is not a possibility if the demand was analyzed only at the entire neighborhood level.

Fig. 14 illustrates the demand profiles of the clusters for a random Day in the year 2016. Here, it is possible to comprehend the direct energy sharing possibility. Minus value in Fig. 14 indicates the dwellings are overproducing and electricity is fed back to the grid. During the overproducing period, dwellings in Cluster-2 seem
to be consuming energy. So, this period demonstrates how energy is directly shared between the houses.

This section discusses the overall sharing potential after introducing the battery storage system. Fig. 15 illustrates the Pareto-frontiers obtained for the conventional situation and the futuristic sustainable case for the years 2016 and 2017. In the figure, a comparison has been made with accumulated clustered predictions and accumulated actual individual consumptions of the buildings. In each case, \( \alpha = 0.5 \) is marked where the CO2 emissions and costs are given similar priority. The value of \( \alpha \) in the sub-graphs in Fig. 15 varies from one to zero from left to right. In case-1, where the battery storage prices are high, it is visible that when \( \alpha = 1 \) which is the cost-optimal case, no battery storage is included. However, in case-2 battery storage is acknowledged even with the cost-optimal case.

After obtaining the Pareto solutions, the aim is to select a suitable battery capacity to be installed in the analyzed neighborhood. Instead of choosing one capacity value, the battery capacities when \( \alpha = 0.25, 0.5, \) and 0.75 have been compared and the corresponding self-sufficiency increment has been recorded. Table 4 demonstrates the overall self-sufficiency increment of the entire neighborhood and the surplus energy fed-back to the grid after introducing the battery storage system for the different \( \alpha \) values.

The behavior of the storage system for two random days is illustrated in Fig. 16 and Fig. 17. Demand in the figures (TotalGrid) illustrates the accumulated profile of the entire neighborhood. All symbols presented in the figures resemble the symbols defined in section 2. For the low PV electricity production day (Fig. 16), all the excess energy could be stored in the battery system. In contrast to Fig. 16, in the high PV production day (Fig. 17), it is possible to observe that the battery could handle only a portion of the

---

**Table 2**

Accumulated values comparison.

| Total\(_{\text{case}}\) (kWh) | Actual | Individual Prediction | Clustered Prediction |
|------------------------------|--------|-----------------------|----------------------|
| Peak                         | 202    | 177                   | 187                  |
| Maximum fed-back             | −280   | −271                  | −273                 |

**Table 3**

Direct sharing summary.

| Year | Direct sharing potential | Self-sufficiency increment |
|------|--------------------------|---------------------------|
|      | Predicted | Actual | Predicted | Actual |
| 2016 | 24 MWh    | 20 MWh | 1.7%      | 1.4%    |
| 2017 | 30 MWh    | 28 MWh | 2.0%      | 1.9%    |

---

**Fig. 13.** Day-ahead predictions for clustered demand profiles.

**Fig. 14.** Demand profiles of the clusters for a random Day in the year 2016.
excessively produced electricity. Therefore, most of the produced electricity is yet fed back to the grid (Symbol: GridIn).

Both these figures correspond to the year 2016 sustainable scenario with an ‘$\alpha$’ value of 0.5 in the objective function. In Fig. 17 it can also be observed, at the beginning of the day the battery system is discharging. This is because of the energy stored by the battery storage during the previous day.

3.4. Excess energy

Fig. 18 illustrates the consumed energy from the grid and the fed back energy to the grid, for the CO2 optimal case ($\alpha = 0$) for the entire neighborhood (all 70 houses). In the CO2 optimal case, the maximum battery storage capacity is utilized. Nevertheless, it is visible from the figure, a significant amount of produced energy is still fed back to the grid without having the opportunity to be consumed inside the neighborhood. From the analysis, it was identified about four months out of the twelve are 100% energy sufficient for both the years 2016 and 2017.

4. Conclusion

This paper discussed the improvement of energy self-sufficiency of a renovated residential neighborhood that utilizes HP and PV systems. The presented method incorporates a futuristic smart-grid application with an efficient unsupervised clustering method and a neighborhood-level energy management system (NEMS).

It was possible to observe that clustering similar demand patterns significantly reduced the data burden for prediction and computational time required while increasing the prediction accuracy. Although the algorithms; k-means and random forest have been proved to have better performances in this paper, they may not be suitable for other scenarios. A sensitivity analysis is needed on some important variables such as electricity tariff. Other than
that, this study assumes perfect foresight of PV generation and this cannot resemble the actual operation. So, it is recommended as future work to compare different clustering and prediction algorithms including variable tariff structures and PV prediction, as further advancement of the proposed method.

To accommodate all-electric heating systems, it is important to
increase the self-sufficiency of buildings. This study showed with the involvement of data-driven prediction and control algorithms this target can be achieved during the spring/summer months. However, a point highlighted in the study is the amount of inconsumable excess energy even after introducing the battery storage systems. To further improve self-sufficiency, the proposed method can also be used for other types of applications such as introducing electric vehicle charging points at the neighborhood level. As many elements across the grid electrify, improvement of self-sufficiency is important to meet the total electricity demand through clean energy sources.

For the practical realization of such proposed NEMS, the traditional business processes, regulations, and energy markets need to be redesigned. Such concepts will be valuable in a smart grid either in the low-voltage AC or DC context and will benefit the service providers to attain considerable efficiency gains. With the presented knowledge in this paper, one can develop new ideas based on this verified method. For example, to show the quality of prediction and the ultimate cost of the system. In conclusion, this paper showed one of the many possibilities for the clever utilization of renewable electricity for a better alignment of supply and demand to confront a future-proof smart energy system at the neighborhood level.

**Declaration of competing interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Acknowledgment**

The research work is funded by Netherlands Organisation for Scientific Research (NWO) Perspective program TTW Project B (14180) – “Interactive energy management systems and lifecycle performance design for energy infrastructures of local communities” (https://ses-be.tue.nl/). We wish to express our gratitude to all the organizations which have been a part of this program. Special thank is given to BAM Bouw en Techniek for helping with the data collection process.

**Annex**

**A. Cleaning of data**

Prior to the development of the clustering and prediction methods, the collected data needed to be preprocessed [22]. The preprocessing is performed to modify the erroneous outliers and fill the missing data points. In this case, offline meters and the unavailability of machine to machine connections are found to be the reasons for these outliers and missing points. The provided data contained some unrealistic outliers, valued a hundred times larger than the mean. The data is inspected visually to see where the threshold for a so-called outlier should be. According to this visual inspection, a threshold for outliers is determined for each data set. Every value above this is replaced by a realistic replacement which is its preceding value.

For the sake of quality and reliability, missing values of the data set also should be filled. These houses have missing values in the ‘consumption’ variable, ranging from 5.6% to 19.7% (Total data points: 68,832 per house). These missing values are firstly filled by the previous time-stamp values. This first step does not fill all the missing data points. Therefore, the next step is to fill the missing points with the identical time-stamp consumption values of the previous day. These two steps reduced the number of missing data points drastically. The final step, which is the employment of moving average was able to mostly fill the rest of the absent values while smoothening the data variation. The whole process is illustrated in Fig. 19.
B. K-means clustering

K-means clustering is a convenient approach to divide a dataset into k distinct groups [21]. The main principle for the k-means clustering approach is to minimize the ‘within-cluster variation’ [21]. Taking into account two important properties; (i) each observation (in this case: Total_load), belongs to one of the k clusters, (ii) no observation belongs to more than one cluster, thus, non-overlapping clusters. The problem that has to be solved to create good clusters according to the k-means method is represented by equation (19) [20], where k is the number of clusters, and \((C_k)\) is a measure of the amount by which the observations within the clusters differ from each other, named; within-cluster variation.

\[
\min_{C_1, \ldots, C_k} \sum_{k=1}^{k} W(C_k) \tag{19}
\]

The within-cluster variation can be expressed in a number of ways, however, by far the most used method involves the squared Euclidean distance. The enthusiastic reader is referred to Ref. [20] to understand more about Euclidean distance. In this method, the ‘k’ value should be chosen manually. This is a limitation [34] of this method.

C. Random-forest algorithm

A recent study of the authors which was published elsewhere [22] discussed the accuracy of different machine learning algorithms and concluded random forest has better performances compared to the other machine learning techniques. Random forest is an ensemble learning method [20], consisting of a collection of regression trees. Since it consists of multiple trees, the prediction value is an average of all the constructed trees. The goal of this is to de-correlate the individual trees and make the prediction more reliable. However random forest doesn't predict well beyond the range in the training data [35]. This may overfit data sets that are particularly noisy and prevent making accurate predictions.

Author statement

Shalika Walker – Conceptualization, Methodology, Software, Validation, Writing - Original Draft, Visualization. Vince Bergkamp – Software, Methodology, Visualization. Dujuan Yang – Supervision, Methodology, Review. Katarina Katic – Supervision. T.A.J. van Goch – Resources, Funding acquisition. Wim Zeiler – Supervision, Review, Funding acquisition.

References

[1] Adoption of the Paris agreement. 2015. https://unfccc.int/resource/docs/2015/cop21/eng/090101.pdf. [Accessed 12 December 2019].

[2] Data tables – Data & Statistics - IEA n.d. https://www.iea.org/data-and-statistics/data-tables?country=WORLD (accessed May 18, 2020).

[3] Kamerbrief over stimulerings duurzame energieproductie (SDE+) 2019. Minst van Econ Zaken En Klim; 2018. http://www.nlstatieholland.nl/ministeries/ministerie-van-economische-zaken-en-klimaat/documenten/kamerstukken/2018/12/21/kamerbrief-over-stimulerings-duurzame-energieproductie-sde-2019. [Accessed 12 December 2019].

[4] Walker S, Labeodan T, Boxen G, Maassen W, Zeiler W. An assessment methodology of sustainable energy transition scenarios for realizing energy neutral neighborhoods. Appl Energy 2018;228:2346–60. https://doi.org/10.1016/j.apenergy.2018.06.149.

[5] Moraga J, Mulder M. Electrification of heating and transport: a scenario analysis of The Netherlands up to 2050. 2018.

[6] Fisher D, Madani H. On heat pumps in smart grids: a review. Renew Sustain Energy Rev 2017;70:342–57. https://doi.org/10.1016/j.rser.2016.11.182.

[7] Electricity and heat statistics. Eurostat; 2017. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Electricity_and_heat_statistics. [Accessed 12 December 2019].

[8] Walker S, Katic K, Maassen W, Zeiler W. Multi-criteria feasibility assessment of cost-optimized alternatives to comply with heating demand of existing office buildings – a case study. Energy 2019;187:115968. https://doi.org/10.1016/j.energy.2019.115968.

[9] Koirala BP, van Oost E, van der Windt H. Community energy storage: a responsible innovation towards a sustainable energy system? Appl Energy 2018;231:570–85. https://doi.org/10.1016/j.apenergy.2018.09.163.

[10] Leeuwen R van. Towards 100% renewable energy supply for urban areas and the role of smart control. University of Twente; 2017. https://doi.org/10.3990/1.5785036543460.

[11] Turecek A, Nielsen P, Madsen H. Electricity consumption clustering using smart meter data. Energies 2018;11:859. https://doi.org/10.3390/en11040859.

[12] McLaughlin F, Duffy A, Conlon M. A clustering approach to domestic electricity load profile characterisation using smart metering data. Appl Energy 2015;141:190–9. https://doi.org/10.1016/j.apenergy.2014.12.039.

[13] Liu C, Gao X, Ma W, Chen X. Research on regional differences and influencing factors of green technology innovation efficiency of China's high-tech industry. J Comput Appl Math 2020;369. https://doi.org/10.1016/j.cam.2019.112597.

[14] Mocanu E, Nguyen PH, Gibescu M, Kling WL. Deep learning for estimating building energy consumption. Sustain Energy and Grids Netw 2016;6:91–9. https://doi.org/10.1016/SECAN.2016.02.005.

[15] Darbellay GA, Slama M. Forecasting the short-term demand for electricity: do neural networks stand a better chance? Int J Forecast 2000;16:71–83. https://doi.org/10.1016/S0169-2070(99)00045-X.

[16] Seyedzadeh S, Rahimian FP, Glesk I, Roper M. Machine learning for estimation of building energy consumption and performance: a review. Vis Eng 2016;6:5. https://doi.org/10.1515/viseng-2016-00046.

[17] Shin C, Lee E, Han J, Yun J, Rheew L, Lee H. The ENERTALK dataset, 15 Hz electricity consumption data from 22 houses in Korea. Sci Data 2016;5:193. https://doi.org/10.1038/s41597-016-0212-5.

[18] Yilmaz S, Chambers J, Patel MK. Comparison of clustering approaches for domestic electricity load profile characterisation – implications for demand side management. Energy 2019;180:665–77. https://doi.org/10.1016/j.energy.2019.05.124.

[19] Yu Shchetinin E. Cluster-based energy consumption forecasting in smart grids. J Phys Conf Ser 2019;1205. https://doi.org/10.1088/1742-6596/1205/1/012051. Institute of Physics Publishing.

[20] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Springer; 2013. https://doi.org/10.1007/978-1-4614-7138-7.

[21] Jin L, Lee D, Sim A, Borgeson S, Wu K, Anna Spurlock C, et al. Comparison of clustering techniques for residential energy behavior using smart meter data. Artif Intel Smart Grids Smart Build 2017:260. https://doi.org/10.1007/978-1-4614-7138-7.

[22] Walker S, Khan W, Katic K, Maassen W, Zeiler W. Accuracy of different machine learning algorithms and added-value of predicting aggregated-level energy performance of commercial buildings. Energy Build 2019. https://doi.org/10.1016/j.enbuild.2019.10.075.09.

[23] Alexander DJ, Tropsha A, Winkler DA. Beware of \( R^2 \) : simple, unambiguous assessment of the prediction accuracy of QSAR and QSPP models. J Chem Inf Model 2015;55:1316–22. https://doi.org/10.1021/acs.jcim.5b00206.

[24] Guideline A. Measurement of energy and demand savings. 2002.

[25] Hoppmann J, Volland J, Schmidt T, Hoffmann V. The economic viability of battery storage for residential solar photovoltaic systems – a review and a simulation model. Renew Sustain Energy Rev 2014;39:1101–18. https://doi.org/10.1016/j.rser.2014.07.068.

[26] Koirala BP, Kolius E, Friege J, Hakvoort RA, Herder PM. Energetic communities
for community energy: a review of key issues and trends shaping integrated community energy systems. Renew Sustain Energy Rev 2016;56:722–44. https://doi.org/10.1016/j.rser.2015.11.080.

[27] Ashouri A, Fux SS, Benz MJ, Guzzella L. Optimal design and operation of building services using mixed-integer linear programming techniques. Energy 2013;59:365–76. https://doi.org/10.1016/j.energy.2013.06.053.

[28] Murray P, Orehounig K, Grosspietsch D, Carmeliet J. A comparison of storage systems in neighbourhood decentralized energy system applications from 2015 to 2050. Appl Energy 2018;231:1285–306. https://doi.org/10.1016/j.apenergy.2018.08.105.

[29] Piernas Muñoz MJ, Castillo Martínez E. Introduction to batteries. Springer Briefs Appl Sci Technol 2018;1–8. https://doi.org/10.1007/978-3-319-91488-6_1.

[30] Marler RT, Arora JS. The weighted sum method for multi-objective optimization: new insights. Struct Multidiscip Optim 2010;41:853–62. https://doi.org/10.1007/s00158-009-0460-7.

[31] Asaee SR, Uğursal VI, Beausoleil-Morrison I. Techno-economic assessment of solar assisted heat pump system retrofit in the Canadian housing stock. Appl Energy 2017;190:439–52. https://doi.org/10.1016/j.apenergy.2016.12.053.

[32] Home | EPEX SPOT n.d. https://www.epexspot.com/en (accessed February 12, 2020).

[33] Moro A, Lonza L. Electricity carbon intensity in European Member States: impacts on GHG emissions of electric vehicles. Transp Res Part D Transp Environ 2017. https://doi.org/10.1016/j.trd.2017.07.012.

[34] Ahmed M, Seraj R, Islam SMS. The k-means algorithm: a comprehensive survey and performance evaluation. Electronics 2020;9:1295. https://doi.org/10.3390/electronics9081295.

[35] Bernard S, Heutte L, Adam S. Towards a better understanding of random forests through the study of strength and correlation. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2009;5755:536–45. https://doi.org/10.1007/978-3-642-04020-7_57. LNBI, Springer, Berlin, Heidelberg.