KU LEUVEN

FACULTEIT INGENIEURSWETENSCHAPPEN

Optimisation of district energy simulations by means of a tailor-made clustering approach

Chadija Callebaut

Thesis voorgedragen tot het behalen van de graad van Master of Science in de ingenieurswetenschappen: architectuur

Promotor:
Prof. dr. ir. arch. Dirk Saelens

Academiejaar 2017 – 2018

Master of Science in de ingenieurswetenschappen: architectuur

KU LEUVEN

FACULTEIT INGENIEURSWETENSCHAPPEN

Optimisation of district energy simulations by means of a tailor-made clustering approach

Chadija Callebaut

Thesis voorgedragen tot het behalen van de graad van Master of Science in de ingenieurswetenschappen: architectuur

Promotor:
Prof. dr. ir. arch. Dirk Saelens

Assessor:
Prof. dr. ir. Philipp Geyer

Begeleiders:
Ir. arch. Ina De Jaeger
Dr. ir. Glenn Reynders

Academiejaar 2017 – 2018

Dankwoord

Het schrijven van deze thesis is een ontzettend leerrijke, maar ook intense ervaring geweest voor mij. Iets wat in het begin een ontzaglijk grote opdracht leek, is dan nu toch op zijn einde gekomen. Tijdens dit proces heb ik kunnen rekenen op de steun en hulp van een aantal mensen, die hierbij dan ook graag zou willen bedanken.

Eerst en vooral wil ik hierbij mijn promotor en begeleiders bedanken. Bedankt professor Saelens, Ina en Glenn, voor al jullie kostbare tijd, aanstekelijk enthousiasme en ongelooflijk goede begeleiding. Ik kon altijd bij jullie terecht met mijn vragen en kreeg hierop ook altijd enorm nuttige antwoorden. Ik heb dit jaar ontzettend veel bijgeleerd, wat ik zonder jullie niet had kunnen verwezenlijken. Bedankt voor jullie kritische blik die ervoor gezorgd heeft dat mijn thesis — zowel het onderzoek als de tekst — elke keer een pak beter werd, voor alle tips en inzichten die mij verder hebben geholpen. Bedankt voor alle interesse, vertrouwen en positieve feedback. Kortom, bedankt voor de beste begeleiding die ik me kon wensen.

Daarnaast wil ik ook mijn ouders bedanken. Bedankt mama en papa, om me zo goed als mogelijk te helpen waar jullie konden. Ik kon elke keer met mijn problemen en vragen naar jullie komen om deze te bespreken. Bovendien hebben jullie me ook ontzettend hard geholpen met het zoeken naar een oplossing voor het feit dat ik in de plaats van mijn vertrouwde Mac een Windows computer nodig had en nadien met de problemen die opdoken bij het installeren van de nodige programma's. Bedankt papa, om altijd vol aandacht mijn teksten na te lezen, grammaticale fouten aan te duiden en verbeteringen voor te stellen, om mijn 'test' te zijn of alle begrippen wel duidelijk genoeg beschreven waren. Bedankt om me zo goed te steunen tijdens het maken van deze thesis.

En tot slot wil ik ook mijn vrienden bedanken, want naast het werken aan deze thesis had ik natuurlijk ook behoefte aan de nodige ontspanning. Bedankt Amélie, Tobias, Tom en Wouter dat ik bij jullie terecht kon met mijn problemen of vragen in verband met mijn thesis, maar zeker en vast ook voor alle gezellige momenten van ontspanning. Onze babbeltjes na de les die toch altijd langer duurden dan gepland, onze middagen in de Alma waar we serieus wat gelachen hebben en onze donderdagavonden die ik nooit meer zal vergeten, hebben van dit jaar echt een leuke periode gemaakt. Samen onze gedachten even verzetten van onze thesissen en dromen over de vooruitzichten aan deze zomer was exact wat ik nodig had om er nadien weer volop tegenaan te kunnen gaan. Ik ga deze momenten zeker missen volgende jaren, wanneer het niet meer zo vanzelfsprekend zal zijn om elkaar bijna dagelijks te zien of te horen.

Chadija

Abstract

As integrated measures become more and more prevalent to accomplish energy-efficient districts, the tools to evaluate newly developed technologies are evolving from the scale of one individual dwelling to the scale of an entire district or even larger. Although smart metering and large-scale monitoring would provide the most accurate view of the actual situation, these data are mostly not available due to privacy issues. Moreover, these models are not well suited to examine new technologies since they cannot be evaluated before the application. Instead, district energy simulations are often employed, making use of data on the building geometries and additional information about amongst others the building envelope qualities. Since simulating every dwelling individually requires a lot of data — which was, until recently, not always available —, various researchers have examined methods to simplify the model. Two of these methods comprise the application of archetypes — such as the European research project TABULA developed for amongst others Belgium — or sample buildings.
As the data availability is expanded and the computational capabilities are increasing, detailed simulations of every dwelling in the district individually have become a feasible approach. Especially, since simulations based on the TABULA archetypes can cause a significant reduction in precision compared to the results of detailed simulations — based on the complete GIS data approach. However, these simulations still require a considerably long calculation time.

A reduction of this calculation time could be a significant incentive for studies based on district energy simulations. Therefore, an improved method is developed and investigated in this work: the tailor-made clustering approach. Based on the available geometric and statistical data, the building stock of the examined district is divided into specific tailor-made clusters of which one representative building is selected and simulated. The results of these simulations are subsequently scaled up to each represent the entire corresponding cluster. The main goal of this work is to examine whether district energy simulations based on tailor-made clusters generate sufficiently accurate simulations compared to the complete GIS data approach without drastically increasing the complexity of the model.
First, the implementation of the approach is examined. Since the potential variables — based on which the clustering can happen— are highly depending on the data availability and the specific conditions of the particular case study, not all previously applied variables in other studies are qualified. The potential variables in this case study are explained and the sets that generate the most accurate results for each defined use case are selected based on a linear multivariate regression of the results of the model based on the complete GIS data approach.
To select the most suited cluster technique, the two most used techniques — k-means clustering and hierarchical agglomerative clustering — are compared to each other. K-means clustering appeared to generate the most accurate results in every examined use case.
Then, the accuracy of the tailor-made clustering approach is analysed for each use case and compared to reported errors in some other studies which are making use of archetypes or sample buildings. The developed approach generates estimations with errors of about ten percent when the stock is still divided into a limited number of clusters. Although this is not proven since the case studies are different, the developed approach seems to generate more accurate estimations of the simulations based on the complete GIS data approach than the approaches that make use of archetypes or sample buildings. The errors made by the developed approach can be placed into perspective by mentioning the errors made by the simulation itself, which are of the same order of magnitude.

Samenvatting

Naarmate geïntegreerde maatregelen steeds vaker voorkomen om energie-efficiënte wijken te realiseren, evolueren ook de *tools* om nieuwe technologieën te beoordelen, van de schaal van één individuele woning naar de schaal van een volledige wijk of zelfs groter. Ondanks dat slimme metingen en monitoring op grote schaal het meest accurate beeld van de werkelijke situatie zouden vormen, zijn deze gegevens meestal niet beschikbaar omwille van de *privacy*. Bovendien zijn deze modellen niet het meest geschikt om nieuwe technologieën te beoordelen, aangezien ze deze niet kunnen evalueren voor de toepassing ervan. In de plaats daarvan worden vaak energiesimulaties van wijken toegepast, waarbij gebruik gemaakt wordt van gegevens over gebouwgeometrieën en bijkomende informatie in verband met onder andere de kwaliteit van de gebouwschil. Aangezien het simuleren van elke woning afzonderlijk een grote hoeveelheid gegevens vereist — die tot voor kort niet altijd beschikbaar waren —, is er veel onderzoek gedaan naar methodes om het model te vereenvoudigen. Twee van deze methodes houden de toepassing in van archetypes — zoals bijvoorbeeld het Europese onderzoeksproject TABULA voor onder andere België ontwikkelde — of van voorbeeldgebouwen.
Aangezien de beschikbaarheid van gegevens uitgebreid is en de verwerkingscapaciteit van computers aan het toenemen is, worden gedetailleerde simulaties van elke individuele woning in de wijk een haalbaardere aanpak. Zeker omdat de simulaties gebaseerd op de TABULA-archetypes een significante reductie van de precisie kunnen veroorzaken in vergelijking met de resultaten van de gedetailleerde simulaties — gebaseerd op de 'volledige GIS data aanpak'. Deze simulaties vereisen echter nog steeds een aanzienlijke rekentijd.

Een vermindering van de rekentijd zou een significante stimulans kunnen betekenen voor studies gebaseerd op energiesimulaties van wijken. Daarom is er een verbeterde methode ontwikkeld en onderzocht in dit werk: de 'op-maat-gemaakte cluster aanpak'. Op basis van de beschikbare geometrische en statistische gegevens, wordt het gebouwenpark van de onderzochte wijk in specifiek gevormde clusters onderverdeeld, waarvan er telkens één representatief gebouw geselecteerd en gesimuleerd wordt. De resultaten van deze simulaties worden vervolgens opgeschaald, zodat ze elk de bijhorende volledige cluster vertegenwoordigen. Het hoofddoel van dit werk is het onderzoeken of energiesimulaties van wijken gebaseerd op specifiek gevormde clusters voldoende accurate simulaties kunnen genereren, vergeleken met de 'volledige GIS data aanpak', zonder de complexiteit van het model drastisch te verhogen.
Eerst wordt de implementatie van de aanpak bestudeerd. Aangezien de potentiële variabelen — op basis waarvan de clustering kan gebeuren — sterk afhankelijk zijn van de beschikbare gegevens en de specifieke condities van de *case study*, komen niet alle variabelen, die eerder gebruikt werden in andere studies, in aanmerking. De potentiële variabelen voor deze *case study* worden toegelicht en de sets die de meest accurate resultaten opleveren voor elke gedefinieerde *use case* worden geselecteerd aan de hand van een meervoudige lineaire regressieanalyse van de resultaten van het model gebaseerd op de 'volledige GIS data aanpak'. Om de meest geschikte clustertechniek te bepalen, worden de twee meest toegepaste technieken — *k-means* clustering en agglomeratieve hiërarchische clustering — met elkaar vergeleken. *K-means* clustering blijkt de meest accurate resultaten te genereren in elke onderzochte *use case*.
Daarna wordt de accuraatheid van de 'op-maat-gemaakte cluster aanpak' geanalyseerd voor elke *use case* en vergeleken met gerapporteerde fouten in een aantal andere studies, die gebruik maken van archetypes of voorbeeldgebouwen. De ontwikkelde aanpak genereert benaderingen met fouten van ongeveer tien procent wanneer het gebouwenpark nog steeds in

een beperkt aantal clusters is onderverdeeld. Ondanks dat het niet bewezen is, omdat de *case studies* verschillen, lijkt de ontwikkelde aanpak een accuratere benadering te genereren van de simulaties gebaseerd op de ‘volledige GIS data aanpak’ dan de aanpak met archetypes of voorbeeldgebouwen. De fouten gemaakt door de ontwikkelde aanpak kunnen in perspectief geplaatst worden door de fouten aan te halen die veroorzaakt worden door het simuleren zelf, welke van dezelfde grootteorde zijn.

Contents

1. Introduction

1.1 Context and problem statement

In the current circumstances of climate change, environmental pollution and depletion of fossil fuel resources, energy efficient dwellings and renewable energy resources have become more and more important topics. The International Energy Agency (2017) reported that between 1971 and 2015 the total final energy consumption more than doubled and that the residential sector is responsible for 22 percent of this consumption (Figure 1). The residential sector is hence the third largest consumer of all sectors. To reach the by politics imposed targets (United Nations Framework Convention on Climate Change, 2015), integrated measures become more prevalent in addition to the previously already established energetic renovations, such as supplementary insulation of individual buildings.

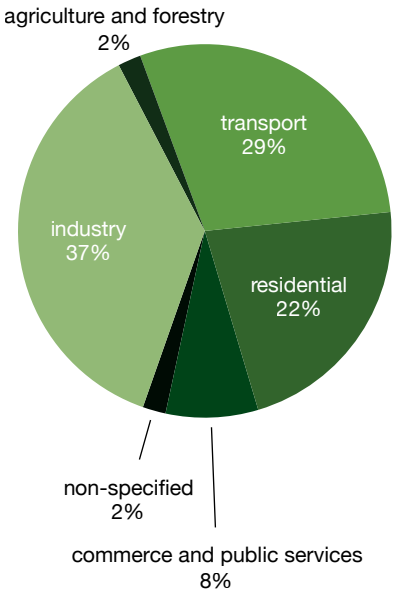


Figure 1. Total final energy consumption by sector in 2015 (International Energy Agency, 2017).

Renewable energy resources such as solar panels or wind energy are increasingly used for both distributed and centralised generation. Furthermore, smart electricity grids or district heating networks, as ways to plan the energy supply, are more and more explored to optimally fulfil the demand at every moment (Sokol, Cerezo Davila and Reinhart, 2017). Smart electricity grids are energy networks relying on digital communication in order to integrate distributed energy storage and demand side flexibility, so that the balance between the fluctuating demand and supply can be optimised. Demonstrations of these smart grids on a large scale are, according to the International Energy Agency (2011), urgently needed to explore solutions that can be implemented on the existing electricity infrastructure. The fourth generation of district heating networks are presented as the thermal equivalent of these smart electricity grids. District heating networks redistribute local heat resources, that otherwise would be wasted, to meet the local heat demands (Werner, 2017). As the temperature is significantly reduced and storage is added in the fourth generation thermal networks, they are able to harvest low-cost and CO_2-neutral production resources. Therefore, these networks show strong potentials to be possible heat supplies in the future world but the assessment and implementation hereof need to be broadened in order to amass global benefits (Werner, 2017). The implementation of integrated measures — shifting from the scale of one individual dwelling to the scale of an entire district or even larger — thus will become impossible to disregard in the cityscapes of the near future. The design operation and optimisation of future energy systems is therefore becoming increasingly complex and the analysis of energetic behaviours is also evolving from the scale of an individual building to the scale of an entire district. Likewise, the tools to evaluate possible new measures will have to follow this evolution in scale.

Formerly, the main focus was on tools to assess individual dwellings during the design phase or refurbishments in order to meet the energy performance regulations of buildings (EPB) (Allacker, 2010). Whereas these tools were sufficient to evaluate the individual measures, the upcoming integrated measures demand more integrated evaluation tools. As Allegrini et al. (2015) explained: “neither the building nor the system can be fully understood in isolation”. Therefore, the dwellings must be considered as elements in the urban energy systems to

evaluate the interactions between the buildings and the system. Integrated district energy simulations are consequently gaining importance. Various studies concentrate on models and modelling approaches to assess district energy systems (Allegrini et al., 2015; Kavgic et al., 2010; Swan and Ugursal, 2008). The techniques to model district energy consumptions consist of different approaches, which can be divided into two categories: top-down and bottom-up techniques (Swan and Ugursal, 2008). Top-down techniques make use of data on the total energy consumption of the residential sector to evaluate the characteristics of the sector as a whole, while bottom-up techniques approach the modelling on the scale of individual dwellings. The results of the energy consumptions of one dwelling or a group of dwellings are then extrapolated to a region or nation. Since top-down techniques concentrate on the interaction between the energy sector and the economy and lack information on technological details, bottom-up techniques are better suited to evaluate technological measures (Kavgic et al., 2010).

These bottom-up techniques can be further divided in statistical models and engineering models or models based on building physics (Kavgic et al. 2010; Swan and Ugursal, 2008). Where statistical models rely on historical data, physical models estimate the energy consumption based on quantitative data of physically measurable variables. Smart metering and large-scale monitoring to gather data on the individual energy consumption would generate the most accurate view of the actual consumption but this data is rarely available due to privacy issues (Ghiassi and Mahdavi, 2016a; b). However, the usage of this data is not always an appropriate approach for evaluations of new measures or technologies, because only the existing situations can be examined but new technologies cannot be evaluated before the application (Swan and Ugursal, 2008). Due to these drawbacks of using measured data, simulations of the integrated scenarios are often employed. The scenarios can be simulated using data on the building geometry with additional data on the quality of the building envelope and the user profiles. Data on the building geometry can be obtained from geospatial data models, which are embedded in geographic information systems (GIS). For example, Caputo, Costa and Ferrari (2013) made use of the “buildings’ map” for information on the geometries of the Italian building stock and the national standards for additional data on amongst others the building envelope materials and buildings’ systems to simulate the impact of energy policies. Another example out of the various studies is the work of Page, Dervey and Morand (2014), for which they are using GIS datasets for information on amongst others the year of construction, the building use and footprint and national building regulations to specify the building fabrics and user profiles.
Making use of GIS data, both static and dynamic district energy simulations can be executed. Where static simulations generate monthly or yearly values of the energy consumption of the dwellings, dynamic energy simulations offer the opportunity to consult the energy consumption in more detail and examine interactions in the system (Remmen et al., 2016). That way the time dimension is also included in the simulations, enabling to examine how long a certain amount of energy is used within a district — using a load duration curve — and when — through the energy demand profile.

To execute bottom-up energy simulations, a paramount quantity of data is required. First, the geometry input data must contain information about the building envelope areas, orientations and inclinations and the window-to-wall ratios. Besides that, additional data can be used to determine the non-geometric building properties. To examine a district, either all the data on the buildings in this district has to be collected or a simplified representation of the district must be introduced. Since the process of data collection for larger amounts of buildings is an intensive step (Reinhart and Cerezo Davila, 2016) and the availability of the necessary

information was inadequate until recently, researchers have developed methods to simplify the representation of the building stock, using for example archetypes or sample buildings. The first method comprises the application of archetypes. Hereby, the building stock is subdivided into groups of buildings that exhibit a similar energetic behaviour, based on a specific set of energy performance-related variables (Swan and Ugursal, 2008). This set of variables differs for each research, depending on the specific conditions and available data. Each group of buildings is exemplified by an archetype — a not-necessarily-existing building or "building definition" (Reinhart and Cerezo Davila, 2016) which approaches the average of all the buildings in the group and represents them all. The energy-related properties of the archetypes are determined and the results hereof are scaled up by multiplying them by the number of buildings in the groups (Swan and Ugursal, 2008). The other method to simplify the building stock contains the usage of sample buildings. Instead of non-existing archetypes, this technique makes use of actual sample buildings to represent the entire building stock (Swan and Ugursal, 2008). The sample buildings are simulated and the results — appropriately weighted — are used to estimate the energy consumption of the regional or national housing stock. Since archetypes are predefined, they can only contain a limited variety. In contrast to the use of sample buildings, with which a higher degree of variety can be reached if the sample size is sufficiently large. Nevertheless, the application of this technique has been limited, according to Swan and Ugursal (2008), since it is data intensive.

An example of the archetype approach, which is employed in Belgium, is the European research project TABULA. TABULA has specified archetypes for 13 different nations, which are operable in the whole country (Cyx et al., 2011). For Belgium, TABULA contains 30 archetypes developed by VITO, which are considered as "typical" for the Belgian context. These archetypes are defined based on six construction periods and five building types of which three single-family dwellings and two apartments (Cuypers et al., 2014). The distinction between these single-family housing types is made based on the number of neighbours, as the archetypes are specified as a detached, semi-detached and terraced house. The distinction between the apartment types is mainly made based on compactness, as Cuypers et al. (2014) stated that the factor with a major influence on the energy consumption of an apartment is the total area through which the apartment is in contact with the outdoor climate. Therefore TABULA distinguishes a strongly enclosed and an exposed apartment. As Protopapadaki, Reynders and Saelens (2014) explained, the properties of each archetype — such as the U value of various components of the building envelope, the floor area, the heated volume, the total loss area and the system properties — are derived from national statistics and should therefore represent the average of the whole country (Cyx et al., 2011).

Since the geospatial databases have been expanded over the last decades and became more accessible to the public (Reinhart and Cerezo Davila, 2016), the availability of the necessary information on the building geometry is expanded. Moreover, the capabilities of computers are increasing rapidly (Swan and Ugursal, 2008), through which detailed calculations based on the entire set of data by now have become a feasible approach. Especially because the application of archetypes can cause a considerable reduction in precision of the district energy simulations compared to these detailed calculations, which include simulations of each building individually. De Jaeger, Reynders and Saelens (2017) reported that the application of the archetypes defined by TABULA underestimated the peak power of the examined district with 26 percent for the detached dwellings and with 95 percent for the terraced buildings. The total energy use was underestimated by three to 80 percent respectively. TABULA has investigated the building stock of the whole country and defined a limited number of typical geometries. However, these geometries do not always correspond to the building geometries of each

particular district. For example, the dimensions of a terraced house in an old city centre are unlikely to be identical to the dimensions of a terraced house in a suburban neighbourhood. Since all buildings are represented as an exact copy of an archetype, simulations based on the archetypes defined by TABULA can be less representative or can result in a distorted view of a particular district. As a large difference in the results of calculations based on the TABULA approach compared to detailed calculations based on the entire set of data is demonstrated (De Jaeger, Reynders and Saelens, 2017), the detailed calculations are preferred above the TABULA approach. Since this complete dataset approach — from here referred to as complete GIS data approach — contains the detailed calculation of each dwelling individually, the disadvantage of this approach is the still considerably long calculation time.

A reduction of the considerably long calculation time with the complete GIS data approach could be a significant incentive for studies based on district energy simulations. An improved method could most likely combine the advantages of both approaches — the complete GIS data approach and the approach with archetypes or sample buildings: by making use of the available GIS and statistical data, specific tailor-made clusters could be determined for each case separately. Clusters are groups of, in this case, dwellings which show similar energetic behaviour within one cluster, whereas the behaviours differ between separate clusters (Manning, Raghavan and Schütze, 2009a). For each of these clusters, one building could be selected to represent the entire cluster, similar to the purpose of an archetype. These representative buildings should match the average of the clusters as good as possible and should therefore be the buildings the closest to the centre of each cluster. That way, only a limited number of buildings would have to be simulated to calculate the energy demand of the entire district, but these buildings would thereby represent the specific building stock of the particular case as accurate as possible.
This approach already has been applied by Ghiassi et al. (Ghiassi and Mahdavi, 2016b) to a district in Vienna, containing 750 dwellings. They used Multivariate Cluster Analysis (MCA) to divide the building stock. Subsequently, from each cluster a representative building was selected, on which detailed energy-related simulations were performed. The results from these simulations were used to represent the entire district by multiplying them by the number of buildings in each corresponding cluster. With this approach, Ghiassi et al. (Ghiassi and Mahdavi, 2016b) developed an interface between the available GIS data and the buildings standards, valid in Vienna, Austria, on the one hand and the explicit requirements for energetic simulations on the other hand. Only a limited number of buildings were specifically selected to represent the entire district. Depending on the applied cluster technique and the particular scenario, Ghiassi (Ghiassi and Mahdavi, 2017) investigated situations where the building stock was divided into six to 21 clusters with errors of the annual heating demand prediction for the entire stock of zero to 17 percent.

The clustering of the building stock must happen based on a number of variables. These variables must represent the energetic performance of the dwellings in such a way that the building stock will be subdivided into groups of dwelling which behave energetically similar. Preceding studies each made use of a different set of variables to make the distinction between the defined archetypes. The sets of variables are depending on the specific conditions and the availability of the data and on which variables have a major influence — according to the researchers — on the energetic behaviour in the particular case study. The European research project TABULA, like previously explained, made use of the construction period and the number of neighbours or the compactness to demarcate the different archetypes for Belgium. Ghiassi et al. (2015) determined the set of variables that they used, based on physical and contextual properties of the dwellings. The main terms in the heat

balance of the buildings were looked at to determine the most relevant variables. Besides those variables, they pre-divided the building stock based on the user profiles. Jones et al. (2001, in: Ghiassi et al., 2015; Swan and Ugursal, 2008) studied a county borough in the United Kingdom and for that they used age groups and several geometrical variables — like the built form, footprint, floor area, exposed end areas, height and window-to-wall ratio — but also information on the HVAC systems, the user profiles and the location to make the distinction between the archetypes. Huang and Broderick (2000, in: Ghiassi et al., 2015; Swan and Ugursal, 2008) distinguished 16 different regions in their study in the USA and made use of the construction periods, housing typologies, user profiles and the different climate zones to delimit their archetypes. Shimoda et al. (2004, in: Swan and Ugursal, 2008) defined dwelling types and household types to determine archetypes for Osaka, Japan. For another study executed in Asia — in Hong Kong —, Wan and Yik (2004, in: Swan and Ugursal, 2008) only delineated one archetype. To introduce variety, they rearranged the floor plan layout and orientation and applied various family types and user profiles to this dwelling. The use of various variables in each study shows that the selection of the variables is not unambiguously and is strongly depending on the specific conditions of the particular study but also on the availability of the data.

To execute the clustering itself, multiple techniques exist. Since the energetic behaviour of dwellings depends on numerous variables, it is necessary that the applied cluster technique is able to incorporate multiple criteria based on which the clusters will be divided (Ghiassi, 2017). Moreover Ghiassi (2017) desired to maintain control over the number of clusters and therefore employed Multivariate Cluster Analysis (MCA), which includes various techniques that divide a multivariate data space into separate homogeneous groups. Although these techniques are widely used in various scientific fields, Ghiassi (2017) stated that the potential of clustering towards building stock classification has not sufficiently been explored. For this specific application of dividing the building stock into coherent groups of which the dwellings show an energetically similar behaviour, the optimal number of clusters cannot be known in advance. Therefore cluster techniques for which the number of clusters does not need to be specified beforehand seem more interesting. However, the number of clusters will then have to be determined afterwards. After the selection and simulation of the representative buildings or archetypes, the results of these buildings must be scaled up to represent the entire building stock. This can happen by simply multiplying the results of each representative building by the number of buildings in that cluster or by taking the geometry of the individual buildings into account. In this way, the variations in energetic behaviour of the dwellings in one cluster might be met by diversifying the results by means of the building geometries.

1.2 Research questions

To reduce the considerably long calculation time for district energy simulations based on the complete GIS data approach, the improved method that combines the advantages — called the tailor-made clustering approach — is examined. The main research question is formulated as follows:

> *Can district energy simulations based on tailor-made clusters generate sufficiently accurate simulations compared to the complete GIS data approach without drastically increasing the complexity of the model?*

To be able to implement this tailor-made clustering approach, following aspects need to be investigated:
The variables based on which the clustering of the building stock will happen are depending on

both the specific conditions and the availability of the data. Therefore, following questions have been explored:

> *Which variables are used in previous studies? Are these variables also applicable with the available datasets in Belgium? And which set of variables yields the most accurate estimation of the model based on the complete GIS data approach?*

To determine which of the various cluster techniques is to be applied and if the geometry of the buildings needs to be taken into account to diversify the results, following questions are examined:

> *Which cluster technique yields the most accurate estimation of the detailed model?*
> *Which way of upscaling diversifies the results in the best way?*

To validate the approach, the accuracy of a model based on the tailor-made clustering approach needs to be examined. Therefore, the errors made by the tailor-made clustering approach compared to the complete GIS data approach are quantified exactly and following questions are posed:

> *What is the accuracy of the model based on the tailor-made clustering approach compared to the model based on the complete GIS data approach? And is this tailor-made clustering approach more accurate than a random clustering and selection of representative buildings?*

1.3 Overview

The chapters of this work follow the arrangement of the foregoing research questions.

In chapter 2, the methodology is first described. To be able to assess the developed tailor-made clustering approach, the results of this approach are compared to the results of the complete GIS data approach. A couple of scenarios or use cases are considered, which are assuming an interest in different energy key performance indicators (KPIs). The first scenario considers an interest in all three defined energy KPIs: peak power as well as energy demand for SH as specific energy demand. The second use case suggests a scenario with interest only in peak power and the third use case only in energy demand for SH. The workflow of both the complete GIS data approach and the tailor-made clustering approach are described in detail. Finally, the reference scenario is clarified. The applied data sources in this work are explained and the investigated case study of the Boxbergheide in Genk is introduced.

In chapter 3, the selection of the set of variables for each examined use case is explained. To determine the set of variables, first a literature study was performed to examine which variables have been used in previous studies. Then, the available datasets in Belgium and the specific conditions of the particular case study are reviewed. Subsequently, linear multivariate regression of the results based on the complete GIS data approach is used to select the most optimal set of variables for the defined scenarios. As verification of the election, the selected set of variables for each scenario is compared with the variances of the defined variables in the clusters when clustered based on the energy KPI(s).

In chapter 4, the clustering of the building stock and the upscaling of the results to the entire stock are explained. To cluster the building stock, two cluster techniques are most obvious to deliberate according to a literature study: k-means and hierarchical clustering. First, these techniques are described in more detail. Then, the results of simulations based on each of these techniques are compared to each other and the technique that generates the most

accurate estimation of the detailed reference scenario — being k-means clustering — is determined. Subsequently, the most accurate way of scaling the results up to the entire stock is investigated. The results of the representative buildings can be scaled up in several ways: simply by multiplying them by the number of buildings in each cluster or scaled with a geometric property of the buildings, such as the volume of the building, the floor area or the footprint. Which way of upscaling generates the most accurate approach, depends on the energy KPI itself, the use case and the number of clusters.

In chapter 5, the accuracy of the tailor-made clustering approach is discussed. The errors made by the tailor-made clustering approach are calculated as the deviations of the results based on this approach compared to the results of the complete GIS data approach. These deviations are partly depending on the 'errors' made by the generalisation of the building stock because of the clustering. Besides that, the deviations are also depending partly on the fact that the selected set of variables is only estimating the energetic behaviour of the dwellings. The fraction of the errors that is due to the clustering is examined by executing a clustering based on the results — the energy KPIs — of the complete GIS data approach. The additional fraction that is caused by the estimation of the energetic behaviour by the set of variables is determined by a comparison of these errors with the errors of the clustering based on the variables. Whether the tailor-made clustering approach is a useful approach, is verified by comparing the results with the results of a random clustering of the building stock and a random selection of representative buildings. Thereafter, the errors made by the developed approach are compared to errors made in other studies using archetypes or sample building and by detailed simulations relative to the actual situations.

Finally in chapter 6, the conclusions of this work are recapitulated and the aspects that have to be further investigated are formulated.

1.4 Scope

Since district energy simulations can involve various distinct aspects and focuses, not all of the possible elements are taken into account. The scope of this work is limited to only the residential buildings of the examined district — the Boxbergheide in Genk, Belgium —, as the residential sector is the third largest energy consumer of all sectors (International Energy Agency, 2017). Additionally, to simplify the simulations, only the energy demand for space heating is considered in this work — since heating comprises roughly half of the total energy consumption in Belgium (Heat Roadmap Europe, 2017) — and the presence of ideal heating systems is adopted. To be able to focus on the development of the approach, varying user profiles are not taken into account, but a standard occupant is simulated in each dwelling, according to the ISO 13790 standards (International Organization for Standarization, 2008). Finally, it must be stated that just one case study is examined, concentrating on three defined energy KPIs — peak power, energy demand for space heating (SH) and specific energy demand (see chapter 2). Therefore, no general conclusions can be proven that are valid for every other district or energy KPI. Moreover, the case study is impacting the findings of the steps in the implementation of the approach, like for example at the selection of the sets of variables. This will have to be taken into account when drawing conclusions.

2. Methodology

The developed tailor-made clustering approach consists of sequential steps, illustrated in the figure below (Figure 2). To divide the building stock into specific tailor-made clusters, the approach uses a set of variables derived from the available geospatial dataset and additional statistical data. When the clusters are determined, from each cluster a representative building — which is located the closest to the centre of the cluster — is selected and dynamically simulated. Finally the results of the representative buildings are scaled up to represent to entire examined district — which contains in this case study only the residential single-family houses of the analysed neighbourhood of the Boxbergheide in Genk.
The tailor-made clustering approach is assessed by comparing the results based on this approach with the results based on the detailed simulations of the complete GIS data approach. This assessment has been executed based on a couple of scenarios or use cases. These use cases assume different interests in defined energy-related key performance indicators (energy KPIs). To investigate several divergent aspects of the energetic behaviour of the dwellings, following energy KPIs have been defined: peak power [kW] — being the maximum energy demand of a dwelling needed for space heating at any specific moment —, energy demand for space heating (SH) [kWh] — being the total demand of energy required for space heating during the simulated time period of one year — and specific energy demand [kWh/m^2] — being the energy demand for space heating per surface unit of the floor area of each dwelling in order to eliminate the dependency on the size of the dwelling. The first defined scenario depicts a general study which assumes an interest in all three defined energy KPIs. The second scenario portrays a more specific study and only assumes an interest in peak power and the third only in energy demand for SH.

The following sections in this chapter explain both applied approaches in detail: first the workflow of the complete GIS data approach — which is used to calculate the results of the reference scenario — is described and then the workflow of the developed tailor-made clustering approach. Subsequently, the reference scenario is clarified. The applied data sources for the particular case study are reviewed and the ascribed user profiles are explained. Additional data on the construction periods is reported. And finally, the investigated neighbourhood is shortly presented and the results of the reference scenario are displayed.

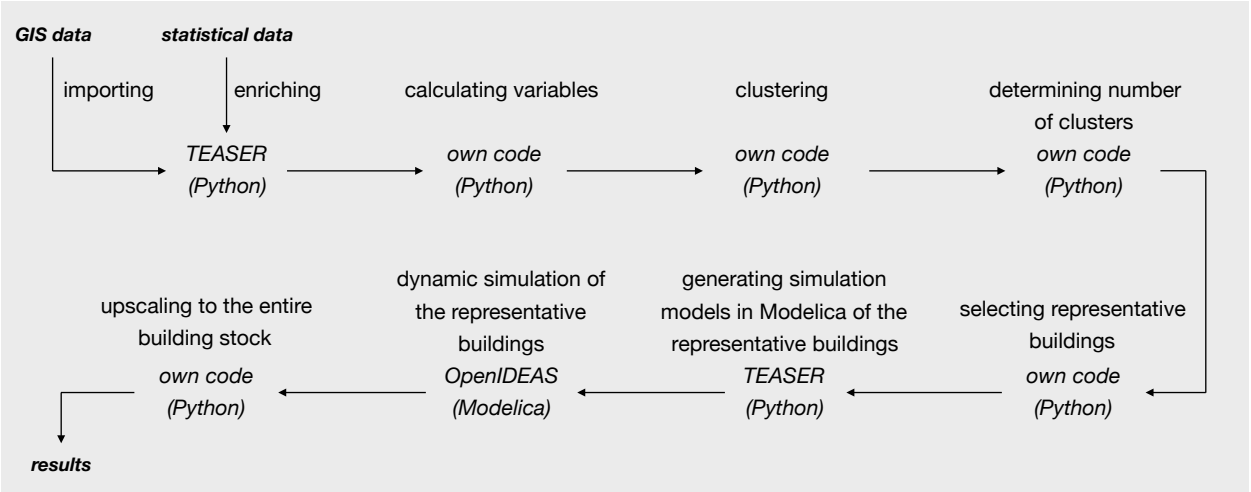


Figure 2. Workflow of the tailor-made clustering approach.

2.1 Complete GIS data approach

The complete GIS data approach comprises a detailed simulation of each individual dwelling in the building stock, therefore making use of the complete GIS dataset. The workflow of this approach is illustrated in the figure below (Figure 3) and is mainly relying on the open-source program TEASER (Remmen et al., 2016). The software tool, called “Tool for Energy Analysis and Simulation for Efficient Retrofit”, is developed to execute the conversion process from CityGML datasets to ready-to-use building energy simulation models in Modelica. As the available GIS dataset in Belgium contains information on building geometries written according to CityGML, this tool is used to translate the information on the geometries supplemented with additional statistical data to simulation models. However, some adjustments have been made to prepare the tool for application in Belgium (De Jaeger, Reynders and Saelens, 2017). The previously German statistical data have been replaced by Belgian data. In addition, the common walls of terraced and semi-detached dwellings are identified and further included in the calculations to avoid major overestimation of the energy demand of these dwellings. Finally, the export to the IDEAS library — which is explained in the next paragraph — is implemented since this library is deployed.

As the first step in the workflow of the approach, the GIS dataset is imported and enriched with statistical data on the quality of the building envelope and the user profiles. Subsequently, TEASER generates a building energy simulation model in Modelica of each dwelling individually, which is used to execute the dynamic simulations in OpenIDEAS. OpenIDEAS is an open framework developed for integrated district energy simulations (Baetens et al., 2015). The building model in IDEAS contains six main object types: a zone and five types of building components — a window, an outer wall, an inner wall, a boundary wall and a slab on the ground (Jorissen et al., 2018). For each of these object types, IDEAS has written fundamental heat transfer equations and includes variables on the characteristics of the component. Through ‘ports’, components are connected and can interact with each other. In this approach, each dwelling is modelled as a two-zone building: the ground floor is assumed to be the day zone and the upper floors the night zone. Then, OpenIDEAS is making use of the simulation program Dymola to solve the equations. For each zone, the demanded inside-temperature profile is simulated based on the ascribed user profile — which is identical for each dwelling, since the same deterministic occupant behaviour and the presence of an ideal heating system is adopted. The outside-temperature profile is derived from the average temperature profile of the specific location. With these temperature profiles and taking the building envelope characteristics into account through the variables included in the IDEAS components, the energy demand profile is calculated for each dwelling individually.
Finally, the results of these simulations are interpreted and converted into multiple key performance indicators (KPIs). The results comprise values for every ten minutes of the energy demand [kW] for space heating of the dwelling, presented as a graph showing the energy demand in function of the time. Following KPIs, related to the energy performance, are

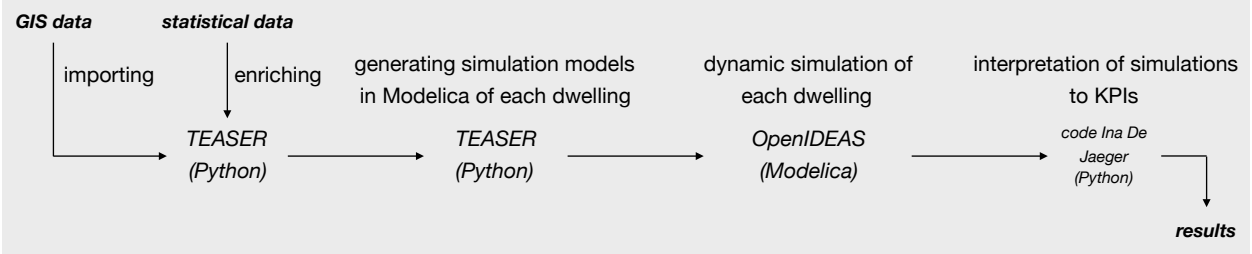


Figure 3. Workflow of the complete GIS data approach.

calculated: peak power [kW], energy demand for SH [kWh], specific energy demand [kWh/m^2] and overheating risk [Kh]. The maximum value of the results is saved as the KPI peak power. To determine the energy demand for SH, the function is integrated over the entire range of the time. Specific energy demand is calculated as the energy demand divided by the floor area of each dwelling and overheating risk is evaluated as the time that the indoor temperature exceeded 25 degrees (De Jaeger, Reynders and Saelens, 2017).

2.2 Tailor-made clustering approach

The workflow of the tailor-made clustering approach is illustrated in the figure in the introduction (Figure 2). First, the GIS data is imported and enriched with statistical data on the building envelope characteristics and the user profiles, using TEASER (Remmen et al., 2016) similar to the workflow of the complete GIS data approach. Then, some additional steps are integrated in the workflow. The set of variables based on which the clustering will happen — selected as explained in chapter 3 —, is calculated and subsequently applied to execute the clustering. To finalise the clustering, the number of clusters has to be specified and is therefore determined in function of the desired degree of accuracy — as described in chapter 5. Subsequently, from each cluster a representative building — the building the closest to the centre of the cluster — is selected. From these limited number of dwellings TEASER generates a building energy simulation model in Modelica as in the previously described workflow, which is then used to dynamically simulate the representative building. Finally, the results of the simulations of all the representative buildings are interpreted and converted into energy KPIs to then scale them up to the entire building stock.

2.3 Reference scenario

2.3.1 Applied data sources

The *Grootschalig Referentiebestand* (GRB) consists of a geographical dataset with geometric information on the building stock in Level of Detail 1 (LOD1) for the whole of Flanders, Belgium. LOD1 means that the surface of the ground floor is extruded over the average height of the building (Figure 4). For some cities, like for example Genk, there already exists a model in LOD2. This is a more detailed dataset in which the rear buildings and the pitched roofs are also modelled (Figure 4). Naturally, with this dataset a more accurate model of the real situation can be simulated. Besides this geometric information, the buildings have some attributes, such as type — main building or outbuilding —, street name or house number (De Wolf, 2017).

Figure 4. Representation of a dwelling in different Levels of Detail (LOD). (Vlaamse overheid, 2018)

Additional information is derived from statistical data. Based on the prevailing Belgian standards, TABULA defined the compositions of each construction element — like the roof, outer and inner walls, slabs on the ground and windows — for every construction period

separately (Cuypers et al., 2014). This information is subsequently used to compose the applied IDEAS library (Jorissen et al., 2018). The distinguished construction periods are from 0 until 1945, from 1946 until 1970, from 1971 until 1990, from 1991 until 2005 and from 2006 until 2011. For example, an outer wall from a dwelling built between 1946 and 1970 is defined as: eight centimetres façade masonry, ten centimetres air cavity, 14 centimetres load bearing masonry and two centimetres plaster.
As there is no data available on the actual occupant behaviours and for the sake of simplicity since varying user profiles are not the focus of this work, the user profiles are assigned based on the standards. In every dwelling, the same occupant behaviour is modelled, according to the ISO 13790 standards. Since only space heating is considered in this case study, the occupant behaviour only includes requirements on the temperature set points and the internal gains. These temperature set points consist of a desired indoor air temperature of 21 degrees in occupied periods during the days, 18 degrees during the nights and 16 degrees in unoccupied periods. The internal gains include the heat generated by other internal sources than those intentionally used for space heating — such as metabolic heat from occupants, dissipated heat from appliances and lighting devices, heat dissipated form or absorbed by water or sewage systems and by heating, cooling or ventilation systems and finally, heat from or to processes and goods (International Organization for Standarization, 2008).
Provisionally, the GRB does not contain any information about the construction period of the dwellings. The construction periods have a substantial impact on the energetic performances, since the additional information, derived from the statistical data, is assigned based on these construction periods. Thus far, every dwelling was assigned the same construction period and the construction year was set at the standard value of 1980. Because of the substantial impact on the energetic behaviour, the actual construction periods of the dwellings in the district of the case study were checked manually (Map 1).

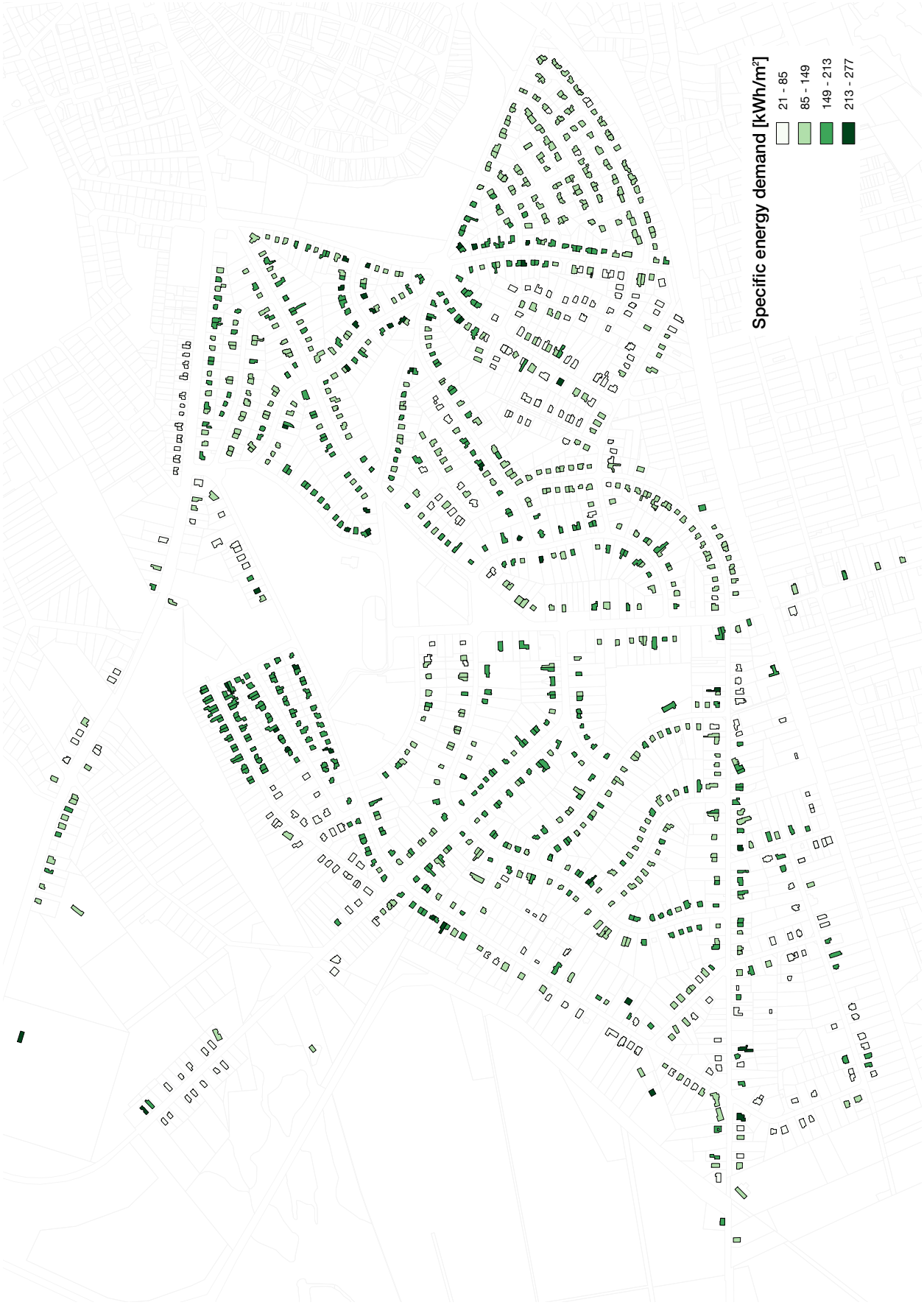
2.3.2 Case study: Boxbergheide, Genk

The case study includes the central part of the Boxbergheide in Genk. This is a residential district, mainly consisting of single-family dwellings, which is the focus of this work. Therefore, only the single-family dwellings in this neighbourhood are taken into account in this study. The case study covers 1230 dwellings with sufficient diversified orientations and all existing data of these dwellings is available. The Boxbergheide was established as a garden district shortly after the second World War to provide better living conditions for mine workers (Molemans, 1998). Because of these circumstances, most dwellings in the district were erected in the same period — between 1947 and 1970 — and have a very similar appearance. During the period of 1971 until 1990, a number of streets were completed. In some other streets, there are more modern dwellings from the latest construction period and even some parcels unbuilt (Map 1) (Verhelst, 2016; Vlaamse overheid, 2017).
The district has been simulated using the complete GIS data approach as reference scenario. The results of this simulation, interpreted to KPIs, can be seen in following maps (Map 2 to 4).

Map 1. Construction periods of single-family dwellings in the Boxbergheide.

Map 2. Peak power [kW] for space heating of single-family dwellings in the Boxbergheide.

Map 3. Energy demand [kWh] for space heating of single-family dwellings in the Boxbergheide.



Map 4. Specific energy demand [kWh/m²] for space heating of single-family dwellings in the Boxbergheide.

3. Variables

As stated in the introduction, a lot of research initiatives already used several variables to demarcate their archetypes. But since each initiative used a different set of variables, the selection of the set of variables to cluster the building stock is not unambiguously defined. The method applied in this work to select the most accurate set of variables out of the possible options for each examined use case, is explained in this chapter.
First, a literature study was performed and a summary of the applied variables in a collection of relevant studies is given to determine all the possible variables. Subsequently, the available datasets in Belgium and the specific conditions of the particular case study in a suburban Flemish neighbourhood are reviewed and the potential variables for this case study are described. Then, linear multivariate regression of the energy KPIs of the reference scenario — calculated with the complete GIS data approach — is executed to determine a ranking of the defined variables for each KPI individually. The striking aspects in the ranking of the variables are subsequently discussed. Since the optimal number of variables cannot be deduced from the linear regression, a number of sets of variables, derived from this linear regression, are compared to each other by stacking out the root mean square errors in function of the number of clusters. Finally, the selected sets of variables for each use case are verified by comparing the variances of the variables in the clusters from a clustering based on the energy KPIs of the reference scenario with the variances of the variables in the entire building stock. Generally, the variables in the sets, based on which the clustering will happen, tend to have a smaller distribution in the clusters from the clustering based on the energy KPIs, which is confirming the influence of these variables on the KPIs.

3.1 Literature study

Several researchers have already defined archetypes based on a set of variables. These sets of variables were different in each study, adequate for the specific conditions and case study. For example, the European research project TABULA made, for the Belgian building stock, the distinction between the archetypes based on construction periods — with corresponding insulation standards — and typologies (Cuypers et al., 2014). Ghiassi et al. (2015) opted for the following five variables, based on a study of heat balances of dwellings: effective average envelope U value, effective window-to-wall ratio, thermal compactness, heated volume and effective floor height. Huang and Broderick (2000, in: Swan and Ugursal, 2008) defined archetypes based on the housing typology, construction period, use and climate zone. Caputo, Costa and Ferrari (2013) used more variables to demarcate the archetypes: the construction period, compactness, heated volume, total loss area, floor area, number of stories, window area and the use. While Page, Dervey and Morand (2014) limited the variables to the construction period and the use. In this way, 32 relevant studies have been enumerated by Ballarini, Corgnati and Corrado (2014), Ghiassi et al. (2015) and Swan and Ugursal (2008).
All the applied variables in these studies can be subdivided into a number of categories (Table 1). Most of these variables are geometric, but not all of these are directly derivable from GIS data. For instance, when working with an LOD2 model, the window-to-wall ratio is not only depending on the geometric variable total loss area, but is also determined by the construction period based on statistical data. Besides geometric variables, researchers also used environmental factors — such as climate zone or exposure —, operational parameters — dealing with the use and occupancy —, present heating and cooling systems, thermal quality of the building envelope and other variables — such as the construction period and function of the ground floor. Most frequently applied variables in the reviewed studies are typology or built form, construction period, use and climate zone (Table 1).

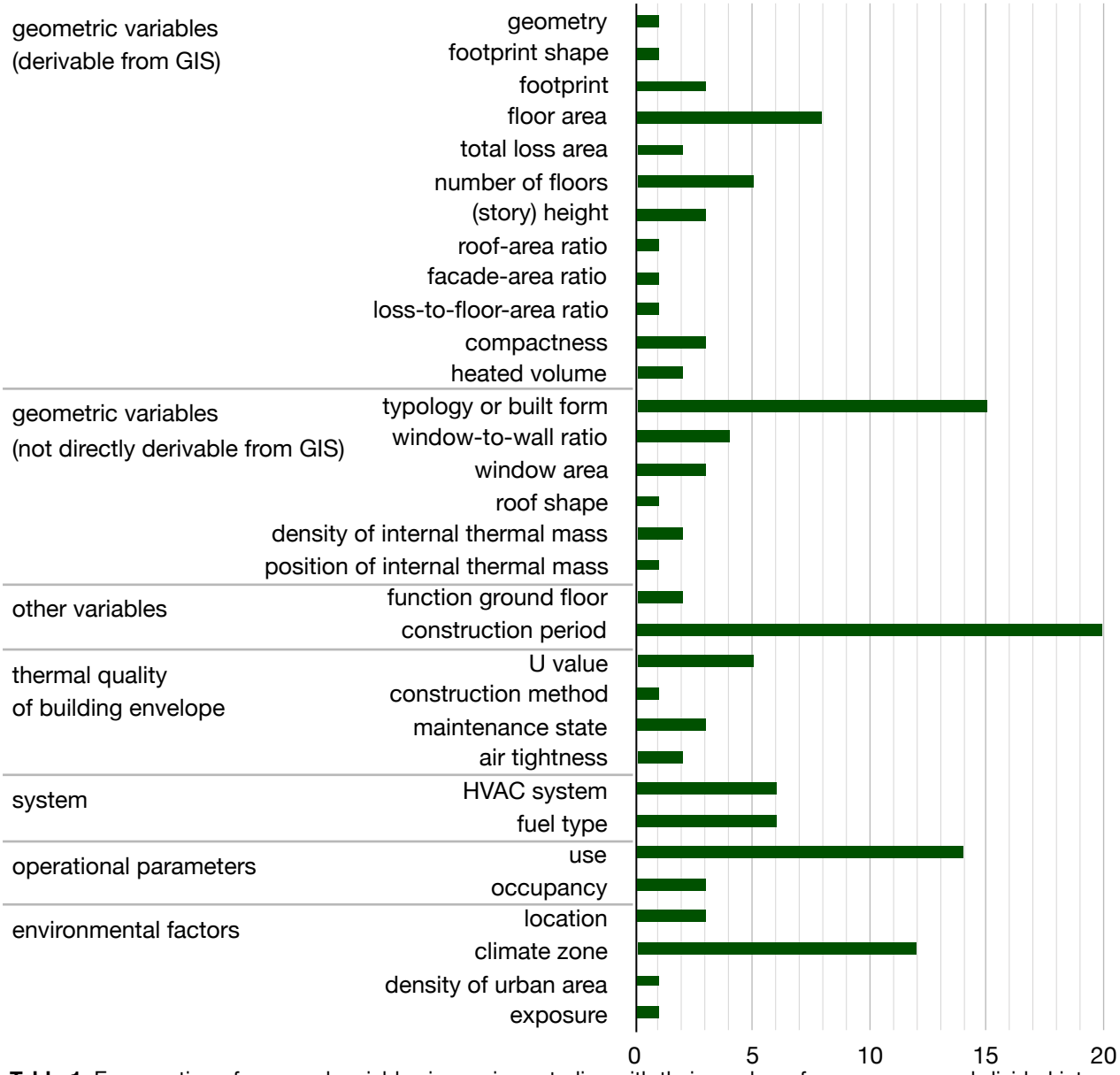


Table 1. Enumeration of occurred variables in previous studies with their number of appearances, subdivided into categories. (Ballarini, Corgnati and Corrado, 2014; Ghiassi et al., 2015; Swan and Ugursal, 2008)

3.2 Selection of sets of variables

Due to the restricted availability of the data and the environmental conditions, not all variables applied by other researchers can be used in this particular case. As the scope of this work is limited to modelling the same occupant behaviour in every dwelling following the ISO 13790 standards (International Organization for Standardization, 2008), the operational parameters — use and occupancy — cannot be applied. The environmental factors can also be disregarded, since there is no distinction in separate climate zones in Belgium and the density of the urban area and the exposure is negligible in this case study, which is typical for suburban neighbourhoods in Belgium. Since only the energy demand required for space heating is taken into account and ideal systems are adopted in this work, the variables dealing with the systems do not apply. Furthermore, there is no information available about the maintenance state and air tightness of the buildings. As the focus of this work is on single-family dwellings, all buildings have the same function of the ground floor — private dwellings — and construction method — typical Flemish brick walls which only differ in composition depending on the various construction periods. The models in LOD2 do not contain information about the density or position of the internal thermal mass and all stories are estimated to have the same height.

Based on the remaining variables, following eleven potential variables have been defined:

Footprint, which contains the surface area of the ground floor of the dwelling;

Floor area, which is the surface area of all the floors of the dwelling accumulated;

Total loss area, the total area of the building envelope through which the dwelling is in contact with the outside climate;

Number of stories, calculated as the height of the dwelling divided by three meters — the average height of one story —, if this variable is not specified in the GIS dataset;

Heated volume, the entire inside volume of all the dwelling parts, since there is no information on which spaces are heated and insulated and which not;

Compactness, calculated as the total loss area divided by the heated volume;

Loss-to-floor-area ratio, another way to express the compactness of the dwellings, calculated as the total loss area divided by the floor area;

Typology, which makes a distinction between detached, semi-detached or terraced houses, in this case determined by the number of neighbours;

Window area, which is the surface area of all window openings in the entire building envelope accumulated;

Window-to-wall ratio, calculated as the window area divided by the total loss area;

Construction period, which contains the periods defined by TABULA.

To determine which combinations of these defined variables will generate the most accurate model for the three use cases, two aspects have been investigated. First, linear multivariate regression is used to compose a ranking in importance of the variables for each energy KPI. Since this method cannot be used to conclude the optimal number of variables for the use cases, the optimal number is determined by comparing the root mean square errors in function of the number of clusters for each possible number of variables.

3.2.1 Linear multivariate regression

The first step consists of a linear multivariate regression analysis for each energy KPI individually. Applying the method of ordinary least squares (OLS), the model is given by the equation below (Hutcheson, 2011):

$$y_i = \sum_{p=1}^{n} a_p x_{ip} \ + \epsilon_i$$

with a_p the regression coefficient for every predictor p or variable, x_{ip} the value of the variables for dwelling i, ϵ_i the unobserved scalar random variables or deviations for dwelling i and y_i the estimated value of the investigated energy KPI for dwelling i.
Subsequently, the presumable accuracy of all possible models is weighted relative to the other models for that energy KPI, making use of the Akaike Information Criterion (AIC), converted by Banks and Joyner (2017) to be combined with OLS, according to following equation:

$$AIC = 2n + N \ln\left(\frac{RSS}{N}\right)$$

with n the "number of independently adjusted parameters within the model" (Akaike, 1974) or the number of variables, N the number of dwellings in the building stock and the Residual Sum of Squares (RSS) calculated with following equation:

$$RSS = \sum_{i=1}^{N} (KPI_i - y_i)^2$$

where KPI_i is the actual value of the investigated energy KPI and y_i the estimated value by the model for dwelling i.
Prior to the calculation of OLS and AIC, the mutual correlations between the variables are examined using a correlation matrix. The matrix shows the linear relationship between each two variables, calculated as Pearson correlation coefficients. When two variables show a high correlation ($\rho > 0{,}9$), only one of them is taken into account in the multivariate regression. *Heated volume* showed a high correlation with *floor area* ($\rho = 0{,}89$) and *total loss area* ($\rho = 0{,}91$) and is therefore not taken into account.

The graphs indicate that the accuracy of the models is growing with an increasing number of variables until eight variables, as the AIC is decreasing. At eight variables, the AIC reaches a minimum for every KPI, although this only comprises a very small difference of about 1 on the order of magnitude of 1000 or 10000 (Table 2). However, the graphs show a clear 'elbow' at three, four or five variables. Hence, after that certain number of variables, adding an extra variable will not decrease the AIC a lot and is argued not to be that interesting anymore. Although these graphs give the impression that the sets of eight variables will generate the most accurate model, the next paragraph will explain that there is a lower optimum number of variables. Therefore, linear multivariate regression can only be used to compose a ranking in importance of the variables for each energy KPI and not to unambiguously select the optimal set of variables for each use case. The ranking of the variables is made based on which variable is added to the most accurate combination of each number of variables and is shown in following table for each KPI (Table 2).

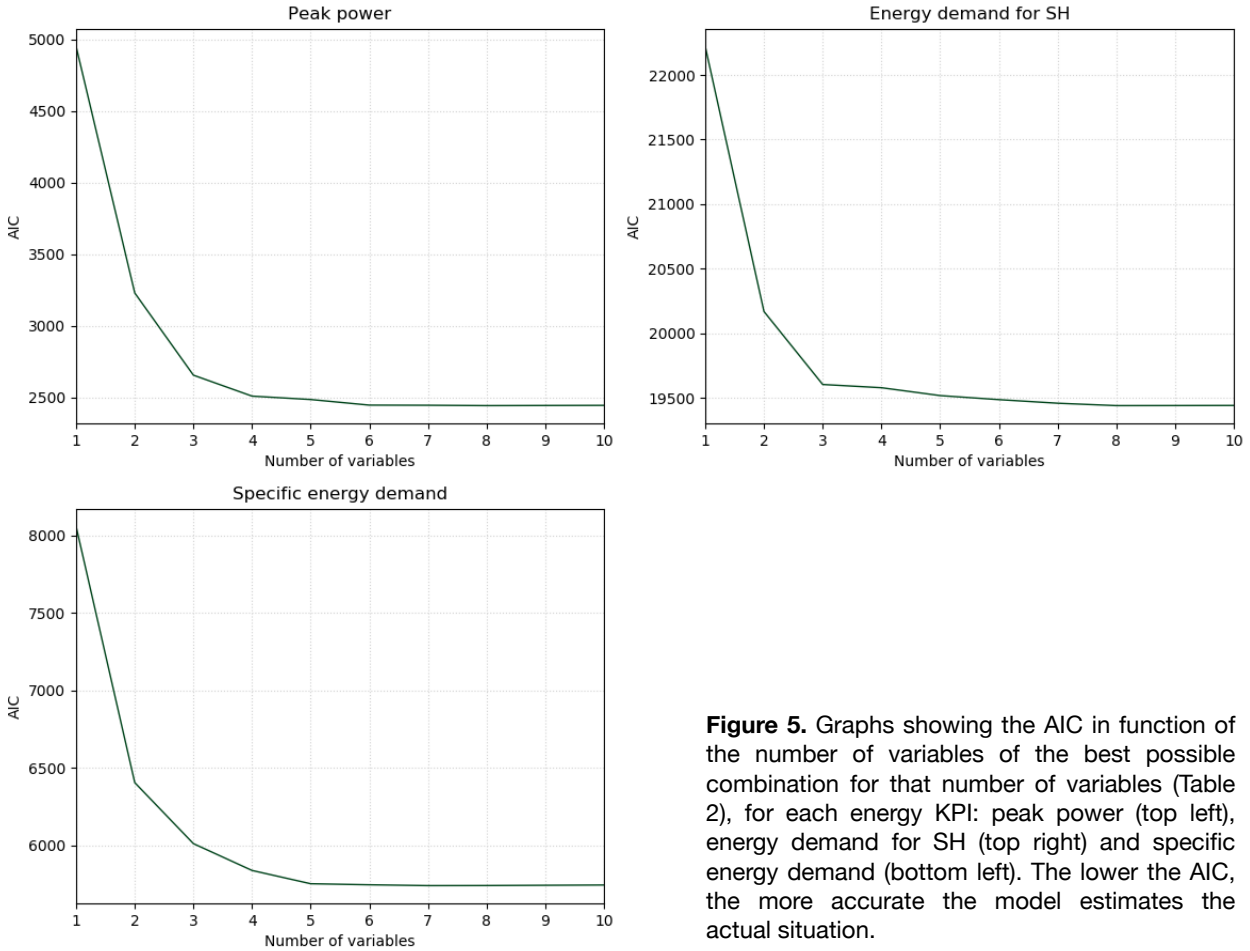


Figure 5. Graphs showing the AIC in function of the number of variables of the best possible combination for that number of variables (Table 2), for each energy KPI: peak power (top left), energy demand for SH (top right) and specific energy demand (bottom left). The lower the AIC, the more accurate the model estimates the actual situation.

number of variables (n)	best combination of variables		
	peak power	energy demand for SH	specific energy demand
1	*total loss area* (4951)	*construction period* (22221)	*construction period* (8057)
2	total loss area and *construction period* (3229)	*total loss area* and construction period (20168)	*loss-to-floor-area ratio* and construction period (6404)
3	*floor area*, total loss area and construction period (2655)	*footprint*, total loss area and construction period (19603)	*footprint*, loss-to-floor-area ratio and construction period (6010)
4	*footprint*, floor area, total loss area and construction period (2509)	footprint, *floor area*, total loss area and construction period (19578)	footprint, *total loss area*, loss-to-floor-area ratio and construction period (5836)
5	footprint, floor area, total loss area, *window-to-wall ratio* and construction period (2485)	footprint, total loss area, *window-to-wall ratio*, *window area* and construction period (19517)	footprint, total loss area, loss-to-floor-area ratio, *compactness* and construction period (5751)
6	footprint, floor area, total loss area, window-to-wall ratio, *window area* and construction period (2446)	footprint, *floor area*, total loss area, window-to-wall ratio, window area and construction period (19485)	footprint, total loss area, loss-to-floor-area ratio, compactness, *window area* and construction period (5744)
7	footprint, floor area, total loss area, *number of stories*, window-to-wall ratio, window area and construction period (2445)	footprint, floor area, total loss area, *compactness*, window-to-wall ratio, window area and construction period (19458)	footprint, total loss area, *number of stories*, loss-to-floor-area ratio, compactness, window area and construction period (5739)
8	footprint, floor area, total loss area, number of stories, *loss-to-floor-area ratio*, window-to-wall ratio, window area and construction period (2442)	footprint, floor area, total loss area, *number of stories*, compactness, window-to-wall ratio, window area and construction period (19439)	footprint, total loss area, number of stories, loss-to-floor-area ratio, compactness, *typology*, window area and construction period (5739)
9	footprint, floor area, total loss area, number of stories, loss-to-floor-area ratio, *compactness*, window-to-wall ratio, window area and construction period (2443)	footprint, floor area, total loss area, number of stories, compactness, *typology*, window-to-wall ratio, window area and construction period (19440)	footprint, *floor area*, total loss area, number of stories, loss-to-floor-area ratio, compactness, typology, window area and construction period (5741)
10	footprint, floor area, total loss area, number of stories, loss-to-floor-area ratio, compactness, *typology*, window-to-wall ratio, window area and construction period (2444)	footprint, floor area, total loss area, number of stories, *loss-to-floor-area ratio*, compactness, typology, window-to-wall ratio, window area and construction period (19440)	footprint, floor area, total loss area, number of stories, loss-to-floor-area ratio, compactness, typology, *window-to-wall ratio*, window area and construction period (5743)

Table 2. Combinations of variables with the lowest AIC for every number of variables, for each energy KPI. The additional variable(s) with respect to the previous combination of variables is (are) noted in italic. The value of the AIC —rounded to a unit— for these combinations is shown between brackets.

Construction period, *total loss area* and *footprint* appear to be important variables for all three energy KPIs (Table 2). The importance of the construction period for the energetic behaviour of a dwelling is straightforward, since several properties of the dwellings which are not included in the GIS dataset — such as the quality of the building envelope or the window area — are assigned based on the construction period. Additionally, the total loss area is another variable with an important influence on the energetic behaviour. This variable, together with the quality of the envelope, is responsible for the amount of heath being lost to the outside environment. Total loss area is somewhat less important for the specific energy demand, since the dependency on this KPI of the size of the dwellings has been eliminated by dividing by the floor area. The correlation between total loss area and floor area amounts to $\rho = 0{,}83$. As the dependency on the floor area has been eliminated, the importance of the total loss area is hence decreased. Finally, footprint emerges as a 'pseudo-variable' since this variable cannot offer an unambiguous explanation for the energetic behaviour but contains correlations with multiple other variables such as floor area ($\rho = 0{,}48$), total loss area ($\rho = 0{,}61$) and heated volume ($\rho = 0{,}56$).

The variable *floor area* has an important influence on both peak power and energy demand for SH, but not on specific energy demand. The floor area is an unambiguous indication of the size of the dwellings, which plays a crucial role in the determination of peak power and energy demand for SH. As mentioned before, the dependency on this variable has been eliminated for specific energy demand.
The variable *loss-to-floor-area ratio* appears to have an important influence in the determination of the specific energy demand, but not so much of peak power and energy demand for SH. Moreover, compactness seems as well to be more important for specific energy demand than for the two other KPIs. This can be explained by the meanings of the KPIs: where peak power and energy demand for SH depend highly on the size of the dwelling, this dependency has been eliminated for the specific energy demand. Instead, the composition of the dwelling becomes more important for the determination of this KPI.

On the other hand, both *typology* and *number of stories* appear to have little impact on none of the three KPIs. These variables tend to give insufficient information about the energetic behaviour of the dwellings. Typology — here defined as number of neighbours — does not give any indication about the size of the dwelling nor about the composition of the plan dealing with the form of the dwelling. As the shared walls are no part of the total loss area since they are considered to be adiabatic, the number of neighbours is not adding any information. Although number of stories is supposed to predict some information about the size, the correlation between number of stories and floor area amounts to $\rho = 0{,}48$ and between number of stories and heated volume only to $\rho = 0{,}23$. This variable does not seem to predict that much about the size, which can be explained by the variance in the building stock: most buildings have two or three stories and only a few one or four. As the variance of this variable might be more spread out in other case studies, the ability to represent the size of a dwelling might augment and hence the importance might increase.

3.2.2 Determination of optimal number of clusters

The impression that the sets of eight variables are generating the most accurate models — caused by the minimal values of the AIC —, appears to be incorrect when these sets are applied to cluster the building stock, as both techniques are working differently. AIC selects the combination of variables which generates the multidimensional 'plane' with the lowest

distances between the 'plane' and the actual data points. The more variables, the more nuances can be placed in the formula of OLS and therefore the more possibilities to lower the distances between the 'plane' and the data points. Although the multivariate regression denotes the sets with eight variables as those that will generate the most accurate model, not all the variables in the sets contribute as much as the others to the accuracy of the model. The variables lower on the ranking assist less to the explanation of the energetic behaviour of the dwellings than those at the top. However, the variables are not weighted when applied for clustering and all variables are considered accordingly equivalent. Consequently, when several dwellings possess values for a less influential variable that are extremely similar, these dwellings might end up in the same cluster without taking the more significant variables into account. Thus, the optimal number of clusters cannot be concluded using only multivariate regression, but also depends on the mutually relative importance of the variables.
The optimal number of clusters achieves a balance between the improvements of the model by adding an extra variable and the relative importance of the variables in terms of the contribution to the explanation of the energetic behaviour. This optimal number of variables is determined for each use case by executing the clustering for several numbers of variables, using the best combination for that number of variables (Table 2), and plotting the root mean square errors (RMSEs) in function of the number of clusters for all these combinations.

Use case Peak power

When only the energy KPI peak power is important, the set of variables can be chosen based on only the ranking of variables for this KPI. With the RMSEs in function of the number of clusters staked out in graphs, a clear preference for the set of three variables can be noticed (Figure 6). For both ways of upscaling — with the number of buildings in the clusters or scaled with the floor area of the buildings (see chapter 4) — the clustering with the set of three variables generates the lowest errors. Therefore, the set of variables applied for use case Peak power comprises next variables:

footprint, total loss area and construction period

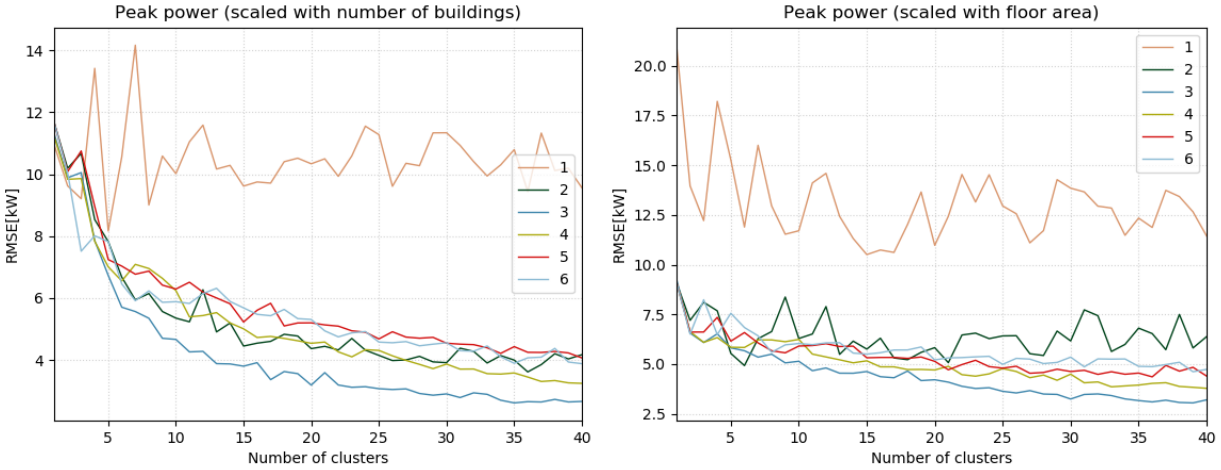


Figure 6. Graphs showing the root mean square errors in function of the number of clusters for sets of several numbers of variables for the use case Peak power. The results are scaled up with either the number of buildings (left) or scaled with the floor area of the buildings (right).

Use case Energy demand for SH

If there is only interest in energy demand for SH, the set of variables can again be chosen based on the ranking of variables for this KPI. When the various sets of variables are stacked out relative to each other, no definite preference is observed (Figure 7). If the results are scaled

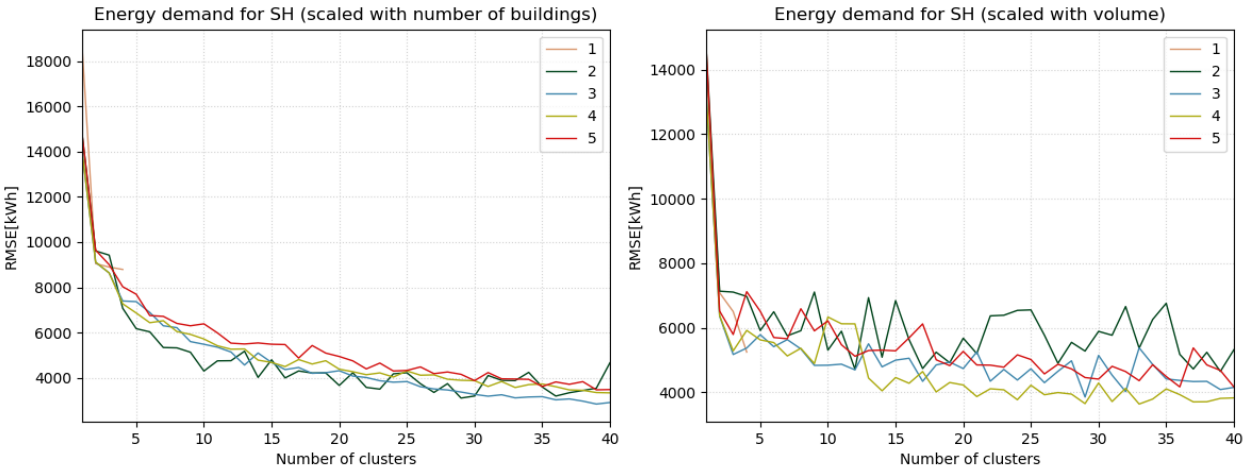


Figure 7. Graphs showing the root mean square errors in function of the number of clusters for sets of several numbers of variables for the use case Energy demand for SH. The results are scaled up with either the number of buildings (left) or scaled with the volume of the buildings (right).

up with the number of buildings, the set with two variables shows the lowest RMSEs but also an erratic course (Figure 7, left). Therefore, the set with three variables appears to be the most reliable set with the lowest RMSEs. If less than 15 clusters are required, scaling with the volume of the buildings will lower the RMSEs (see chapter 4). Also in this case the set of three variables appears to generate the most accurate and reliable model (Figure 7, right). Thus, the set of variables applied for use case Energy demand for SH comprises next variables:

floor area, total loss area and construction period

Use case Peak power, energy demand for SH and specific energy demand

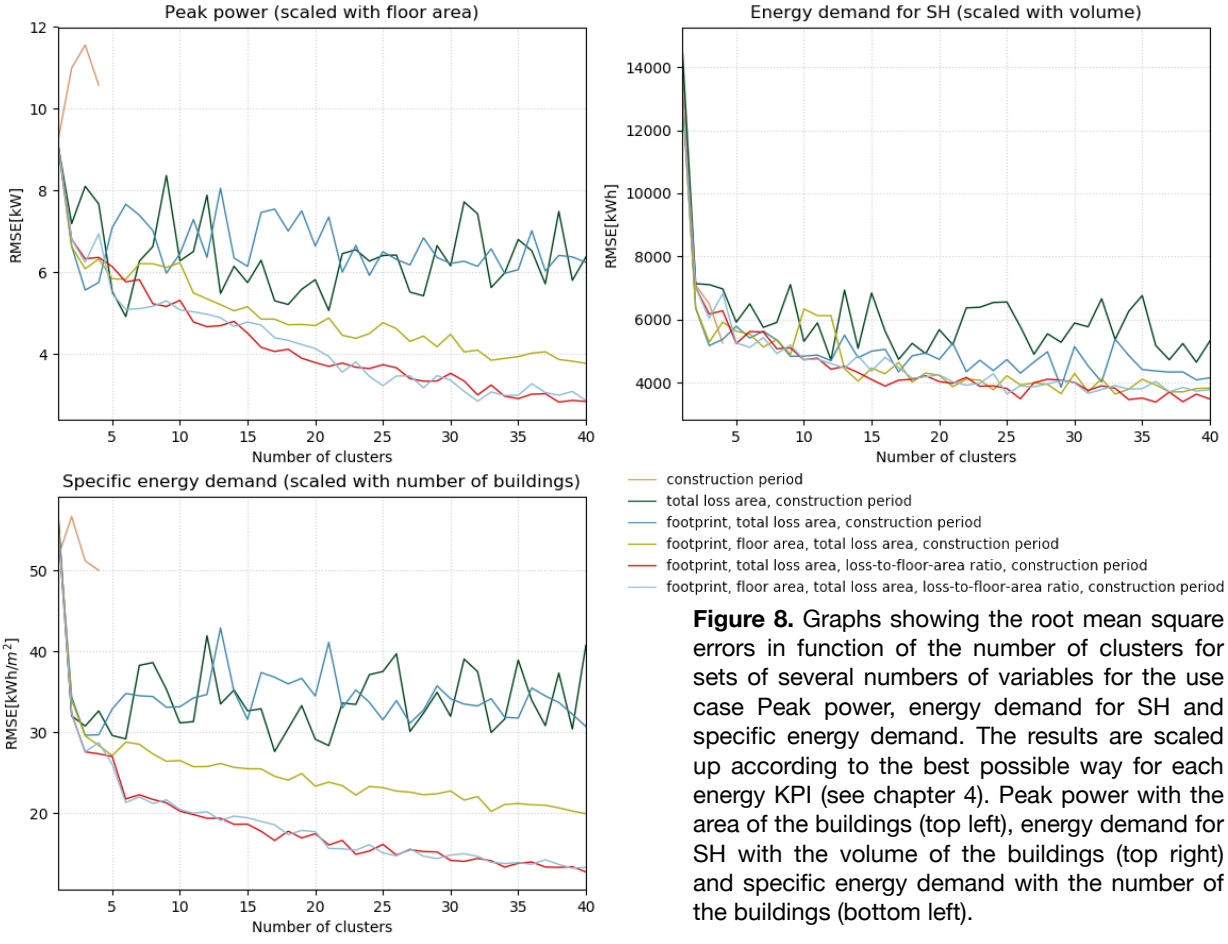


Figure 8. Graphs showing the root mean square errors in function of the number of clusters for sets of several numbers of variables for the use case Peak power, energy demand for SH and specific energy demand. The results are scaled up according to the best possible way for each energy KPI (see chapter 4). Peak power with the area of the buildings (top left), energy demand for SH with the volume of the buildings (top right) and specific energy demand with the number of the buildings (bottom left).

When all three energy KPIs are important for the research, the set of variables must be a combination of the ranking of the variables for peak power as well as energy demand for SH as specific energy demand. Peak power and energy demand for SH are both mainly depending on the size of the dwellings and therefore a similar ranking of the variables is manifested. Specific energy demand is depending more on the composition of the dwellings and introduces slightly different variables as most important. When the results for several sets of variables are stacked out, a set of four variables — which combines both the size and the composition of the dwellings — and one of five variables come forward as the most accurate sets for all three energy KPIs (Figure 8). Since adding the fifth variable does not generate lower RMSEs, the set of four variables is selected for the use case Peak power, energy demand for SH and specific energy demand and comprises next variables:

footprint, total loss area, loss-to-floor-area ratio and construction period

3.3 Verification by means of comparing variances of variables

Whether the variables which come forward as most influential based on the linear multivariate regression for each use case appear to be also most dominant in a clustering based on the results, can be verified by examining the variances of the variables in the clusters of a clustering based on the energy KPI(s) relative to the variances in the entire building stock. Since the rankings deduced from the AICs suggest which variables have the most influence on the energy KPIs, it is to be expected that the clusters of a clustering based on these energy KPIs show a narrower distribution of the most important variables. How much narrower the distribution in each cluster is relative to the distribution in the entire building stock, is calculated for each variable as follows:

$$\sigma_{ratio} = \frac{\sigma_{cluster}}{\sigma_{stock}}$$

with $\sigma_{cluster}$ the standard deviation in each cluster, σ_{stock} the standard deviation in the entire building stock and σ_{ratio} the ratio of these two standard deviations. The lower this ratio is, the smaller the distribution in the cluster.

Additionally, a ratio of the differences in means is determined for each variable according to the formula below:

$$\Delta\mu_{ratio} = \frac{\mu_{stock} - \mu_{cluster}}{\sigma_{stock}}$$

where μ_{stock} is the mean of the entire stock, $\mu_{cluster}$ the mean of each cluster individually and σ_{stock} the standard deviation of the entire stock. In this way, $\Delta\mu_{ratio}$ contains the ratio of how far the mean of a cluster deviates from the mean of the entire stock divided by the average deviation in the entire stock. The more this ratio diverges from zero, the more the distribution in the cluster is located in the extreme values of the distribution of the entire stock —below zero, at the maximum values and above zero, at the minimum values.

Which variables have the most divergent distributions can be determined using the t-test by stating as null hypothesis — which should be contradicted — that the mean of $\sigma_{ratio} - \mu_0 -$ equals one. The actual mean of the σ_{ratio}s is determined as follows:

$$\mu = \frac{1}{N}\sum_{i=1}^{N}\sigma_{ratio,i}$$

with N the number of dwellings in the entire building stock and $\sigma_{ratio,i}$ the ratio of the standard deviation for the cluster to which dwelling i belongs. In this way, the number of dwellings in each cluster is taken into account, so that more weight is given to the larger clusters.

Subsequently, the t-value is calculated for each variable according to following formula:

$$t = \frac{\mu - \mu_0}{\sigma}\sqrt{N}$$

with σ the standard deviation of all the $\sigma_{ratio,i}$s. The lower the t-value, the less likely the null hypothesis is valid. This means that the mean of σ_{ratio} is lower than one — the lower the t-value, the lower the σ_{ratio}s and thus the more the distributions in the clusters are smaller than the distribution in the entire building stock. Therefore, the variables with the lowest t-values are the most dominant in the clustering based on the energy KPI(s).

For each use case and a certain number of clusters, the ratio of the standard deviation and the ratio of the difference in mean has been calculated per cluster for every defined variable (Figure 9 to 11). Then, the most dominant variables were searched using the t-test (Table 3 to 5). The findings with these results are more elaborately discussed in the following paragraphs.

3.3.1 Use case Peak power

The preference for the set of three variables is quite apparent in this use case. When the stock is divided into 15 clusters, the difference in performance of the investigated sets is clearly noticeable (Figure 6). Therefore, the selected set of three variables is verified for a division of the stock in 15 clusters. When the σ_{ratio}s are stacked out, ordered by a decreasing number of buildings in the clusters, it can be noted that the most of the clusters have a σ_{ratio} significantly lower than one for the following variables: floor area, total loss area, heated volume, window area and construction period (Figure 9). The largest clusters show the lowest σ_{ratio}s for the variables floor area and window area. For the variables footprint, floor area, total loss area, heated volume and window area the $\Delta\mu_{ratio}$ is diverging the most from zero but for clusters of average to small sizes. The variables with the lowest t-values are floor area, total loss area, heated volume — which is related to floor area ($\rho = 0{,}89$) and total loss area ($\rho = 0{,}91$) — and construction period (Table 3). This finding largely confirms the ranking based on AICs and the selected set of variables.

3.3.2 Use case Energy demand for SH

The preference for the set of three variables in this use case is not as apparent as the previous use case. But when the stock is divided into 25 clusters, this set generates noticeably lower RMSEs than the other investigated sets (Figure 7). The selected set of three variables is therefore verified at this number of clusters. The variables with a σ_{ratio} significantly lower than one are floor area, total loss area, heated volume, window area and construction period (Figure 10). The variables with a $\Delta\mu_{ratio}$ diverging the most from zero are footprint, floor area, total loss area, heated volume, window area and construction period. The distributions show a very similar course as the distributions for the use case Peak power, which is corresponding with the similar ranking of variables (Table 2). Construction period appears to have a much larger influence on the energy demand for SH than on the peak power, what again can be found back in the higher position of construction period in the ranking for energy demand for SH than in

this for peak power. The variables with the lowest t-values are floor area, total loss area, heated volume and construction period (Table 4). Although footprint is one of the variables in the set, the t-value is rather high compared to the other variables. This could be due to the fact that footprint emerges as a 'pseudo-variable' and contains correlations with other variables, such as floor area, total loss area and heated volume, which do have a very low t-values.

3.3.3 Use case Peak power, energy demand for SH and specific energy demand

The RMSEs of the clustering based on the selected set of four variables do not differ that much from those of the clustering based on the set of five variables. Therefore, the additional variable in the set of five variables — namely floor area — will most likely also come forward as an important variable. When the building stock is divided into 15 clusters, the clustering based on the selected set generates the lowest RMSEs (Figure 8). Therefore, the distribution of the variables is again verified for a division of the stock based on all the energy KPIs into 15 clusters. The variables for which the σ_{ratio}s are significantly lower than one are floor area, total loss area, loss-to-floor-area ratio, heated volume and construction period (Figure 11). For the variables footprint, floor area, total loss area, loss-to-floor-area ratio, heated volume and window area the $\Delta\mu_{ratio}$ is diverging the most from zero. The variables with the lowest t-values are floor area, total loss area, loss-to-floor-area ratio and construction period (Table 5). This finding is again largely confirming the selected set of variables based on the ranking made by the linear multivariate regression. Besides variables containing information about the size of the dwellings, also loss-to-floor-area ratio appears as a dominant variable for this use case and the dependency on the heated volume is less important. Again, the variable footprint has a relatively high t-value and seems to be not very dominant, but the same annotation can be made that footprint emerges as a 'pseudo-variable' and contains correlations with other variables dealing with the dwelling size.

variable	t-value
footprint	-14,31
floor area	-123,90
total loss area	-107,05
number of stories	-43,28
loss-to-floor-area ratio	-29,16
compactness	-6,36
typology	-23,66
heated volume	-102,55
window-to-wall ratio	-17,50
window area	-59,37
construction period	-62,52

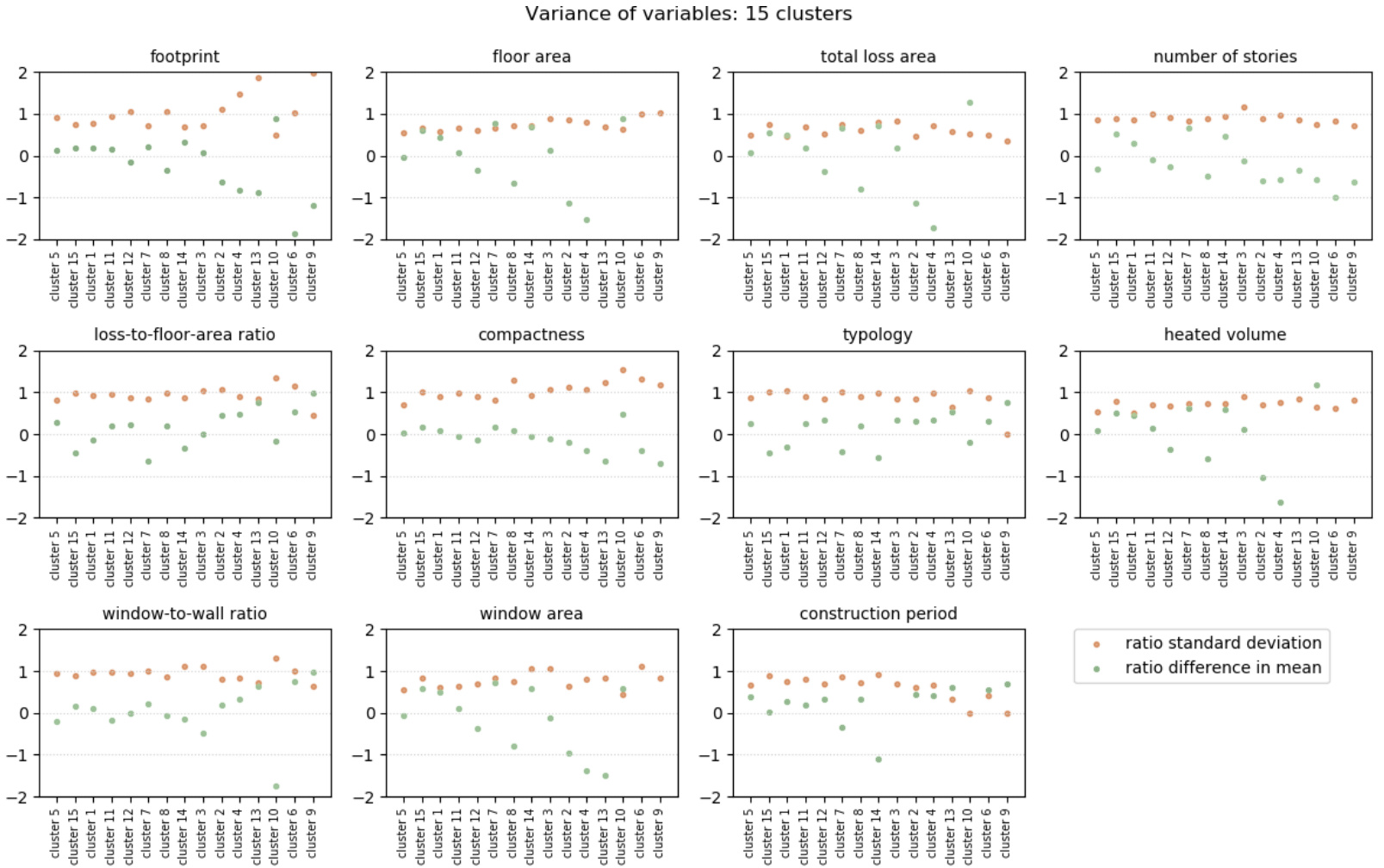
Table 3. T-values, rounded to two decimal digits, for every defined variable when clustered based on the energy KPI peak power.

variable	t-value
footprint	-12,85
floor area	-28,66
total loss area	-41,34
number of stories	-23,49
loss-to-floor-area ratio	-7,24
compactness	-7,20
typology	-13,22
heated volume	-31,43
window-to-wall ratio	-16,26
window area	-29,03
construction period	-96,11

Table 4. T-values, rounded to two decimal digits, for every defined variable when clustered based on the energy KPI energy demand for SH.

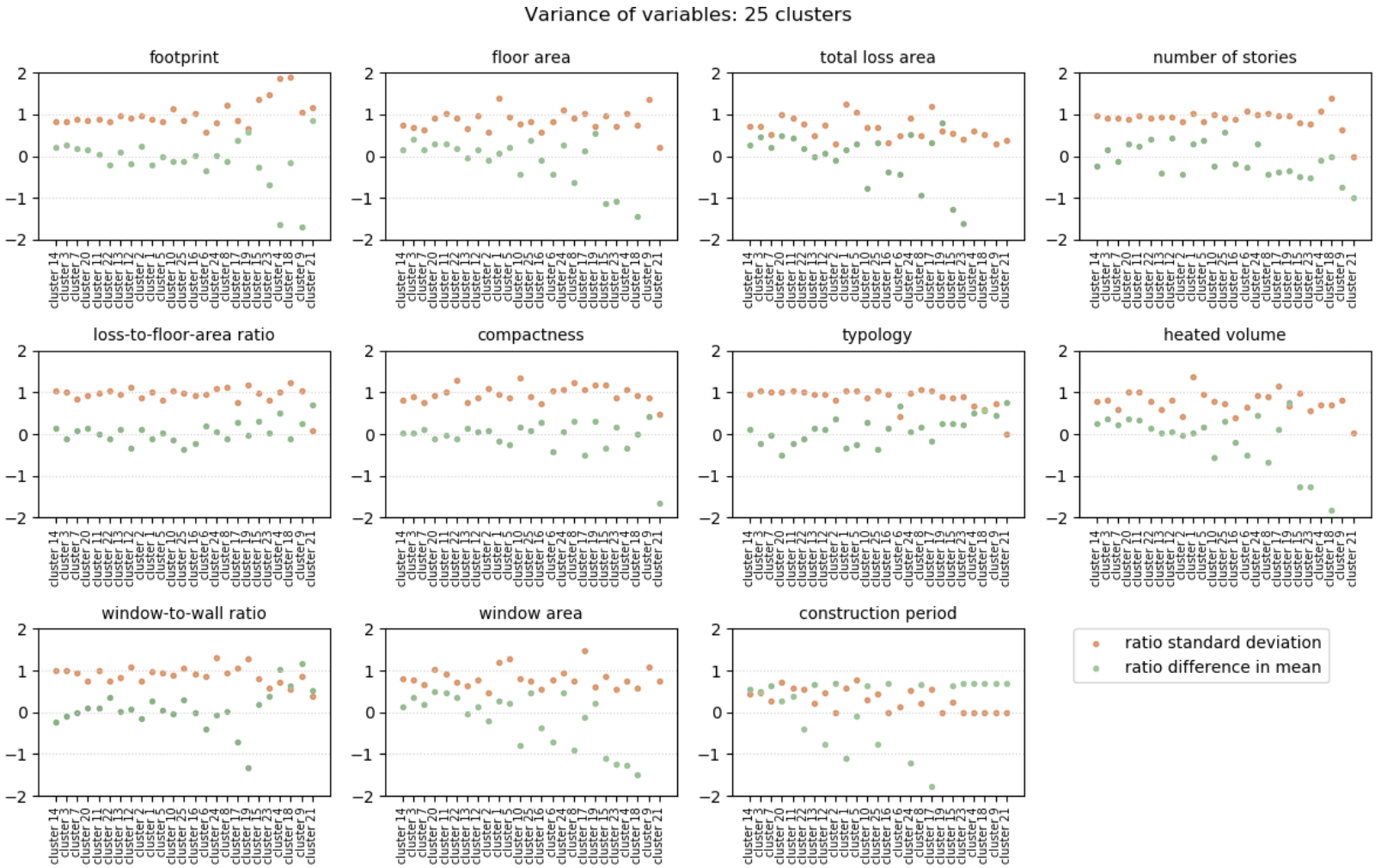
variable	t-value
footprint	-20,46
floor area	-72,86
total loss area	-86,31
number of stories	-49,26
loss-to-floor-area ratio	-84,08
compactness	-22,52
typology	-24,84
heated volume	-66,44
window-to-wall ratio	-21,32
window area	-71,12
construction period	-224,69

Table 5. T-values, rounded to two decimal digits, for every defined variable when clustered based on the energy KPIs peak power, energy demand for SH and specific energy demand.



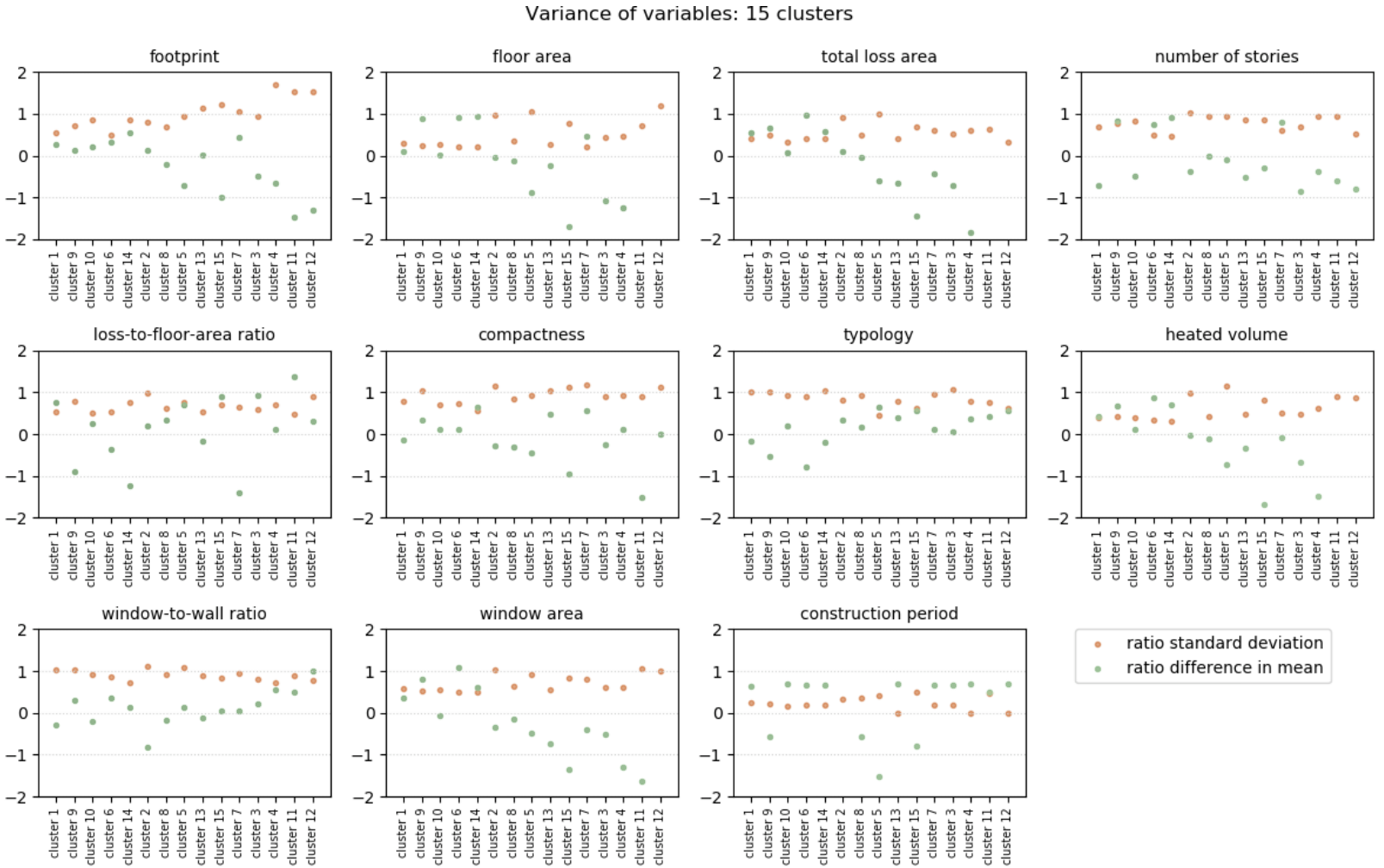
cluster-number	number of buildings
cluster 5	164
cluster 15	148
cluster 1	145
cluster 11	140
cluster 12	129
cluster 7	126
cluster 8	105
cluster 14	79
cluster 3	54
cluster 2	51
cluster 4	34
cluster 13	28
cluster 10	13
cluster 6	9
cluster 9	5

Figure 9. Examination of the variances in the 15 clusters of the clustering based on the energy KPI peak power. The ratio of the standard deviation and of the difference in mean are stacked out, ordered by a decreasing number of buildings in the clusters (right), for each defined variable.



cluster-number	number of buildings
cluster 14	107
cluster 3	100
cluster 7	98
cluster 20	75
cluster 11	75
cluster 22	68
cluster 13	66
cluster 12	64
cluster 2	59
cluster 1	59
cluster 5	57
cluster 10	57
cluster 25	50
cluster 16	47
cluster 6	44
cluster 24	35
cluster 8	32
cluster 17	29
cluster 19	28
cluster 15	25
cluster 23	19
cluster 4	16
cluster 18	11
cluster 9	7
cluster 21	2

Figure 10. Examination of the variances in the 25 clusters of the clustering based on the energy KPI energy demand for SH. The ratio of the standard deviation and of the difference in mean are stacked out, ordered by a decreasing number of buildings in the clusters (right), for each defined variable.



cluster-number	number of buildings
cluster 1	152
cluster 9	143
cluster 10	142
cluster 6	136
cluster 14	102
cluster 2	95
cluster 8	84
cluster 5	77
cluster 13	67
cluster 15	62
cluster 7	52
cluster 3	51
cluster 4	38
cluster 11	19
cluster 12	10

Figure 11. Examination of the variances in the 15 clusters of the clustering based on the energy KPIs peak power, energy demand for SH and specific energy demand. The ratio of the standard deviation and of the difference in mean are stacked out, ordered by a decreasing number of buildings in the clusters (right), for each defined variable.

3.4 Conclusion

To demarcate archetypes, a wide variety of variables has been used by various researchers. The selection of the set of variables that generates the most accurate estimation of the reference scenario is thus not unambiguously defined. Which variables are qualified to be used, depends on both the availability of the data and the specific conditions of the particular case study. Since the examined district in this work is a Flemish suburban residential neighbourhood, variables such as the density of the urban area, exposure and climate zones cannot be used to cluster the building stock. Moreover, only the energy demand required for space heating is taken into account and ideal heating systems are adopted, thus the variables dealing with the present heating and cooling systems are also not qualified. In this way, eleven potential variables have been specified out of the enumerated variables which were formerly used in other relevant studies: footprint, floor area, total loss area, number of stories, loss-to-floor-area ratio, compactness, typology, heated volume, window-to-wall ratio, window area and construction period. Naturally, if the conditions of the case study were different or if the available data contained other or additional information, other potential variables would have been possible.

The most influential of these eleven potential variables for each of the three defined energy KPIs — peak power, energy demand for SH and specific energy demand — are determined making use of a linear multivariate regression analysis of these energy KPIs. The linear regression is proposing a ranking in importance of the potential variables. The construction period is for all three KPIs ranked as one of the most important variables. Additionally, the variables dealing with the size of the dwellings appear for both peak power and energy demand for SH to be most important — such as footprint, floor area and total loss area —, whereas variables such as compactness and loss-to-floor-area ratio are ranked much lower. For specific energy demand, this turns out opposite. The variables dealing with the composition of the dwellings — like loss-to-floor-area ratio — have much more influence since the dependency on this KPI on the size of the buildings has been eliminated by dividing by the floor area. The variables number of stories and typology are listed lowest in the ranking, since they tend to contain insufficient information about the size or the composition of the dwellings. The variances of the number of stories in this case study are rather limited, since most dwellings have either three or four stories. Therefore, this variable does not indicate enough about the dwelling size to be an important variable for the clustering. But if the variances of this variable would be larger in another case study, the ranking might be different. Variables with a smaller distribution might be eliminated out of the ranking and others might occur higher on the list. To determine an absolute ranking of the potential variables for each energy KPI, other case studies with different variances will have to be examined.

The linear multivariate regression proposes the sets of eight variables as most accurate, but this turns out to be incorrect when these sets are applied for the clustering, as both techniques are working differently. Since the variables are not weighted when applied to execute the clustering and not all variables are equally responsible for the explanation of the energetic behaviour, the variables lower on the ranking should not be included in the selected sets. To determine the optimal number of variables for each use case, several sets of different numbers of variables are compared to each other and sets of three or four variables occur as the sets that generate the most accurate estimation of the reference scenario. For the use case Peak power, the most accurate set contains following variables: floor area, total loss area and construction period. For the use case Energy demand for SH, the set is very similar: footprint, total loss area and construction period. For the use case Peak power, energy demand for SH

and specific energy demand, the set contains a combination of the variables, highest in ranking, of the three KPIs: footprint, total loss area, loss-to-floor-area ratio and construction period.
Finally, the selected sets of variables for each use case are verified by comparing the variances of the variables in clusters, made by a clustering based on the energy KPIs which are important for the use case, with the variances of the variables in the entire building stock. The findings of these verifications tend to confirm the selected sets and the rankings suggested by the linear regression, since the variables in the sets show a smaller distribution in the clusters than in the entire stock. Therefore, they are most likely connected with the KPIs based on which the clustering happened. Only the variable footprint did not show a significantly smaller distribution, but this can be explained by understanding this variable as a 'pseudo-variable'. This means that this variable does not unambiguously define the energetic behaviour but contains correlations with other variables, like floor area, total loss area and heated volume.

Although only limited use cases and energy KPIs are examined in this work, this chapter presents a generally usable method to determine a ranking of the defined potential variables for any energy KPI. Based on this ranking, the optimal set of variables can easily be established by comparing several options, taking the ranking of all the KPIs of interest into account. As previously explained, the potential variables and the rankings of the variables might differ when examining another case study. To make general conclusions on the most accurate set of variables based on which the clustering will happen for several energy KPIs, other case studies will have to be examined.
Additionally, the scope of the approach can be expanded to the application of varying user profiles, by making use of the StROBe module — Stochastic Residential Occupancy Behaviour. This module is generating simulations of residential human behaviour based on stochastic data in order to implement a divergent user profile in each dwelling (Baetens et al., 2015). Before this variable is qualified to be applied to cluster on, the profiles must be summarised into one value which can be determined without the knowledge of the results of the energy simulations. Then, the importance of this additional variable relative to the other defined variables can be examined using the method presented in this chapter.

4. Clustering of the building stock

Since clustering of data is a widely applied method to gain insight into the characteristics of the data, various techniques to execute the clustering have been developed. The techniques that are qualified to be applied for the clustering of a building stock for the purpose of this work, must be able to cluster based on multiple variables, since the energetic behaviour of a dwelling is depending on several aspects. Therefore Ghaissi (2017) made use of Multivariate Cluster Analysis (MCA), which comprises various techniques that divide a multidimensional data space into homogeneous groups.
First, a literature study was executed to determine which techniques are qualified for this application. To gain insight in the process of the techniques, the two most applied and therefore most elaborated techniques — k-means and agglomerative hierarchical clustering — are explained. Then, to prevent undesirable weighting of the variables due to the different orders of magnitudes of the values, the data is either standardised — using the z-score formula — or normalised — converted to a range between zero and one — before the clustering is executed. Subsequently, the most accurate technique and method to standardise or normalise the data is selected based on a comparison of the root mean square errors of all four options.

After the clustering and the simulation of one representative building of each cluster — the building the closest to the centre of the cluster —, the results of these representative buildings must be scaled up to each represent the entire cluster. The upscaling of the results can happen in several ways: either by multiplying them by the number of dwellings in each cluster or taking a geometrical property of the dwellings into account and diversifying the results by scaling them with the property.
Which way of upscaling generates the most accurate results is depending on the energy KPI, but also on the use case and the number of clusters. After a certain number of clusters the scaling of the results by a geometrical property is not recommended anymore, since the errors are higher than the errors made by multiplying by the number of buildings in the clusters. This is discussed in detail in the penultimate section of this chapter.

4.1 Literature study

Clustering of objects is a technique that is widely used to understand the characteristics of the objects and identify them with a type and is applied in various scientific disciplines (Rokach and Maimon, 2010). The purpose of clustering implies that objects — or in this application dwellings — with similar features are grouped in one cluster so that the clusters are coherent internally and that objects with dissimilar features belong to different clusters (Manning, Raghavan and Schütze, 2009a). Since the dwellings in one cluster are supposed to exhibit similar energetic behaviour, it is acceptable to represented all the dwellings in one cluster by one dwelling which is located the closest to the middle — or 'centroid' — of this cluster.
Since the notion of a cluster is not unambiguously defined, many different cluster techniques have been developed (Rokach and Maimon, 2010). However, most used cluster techniques can be divided into two groups: centroid-based or partitioning cluster methods and connectivity-based or hierarchical cluster methods. The main difference between these two groups contains the moment of deciding the number of clusters and thus how the centroids of the clusters are elected. Centroid-based cluster techniques depart from randomly selected centroids and thus the number of clusters must be defined in advance to determine how many centroids must be selected. These techniques can also be appointed as "flat clustering techniques", since they create "a flat set of clusters without any explicit structure that would relate clusters to each other" (Manning, Raghavan and Schütze, 2009a). Whereas with connectivity-based techniques

the number of clusters is decided after the clustering and therefore the positions of the centroids are intrinsic to the data collection. Clustering with these techniques creates a hierarchy between the separate clusters (Manning, Raghavan and Schütze, 2009a). Ghiassi (2017) describes this distinction between the groups as “whether or not the identified clusters are nested”.

The most widely used en thus elaborated centroid-based cluster technique is k-means clustering because of its simplicity and efficiency (Manning, Raghavan and Schütze, 2009a). This technique is explained in the next paragraph. Hierarchical clustering comprises methods that approach the clustering either top-down — divisive hierarchical clustering — or bottom-up — agglomerative hierarchical clustering (Rokach and Maimon, 2010), which is explained in the subsequent paragraph.

4.1.1 K-means clustering

K-means clustering starts with the random selection of k centroids, with k being the predetermined number of clusters. Each data point is assigned to the nearest centroid and the Residual Sum of Squares (RSS) is calculated for each cluster according to following equation (Geyer, Schlüter and Cisar, 2016; Manning, Raghavan and Schütze, 2009a):

$$RSS_i = \sum_{x \in K_i} (x - \overline{x}_i)^2$$

with $x \in K_i$ being every data point in cluster K_i and $\overline{x}_i$ the centroid of that cluster. Then, the overall RSS is computed:

$$RSS = \sum_{i=1}^{k} RSS_i$$

The RSS is a measure to determine how well the centroids represent all the data points in the clusters (Manning, Raghavan and Schütze, 2009a): the smaller RSS, the smaller the distances between each data point in the cluster and the centroid and thus the better the clustering.

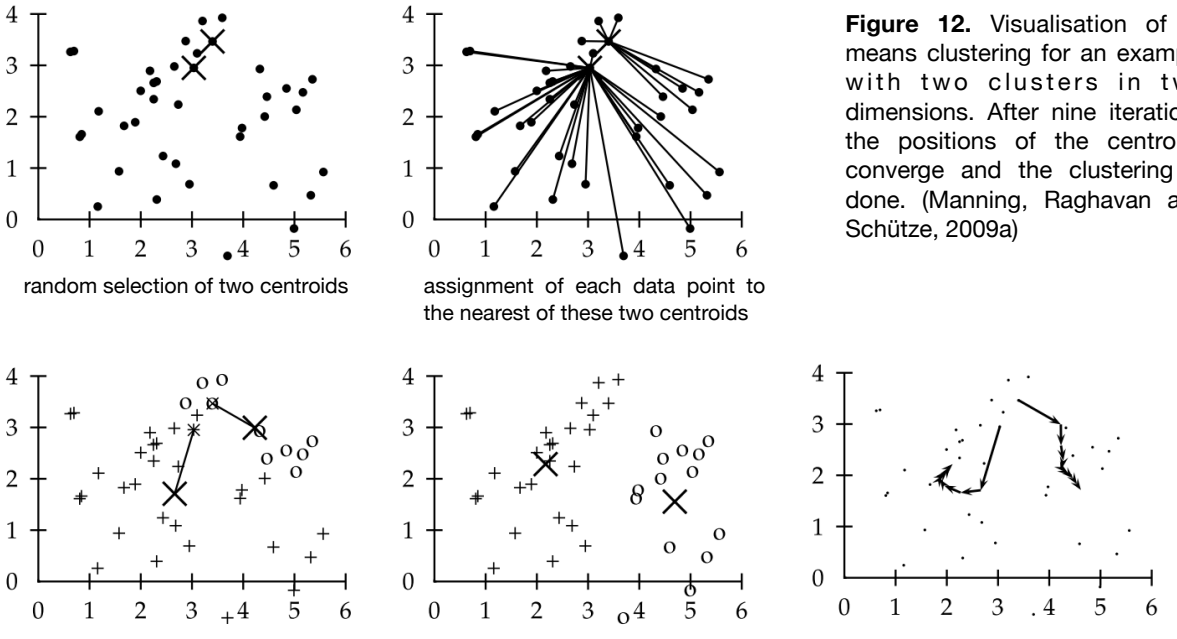


Figure 12. Visualisation of k-means clustering for an example with two clusters in two dimensions. After nine iterations the positions of the centroids converge and the clustering is done. (Manning, Raghavan and Schütze, 2009a)

Therefore, the goal of k-means clustering is to minimise this RSS and since k is fixed, this comes to the same as minimising the average inner squared Euclidean distance.
When the RSSs of each cluster have been calculated, the positions of the centroids are redetermined to minimise the RSSs. Subsequently, the data points are again assigned to the closest centroid to then recalculate the RSS. This method is iteratively repeated until the final distribution has been reached (Figure 12).
K-means clustering is reported as a method with complexity of *O(NKd)*, where *N* is the number of buildings, *K* the number of clusters and *d* the dimensions. This means that the calculation time will increase linearly with the number of buildings and therefore k-means emerges as a suitable method for clustering a large building stock (Geyer and Schlüter, 2017).

4.1.2 Agglomerative hierarchical clustering

Agglomerative hierarchical clustering does not need a predetermined number of clusters, since this method starts with assigning a cluster to each data point individually. Then, the two closest clusters are merged together. This similarity measure can be calculated in several ways: using single-link, complete-link or average-link clustering (Manning, Raghavan and Schütze, 2009b; Rokach and Maimon, 2010). For single-link clustering, the distance between two clusters is determined by the — in this case — Euclidean distance between the two nearest data points (Figure 13, left). For complete-link clustering, the distance is calculated in the opposite way: the distance between two clusters is equal to the longest distance between any point of one cluster and any point of the other cluster (Figure 13, middle). Finally, the distance for average-link clustering is calculated as the average distance of every member of one cluster to every member of the other cluster (Figure 13, right). The way of determining the distances at complete-link clustering generates the most compact clusters (Rokach and Maimon, 2010) and offers therefore the most interesting perspectives for this application.
The two clusters with the smallest mutual distance are merged together. This step is repeated until only one cluster remains. The steps are usually displayed in a dendrogram, where each horizontal line represents a merging of two clusters with the value at the vertical axis as the similarity between those two clusters (Figure 14). In this way, agglomerative hierarchical clustering automatically goes through every possible number of clusters and the desired number of clusters can be defined by 'cutting' the dendrogram at a certain height — the height that represents the desired inner distance of the clusters.
The computational complexity is a disadvantage of this method, since this amounts to $O(N^2)$ with *N* the number of buildings. This means that the calculation time will increase quadratically with the number of buildings and this method is therefore less suited to cluster large building stocks (Geyer and Schlüter, 2017).

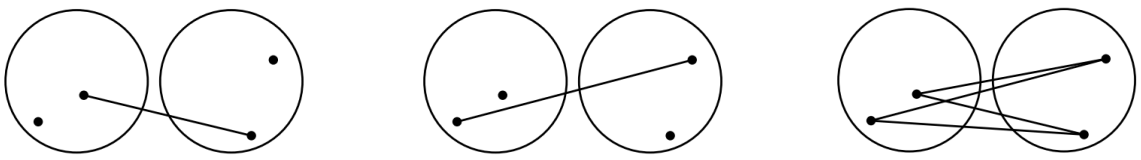


Figure 13. Visualisation of the ways of calculating the distance between two clusters: single-link (left), complete-link (middle) and average-link clustering (right). (Manning, Raghavan and Schütze, 2009b)

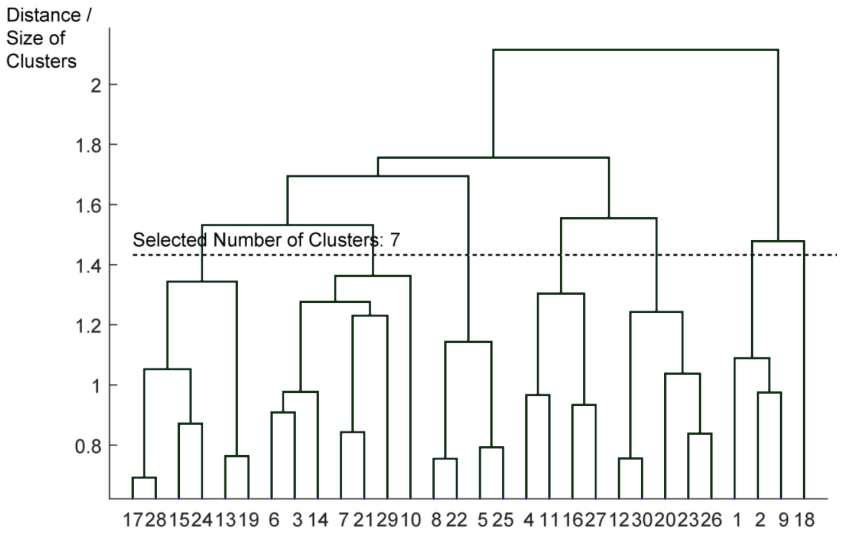


Figure 14. Dendrogram of an agglomerative hierarchical clustering, with the indices of the data points at the horizontal axis and the distances at the vertical axis. The clustering with the desired number of clusters, for example seven, can be determined by 'cutting' the dendrogram. (Geyer, Schlüter and Cisar, 2016)

4.2 Selection of cluster technique

Since the variables do not all have the same order of magnitude, the values of the variables have to be rescaled before the clustering can happen to prevent undesirable weighting. The most common methods to rescale the values are either standardising or normalising the data.
Standardisation of the values will convert the mean to zero and the standard deviations to minus and plus one and can be executed using the z-score formula (Steinley and Brusco, 2008):

$$x_{ij,stand} = \frac{x_{ij} - \overline{x}_j}{\sqrt{Var(x_j)}}$$

where x_{ij} is the value of the variable j for dwelling i, $\overline{x}_j$ the mean of variable j, $x_{ij,stand}$ the standardised value of variable j for dwelling i and $\sqrt{Var(x_j)}$ the standard deviation of variable j, with $Var(x_j)$ calculated as follows:

$$Var(x_j) = \frac{\sum_{i=1}^{N} (x_{ij} - \overline{x}_j)^2}{N}$$

with N the number of dwellings of the entire building stock.

Normalising the data implies that the values are converted to a range between zero and one and can be determined using following formula:

$$x_{ij,norm} = \frac{x_{ij} - x_{j,min}}{x_{j,max} - x_{j,min}}$$

where $x_{j,min}$ is the minimum value of variable j for all dwellings, $x_{j,max}$ the maximum value of variable j and $x_{ij,norm}$ the normalised value of variable j for dwelling i.

Both cluster methods — k-means and agglomerative hierarchical clustering — are executed using standardisation as well as normalisation and are evaluated compared to each other. To perform k-means clustering, the Python package Scikit-learn has been used (Pedregosa et al., 2011) and to execute the hierarchical clustering the module *hierarchy* from the clustering

package of Scipy (*scipy.cluster*) has been applied (Scipy community, 2018). The clustering is executed for every number of clusters in a range from one to 40 clusters and the root mean square errors (RMSEs) are stacked out in function of the number of clusters (Figure 15 to 17). If an improvement would occur, the results of the energy KPIs are scaled with the geometrical properties of the dwellings which generates the lowest RMSEs for that energy KPI, clustering method and use case to make sure that in every case the best possible results are compared.

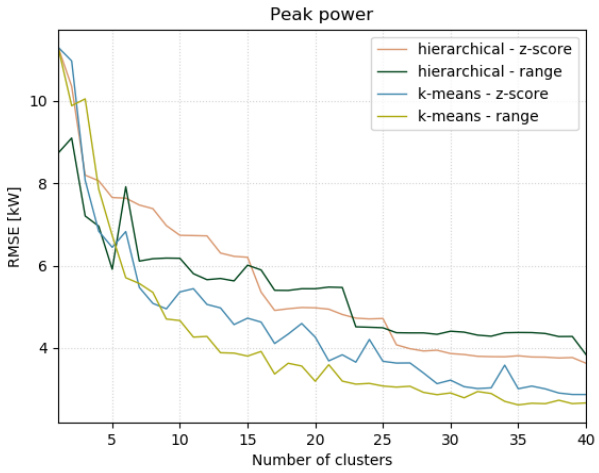


Figure 15. Graph showing the root mean square errors in function of the number of clusters for the energy KPI peak power in the use case Peak power, comparing both cluster methods and standardisation versus normalisation. The results generated with hierarchical clustering and normalisation are scaled with the volume of the dwellings.

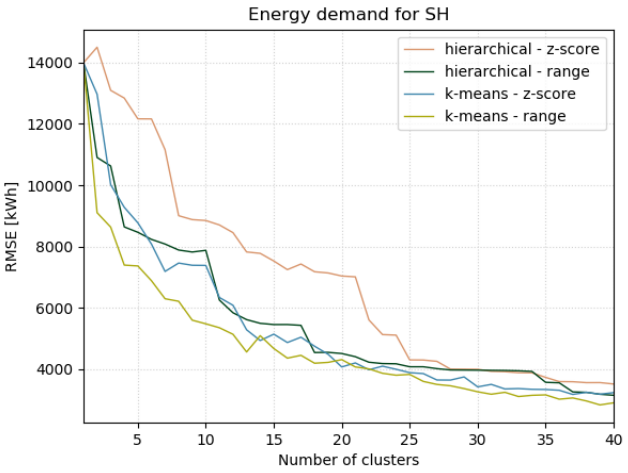
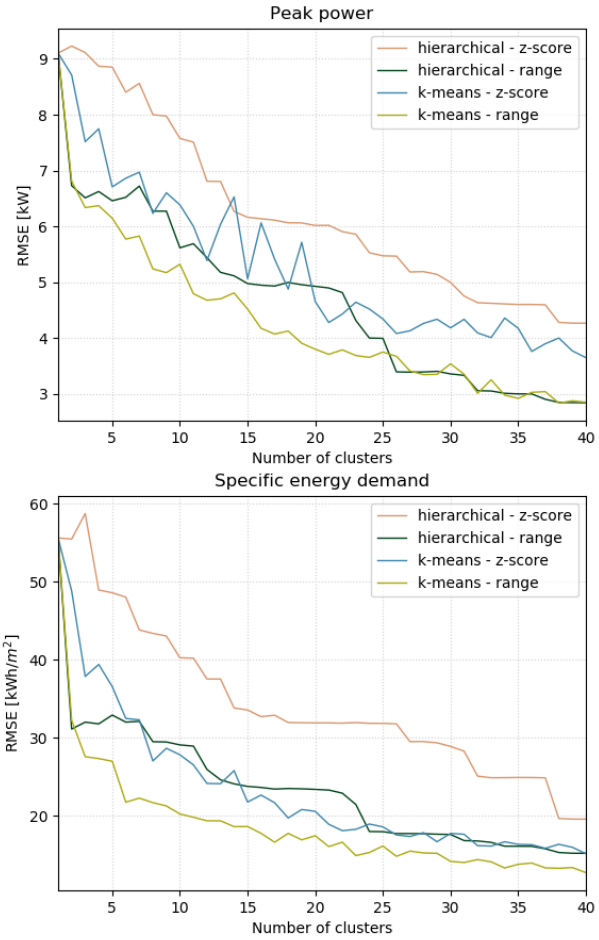


Figure 16. Graph showing the root mean square errors in function of the number of clusters for the energy KPI energy demand for SH in the use case Energy demand for SH, comparing both cluster methods and standardisation versus normalisation. None of the results are scaled with a geometrical property of the dwellings.



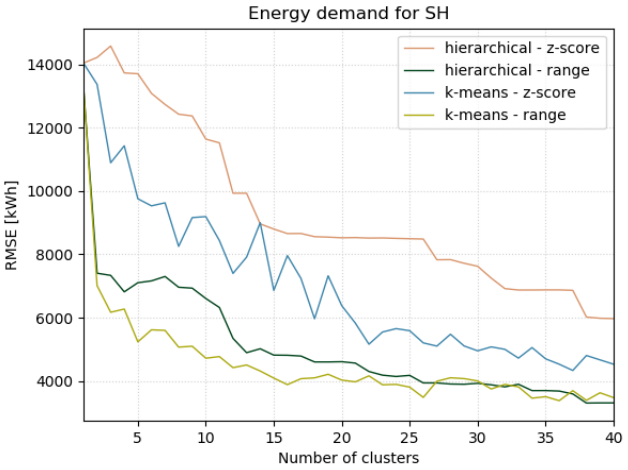


Figure 17. Graphs showing the root mean square errors in function of the number of clusters for the energy KPIs peak power (top left), energy demand for SH (top right) and specific energy demand (bottom left) in the use case Peak power, energy demand for SH and specific energy demand, comparing both cluster methods and standardisation versus normalisation. The results for the energy KPI peak power are scaled with the floor area of the dwellings and the results for the energy KPI energy demand for SH generated with standardisation are also scaled with the floor area and with normalisation are scaled with the volume of the dwellings.

For all use cases — Peak power (Figure 15), Energy demand for SH (Figure 16) and Peak power, energy demand for SH and specific energy demand (Figure 17) — k-means clustering in combination with normalisation appears to be the method that generates the lowest RMSEs for all energy KPIs which are of interest in the particular use case — up to about one kW lower than the second best method for peak power, about 1900 kWh lower for energy demand for SH and about ten kWh/m^2 lower for specific energy demand.

The better performance of normalisation, compared to standardisation, can be comprehended from the definitions. Whereas normalisation transforms the values of every variable to a range exactly from zero to one, standardisation converts the values in the range of the standard deviations to the exact range from minus to plus one. This means that the standardised values of the data points smaller or larger than the standard deviations are lower or higher than minus or plus one. But in this way, standardising the variables grants unintentionally more weight to the variables with these larger values and these are not necessarily the variables who are the most influential in the determination of the energy KPIs. Moreover, standardisation with the z-score assumes a normal distribution of the data and thus makes a mistake with the variables that do not have this normal distribution, while normalisation does not make this assumption.

4.3 Upscaling to the entire building stock

After the clustering, one building of every cluster is selected as the representative building and is then simulated. Since these buildings are each supposed to represent the entire cluster, the interpreted results of these simulations can be applied to all the dwellings in every cluster and thus are scaled up to the entire building stock by multiplying them by the number of dwellings in each cluster. Since the energy KPIs are partly depending on geometrical properties of the dwellings, it could be that scaling the results with a geometrical property — such as the floor area, the volume or the footprint of the building — might improve the overall results. To scale the results, the results of the representative buildings are divided by the geometrical property of these representative buildings and then multiplied by the property of the actual buildings. For every energy KPI, it has been examined whether the scaling with one of these last-mentioned geometrical properties decreases the RMSEs or just the upscaling by multiplying by the number of buildings generates the lowest errors. This has been investigated for both the clustering based on the energy KPIs themselves as based on the defined sets of variables (see chapter 3).

When the building stock is clustered based on the energy KPIs, scaling the results with any geometrical property does not improve the RMSEs (Figure 18 to 20). This phenomenon can conveniently be comprehended by considering what the clustering based on the energy KPIs comprises. The dwellings with the most similar results are grouped together in a cluster, regardless the properties of the dwellings which are explanatory for the energetic behaviour of the buildings. This implies that dwellings with both a large and a small floor area might end up in the same cluster while still having very similar results. When these results are scaled with for example the floor area, they are adjusted in a way that worsens the results and therefore enlarges the RMSEs. This can also be seen in the graphs on the next page (Figure 18 to 20). Scaling the results up by multiplying by the number of buildings generates in every case the lowest RMSEs and the most smooth course. Scaling the results with either the floor area or the volume of the buildings enlarges the RMSEs and generates a more erratic course, while scaling with the footprint of the buildings causes an extremely erratic course and even larger RMSEs. This indicates that there is no direct connection at all between the geometric property *footprint* and the specified energy KPIs.

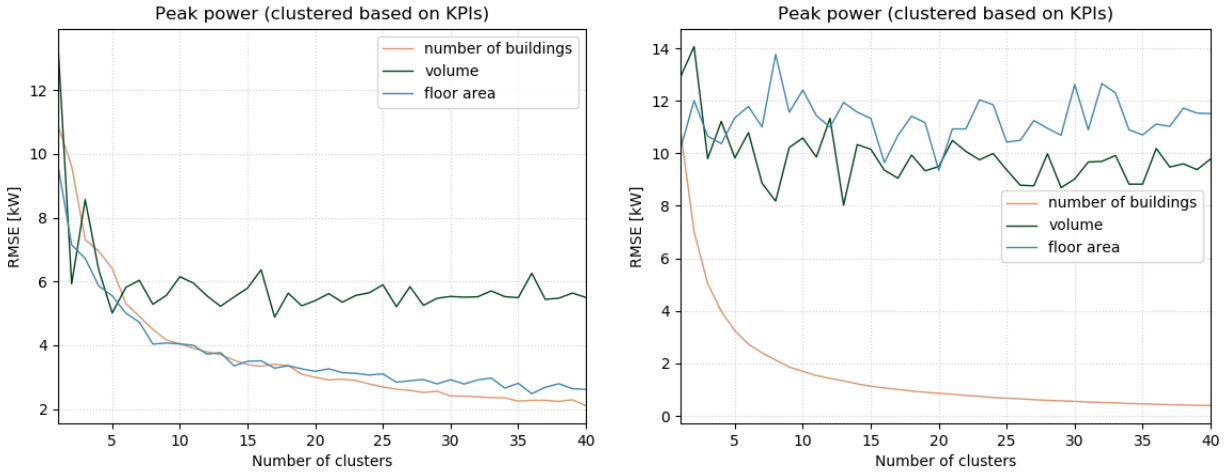


Figure 18. Graphs showing the root mean square errors in function of the number of clusters for the energy KPI peak power in use case Peak power, energy demand for SH and specific energy demand (left) and use case Peak power (right). The results are scaled up with either the number of buildings or scaled with a geometrical property of the buildings: floor area, volume or footprint. The RMSEs of the results scaled with footprint in both use cases are higher than the range of the graphs.

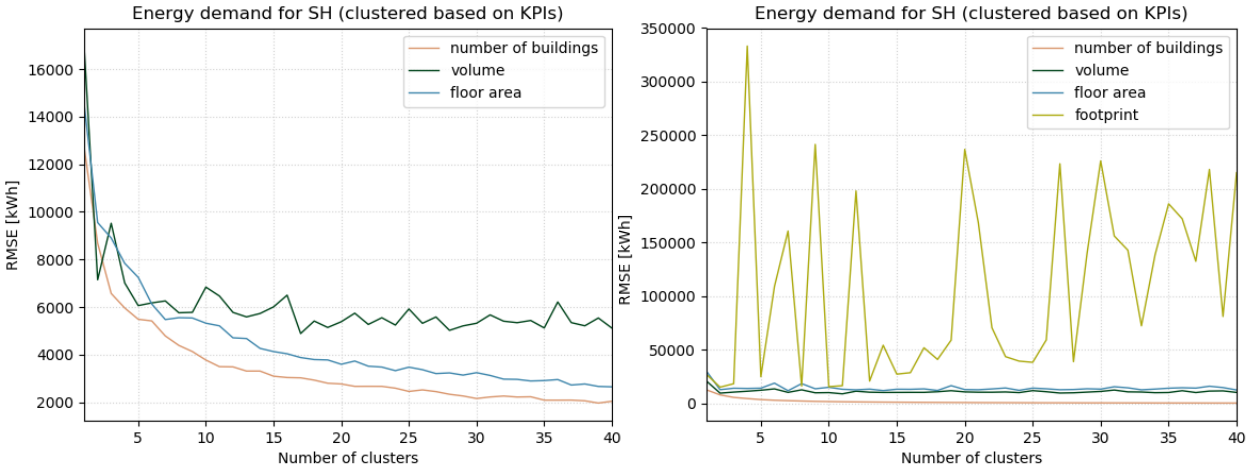


Figure 19. Graphs showing the root mean square errors in function of the number of clusters for the energy KPI energy demand for SH in use case Peak power, energy demand for SH and specific energy demand (left) and use case Energy demand for SH (right). The results are scaled up with either the number of buildings or scaled with a geometrical property of the buildings: floor area, volume or footprint. The RMSEs of the results scaled with footprint in use case Peak power, energy demand for SH and specific energy demand (left) are higher than the range of the graph.

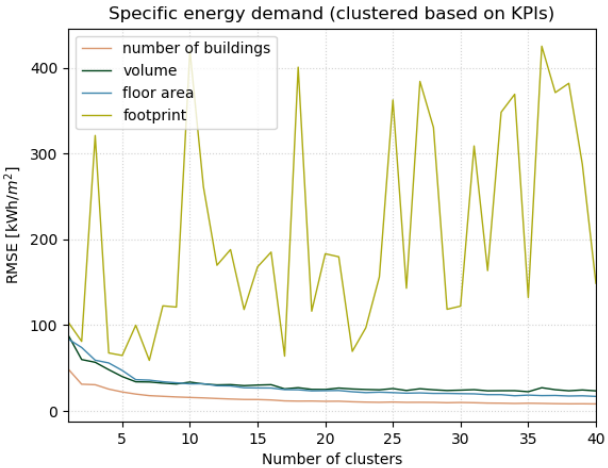


Figure 20. Graphs showing the root mean square errors in function of the number of clusters for the energy KPI specific energy demand in use case Peak power, energy demand for SH and specific energy demand. The results are scaled up with either the number of buildings or scaled with a geometrical property of the buildings: floor area, volume or footprint.

When the building stock is clustered based on the defined sets of variables, scaling the results with a geometrical property does improve the RMSEs. Which way of upscaling generates the best results is depending on the energy KPI, the use case and the number of clusters and is discussed in the following paragraphs for each energy KPI individually.

4.3.1 Peak power

The RMSEs of the energy KPI peak power in the use case Peak power, energy demand for SH and specific energy demand can be decreased by scaling the results with the floor area of the buildings (Figure 21, left). Scaling with the volume of the buildings does also improve the RMSEs, but from 16 or more clusters, this scaling factor generates higher RMSEs than the factor *floor area*. In the use case Peak power, scaling with the factor *floor area* does also decrease the RMSEs until a certain number of clusters — in this case study until five clusters (Figure 22, right). Although the RMSEs are lower for two and three clusters when scaled with the volume of the buildings, this scaling factor shows a more erratic course. Therefore, scaling with the factor *floor area* achieves a more reliable improvement.

From six or more clusters, scaling with a geometrical property does not improve the results anymore in the use case Peak power (Figure 21, right). This could be explained by the following hypothesis. Scaling the results causes both improvements and additional errors, since the energy KPI peak power is depending on the floor area but not only on this property. Dividing the building stock in more clusters causes the clusters with a wide distribution of the floor area to split, since *floor area* is one of the variables in the set based on which the clustering happens (see chapter 3). The clusters that are not further split up most likely have a small distribution of the variables in the set, like *floor area*. Therefore, scaling with the floor area in these clusters does not improve the results, since the differences in the energy KPI peak power in these clusters are not depending on the floor area anymore but on other factors. The more clusters there are, the more clusters with a small distribution of the floor area exist and the more clusters for which there are caused more additional errors than improvements by scaling with the floor area. From a certain number of clusters —in this case from six— the additional errors are larger than the improvements and from then scaling with the factor *floor area* is not recommended anymore.

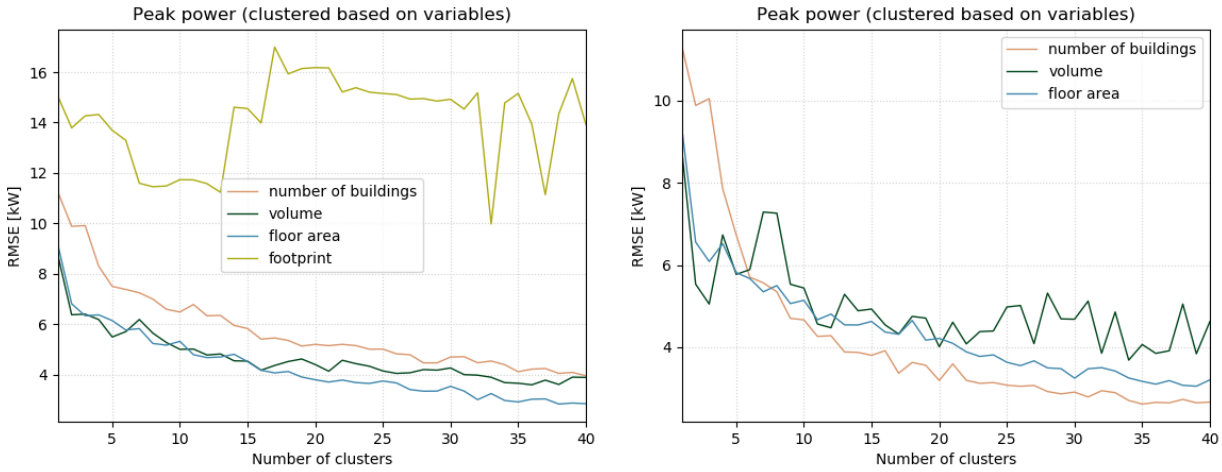


Figure 21. Graphs showing the root mean square errors in function of the number of clusters for the energy KPI peak power in use case Peak power, energy demand for SH and specific energy demand (left) and use case Peak power (right). The results are scaled up with either the number of buildings or scaled with a geometrical property of the buildings: floor area, volume or footprint. The RMSEs of the results scaled with footprint in use case Peak power (right) are higher than the range of the graph and therefore not visible on the figure.

This hypothesis can be investigated by examining the distributions of the floor area in the clusters of the clusterings based on the defined set of variables, taking the number of dwellings in each cluster into account. Therefore, the ratios of the distributions of each variable in the clusters relative to the distributions of each variable in the entire building stock are calculated in the same way as described in chapter 3, using following formula:

$$\sigma_{ratio} = \frac{\sigma_{cluster}}{\sigma_{stock}}$$

The lower the σ_{ratio}, the smaller the distribution in the cluster compared to the distribution in the entire stock. To determine how many dwellings are part of a cluster for which the distribution of the floor area is still sufficiently wide, the number of dwellings in the clusters with a σ_{ratio} larger than 0.75 are counted. These clusters contain dwellings with a distribution of the floor areas up to 75 percent of the distribution in the entire stock and will therefore most likely introduce larger improvements by the scaling with the floor area than additionally made errors.

When the building stock is divided into five clusters, clusters 2, 3 and 5 have a σ_{ratio} larger than 0.75 (Figure 22, left). In this way, 409 dwellings are part of clusters with a rather wide distribution of the floor area (Table 5, see page 43). When the stock is divided into 15 clusters, scaling the results with the floor area is not improving the RMSEs anymore. Therefore, the number of dwellings in clusters with a wide distribution of the floor area should be significantly lower. Only clusters 6 and 4 have a σ_{ratio} higher than 0.75 (Figure 22, right), which makes that only 66 dwellings are part of clusters with a rather wide distribution of the floor area (Table 6, see page 43).
Since the improvement by scaling the results with the floor area in the use case Peak power, energy demand for SH and specific energy demand is still valid when the building stock is divided into for example 25 clusters, the number of dwellings in clusters with a wide distribution should still be sufficiently high when the stock is divided into 25 clusters. The number of dwellings in clusters with a σ_{ratio} higher than 0.75 equals 437 when the building stock is divided into five clusters — clusters 4, 3 and 2 (Figure 23, left and Table 7, see page 43)— and 162 when the building stock is divided into 25 clusters — clusters 4, 17, 9, 18, 19, 21, 6 and 24 (Figure 23, right and Table 8, see page 43). Although the building stock is divided

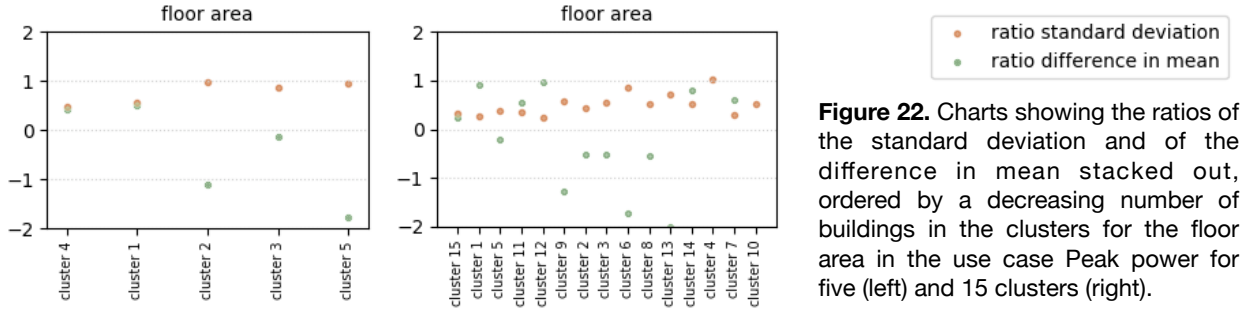


Figure 22. Charts showing the ratios of the standard deviation and of the difference in mean stacked out, ordered by a decreasing number of buildings in the clusters for the floor area in the use case Peak power for five (left) and 15 clusters (right).

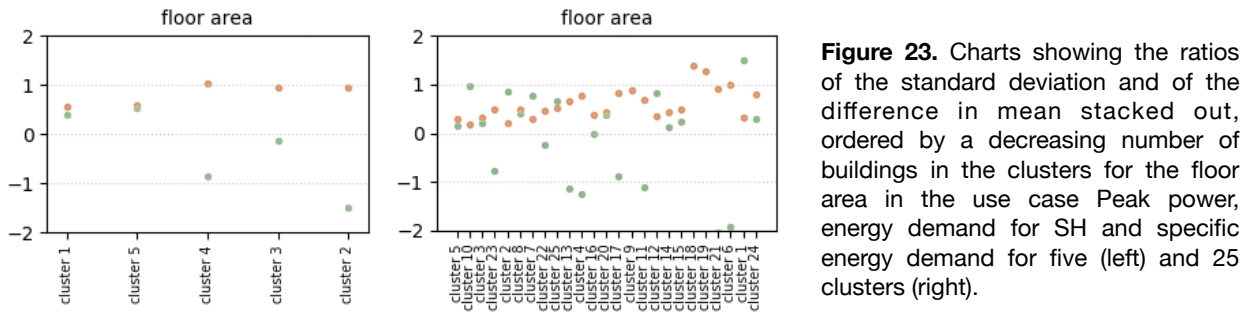


Figure 23. Charts showing the ratios of the standard deviation and of the difference in mean stacked out, ordered by a decreasing number of buildings in the clusters for the floor area in the use case Peak power, energy demand for SH and specific energy demand for five (left) and 25 clusters (right).

in more clusters than the situation where the stock was divided into 15 clusters in the use case Peak power, still more dwellings are part of a cluster with a rather wide distribution of the floor area. Therefore, these findings tend to confirm the posed hypothesis.

4.3.2 Energy demand for SH

The results of the energy KPI energy demand for SH can also be improved in the use case Peak power, energy demand for SH and specific energy demand by scaling them with the volume of the buildings (Figure 24, left). Again, in the use case that concentrates only on the energy demand for SH, the results can be decreased by scaling them with the volume of the buildings but only up to a certain number of clusters — in this case study up to 12 clusters (Figure 24, right).

The same hypothesis as for the KPI peak power is checked for the KPI energy demand for SH: the number of dwellings in clusters with a σ_{ratio} for the variable *volume* larger than 0.75 are counted for several situations.

For the use case Peak power, energy demand for SH and specific energy demand, the improvement of the scaling with the volume is still valid when the building stock is divided into for example 25 clusters and thus the number of dwellings in clusters with a rather wide distribution of the volume should still be sufficiently high. When the stock is divided into five clusters, the clusters 4, 3 and 2 show a σ_{ratio} larger than 0.75 (Figure 25, left). This makes that 437 dwellings are part of a cluster with a rather wide distribution of the volume, for which the results will most likely be improved by a scaling with this dwelling property (Table 7, see page 43). When the stock is divided into 25 clusters, still 120 dwellings are part of a cluster with a σ_{ratio} larger than 0.75 — clusters 17, 9, 18, 19, 21, 6 and 24 (Figure 25, right and Table 8, see page 43). Again, the hypothesis seems to be correct.

For the use case Energy demand for SH, scaling with the volume is decreasing the errors when the building stock is divided into five clusters but not anymore when the stock is divided into 25 clusters. Therefore, the number of dwellings in clusters with a rather wide distribution of the volume should be significantly lower when the stock is divided into 25 clusters. When divided into five clusters, clusters 3, 4 and 5 show a σ_{ratio} larger than 0.75 (Figure 26, left) and 437 dwellings are part of these clusters (Table 9, see page 43). When the stock is divided into 25 clusters, clusters 19, 6, 17, 20, 23, 7, 25, 12 and 13 have a σ_{ratio} above 0.75 (Figure 26, right)

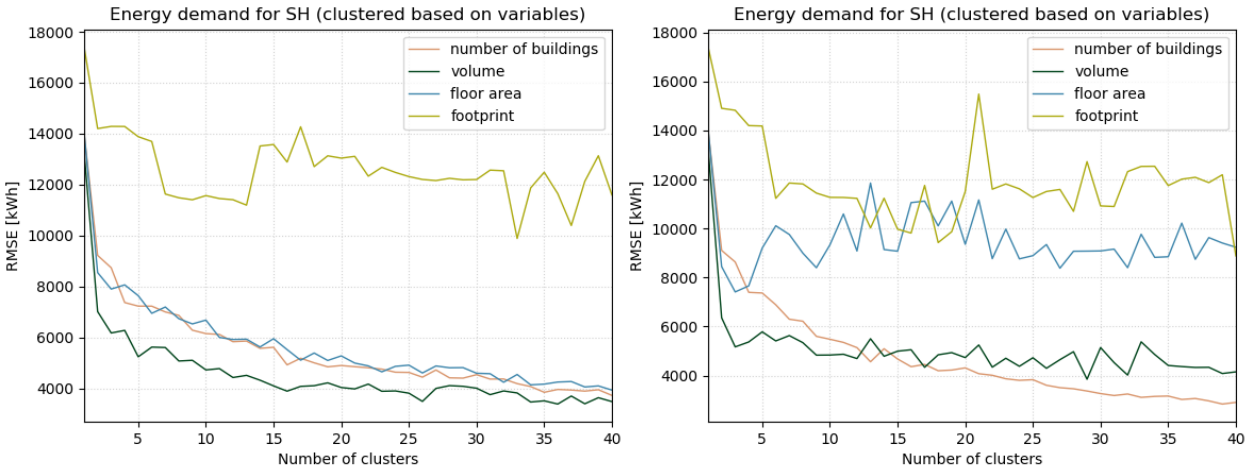


Figure 24. Graphs showing the root mean square errors in function of the number of clusters for the energy KPI energy demand for SH in use case Peak power, energy demand for SH and specific energy demand (left) and use case Energy demand for SH (right). The results are scaled up with either the number of buildings or scaled with a geometrical property of the buildings: floor area, volume or footprint.

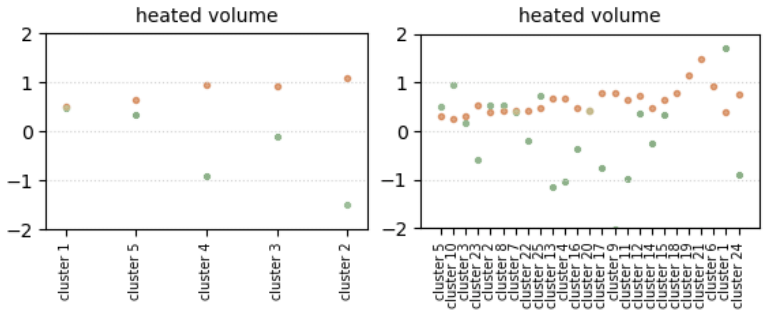


Figure 25. Charts showing the ratios of the standard deviation and of the difference in mean stacked out, ordered by a decreasing number of buildings in the clusters for the volume in the use case Peak power, energy demand for SH and specific energy demand for five (left) and 25 clusters (right).

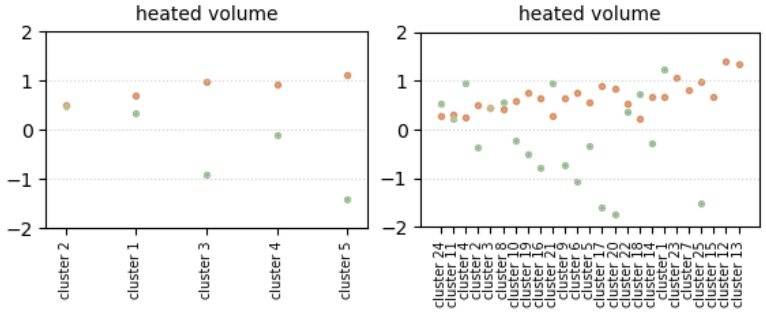


Figure 26. Charts showing the ratios of the standard deviation and of the difference in mean stacked out, ordered by a decreasing number of buildings in the clusters for the volume in the use case Energy demand for SH for five (left) and 25 clusters (right).

ratio standard deviation
ratio difference in mean

and contain 189 dwellings accumulated (Table 10, see page 43). In contrast with the previously discussed energy KPI peak power, the number of dwellings is still quite large. The posed hypothesis does not apply in this situation.
Whether the hypothesis is therefore false or rather the given explanation is not the only influence to determine if the scaling is still improving the results or not, is not clear. In contrast with the situation for the KPI peak power, the scaling factor volume is not included in the set of variables based on which the clustering has happened. This might involve an explanation of the difference but has to be examined further. Moreover, why exactly these dwelling properties are improving the results and when this improvement is not valid anymore, should also be investigated further to be able to draw general conclusions.

4.3.3 Specific energy demand

In contrast to the other KPIs, the RMSEs of the energy KPI specific energy demand cannot be decreased by scaling the results with any geometrical property of the dwellings (Figure 27). Simply multiplying the results by the number of buildings in each cluster generates the lowest RMSEs. This confirms the previously mentioned statement that the energy KPI specific energy demand is not directly depending on the size of the dwellings but rather on the composition of the plan (see chapter 3). Therefore, scaling the results with properties related to the size will only introduce additional errors and no improvements.

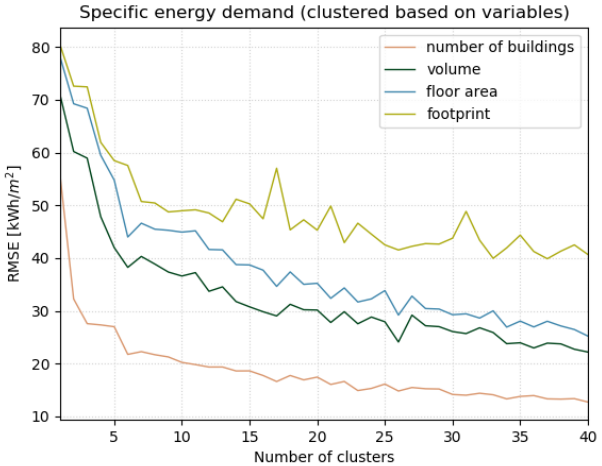


Figure 27. Graphs showing the root mean square errors in function of the number of clusters for the energy KPI specific energy demand in use case Peak power, energy demand for SH and specific energy demand. The results are scaled up with either the number of buildings or scaled with a geometrical property of the buildings: floor area, volume or footprint.

cluster-number	number of buildings
cluster 4	590
cluster 1	231
cluster 2	172
cluster 3	144
cluster 5	93

Table 5. Use case Peak power — five clusters

cluster-number	number of buildings
cluster 15	277
cluster 1	216
cluster 5	168
cluster 11	96
cluster 12	86
cluster 9	75
cluster 2	68
cluster 3	50
cluster 6	40
cluster 8	39
cluster 13	34
cluster 14	26
cluster 4	26
cluster 7	25
cluster 10	4

Table 6. Use case Peak power — 15 clusters

Tables 5 to 10. Number of dwellings in each cluster for several use cases and number of clusters, ordered by a decreasing number of dwellings. The use case and the number of clusters is specified at each table.

cluster-number	number of buildings
cluster 1	572
cluster 5	221
cluster 4	190
cluster 3	135
cluster 2	112

Table 7. Use case with all energy KPIs — five clusters

cluster-number	number of buildings
cluster 5	172
cluster 10	131
cluster 3	97
cluster 23	90
cluster 2	88
cluster 8	82
cluster 7	68
cluster 22	62
cluster 25	48
cluster 13	45
cluster 4	42
cluster 16	41
cluster 20	37
cluster 17	37
cluster 9	31
cluster 11	30
cluster 12	26
cluster 14	23
cluster 15	21
cluster 18	13
cluster 19	12
cluster 21	11
cluster 6	11
cluster 1	7
cluster 24	5

Table 8. Use case with all energy KPIs — 25 clusters

cluster-number	number of buildings
cluster 2	576
cluster 1	217
cluster 3	186
cluster 4	138
cluster 5	113

Table 9. Use case Energy demand for SH — five clusters

cluster-number	number of buildings
cluster 24	175
cluster 11	134
cluster 4	130
cluster 2	98
cluster 3	94
cluster 8	79
cluster 10	73
cluster 19	49
cluster 16	49
cluster 21	48
cluster 9	47
cluster 6	46
cluster 5	39
cluster 17	31
cluster 20	27
cluster 22	20
cluster 18	20
cluster 14	16
cluster 1	11
cluster 23	9
cluster 7	9
cluster 25	8
cluster 15	8
cluster 12	6
cluster 13	4

Table 10. Use case Energy demand for SH — 25 clusters

4.4 Conclusion

Out of the numerously developed multivariate cluster techniques, the two most applied and therefore also best elaborated techniques — k-means and agglomerative hierarchical clustering — have been evaluated compared to each other. Additionally, to avoid undesirable weighting of the variables because of the differences in order of magnitude, the data has to be rescaled before the clustering is executed. The data can be either standardised — using the z-score formula — or normalised — converting the data points to the exact range from zero to one. The four possible options — combining the both cluster methods and the options to rescale the data — have been tested and the clustering of normalised data using the k-means technique appeared to generate in every use case the most accurate results for all the energy KPIs of interest. The better performance of the normalisation is most likely due to the rescaling of the data by this method to the exact same range for every variable, while rescaling with the z-score formula the range is depending on how much the extrema diverge from the standard deviations. In this way, the variables with larger extrema are weighted as more important although this might not be the most influential variables.

When the clustering is executed and the representative buildings are selected and simulated, the results of these simulations have to be scaled up to represent the entire examined building stock. This upscaling can happen by simply multiplying the results of each representative building by the number of dwellings in the corresponding cluster or by taking a geometrical property, dealing with the size of the dwellings, into account. Since the energetic behaviour of a dwelling is depending on the geometry, scaling the results with these properties might decrease the errors. The investigated situations demonstrate that the peak power and the energy demand for space heating (SH) both can be improved by scaling with respectively the floor area or the volume of the buildings. This finding seems to be explainable since both these energy KPIs are mainly depending on the size of the dwellings (see chapter 3). The specific energy demand for space heating on the other hand cannot be improved by scaling with any of the investigated dwelling properties. This is confirming the previously described explanation that the dependency on specific energy demand of the size of the dwelling is eliminated by dividing by the floor area. Whether the specific energy demand might be improved making use of another dwelling property, which is more related to the composition of the dwelling, will have to appear out of further research.

However, the improvement by the scaling of the results for the energy KPIs peak power and energy demand for SH depends on the number of clusters but also on the use case. For the use case Peak power, energy demand for SH and specific energy demand, the errors of both KPIs are decreased by the scaling for each number of clusters from one to 40. But for the use cases with only interest in one KPI, the improvement only applies for a limited number of clusters. After a certain number of clusters, the most accurate way to scale the results up is just by multiplying by the number of dwellings in each cluster. Following hypothesis has been posed: since the energy KPI not only depends on one property, scaling the results with a geometrical property both implies improvements and additional errors. Due to the limited amount of dwellings in clusters with a rather wide distribution of the property after a certain number of clusters, the improvement by scaling the results can only be made for this limited number of dwellings and the improvements no longer compensate the additional errors. This hypothesis has neither been demonstrated nor rejected. For the KPI peak power the reasoning appeared to be correct, but not for the KPI energy demand for SH. To fully understand which way of upscaling and until which number of clusters the scalings might improve the results, this technique of diversifying the results has to be further investigated in detail.

5. Accuracy of the tailor-made clustering approach

To investigate the accuracy of the developed tailor-made clustering approach, the root mean square errors (RMSEs) of the results based on this approach compared to the results of the reference scenario — based on the complete GIS data approach — are analysed. Part of those errors is caused by the clustering and the generalisation of the results of the representative buildings to all the dwellings in each corresponding cluster. This fraction can be examined by analysing the RMSEs that occur when the building stock is clustered based on the energy KPIs themselves. Additionally, the other part of the errors is due to the confined estimation of the energetic behaviour of the dwellings by the selected set of variables. This fraction can be derived from the difference between the RMSEs of the clustering based on the energy KPIs and the RMSEs of the clustering based on the set of variables. In order to determine the distinction between the use case with interest in all three energy KPIs and the case where only one KPI is considered, the three use cases are consecutively discussed, taking the possible improvements by a scaling with a geometrical property into account (see chapter 4). Subsequently, to validate the tailor-made clustering approach, the errors made by this approach are compared to the errors made by a random clustering of the building stock and a random selection of a representative building out of each cluster.

Then, the accuracy of the tailor-made clustering approach is summarised and compared to the accuracy of district energy simulations based on archetypes or sample buildings. The RMSEs made in this work are compared with the reported errors in several previous studies. Additionally, to put the errors made by the tailor-made clustering approach relative to the reference scenario — based on the complete GIS data approach — in perspective, the reported errors made by a detailed calculation based on the GIS dataset in some previous studies are mentioned. Although no general conclusions can be made because the works do not comprise the same case study, the tailor-made clustering approach seems to be able to generate quite accurate estimations of the reference scenario in this work.

5.1 Use case Peak power, energy demand for SH and specific energy demand

In following paragraphs, the deviations of the results based on the tailor-made clustering approach compared to the results of the reference scenario based on the complete GIS data approach are analysed for the use case Peak power, energy demand for space heating (SH) and specific energy demand. First, to determine the fraction of the errors due to the generalisation by the clustering, the RMSEs of the clustering based on the energy KPIs are discussed. Subsequently, the RMSEs of the clustering based on the selected set of variables (see chapter 3) are compared with the first RMSEs. For two of the three energy KPIs the results can be improved by scaling them with a geometrical property of the dwellings — being the floor area for the KPI peak power and the volume of the buildings for the KPI energy demand for space heating (SH) (see chapter 4). Finally, the tailor-made clustering approach is being validated by a comparison of the RMSEs with those of a clustering based on a random selection of clusters and representative buildings. To be able to compare the RMSEs of the different KPIs, they are plotted in percentages (Figure 28 to 30). These percentages are calculated by dividing the RMSE with the average value over the entire building stock of each energy KPI — which are 34.64 kW for peak power, 32373.44 kWh for energy demand for SH and 130.40 kWh/m^2 for specific energy demand.

5.1.1 Clustering based on the energy KPIs

When the clustering is executed based on the energy KPIs themselves and the results are scaled up with the number of buildings, the RMSEs show the same course for all three KPIs (Figure 28 to 30, left). When the building stock is divided into 15 clusters, the estimation of the KPIs exhibits an error of approximately ten percent. When the stock is split into 40 clusters, the RMSEs remain just above five percent for each KPI. This same course is logical, since equal value is granted to each KPI as they are normalised and not weighted to perform the clustering.

Although scaling the results of the clustering based on the energy KPIs with the floor area or the volume of the dwellings only introduces additional errors (see chapter 4), the RMSEs do not deteriorate enormously by these scalings. The RMSEs of the KPI peak power still show a smooth course that is about one percent higher than the RMSEs when the results are not scaled (Figure 28, right). The RMSEs of the KPI energy demand for SH are somewhat more erratic and are about five percent worse than the RMSEs of the clustering based on the set of variables (Figure 29, right). However, the errors do not increase enormously which indicates that the clustering based on the KPIs contains somehow clusters where the results are similar but can still be improved by scaling them with the geometrical property. This is most likely due to the fact that the clustering is concentrating on three KPIs which ensures that the dwellings in a cluster include a variation of the actual values of the results. This variation can then be estimated by the scaling of the results.

5.1.2 Clustering based on the set of variables

The clustering based on the set of variables generates a similar course for peak power and energy demand for SH (Figure 28 to 29, left). When the building stock is divided into ten clusters, the RMSEs amount to somewhat less than 20 percent. At 20 clusters it amounts to approximately 15 percent and at 40 clusters to about 12 percent. With this, the additional error made by the estimation of the energetic behaviour with the set of variables amounts to about seven percent for these two energy KPIs. The results of specific energy demand perform slightly better (Figure 30). With the building stock split into ten clusters, the RMSE of specific energy demand for space heating amounts to approximately 15 percent. At 20 clusters it comes down to 13 to 14 percent and at 40 clusters to about ten percent. Thus, the additional error made by the set of variables for specific energy demand is four percent. This implicates that the selected set of variables approaches the behaviour of the specific energy demand somewhat better than those of the two other energy KPIs.

However, the error of peak power and energy demand for SH can be improved by scaling the results with a geometrical property (see chapter 4). When scaled with the floor area of the dwellings, the RMSEs for the energy KPI peak power can be partially compensated (Figure 28, right). When the stock is divided into ten clusters, the RMSE is on approximately 15 percent. At 20 clusters it amounts to 11 to 12 percent and at 40 clusters to about eight percent. The improvement decreases with the number of clusters, but overall the scaling reduces the RMSEs with five to two percent. Therefore, the scaling ensures that peak power is slightly better estimated than specific energy demand, since this KPI cannot be improved by a scaling.

The performance for the energy KPI energy demand for SH can also be improved by a scaling with the volume of the dwellings (Figure 29, right). If the stock is split into ten clusters, the RMSE amounts to approximately 15 percent. At 20 clusters it amounts to about 13 percent and at 40 clusters to somewhat more than ten percent. Again, the improvement decreases with the number of clusters, even faster than with the KPI peak power. At 40 clusters, the improvement is almost negligible, but still this improvement ensures that energy demand for SH is as well estimated as the specific energy demand.

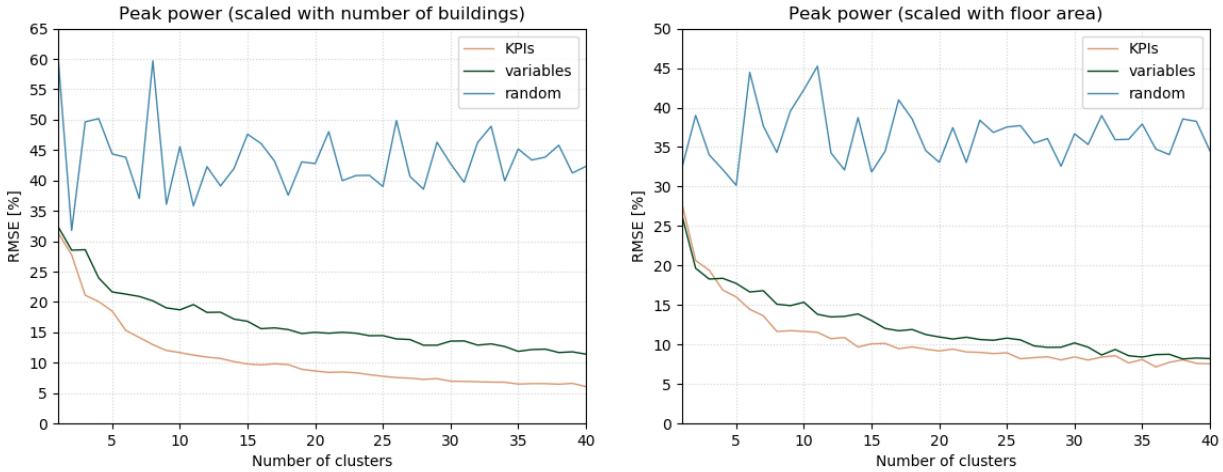


Figure 28. Graphs showing the root mean square errors in percentages in function of the number of clusters for the energy KPI peak power in use case Peak power, energy demand for SH and specific energy demand. The results are scaled up with either the number of buildings (left) or scaled with the floor area of the buildings (right).

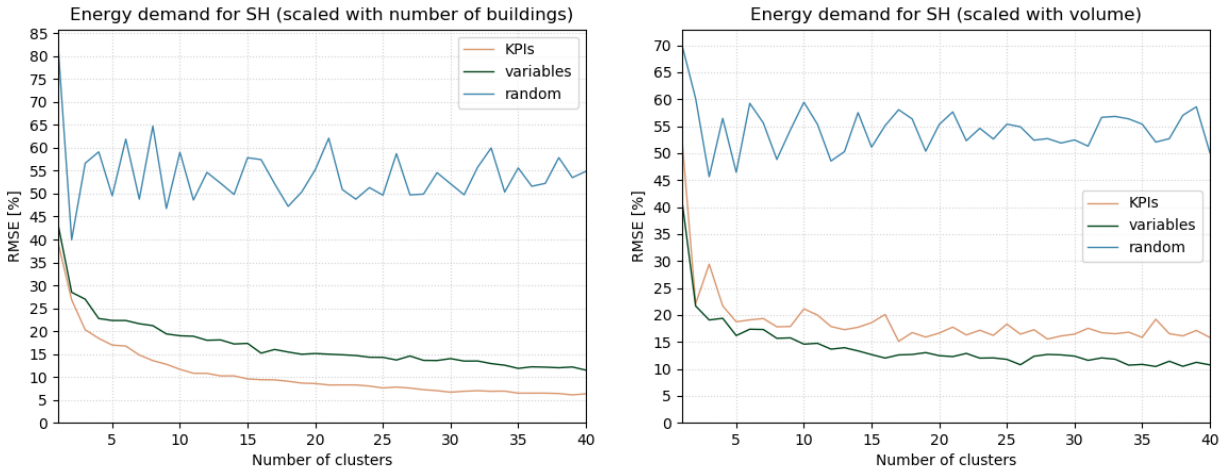


Figure 29. Graphs showing the root mean square errors in percentages in function of the number of clusters for the energy KPI energy demand for SH in use case Peak power, energy demand for SH and specific energy demand. The results are scaled up with either the number of buildings (left) or scaled with the volume of the buildings (right).

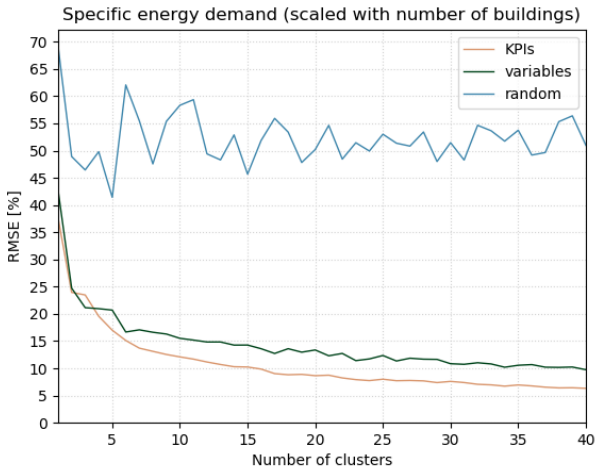


Figure 30. Graphs showing the root mean square errors in percentages in function of the number of clusters for the energy KPI specific energy demand in use case Peak power, energy demand for SH and specific energy demand. The results are scaled up with the number of buildings.

5.1.3 Validation by a random selection

When the building stock is randomly divided into clusters and all the dwellings in each cluster are represented by a randomly selected building of that cluster, the RMSEs show a very erratic course (Figure 28 to 30). This means that the better results for certain numbers of clusters — where a lower RMSE occurs — are rather coincidental and therefore these results are not really reliable. Moreover, the RMSEs of the random clustering are a lot higher than those of the clustering based on the set of variables. The errors are fluctuating around 40 to 50 percent for the KPI peak power (Figure 28, left), around 50 to 60 percent for energy demand for SH (Figure 29, left) and around 50 to 55 percent for specific energy demand (Figure 30). The results for the KPI peak power can be slightly improved by scaling them with the floor area of the dwellings so that they fluctuate between 35 and 40 percent (Figure 28, right). But also in this way the errors remain significantly higher than the errors made by the tailor-made clustering approach. Therefore, a random clustering does not generate reliable results for the district simulated in this work, while the tailor-made clustering approach produces smaller errors. To be able to generally conclude that the developed approach is more accurate than a random clustering, other districts will have to be examined. The actual accuracy of this approach depends on the number of clusters and can thus partially be chosen in function of the needs of the study.

5.2 Use case Peak power

The results of one energy KPI can be further improved by focussing on just this KPI. Since the results of the other KPIs are somewhat deteriorated, this only has to be considered if only this KPI is of interest. In the following paragraphs, the results of the use case Peak power will be discussed, similarly to the results of the previous use case. First, the clustering based on the KPI itself is reviewed to examine the error made by the clustering. Then, the additional error made by the estimation of the energetic behaviour with the set of variables is investigated by compering the results of the clustering based on the energy KPI with the results of the clustering based on the selected set of variables. Again, the RMSEs are plotted in percentages (Figure 31 to 32).

5.2.1 Clustering based on the energy KPI

When the building stock is clustered based on the energy KPI peak power, the RMSE curve of this KPI — scaled up with the number of buildings — approaches zero with an increasing number of clusters (Figure 31, left). Already at five clusters the RMSE drops below ten percent and at ten clusters it reaches five percent. If there is only interest in this one KPI, the clustering can reach an optimal division of the building stock to approach the results of this KPI. The variations of the KPI peak power for the dwellings in one cluster, that occurred in the previously discussed use case, are no longer present since the cluster technique can focus on only one input factor to divide the building stock.
However, the other KPIs show a much worse course of the RMSEs. The errors for both energy demand for SH (Figure 32, left) and specific energy demand for space heating (Figure 32, right) made by the clustering based on the KPI peak power are even higher than the errors made by the clustering based on the selected set of variables. Although this set is selected only based on the ranking for the KPI peak power, the variables still contain some ability to explain the results of the two other KPIs since they likewise depend on these variables to greater or lesser extent. Where the course of the RMSEs for specific energy demand approximately approaches the course of the RMSEs of the random clustering, the errors made for energy demand for SH are a lot lower — fluctuating around 25 percent. This indicates that the KPI energy demand for SH is somehow linked to the KPI peak power, while this is absolutely not the case for the KPI

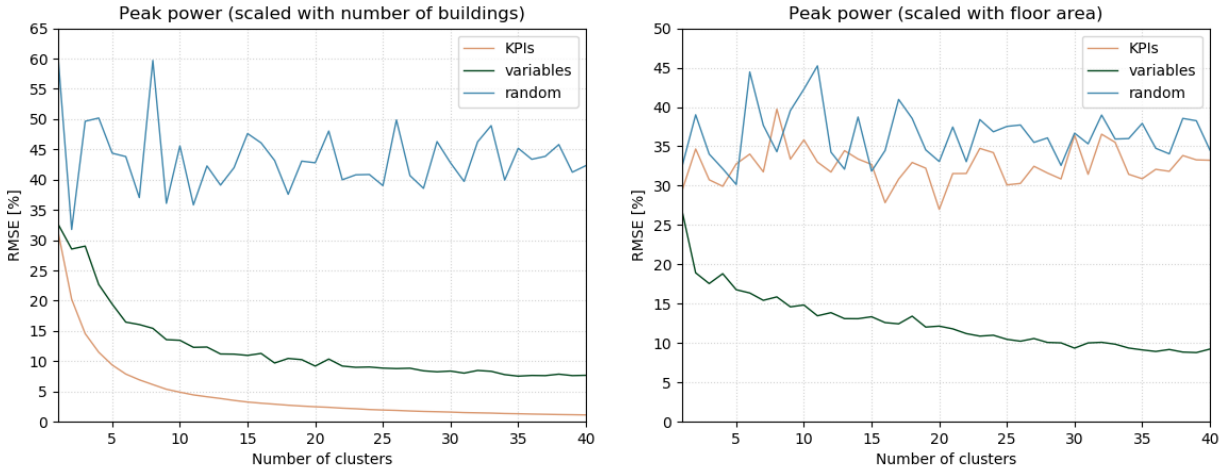


Figure 31. Graphs showing the root mean square errors in percentages in function of the number of clusters for the energy KPI peak power in use case Peak power. The results are scaled up with either the number of buildings (left) or scaled with the floor area of the buildings (right).

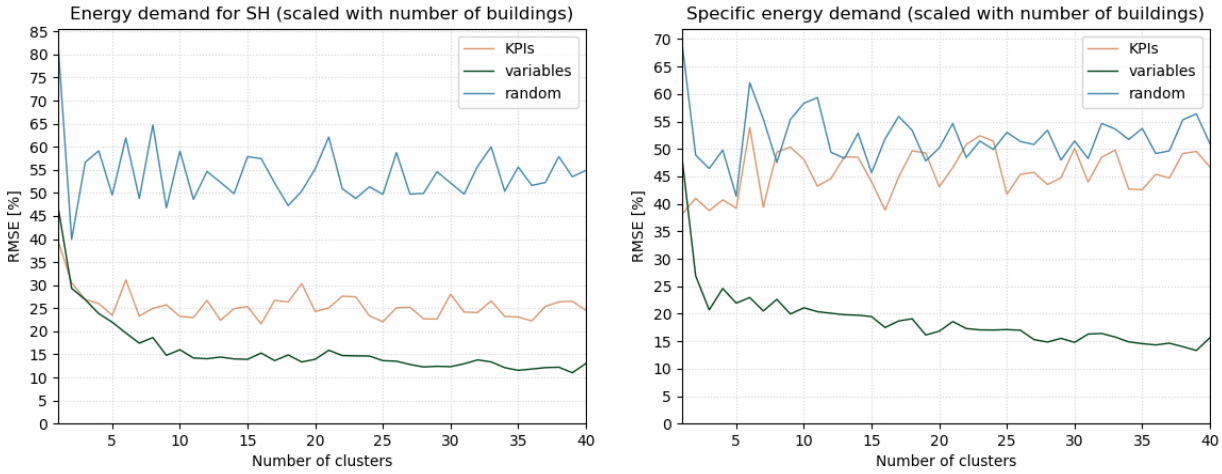


Figure 32. Graphs showing the root mean square errors in percentages in function of the number of clusters for the energy KPIs energy demand for SH (left) and specific energy demand (right) in use case Peak power. The results are scaled up with the number of buildings.

specific energy demand. This is also apparent from the ranking of the variables (see chapter 3): the most important variables for peak power are very similar to those for energy demand for SH and concentrate mainly on the size of the dwellings — *floor area* and *footprint* —, while those for specific energy demand emphasise also on the composition of the plan — *loss-to-floor-area ratio*.

5.2.2 Clustering based on the set of variables

The clustering based on the selected set of variables generates additional errors from about seven to eight percent (Figure 31, left). When the building stock is divided into ten clusters, the RMSE amounts to approximately 13 percent. At 20 clusters it amounts to about ten percent and at 40 clusters to somewhat less than eight percent. The additional errors are similar to those for the previous use case, which means that the set of variables selected for the use case Peak power estimates the behaviour of this energy KPI as well as the set for the previous use case is estimating the behaviour of all the KPIs combined. However, the improvement by only concentrating on peak power decreases with the number of clusters: at ten clusters this improvement is about seven percent, at 20 clusters about five percent and at 40 clusters only about two percent. Thus, if the desired degree of accuracy is large and more than 40 clusters

are needed, concentrating on only the KPI peak power will not improve the results anymore.
Up to five clusters, the RMSEs can be decreased by scaling the results with the floor area of the dwellings. This reduces the RMSE at two clusters from about 29 percent to only 18 percent. But since this reduction of the RMSEs can only be applied up to a certain number of clusters, the improvement of the results — scaled with the floor area — by concentrating on one KPI compared to the results — also scaled with the floor area — of the previous use case diminishes. Yet, concentrating on only the KPI peak power still generates somewhat lower RMSEs, but only less than one percent.
The estimation of the KPIs that are not of interest in this use case are, logically, somewhat worse than in the previous use case. The RMSEs of energy demand for SH — scaled up with the number of buildings — are about one to two percent higher than those for the previous use case — scaled with the volume of the buildings —, despite the somewhat more erratic course (Figure 32, left). The errors of specific energy demand — both scaled up with the number of buildings — are about five percent higher (Figure 32, right). Again, this indicates the link between the KPI peak power and energy demand for SH and the discrepancy between peak power and specific energy demand. The curves do not show a similar course as the RMSEs of a random clustering since the set of variables is still including some explanation of the two other KPIs.

5.3 Use case Energy demand for SH

Besides the use case Peak power, the scenario of interest in only the KPI energy demand for space heating (SH) is also investigated. In the following paragraphs, the results of this use case will be discussed like the previous use cases. The results of the clustering based on the energy KPI itself are first examined to investigate the error made by the generalisation of the clustering. Then, the results of the clustering based on the set of variables are discussed to identify the additional error made by the estimation of the set. Here too, the RMSEs are plotted in percentages (Figure 33 to 34).

5.3.1 Clustering based on the energy KPI

The RMSEs of the KPI energy demand for SH made by the clustering based on the KPI itself — scaled up with the number of buildings — also approach zero when the number of clusters increases (Figure 33, left). Again, the error amounts to somewhat more than ten percent when the building stock is divided into five clusters and to about five percent at ten clusters. This is the same course as the RMSEs from the KPI peak power in the previous use case. The clustering is again reaching an optimal division of the building stock to estimate the results of the energy KPI energy demand for space heating (SH).
The RMSEs of the other energy KPIs are higher and show a much more erratic course (Figure 34). However, the errors for both KPIs lie under the errors of a random clustering and therefore both KPIs are somehow related to the energy demand for SH. As explained previously, both peak power and energy demand for SH mainly depend on the size of the dwellings. The clustering based on the KPI energy demand for SH is therefore indirectly also taking peak power into account. Since specific energy demand is determined based on the energy demand for SH (see chapter 2), the clustering based on the energy demand for SH is again indirectly including a connection with the specific energy demand. The RMSEs for peak power — fluctuating between 15 and 25 percent — are smaller than those for specific energy demand — which fluctuate between 30 and 40 percent. The link between the KPIs energy demand for SH and peak power is thus larger than the link between energy demand for SH and specific energy demand.

5.3.2 *Clustering based on the set of variables*

The additional errors made by the estimation of the energetic behaviour by the set of variables amount to about nine to ten percent (Figure 33, left). When the building stock is divided into ten clusters, the RMSE is approximately 17 percent. At 20 clusters the error amounts to about 14 percent and at 40 clusters to circa nine percent. With this, the estimation of the energy demand for SH by the selected set of variables is somewhat less accurate than the estimations in the other use cases, where the additional errors amount to about seven percent. Again, the improvement by only concentrating on energy demand for SH is decreasing with the number of clusters: at ten clusters the improvement is about three percent and from 20 clusters on, this improvement only amounts to one percent. The improvement of the results for the KPI energy demand for SH by only concentrating on this KPI is lower than the improvement for the KPI peak power — as described in the previous section. This illustrates the somewhat lesser estimation of the KPI energy demand for SH by the selected set of variables for this use case than the estimation of peak power by the corresponding set of variables for the previously discussed use case.

Up to 12 clusters, the RMSEs can be decreased by scaling the results with the volume of the dwellings. Here, the same determination applies as in the use case Peak power: the improvement of the results — scaled with the volume — for the use case with only interest in

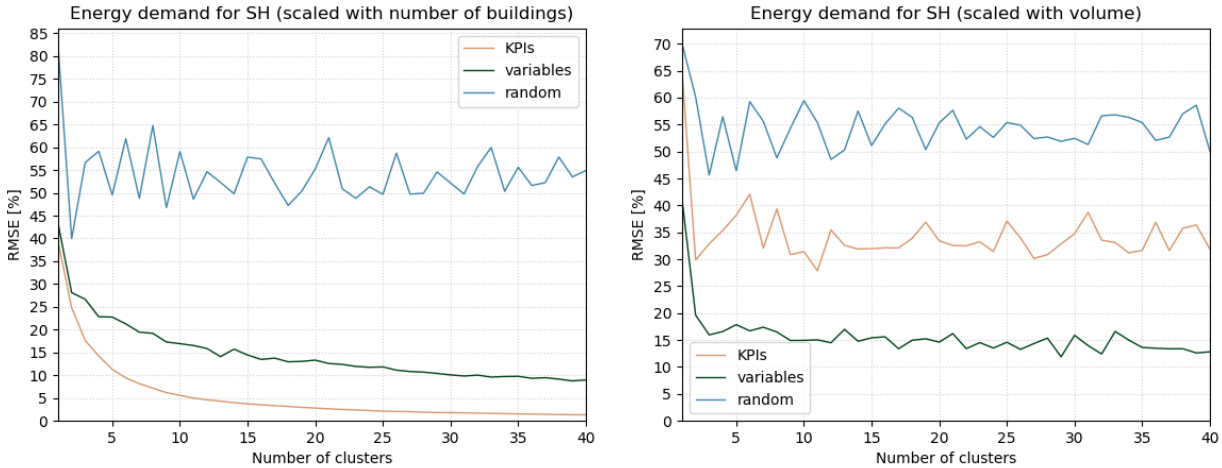


Figure 33. Graphs showing the root mean square errors in percentages in function of the number of clusters for the energy KPI energy demand for SH in use case Energy demand for SH. The results are scaled up with either the number of buildings (left) or scaled with the volume of the buildings (right).

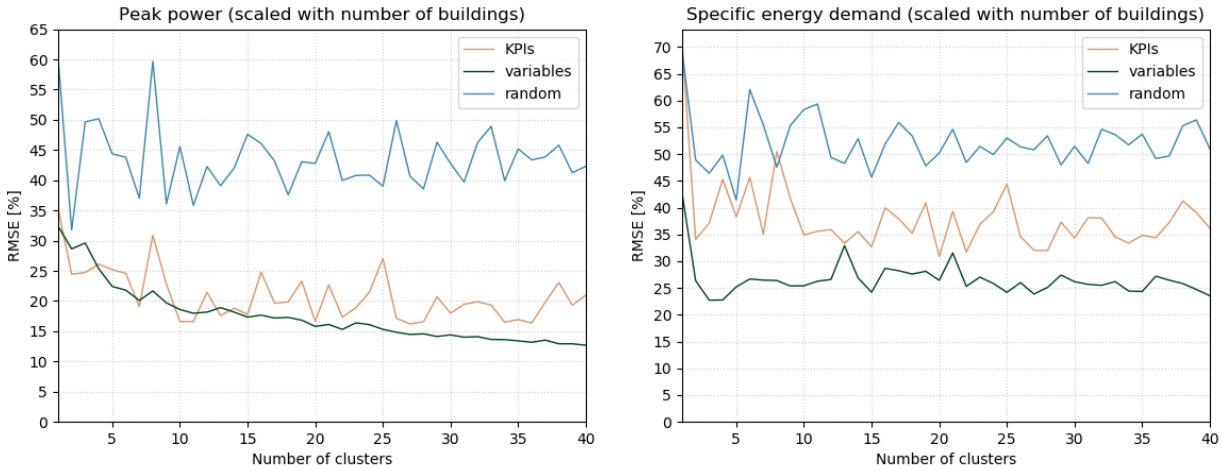


Figure 34. Graphs showing the root mean square errors in percentages in function of the number of clusters for the energy KPIs peak power (left) and specific energy demand (right) in use case Energy demand for SH. The results are scaled up with the number of buildings.

the energy demand for SH relative to the results — also scaled with the volume — for the use case with interest in all three energy KPIs decreases. This is due to the fact that the improvement by the scaling in the use case Energy demand for SH can only be applied up until 12 clusters and in the use case Peak power, energy demand for SH and specific energy demand this improvement is still possible at 40 clusters. But concentrating on only the energy demand for SH generates still somewhat lower RMSEs —of about one percent — for this KPI than the first discussed use case, which combines an interest in all three energy KPIs. However, when a high degree of accuracy is desired and more than 40 clusters are needed, the slight improvement of the results for the energy demand for SH by only concentrating on this KPI might not be interesting enough to consider this use case, since in this way the results of the other KPIs are less accurate.
The RMSEs of peak power — when scaled up with the number of buildings — are about one percent higher than in the use case with interest in all three KPIs — also scaled with the number of buildings (Figure 34, left). The ranking of variables for energy demand for SH and for peak power are very similar. Therefore, the KPI peak power is indirectly also taken into account in this use case. However, the results of the KPI peak power can be improved in the use case Peak power, energy demand for SH and specific energy demand by scaling them with the floor area of the dwellings. Then, that use case is performing significantly better — about five percent. The RMSEs of specific energy demand fluctuate around 25 percent and are thus unreliable. Although the clustering based on the KPI proved that energy demand for SH and specific energy demand are related, this does not appear in the clustering based on the set of variables.

5.4 Summary of the accuracy and comparison with previous studies

The number of clusters required to achieve a certain degree of accuracy is summarised for each energy KPI in the table below (Table 11). The best possible way of upscaling for each KPI in each use case has been taken into account. An estimation of the results of the complete GIS data approach with a root mean square error (RMSE) of less than 20 percent is achieved with already two or three clusters for the KPIs peak power and energy demand for space heating (SH) in both use cases with an interest in these KPIs. The RMSEs of the KPI specific energy demand are declining less rapidly at the lower number of clusters and thus six clusters are required to reach an error less than 20 percent. The use cases with interest in only one KPI will achieve the potential desired degree of accuracy at a lower number of clusters than the use case Peak power, energy demand for SH and specific energy demand. Although the exact RMSEs of the KPIs between both of the use case with an interest in the KPI are not very

energy KPI	use case	≤ 20%	≤ 15%	≤ 10%
peak power	Peak power, energy demand for SH and specific energy demand	2 clusters	11 clusters	31 clusters
	Peak power	2 clusters	9 clusters	22 clusters
energy demand for SH	Peak power, energy demand for SH and specific energy demand	3 clusters	10 - 11 clusters	> 40 clusters
	Energy demand for SH	2 clusters	9 - 11 clusters	30 - 35 clusters
specific energy use	Peak power, energy demand for SH and specific energy demand	6 clusters	12 - 13 clusters	40 clusters

Table 11. Summary of the number of clusters required when a certain degree of accuracy is desired, for each energy KPI in all possible use cases.

distinct, the use cases with only one KPI of interest reach a maximum error of ten percent at a significantly lower number of clusters. This is due to the reducing decrease in RMSE with an increasing number of clusters, which generates a flattening curve. Estimating the KPI energy demand for space heating (SH) with high degree of accuracy requires more clusters than the two other KPIs. Moreover, in the previous section the lesser approach of the KPI energy demand for SH by the selected set of variables is remarked. Therefore, it can be concluded that this KPI is harder to estimate than both other KPIs.

To validate the errors made by the tailor-made clustering approach, the reported errors of some previous studies are discussed. Shimoda et al. (2004) classified the building stock of Osaka, Japan, into 55 family types, two dwelling types and ten floor area groups. They summarised the family types into 23 categories and demarcated 20 archetypes based on the floor area groups and the dwelling types. As the residential building stock of Osaka contains 1128 dwellings, the defined archetypes consist of 1.77 percent of the stock. Converted to the examined stock in this case study, would mean that the building stock is divided into 22 clusters. Shimoda et al. (2004) reported that the statistical value of the total energy consumption — taking electricity for heating, cooling, domestic hot water, lighting, kitchen and other appliances into account — was underestimated with 18 percent. When the building stock in this case study is divided into 22 clusters, the root mean square error (RMSE) relative to the reference scenario — based on the complete GIS data approach — in the use case energy demand for space heating (SH) amounts to 12.4 percent. However, the comparison of these percentages must be considered with some caution, since the case studies contain different conditions and data availabilities, distinct scopes of the calculation of the energy demand and Shimoda et al. (2004) use various family types while this case study applied the standard user profile in each dwelling. Above all, the verification of the results by Shimoda et al. (2004) happened with a comparison with the statistical values and in this work with the values from the complete GIS data approach as statistical values are not available.
Dall'O', Galante and Torri (2012) examined the residential building stock of Carugate, Italy, which contains approximately 1230 apartment buildings. They classified the building stock based on the construction period and selected a certain number of sample buildings out of each class, depending on the fraction of apartment buildings belonging to the class. The total number of sample buildings amounts to 93, which comes down to 7.5 percent of the building stock. Converted to a building stock which contains 1230 dwellings, this would mean that the stock is divided into 92 clusters. Dall'O', Galante and Torri (2012) reported an accuracy of the developed method of approximately ten percent for the calculation of the total energy demand for heating and domestic hot water. This accuracy was determined relative to the actual energy consumption supplied by the gas distributor. An accuracy degree of RMSEs less than ten percent for energy demand for SH in this work is reached at a much lower number of clusters in both use cases. However, the comparison of the accuracies must again be considered with caution, since the case studies are not equal and the errors in the method of Dall'O', Galante and Torri (2012) are reported relative to the actual data, while in this work, they are relative to the results of the complete GIS data approach.

Some researchers documented the errors by simulations based on detailed 3D city models of the building stock. Orehounig et al. (2011, in: Ghiassi et al., 2015) for example estimated the energy consumption of a village of about 100 dwellings based on two approaches. They categorised the building stock based on the use and the construction periods and generated archetypes. This simplified approach resulted in an estimation of the actual values with a deviation of 25 percent. Then, they modelled each dwelling in detail and simulated them. This approach generated an estimation with approximately eight percent of deviation from the

actual values. This means that a simulation based on a simplified archetype approach generates about 17 percent higher deviations than a detailed calculation.
Nouvel et al. (2013) compared detailed calculations of two districts in Germany, using an LOD1 model for the case study of Grünbühl in Ludwigsburg and an LOD2 model for the case study of Rintheim in Karlsruhe. Taking the energy demand for space heating and domestic hot water into account, they reported an overestimation by the simulations of the actual values of 21 percent when the building stock is modelled in LOD1. Moreover, they indicate high uncertainties of the results on the level of an individual building. When the stock is modelled in LOD2, the overestimation of the total simulated heating demand of the entire district amounts to only seven percent.
Although the case studies are not comparable, these reported deviations between detailed calculations and actual measured values might place the errors made by the tailor-made clustering approach into perspective. As a detailed calculation generates a deviation of about seven to eight percent of the actual situation, an error of less than ten percent made by the tailor-made clustering approach might be also acceptable. Especially when a model of LOD1 is used, the errors of the tailor-made clustering approach are less influential than those of the simulation itself. However, it is not examined whether the errors of the tailor-made clustering approach are complementary to those of the complete GIS data approach or if they partly compensate these. To exactly determine the errors made by the developed tailor-made clustering approach relative to the actual situation, further research and an extension of the data availability is required.

5.5 Conclusion

The measured root mean square errors (RMSEs) of the results based on the tailor-made clustering approach compared to the results based on the complete GIS data approach are investigated and discussed. The errors are partly depending on the generalisation of the results of the representative buildings by the clustering and partly on the estimation of the energetic behaviour by the sets of variables. Examining the errors of the clusterings based on the KPIs themselves in the three defined use cases showed that clustering based on one factor — one KPI in this case — is creating optimal clusters to estimate the results, reaching an RMSE of nearly zero percent when the building stock is divided into 30 or 40 clusters. When the clustering happened based on three KPIs, the minimum possible RMSE achieved with a limited number of clusters increases from zero to five percent. This is easily understood since the distribution of the values differ for each KPI and thus the clustering cannot focus on only one factor anymore. Since this is exactly the purpose when the clustering happens based on the set of variables, this is absolutely no disadvantage.
The additional errors made by the estimation of the energetic behaviour by the sets of variables are variating with the number of clusters and amount to about eight percent for the KPI peak power in the use case Peak power and to about ten percent for the KPI energy demand for SH in the use case Energy demand for SH. In the use case Peak power, energy demand for SH and specific energy demand, the additional errors are lower since the RMSEs of the clustering based on the KPIs amount to more. The distribution of the focus of the clustering when clustered based on the KPIs is partly compensated by the fact that each variable contains some information about every energy KPI. Therefore, each variable will partly vouch for every KPI. This explains the lower additional errors, in the use case with an interest in all the KPIs, of about seven percent for the KPIs peak power and energy demand for SH and about four percent for the KPI specific energy demand.

To reach a certain degree of accuracy for the KPI energy demand for SH, more clusters are required than for the other two KPIs, in both the use case with interest in all the KPIs as in the use cases with an interest in only one KPI compared to each other. Moreover, the improvement of the results by only concentrating on the energy demand for SH instead of on all the KPIs is lower than the improvement when concentrated on the KPI peak power. This indicates that the set of variables selected for the use case Energy demand for SH is estimating the KPI less accurately than the set for the use case Peak power.
More generally, also applies that if the tailor-made clustering approach needs to be further specified to generate more accurate results, investigating the possibilities of several variables is more useful than examining other cluster techniques. The clustering is quite optimally dividing the building stock into groups taking all the specified variables or KPIs into account, while the selected sets of variables are highly depending on the particular case study. The availability of the data might be extended in the future or the varying user profiles can be taken into account. When the case study contains another district, the specific conditions might differ. Therefore, further examination of the sets of variables will most likely yield more improvement of the accuracy of the tailor-made clustering approach.

As the root mean square errors (RMSEs) are diminishing with an increase of the number of clusters, the desired degree for a particular study can be prompted and will regulate the required number of clusters. To be able to determine the required number of clusters based on the prompted degree of accuracy, further research is needed. Since only one case study is investigated, no general conclusions on the link between the number of clusters and the errors of the model can be made. As the approach will have to be able to guarantee the minimum accuracy, a lot more case studies will have to be examined.

The developed tailor-made clustering approach is validated by means of a comparison with the errors of a random clustering of the building stock and a random selection of a representative building for each cluster. Since the RMSEs of the random clustering are significantly higher and the curve of these errors show a much more erratic course, the tailor-made clustering approach is able to estimate the results of the complete GIS data approach much more accurately and moreover, more reliably.
In addition, the errors made by this approach are compared to reported errors in other works, made by an approach based on archetypes or sample buildings. Although it is not proven since the case studies are different, the results of the tailor-made clustering approach create the impression that this approach is generating a more accurate estimation of the results based on the complete GIS data approach than the archetype or sample building approaches.
Finally, the errors of the tailor-made clustering approach compared to the reference scenario — based on the complete GIS data approach — can be placed into perspective by mentioning the errors of models based on a detailed calculation — similar to the complete GIS data approach — and the actual values of energy consumption. The reported errors of detailed calculations amount to about seven percent when the building stock is modelled in LOD2. The errors of the tailor-made clustering approach with a limited number of clusters are of the same order of magnitude. When the stock is modelled in LOD1, the energy consumption is overestimated by 21 percent. Then, the errors of the tailor-made clustering approach are less influential than the ones made by the simulation. Therefore, the errors of the tailor-made clustering approach might also be acceptable. However, as explained in the previous section, these errors might be complementary to those of the simulations. To exactly determine the errors made by the tailor-made clustering approach compared to the actual values, the data availability will have to be extended before further research can happen.

6. Conclusions and future research

As integrated measures become more and more prevalent on the scale of districts or even larger, the application of district energy simulations is steadily ascending. Several researchers have elaborated methods to simplify the modelling of the building stock and with that the simulations themselves. Since the data availability and the computational capabilities are expanding, simulations based on detailed calculations of each dwelling individually — appointed as the complete GIS data approach — are preferable. However, these simulations still require a long calculation time. Therefore, this work investigated the newly developed tailor-made clustering approach and following research question was posed:

> *Can district energy simulations based on tailor-made clusters generate sufficiently accurate simulations compared to the complete GIS data approach without drastically increasing the complexity of the model?*

To answer this question, two aspects have been distinguished: the implementation of the developed approach and the accuracy of the results generated by this approach. First, the implementation of this approach was examined posing two groups of sub-questions. The clustering happened based on specific sets of variables, which have been determined for each defined use case in chapter 3. The selection of the cluster technique itself and the upscaling of the results to the entire building stock are discussed in chapter 4. Subsequently, the accuracy of the tailor-made clustering approach was investigated in chapter 5. The errors of the tailor-made clustering approach relative to the reference scenario — based on the complete GIS data approach — are discussed and validated by comparing them with a random clustering and selection of representative buildings. Finally, to place the errors into perspective, reported errors of previous studies are described.

6.1 Variables

A literature study revealed that researchers each have used several different variables in previous studies to demarcate archetypes. Therefore, the selection of the set of variables based on which the clustering must happen is not unambiguously defined. First, all the applied variables in a collection of previous studies are enumerated. Since the possibility to apply a variable to execute the clustering is highly depending on the case study, the potential variables in this work are determined and explained. Whether or not a variable is qualified to be applied, is depending on both the data availability and the specific conditions of the particular case study. As the case study in this work comprises a suburban Flemish neighbourhood and only the energy demand for space heating is taken into account, variables dealing with the present systems, operational parameters and environmental factors are not eligible. In this way, following eleven potential variables have been defined: footprint, floor area, total loss area, number of stories, loss-to-floor-area ratio, compactness, typology, heated volume, window-to-wall ratio, window area and construction period.

Then, a method to determine which of these potential variables are most influential in the explanation of the energetic behaviour of the dwellings — expressed in three defined energy KPIs, being peak power, energy demand for SH and specific energy demand — is described. To obtain a ranking of the variables for each KPI, a linear multivariate regression of the results of the reference scenario was executed for the KPIs individually. The construction period came forward as an important variable for every KPI. For both the KPIs peak power and energy demand for SH, the most influential variables contain mainly information about the size of the

dwellings — such as floor area, footprint and total loss area —, whereas the KPI specific energy use is rather explained by variables dealing with the composition of the dwellings — like loss-to-floor-area ratio. This is a logical outcome, since the dependency on the KPI specific energy use from the size of the dwelling is eliminated by dividing the energy demand for SH by the floor area of the dwellings.
Although the linear multivariate regression is suggesting the sets of eight variables as most accurate sets to estimate the results of the reference scenario, the most optimal sets to execute the clustering appear not to be the sets of eight variables. Instead, the optimal number of variables, used to execute the clustering, is about three or four, which is determined by comparing several sets of variables with divergent numbers of variables. For each use case, the set of variables, which generates the most accurate results, is selected. The clustering for the use case Peak power happened based on the floor area, total loss area and construction period. For the use case Energy demand for SH, the set contains the footprint, total loss area and construction period. The set of variables for the use case Peak power, energy demand for SH and specific energy demand is compiled out of the rankings for the three KPIs and includes the footprint, total loss area, loss-to-floor-area ratio and construction period.

Finally, which variables are the most influential in each use case is verified by examining the distributions of the variables in the clusters of a clustering based on the energy KPIs of interest in the particular use case. The distribution of every variable is compared to the distribution of the variable in the entire building stock. If the distribution is significantly smaller, the variable contains a strong link with the explanation of the KPIs of interest in the use case. Generally, the findings of this verification tend to confirm the sets selected by the applied method. Only the variable footprint does not show a significantly smaller distribution in the use cases Energy demand for SH and Peak power, energy demand for SH and specific energy demand. This can be explained by understanding this variable as a ‘pseudo-variable’, which does not contain a direct connection with the explanation of the KPIs, but includes correlations about several other variables dealing with the size of the dwellings — like floor area, total loss area and heated volume.

6.2 Clustering of the building stock

The two most applied and thus most elaborated multivariate cluster techniques — being k-means clustering and agglomerative hierarchical clustering — are described and have been applied. Before the clustering can be executed, the data has to be rescaled using either standardisation or a normalisation to avoid undesirable weighting of the variables based on which the clustering will happen. To perform the standardisation, the z-score formula has been used and the normalisation converts the data into a range from exactly zero to one. The results of the four possible methods are compared to each other. Using the k-means cluster technique to cluster normalised data appeared to generate the most accurate results for every KPI of interest in the three examined use cases. The preference for normalisation is due to the fact that this method scales the data into the exact same range and therefore weighing all the specified variables as equal, while standardisation is giving a larger weight to the variables with more spread extrema, which are not necessarily the variables that are mostly determining the energetic behaviour.

When the clustering is executed and the representative buildings are selected and simulated, the results of these simulations have to be scaled up to represent the entire building stock. This upscaling can happen by multiplying the results of each representative building by the number of dwellings in the corresponding cluster or by taking a geometrical property of the dwellings

into account. Since the energetic behaviour depends on the geometry of the dwellings, the results can be diversified by scaling them with a geometrical property dealing with the size of the dwellings. Whether or not this diversification is improving the results, is different for each KPI, use case and number of clusters. The results for the KPI peak power can be improved by scaling them with the floor area of the dwellings and the results for the KPI energy demand for SH by scaling with the volume of the dwellings. Since both KPIs mainly depend on the size of the dwellings — which also appeared from the ranking of the variables —, the improvement by scaling with these geometrical properties is easily understood. The KPI specific energy demand cannot be improved by scaling with any of the examined properties. Since the dependency on this KPI on the size of the dwellings is eliminated, it seems logical that scaling with a property that is dealing with the size is not improving the results.

As mentioned, the improvement by scaling with a geometrical property does not account for every number of clusters in any use case. In the use case Peak power, energy demand for SH and specific energy demand, the results of the KPIs peak power and energy demand for SH are improved by scaling them with the geometrical properties for every examined number of clusters from one to 40. However, in the use cases with interest in only one energy KPI, no improvements are made anymore after a certain number of clusters. In the use case Peak power, the results for the KPI of interest are improved up to six clusters and in the use case Energy demand for SH, up to 12 clusters. A hypothesis is posed that this could be explained by means of the number of dwellings in clusters with a rather wide distribution of these variables. Since the KPIs are not only depending on one variable, both improvements and additional errors are made when the results are scaled with a geometrical property. The clusters with a wide distribution of the geometrical property, used to scale the results, most likely introduce more improvements than additional errors, while this is the opposite in clusters with a narrow distribution. Therefore, when the building stock is divided in a lot of clusters and only limited number of dwellings are part of a cluster with a rather wide distribution of the property, the additional errors are no longer compensated by the improvements and the scaling is no longer recommended. This hypothesis is neither rejected nor proven, since the examination of the KPI peak power tends to confirm this but the KPI energy demand for SH not. To be able to draw general conclusions or to explain this phenomenon with certainty, more research is required. Which way of upscaling generates the most accurate results, how these results can be diversified introducing improvements and up until which number of clusters this improvements applies, has to be further investigated by examining other case studies and use cases. However, as the results of the upscaling by multiplying by the number of dwellings are already quite accurate, the urge of this further research must be nuanced.

6.3 Accuracy of the tailor-made clustering approach

The root mean square errors (RMSEs) of the results based on the tailor-made clustering approach relative to the results based on the complete GIS data approach can be caused by different aspects. A fraction of the errors is due to the generalisation of the results of the representative buildings to the entire stock by the clustering and another fraction is created by the estimation of the energetic behaviour by the selected sets of variables. The errors generated by the clustering of the stock based on one of the energy KPIs showed that the applied cluster technique — k-means clustering — is capable of dividing the building stock in nearly perfect groups to estimate the results of the reference scenario. When more factors — in this case three KPIs — are taken into account, the clustering can no longer focus on one factor, which makes that the groups are not containing dwellings with exactly the same results for one KPI anymore and therefore the errors increase. But since the sets of variables consist

of factors that should all be taken into account, this is exactly what is desired when the clustering is executed based on the sets of variables.
The additional errors caused by the sets of variables are more significant in the use cases with interest in only one KPI than the errors due to the clustering. In the use case Peak power, energy demand for SH and specific energy demand, these additional errors are lower than in the other use cases, since the fraction caused by the clustering is higher. The distribution of the focus when clustered on the three energy KPIs, which generated this higher fraction of the clustering, is partly compensated by the fact that each variable contains some information on every KPI.

As the applied cluster technique manages to divide the building stock with a very high accuracy based on the given factors, the selected sets of variables should be the first consideration of further examination if the accuracy of the tailor-made clustering approach has to be improved. When another case study is investigated, the list of potential variables will be different and therefore, the selected sets of variables will have to be adjusted. Moreover, due to the data availability and specific conditions of the particular case study in this work, a lot of previously applied variables could not be examined. Thus, more progress is possible in the selection of sets of the variables than in the cluster technique. Especially the energy KPI energy demand for SH is estimated the least accurately by the set of variables for the use case that concentrates on only this KPI.

The accuracy of the results is increasing with the number of clusters. Therefore, the desired degree of accuracy for a particular study can be prompted by the researchers and will define the required number of clusters. However, since only one case study and three use cases have been examined, this link cannot be generalised based on these results. To be able to guarantee at least the prompted degree, the link between the accuracy and the therefore required number of clusters has to be further investigated based on a lot more case studies.

The tailor-made clustering approach is validated by a comparison with results of a random clustering and selection of representative buildings. The errors generated by this random clustering are in every situation a lot larger and the curves show an erratic course, which indicates that the results of this random clustering are not reliable. The tailor-made clustering approach generates clearly a more accurate estimation of reference scenario based on the complete GIS data approach.
The reported errors of simulations based on archetypes or sample buildings in previous studies are discussed and compared to the errors of the tailor-made clustering approach. Although this is not proven since the case studies are different, the developed approach seems to generate more accurate estimations of the simulations based on the complete GIS data approach than the approaches that make use of archetypes or sample buildings, when the number of clusters equals the number of archetypes or sample buildings used in the other case studies, taking the differences in number of dwellings in the entire stock into account.
Finally, the errors made by the developed approach are placed into perspective by mentioning the reported errors in some other studies of detailed simulations — similar to the complete GIS data approach — compared to the actual energy consumptions. When the building stock is modelled in LOD2, the errors made by the tailor-made clustering approach are of the same order of magnitude as the errors made by the simulation. When the stock is modelled in LOD1, the errors by the developed approach are a lot less influential than those by the simulations. Thus, the errors by the tailor-made clustering approach might be also acceptable. However, herewith must be stated that the errors by the developed approach might be complementary to those by the simulations.

6.4 General conclusion and future research

As only one particular case study is examined, the scope of this work cannot prove general conclusions on the developed tailor-made clustering approach. However, in order to be able to generalise the findings, the described method to test the approach can easily be used to investigate multiple other case studies and scenarios. The approach did already demonstrate that it is able to generate accurate results compared to the reference scenario — based on the complete GIS data approach — with a strong simplification of the complex model.
When other case studies are investigated, the potential variables will be different since they are strongly depending on the data availability and the specific conditions of the case study. Therefore, the variables will be ranked in a different order for each KPI and other sets of variables might appear as most accurate for the defined use cases. Additionally, it is possible that the proposed ways of upscaling will not be the most accurate in other case studies. Especially, the number of clusters, for which it is not recommended anymore to scale the results using a geometrical property of the dwellings, will almost certainly not be the same as in this work. Besides that, examining other use cases concentrating on other KPIs or other combinations might be interesting to expand the scope of the developed approach.

Although this is not taken into account in this work, the user profiles seem to have a large influence in the demarcation of archetypes by various other researchers. Instead of determining the demanded temperature set points based on the ISO 13790 standards, the StROBe module — Stochastic Residential Occupancy Behaviour — can be used to simulate the residential human behaviour based on stochastic data and in this way implement divergent user profiles (Baetens et al., 2015). Then, further research is required to investigate the importance of this additional potential variable. Moreover, it must be examined how this variable can be translated to one summarising value for each dwelling in order to be able to use this variable to execute the clustering. Of course, this value must be able to be determined without requiring the knowledge of the results. When the varying user profiles are added to the approach, the constraint that every household demands the same temperature set points at the same moments will be eliminated and a more realistic simulation of the actual situation can be modelled.

Thus far, the accuracy of the developed approach is only examined on the scale of the entire district. The errors are calculated for each dwelling individually and then accumulated to discuss the accuracy of the approach in general. To ascertain that the approach is also generating acceptable errors on the scale of the individual dwellings, the errors must be examined for each dwelling separately. Additionally, it is not yet determined whether districts of another scale demand the same number of clusters to achieve the same accuracy or if this is adapting with the number of dwellings in the district. Finally, the investigated district in this work has some very particular characteristics due to the construction situation. Therefore, it still has to be ascertained that the findings in this work also do apply to other — less monotonous — districts.

Bibliography

Akaike, H., 1974. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6), pp.716–723.

Allacker, K., 2010. *Sustainable Building*. Unpublished doctoral thesis, Katholieke Universiteit Leuven.

Allegrini, J., Orehounig, K., Mavromatidis, G., Ruesch, F., Dorer, V. and Evins, R., 2015. A review of modelling approaches and tools for the simulation of district-scale energy systems. *Renewable and Sustainable Energy Reviews*, 52, pp.1391–1404.

Baetens, R., De Coninck, R., Jorissen, F., Picard, D., Helsen, L. and Saelens, D., 2015. OpenIDEAS – an Open Framework for Integrated District Energy Simulations. In: *BS2015, 14th Conference of International Building Performance Simulation Association*. Hyderabad.

Ballarini, I., Corgnati, S.P. and Corrado, V., 2014. Use of reference buildings to assess the energy saving potentials of the residential building stock: The experience of TABULA project. *Energy Policy*, 68, pp.273–284.

Banks, H.T. and Joyner, M.L., 2017. AIC under the framework of least squares estimation. *Applied Mathematics Letters*, 74, pp.33–45.

Caputo, P., Costa, G. and Ferrari, S., 2013. A supporting method for defining energy strategies in the building sector at urban scale. *Energy Policy*, 55, pp.261–270.

Cuypers, D., Vandevelde, B., Van Holm, M. and Verbeke, S., 2014. *Belgische woningtypologie – Nationale brochure over de TABULA woningtypologie*. 2nd ed.

Cyx, W., Renders, N., Van Holm, M. and Verbeke, S., 2011. *IEE TABULA – Typology Approach for Building Stock Energy Assessment*. Mol.

Dall'O', G., Galante, A. and Torri, M., 2012. A methodology for the energy performance classification of residential building stock on an urban scale. *Energy & Buildings*, 48, pp. 211–219.

Geyer, P. and Schlüter, A., 2017. Clustering and Fuzzy Reasoning as Data Mining Methods for the Development of Retrofit Strategies for Building Stocks. In: H. Song, R. Srinivasan, T. Sookoor and S. Jeschke, eds., *Smart Cities: foundations, principles, and applications*. Hoboken: John Wiley & Sons, Inc., pp.437–468.

Geyer, P., Schlüter, A. and Cisar, S., 2016. Application of clustering for the development of retrofit strategies for large building stocks. *Advanced Engineering Informatics*, 31, pp.32–47.

Ghiassi, N., 2017. *An Hourglass Approach to Urban Energy Computing*. Unpublished doctoral thesis, Vienna University of Technology.

Ghiassi, N., Hammerberg, K., Taheri, M., Pont, U., Sunanta, O. and Mahdavi, A., 2015. An enhanced sampling-based approach to urban energy modelling. In: *Proceedings of BS2015: 14th Conference of International Building Performance Simulation Association*. Hyderabad.

Ghiassi, N. and Mahdavi, A., 2016a. A GIS-based framework for semi-automated urban-scale energy simulation. In: *Central Europe towards Sustainable Building 2016*. Prague.

Ghiassi, N. and Mahdavi, A., 2016b. Utilization of GIS data for urban-scale energy inquiries: A sampling approach. In: S. Christodoulou and R. Scherer, eds., *eWork and eBusiness in Architecture, Engineering and Construction*, 2016th ed. Leiden: CRC Press/Balkema, pp. 251–258.

Ghiassi, N. and Mahdavi, A., 2017. Reductive bottom-up urban energy computing supported by multivariate cluster analysis. *Energy and Buildings*, 144, pp.372–386.

Heat Roadmap Europe, 2017. *2015 Final Heating & Cooling Demand in Belgium*. Available at: <http://www.heatroadmap.eu/resources/HRE4-Country_presentation-Belgium.pdf>.

Hutcheson, G., 2011. Ordinary Least-Squares Regression. In: Luiz Moutinho and Graeme Hutcheson, eds., *The SAGE Dictionary of Quantitative Management Research*. Singapore: SAGE Publications Asia-Pacific Pte Ltd, pp.224–228.

International Energy Agency, 2011. *Technology Roadmap – Smart Grids*. Paris. Available at: <https://webstore.iea.org/technology-roadmap-smart-grids>.

International Energy Agency, 2017. *World energy balances: an overview*. Available at: <https://webstore.iea.org/world-energy-balances-2017>.

International Organization for Standardization, 2008. Energy performance of buildings – Calculation of energy use for space heating and cooling (ISO/DIS Standard No. 13790). Available at: <http://www.cres.gr/greenbuilding/PDF/prend/set3/WI_14_TC-draft-ISO13790_2006-07-10.pdf>.

De Jaeger, I., Reynders, G. and Saelens, D., 2017. Impact of spatial accuracy on district energy simulations. In: *11th Nordic Symposium on Building Physics, NSB2017*. Trondheim.

Jorissen, F., Reynders, G., Baetens, R., Picard, D., Saelens, D. and Helsen, L., 2018. Implementation and verification of the IDEAS building energy simulation library. *Journal of Building Performance Simulation*.

Kavgic, M., Mavrogianni, A., Mumovic, D., Summerfield, A., Stevanovic, Z.M. and Djurovic-Petrovic, M.D., 2010. A review of bottom-up building stock models for energy consumption in the residential sector. *Building and Environment*, 45, pp.1683–1697.

Manning, C.D., Raghavan, P. and Schütze, H., 2009a. Flat clustering. In: *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, pp.321–345.

Manning, C.D., Raghavan, P. and Schütze, H., 2009b. Hierarchical clustering. In: *Introduction to Information Retrieval*, Online edi. Cambridge: Cambridge University Press, pp.377–401.

Molemans, J., 1998. *50 jaar Boxbergheide*. Available at: <http://users.telenet.be/paulschepers/50jBoxbergheide.pdf>.

Nouvel, R., Schulte, C., Eicker, U., Pietruschka, D. and Coors, V., 2013. CityGML-based 3D city model for energy diagnostics and urban energy policy support. In: *Proceedings of BS2015: 13th Conference of International Building Performance Simulation Association*. Chambéry, pp.218–225.

Page, J., Dervey, S. and Morand, G., 2014. Aggregating building energy demand simulation to support urban energy design. In: R. Rajan, M. Sanyogita and K. Nirmala, eds., *Proceedings of 30th International PLEA Conference, plea2014*. Ahmedabad: CEPT University.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., Brucher, M., Perrot, M. and Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, pp. 2825–2830.

Protopapadaki, C., Reynders, G. and Saelens, D., 2014. Bottom-up modelling of the Belgian residential building stock: impact of building stock descriptions. In: *9th International Conference on System Simulation in Buildings*. Liège.

Reinhart, C.F. and Cerezo Davila, C., 2016. Urban building energy modeling - A review of a nascent field. *Building and Environment*, 97, pp.196–202.

Remmen, P., Lauster, M., Mans, M., Osterhage, T. and Müller, D., 2016. CityGML import and export for dynamic building performance simulation in Modelica. In: *Proceedings of the 3rd IBPSA-England Conference BSO 2016*. Newcastle.

Rokach, L. and Maimon, O., 2010. Clustering methods. In: O. Maimon and L. Rokach, eds., *Data mining and knowledge discovery handbook*. Boston: Springer, pp.321–352.

Scipy community, 2018. *Hierarchical clustering (scipy.cluster.hierarchy) — SciPy v1.1.0 Reference Guide*. [online] Available at: <https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html#module-scipy.cluster.hierarchy> [Accessed 31 May 2018].

Shimoda, Y., Fujii, T., Morikawa, T. and Mizuno, M., 2004. Residential end-use energy simulation at city scale. *Building and Environment*, 39, pp.959–967.

Sokol, J., Cerezo Davila, C. and Reinhart, C.F., 2017. Validation of a Bayesian-based method for defining residential archetypes in urban building energy models. *Energy and Buildings*, 134, pp.11–24.

Steinley, D. and Brusco, M.J., 2008. A New Variable Weighting and Selection Procedure for K-means Cluster Analysis. *Multivariate Behavioral Research*, 43(43), pp.77–108.

Swan, L.G. and Ugursal, V.I., 2008. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and Sustainable Energy Reviews*, 13(2009), pp.1819–1835.

United Nations Framework Convention on Climate Change, 2015. *Paris Agreement*. Available at: <https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement>.

Verhelst, J., 2016. *Boxbergheide - Erfgoedobjecten - Inventaris Onroerend Erfgoed*. [online] Available at: <https://inventaris.onroerenderfgoed.be/erfgoedobjecten/302364> [Accessed 25 Dec. 2017].

Vlaamse overheid, 2017. *Geopunt Vlaanderen*. [online] Available at: <http://www.geopunt.be/> [Accessed 24 Dec. 2017].

Vlaamse overheid, 2018. *3D GRB*. [online] Available at: <https://overheid.vlaanderen.be/GRB-3DGRB> [Accessed 19 May 2018].

Werner, S., 2017. International review of district heating and cooling. *Energy*, 137, pp.617–631.

De Wolf, L., 2017. Grootschalig referentiebestand, Datastructuur. Brussel: Informatie Vlaanderen. Available at: <https://download.agiv.be/Producten/Detail?id=1&title=GRBgis>.