Graduation CME

Final Report, 21st *March 2012*

Target group clustering for applications of energy effective renovation concerning privately owned dwellings

A Case Study on Eindhoven, The Netherlands.

Committee Chairman TU/e
prof. dr. ir. B. (Bauke) de Vries
b.d.vries@tue.nl
Supervisor TU/e
dr. ir. E.G.J. (Erik) Blokhuis
e.g.j.blokhuis@tue.nl

Mentor "HetEnergiebureau BV"
J. (Jan) Bekkering
j.bekkering@hetenergiebureau.nl

Corresponding Author
Student TU Eindhoven
Pim (P.M.T.) van Loon (0575337)
+31 6 30722046
p.m.t.v.loon@student.tue.nl

Guiding Company
HetEnergiebureau BV
TheEnergyOffice

Contents

Preface

This report is composed at the end of my graduation research. This thesis is a partial fulfillment of the requirements for the degree of Master of Science in Construction management and Engineering. The research was conducted at “HetEnergieBureau” with cooperation of network operator Endinet and the municipality of Eindhoven.

The project in which this research was conducted has the working title Buurt voor Buurt / Eindhoven in which different companies took part. I have really enjoyed the versatility I experienced at the companies I have spent my days. Contact with stakeholders really kept me going searching for the best possible result.

This research attempts to focus on the energy performance of dwellings in Eindhoven by exploring clusters of target groups in a program for energy effective renovation. To reach this goal, housing submarket research and marketing aspects are integrated in a study in which a principal component analysis and cluster analysis are conducted. The results are presented as maps made with geographical information software.

It was a real challenge to get acquainted with the research methods and software used. And parallel to this being able to obtain the necessary data and convince all the parties that useful results were achieved. I really want to thank everybody for their useful input and their confidence that I would succeed.

Pim van Loon

Eindhoven, March 21th 2012

List of abbreviations

BIO	Afdeling Beleidsinformatie & Onderzoek van de Gemeente Eindhoven	Department of Policy Information and Research of the municipality of Eindhoven
BvB/e	Buurt voor Buurt / eindhoven	Neighborhood by Neighborhood Eindhoven
CA		Cluster Analysis
CBC		Centroid Based Clustering
CBS	Centraal Bureau voor de Statistiek	Central statistical bureau
CM		Cluster Method
EE	Energie Effectief/Efficiënt	Energy Effective/Efficient
EPC	Energie Prestatie Coefficient	Energy performance coefficient
EPA	Energie Prestatie Advies	Energy performance advise
EI	Energie Index	Energy index
EPDB		Energy Performance of Buildings
GBA	Gemeentelijk BasisAdministratie	Municipality register of personal information
GIS		Geographical Information Systems
KMO		Kaiser-Meyer-Olkin measure of sampling adequacy
PCA		Principal Component Analysis
PV		Photo Voltaic
RV		Response Variable
SYU	Standaard jaarverbruik	Standard Year Usage
UoA		Unit of Analysis
WOZ	Waardering Onroerende Zaken	Valuation of immovable property
WASD	Gemiddelde gewogen standaard afwijking	Weighted Average Standard Deviation

1 Introduction

In the first section of this master thesis the problem at hand is being introduced. Therefore the context in which this research takes place is sketched. By addressing the actors and factors in this research, which took place in a corporate environment (as a graduation internship), the problem statement is formulated. Out of this the research questions rise, and the expected results and relevance of this research are sketched.

1.1 Problem focus and scope

In the following paragraph it is made clear why this paper is focused on the problems in energy effective renovation (EE-renovation) of the existing dwelling stock. As mentioned before this research took place in a corporate environment and several actors had to be activated and become a stakeholder to be able to evaluate target groups clusters for applications of EE-renovation.

1.1.1 Context

Energy Scenario

All the things we do in and for life on earth consumes energy. Currently fossil fuels are used for the majority of our energy production. Looking at future energy scenarios it becomes clear that fossil fuels need to be replaced by other (renewable) energy sources. There are three strong arguments which support an energy transition:

- The use of fossil fuels produces greenhouse gasses like for example carbon dioxide. It is highly probable (98% certainty) that this causes the reinforced greenhouse effect (climate change);
- Fossil fuels are not unlimited resources; fossil fuels in their cheap form are becoming scarce. Other resources are becoming more viable options during the 21st century;
- Economic/Political dependency; the world relies on unstable regimes for the availability of fossil fuels.

Trias Energetica is a simple and logical concept that stimulates to achieve energy savings, reduce our dependence on fossil fuels, and save the environment.

The three consecutive steps of the Trias Energetica are:

- Trying to reduce the demand for energy by implementing energy saving measures;
- Using renewable energy sources like water, sun and wind for the remaining part;
- Producing- and using fossil energy as effectively and efficiently as possible.

Governments in Europe decided there should be common objectives for all countries in the European Union. This leaded to European energy and sustainability objectives for the Netherlands in 2020 which are listed in *Table 1.1*.

	National	European Union
Reduction of greenhouse gasses compared to 1990	20%	20% - 30%
Share of renewables	14%	16% - 17%
Energy consumption reduction	2% annually	20%

Table 1.1 Sustainability objectives National and for the European Union (Ministry of Infrastructure and Environment, 2011)

To meet these objectives a research done by Wesselink et al. (2008) states that five times as much carbon dioxide needs to be reduced in comparison with the 1990-2005 period in the upcoming 10 years.

In this proposal the term energy effective instead of efficient is used (McDonough & Braungart, 2002). They state that we try to make everything in this world efficient. Even "wrong" technologies, e.g. combustion technologies for fossil fuels, can be efficient. By using energy effective this paper states that it is preferable to use energy as effective as possible and not stimulate efficient use of polluting techniques as a favorable option.

Legislation for Energy use of dwellings

The energy use of new build dwellings is already subject to legislation. Architects are forced by the Dutch Building Regulations to design dwellings with a prescribed EPC (Energy Performance Coefficient). The prescribed coefficient will be set lower in the upcoming years. In 2011, the EPC is tightened to 0.6, in 2015 to 0.4 and in 2020 even to 0. By tightening the EPC, the government aims to reduce energy use in new dwellings.

At the moment there are only minor, subordinate standards in the Dutch Building decree for existing dwellings, these standards only focus on safety and risk. Minimum energy standards only apply to new build dwellings. In contrast much lower standards, with a much lower impact on energy savings, are required for extensive renovation of huge buildings. For the renovation of dwellings, with or without a permit, energy standards do not apply.

With the introduction of the EPDB (Energy Performance of Buildings Directive) in 2002 a European standard took effect. This set of rules tries to stimulate energy saving in the build environment. In article 7 the use of energy labels is introduced. An energy label shows the energetic quality of a dwelling compared to a similar standard dwelling. The label was introduced in 2008 and revised in 2010. Dwellings are categorized for A++ to G, respectively a good towards bad performance on building related energy use. A label is mandatory when a dwelling is sold. All dwellings in the social housing sector should have a label at the end of 2012.

In former years some subsidy schemes for energy effective renovation were applied. Amounts ranging from 300 up to 1500 Euros were available for private homeowners who for example installed decentralized power supplies or gained 2 label steps (from e.g. E to C) with an energy efficient measure for their own dwelling.

Energy use of existing dwelling stock

Considering the first step of the trias, it is wise to look at the energy use of the existing stock of dwellings. At the moment there are 7,219,230 dwellings in the Netherlands (CBS, 2011).

Less than 20,000 of these existing dwellings are withdrawn or demolished per year. Together with the new build dwellings of about 55,000 dwellings per year the stock of dwellings will amount to 8,500,000 dwellings in 2045 (van Duijn & Stoeldraijer, 2011). At the current rate of replacement; leaving the possibility of energy effective renovation out of the equation; our stock of dwellings will be sustainable in 350 years. An answer lies in energy effective renovation of existing dwellings. Together with the implementation of the second step in the trias, where for the remaining part renewables should be used, decentralized generation of energy with PV panels can be an integral part of a renovation.

Looking at the possibility for energy effective renovation of dwellings, energy saving measures need to have certain characteristics concerning practicability. Agentschap NL (former Senternovem) tried to categorize the existing dwellings into different types and construction periods. Of these 30 categories some mean values are known, including original and average current energylabel. Together with this research 2 effective saving packages were designed, these packages contain e.g. thermal insulation of the façade or complete building envelope, HP++ glass, replacement of the heating installation with an HP 107 Combination boiler, and a HRU (heat recovery unit) if mechanical ventilation is already in place. In addition: for decentralized generation of energy PV panels and a solar boiler for domestic hot water seem to be the best options.

1.1.2 Involved actors and factors

Involved actors

Municipality of Eindhoven

The municipality of Eindhoven aims to be energy neutral between 2035 and 2045 (Municipality of Eindhoven, 2008). Energy neutral in this case means that the (remaining) energy demand for the own organization, dwellings, industry and remaining connections is generated with renewables inside the borders of Eindhoven. The existing stock of dwellings consumes about 42 percent of the total energy used in the summation above. Without any supporting arguments it is stated that a reduction of 50% in the in 2040 existing city is a feasible goal.

HetEnergiebureau

The Energy Office tries to start up a consortium which will be implementing a strategy to seduce private homeowners to implement saving measures. This group of companies and supporting authorities tries to obtain a subsidy of 500,000 Euros from NL Agency (Agency of the Ministry of Economic affairs, agriculture and innovation) for the energy effective renovation of at least 2,000 dwellings in the target area. The preliminary title of the program is Buurt voor Buurt / Eindhoven (Neighborhood by Neighborhood,) with the abbreviation BvB/e.

Endinet

Endinet, as a full daughter company of Alliander, is the network operator in Eindhoven. With their goal of fast implementation of smart metering and getting acquainted with the problems involved with the implementation, pilot projects are supported. That is why Endinet is involved in BvB/e. Endinet has access to the usage figures of all connections in Eindhoven for electricity and gas, those figures are interesting for integration in the research.

Private homeowners

Private homeowners need to be convinced that saving energy in their own home is possible, increasing their comfort and fun. Furthermore one should be a trustworthy source by making clear that the risks are low and it is financially feasible over a period of about a decade. People hate it to be confronted with their behavior related to energy-use because of two reasons: 1). they want to have their privacy respected and 2). They want to make their own choices and hate to get a sermon. So what is the best way to approach them?

Involved factors

The involved factors are related to the private homeowner, the so called software. He or she is the decision maker in this problem. Where the other 3 indicated actors are trying to convince the private homeowner of a problem he or she may not even be aware of. At first we are zooming in on the actor. He or she lives in a dwelling which is from now on called the hardware. Different factors can be discussed which are a part of the hardware. Afterwards the software related factors are addressed.

Hardware-related (Dwelling)

- Year of construction

The year the house is constructed says much about the energy performance of a dwelling. Before 1992 there was no legislation on the minimal insulation of newly build dwellings. Therefore the original energy performance of dwellings older than 20 years is quite terrible. That is why the year of construction of a dwelling is a real contributor to the energy performance. However there is always a possibility energy saving measures are already implemented.

- Energy label

The energylabel of a dwelling seems to be a good option to use as a measure for energy performance of a dwelling. It is determined in an Energy Performance Advice (EPA, “Energie Prestatie Advies” in Dutch), the label (A++ up to G) represents a range of values in the Energy Index (EI) which is the ratio of the dwelling related energy use to the standard dwelling related energy use of that type and size of dwelling. This topic is widely addressed in the graduation report of Marczinski (2011, pp.10-13). However the energy label of a dwelling is often not determined (yet), especially in the market for private owned dwellings less than 5 percent of the dwellings is labeled. This is the reason the Energy label of a dwelling is not suitable as a factor in this research.

- WOZ-value (Valuation of Immovable Property)

The municipality rates every object, in this case a dwelling is a WOZ-object when it has one owner and one user/household. So a block apartment contains several WOZ-objects because there are besides the ownership at least different households in it.

The WOZ-value of an object is determined by looking at the sales prices of corresponding houses in the spatial proximity of the object itself. The WOZ-value is assessed every year and of course it is correlated with the WOZ-value of last year. The WOZ-value reflects the real value of a dwelling quite well. House price is an important variable in housing

submarket research and therefore this factor could be an important variable in this research.

- Energy consumption

Almost every dwelling in Eindhoven is connected to the central electricity and gas grid. From the exact gas and electricity usage of a dwelling a huge amount of information can be extracted. Demographic, lifestyle of the inhabitants and the physical state of the dwelling all influence the energy use of a dwelling.

- Typology

Typologies are used to categorize objects, this is done for dwellings too. NL agency published their "Example dwellings 2011 existing stock" ("Voorbeeldwoningen 2011 bestaande bouw") (PRC Bouwcentrum & W/E adviseurs, 2011) . Based on an stocktaking of dwelling types (WoON 2006 by PRC Bouwcentrum & W/E Adviseurs (2006))and existing knowledge from the example dwelling study 2007 seven dwelling types are deduced, built in different periods of time they add up to 36 different groups of dwellings having certain characteristics in common.

The municipality of Eindhoven uses a completely different method for classification purposes. Over 120 different types of dwellings are known in their WOZ-database. The department of "Policy Information and Research" (BIO: afdeling "BeleidsInformatie & Onderzoek") of the municipality of Eindhoven introduced a statistical cluster division of Eindhoven. In a dataset of BIO a simpler dwelling categorization is used. The dataset merges data from the WOZ database with information of the register of personal information (GBA, "Basisadministratie Gemeente" in Dutch). More on this topic is included in *Appendix C Data preparation.*

How can the typology of a dwelling be integrated in this study?

Software-related (Private Homeowners)

- Income

The total income of a private homeowner partly determines how wealthy the household is. Moreover the financial situation is of high influence in decision for participation; the income of a household is not available for privacy reasons. Therefore the income of a household is not likely to be used as a factor in this research.

- Age

The age of the decision maker in a household is of influence in the way aspects such as environmental awareness and financial focus are present. In the graduation report of Nieuwenhuijsen (2010) this aspect is concluded to be of high importance for the design of target group strategies.

- Household composition

The size of a household, i.e. the amount of people who live together in a certain dwelling, and the number of children are an indicator for the household composition. The phase of

life in which, for example, a family is, determines the way they react to a proposal of a trustworthy external party to participate in a program for EE-renovation.

Other aspects such as culture, lifestyle and ethnicity also characterize the households and the decision makers. The problem is that these aspects are intangible. Considering ethnicity it is morally disputable to involve it as a variable in the study.

Decision-related (Participation)

The following list of intangible factors is found in literature (Motivaction (2011) and Punj & Steward (1983)) to be of influence when decisions for EE-renovation are made:

- Trust in messenger;
- Communication of messenger; Content of message;
- Financial aspects;
- Related desires;
- Group dynamics.

The decision related factors cannot be considered as variables/attributes in the research. They are intangible and therefore they should be seen as attributes of how target groups can differ from each other. These factors characterize target groups more than they could be used to determine them.

1.2 Problem statement

Over the past few years developments in the field of energy effective renovation of existing dwellings finally got started. Adjacent countries like Belgium and Germany have a lot more experience with applications of e.g. the "passivhouse" (with almost no building-related energy use) and Photo Voltaic (PV) electricity production on rooftops. In case of PV it is presumed that this is partly due to extensive subsidies on an immature technology provided by the other countries. It was an innovation in a niche, but nowadays domestic PV cell applications have a financial payback time of less than 15 years and a carbon payback time of less than 6 years (depending on the site of production). Still only few domestic houses in the Netherlands are equipped with a PV system, is it time to make a change?

Campaigns of the government to reduce energy use in existing dwellings were not often successful in the past. The participation rate was in most cases not exceeding 5%, of which 3% was not attracted by the campaign but already was intrinsically motivated to compete in a program. Idea owners of such programs at municipal level are in a real need for ways to increase this participation rate. The government just finished a report with best practices for building related energy savings for private owners (Motivaction, 2011), this report has been made as part of the "more with less" program (meer met minder). In the report Do's and Don'ts are formulated. One of the Do's only raises more questions:

Choose the target group and their homes with care. There must be a potential saving in the houses and it is important to focus on a target group. A group is characterized by shared values, needs and ages. All residents of a neighborhood are rarely a target. Make sure your approach fits the target audience. Are they sensitive to comfort, money savings or unburdening? Adjust your approach to it.

Easier said than done, but how do you do such a thing if you have more than 50,000 potential dwellings in, for example, Eindhoven? Some of the questions that rose are listed below; these questions will transform into the research questions:

- How can we make the participation rate of private homeowners grow?
- Do we ask the wrong people (software), those who aren't interested in energy saving measures at all?
- Do we try to improve the wrong houses (hardware)?

How can we select the dwellings with the biggest saving potential bearing the characteristics of the household in mind? Is an evaluation method available which can integrate characteristics of hardware and software?

In programs for energy effective renovation of dwellings it is hard and still not clear how to select
the right target group
regarding the dwellings saving potential (hardware) and decision making private homeowner (software).

In the next section research questions are formulated that will lead to the design of a study and the expected results are introduced.

1.3 Research questions

1.3.1 Research sub questions

- Which variables or factors of all dwellings and households in Eindhoven are available for analysis?
- Which variables or factors influence the decision for participation of private homeowners?
- Is the statistical cluster division of Eindhoven a reliable target group division?

1.3.2 Research question

Is an optimization of target group size and geographical distribution possible, when focusing on maximization of (1) participation and (2) reduced energy demand in a program for energy effective renovation for private owners of dwellings?

1.4 Relevance of research

In a graduation research including an internship both practical and theoretical relevance are important. In the transition towards a more sustainable Eindhoven, i.e. total energy neutrality, an important factor is upgrading the energy performance of the existing stock. Programs to achieve this are rarely successful, since too little is done to target the right dwellings and owners. The practical relevance of this research lies in its usability for BvB/e. The foundation tries to recruit up to 2,000 participants in a program for EE-renovation of privately owned dwellings. To reach this goal 20,000 potential participants and their dwellings are selected. The designed method and output will be used to select the 20,000 dwellings used by BvB/e. More on BvB/e is enclosed in *appendix A Management Summary Project plan BvB/e*.

The method itself further explores the field of housing submarket and target group research in marketing, which is introduced in *Chapter 2 Research Design.* Few case studies on housing submarkets in Europe exist. Also worldwide no research has been found on this topic which suggests that factors representing the energy performance of dwellings are not used to determine submarkets yet. This research explores this aspect for the first time. Moreover it is the first time housing submarket research using cluster analysis is used at the TU/e.

1.5 Expected Results

- A set of decision variables that tries to reflect the actual behavior and characteristics of private owners and their dwellings considering to participate in an energy effective renovation program;
- A cluster analysis (clustering). Resulting in a certain number of clusters (geographical constrained typological target groups) in Eindhoven;
- A dataset of all the dwellings and their owners to use in the communication strategy for BvB/e;
- A map of Eindhoven visualizing the target group clusters for energy saving.

1.6 Reading Guide

As a part of this chapter a reading guide is introduced. In *Chapter 1 Introduction* the problem and its context are sketched. Based on this, research questions are formulated and prejudgment of some possible results is done. Based on a literature study into the housing submarket and target group research, a model of the research design is presented in *Chapter 2 Research Design*. A more extensive elaboration on research, software and the used research methods is given in *Chapter 3 Theoretical Orientation*. Together with *Chapter 0*, *Chapter 4 Case Study on Eindhoven* can be seen as the backbone of this report. The case study is for the most part reported as described in literature on cluster analysis studies. Chapter 5 *Conclusion*, Chapter 6 *Discussion*, and *Chapter 7 Acknowledgements* present standard contents, a reflection and some words of thanks. Some of the appendices are seen as classified and therefore maybe not present in this version of the report.

2 Research Design

To be able to answer the research questions a study is designed. Two established concepts used in research are evaluated, in *paragraph 2.1* different approaches in Housing Submarket analysis are discussed and in *paragraph 2.2* the marketing aspect of the problem statement is discussed. Together these are merged into one research model which is shown in *paragraph 2.3*.

2.1 Housing Submarkets

Neighborhoods are a historically grown, physical presentation of groups of buildings. In which neighborhood a property is located is influenced by administrative decisions of planners and therefore historically determined. In the 1960's research on housing markets started, it was based on a belief that the prices of property are not only defined by its physical location but also structural, demographic and socio-economic characteristics have influence. The most often used definition is given by Bourassa et al. (1999) were a submarket is defined as a set of dwellings that are reasonably close substitutes of one another, but relatively poor substitutes for dwellings in other submarkets.

In a paper of Bates (2006) it is stated that planner-defined geographic areas of analysis (neighborhoods for example) diverge substantially from housing submarkets. This suggests that predefined neighborhoods, while valid for some purposes, do not represent areas for predicting the housing market response to policy. Transferring this towards the challenge for target group clustering in Eindhoven this research indicates we should look further then the "a priori" division of the municipality of Eindhoven for neighborhoods and small clusters.

It is disputable that with the use of housing submarkets, factors related to energy usage are left out of the response variables. Although housing submarket theory is most often used in housing economics it is going to be implemented in this research on target group clustering of dwellings and their private owners. In this research it could be wise to use a statistical submarket housing model, where technical, structural and energetic characteristics of a dwelling are taken in account too. In the following page more is explained on the topic of using cluster analysis for defining housing submarkets. The possibilities of cluster analysis are pointed out with some highlights from two scientific articles.

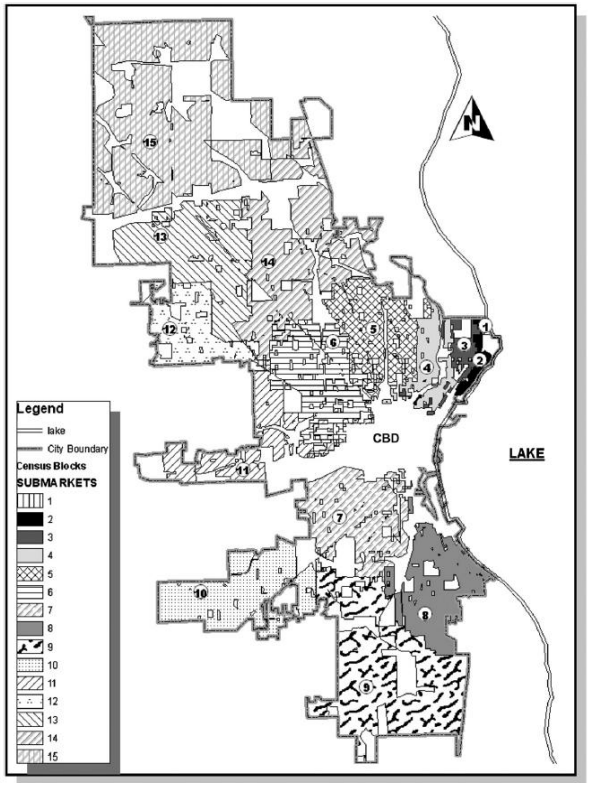


Figure 2.1 The City of Milwaukee divided into 15 housing submarket clusters by Wu & Rashi Sharma (2011)

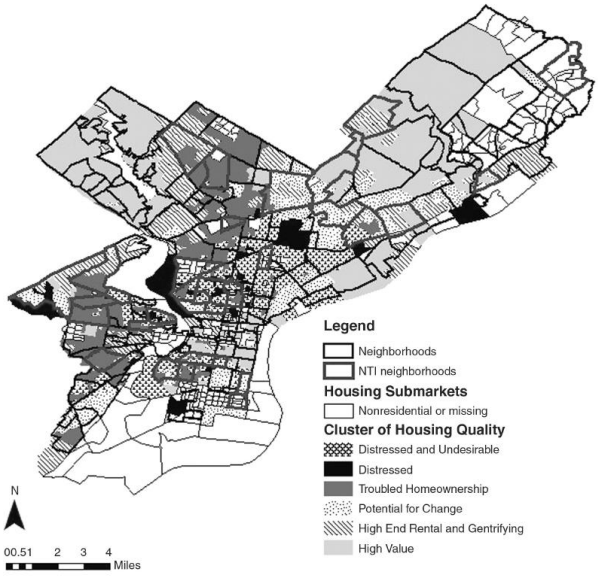


Figure 2.2 The City of Philadelphia divided into 6 socially defined clusters by Bates (2006)

Wu & Rashi Sharma (2011) deals with the topic of housing submarket classification and the role of spatial contiguity, sometimes called nearness or proximity. A spatially constrained data-driven classification methodology is used to deduce spatially integrated housing market segments. In principle two different classification methods can be used. The First method is an “a priori” classification. This term is used for historically grown spatial divisions e.g. neighborhoods which were built simultaneously. By all means no further research is conducted and for example an existing spatial division by a municipality is taken for granted. Secondly some form of further research can be done and a data driven technology is used to determine housing submarkets. These classifications use statistical data analysis for structural, location and demographic variables simultaneously.

Wu & Rashi Sharma divide the single family houses in the city of Milwaukee into 15 submarket clusters with a principal component analysis (PCA) followed by a cluster analysis (CA). Milwaukee has 578,887 inhabitants, for the analysis 86,000 single family houses were used.

The variables that were incorporated can be divided into 4 categories, 1) Value/Cost, 2) Structural attributes, 3) Demography and 4) Location. The structural attributes focus on the number of several types of rooms in a dwelling. Dwelling type, materials used and factors such as whether the building envelope is insulated are not taken into account. In the PCA the component explaining most variance in the data set is dwelling value and demographics, structural attributes and location are of minor influence on the variance.

The clusters found with data-driven classification are compared with the a priori divisions that exist for Milwaukee in two ways, 1) Substitutability, using a hedonic price model and 2)

Similarity using the weighted average standard deviation to check intra-cluster homogeneity. At last the clusters are visually checked on spatial integrity.

This research is extremely useful because it links geographical constrained data into a model where different variables are statistically considered in their coherence. Submarkets are formed with houses more similar to each other based on location and their physical, typological and demographic properties.

It is concluded that the method advocates the utility of spatial submarkets where public and private organizations can identify specific geographic zones of potential growth or with special needs. Is it possible to identify these regions and use them as target groups for energetic effective renovation programs?

Besides this red-hot article of Wu an older article of Bates (2006) is discussed in short in this Housing submarket section. In the paper of bates Ward's method for minimizing the Euclidean distance inside a cluster is used. Philadelphia is divided into six clusters as shown in de map in *Figure 2.2*. It should be noticed that with this cluster analysis the clusters are not limited to be contiguous in space. This could be useful in a research because you are looking for some target groups (clusters) distributed around and throughout the city of Eindhoven. On the other hand you can circumnavigate this by evaluating the city district by district. It is decided to get acquainted with the method by only evaluating one district, we should find spatial contiguous clusters in that district. Because of this it is chosen to use the k-means CA which produces spatial contiguous clusters. By conducting an analysis on all districts in a city it is likely you will find the same sort of clusters throughout a city.

Bourassa was already mentioned in the first paragraph of this section because he is a great contributor in this field of research. The basic literature on this topic is written by him, e.g. Bourassa et al. (2003), Bourassa et al. (2007) and Clapp & Wang (2006). The articles give some more basic information on how to deal with housing submarkets and cluster analysis to achieve the output. The articles are not discussed in this report because they focus on validation of financial aspects too much.

2.2 Marketing: Target groups

In a program for energy effective renovation there is a need for a division of target groups. People are attracted to different aspects of the results of a renovation program and therefore have different grounds for participation.

Cluster analysis is used in market research for selection of possible or preferred consumers, the so called market segmentation. With market segmentation the market is divided into target groups or sub-markets. An older overview for possible application is given by Punj & Steward (1983) and an implementation is conducted by Kuo et al. (2002).

For "Buurt voor Buurt / eindhoven" the plans are that 20.000 owners will be approached for a possible participation in the program. In this research we are aiming for a division of the market into target groups. Beforehand it is hard to estimate what the optimal group size and geographical distribution for those target groups are. Different questions rise when thinking of a distribution into target groups:

- Can we simply rely on the geographical distribution of houses?
- Or should we consider dividing groups into dwelling types?

- Maybe we should even consider dividing groups into different owner types?
- Or should we consider dividing the owners into groups of people with the same shared values?

For the clustering of target groups (especially when large amounts of data are involved) cluster analysis is used. Is it possible to integrate both, let us say combining housing submarkets and target groups into one statistical model?

In a book on research methods of Burns & Burns (2008) simple examples and arguments are given to illustrate the use of cluster analysis for market segmentation. In this extra book chapter it is stated that it is possible to use a variety of interdependent variables to cluster consumer segments often sought for successful marketing strategies. This is exactly what we are looking for in this case.

2.3 Research model

The model shown in Figure 2.3 illustrates the integration of the different fields of research in a cluster analysis. Two studies for a cluster analysis are designed. At the lower right corner the existing statistical cluster division is evaluated. At the lower left corner the target group study is visualized. Of both paths the output is validated and an interpretation is formulated.

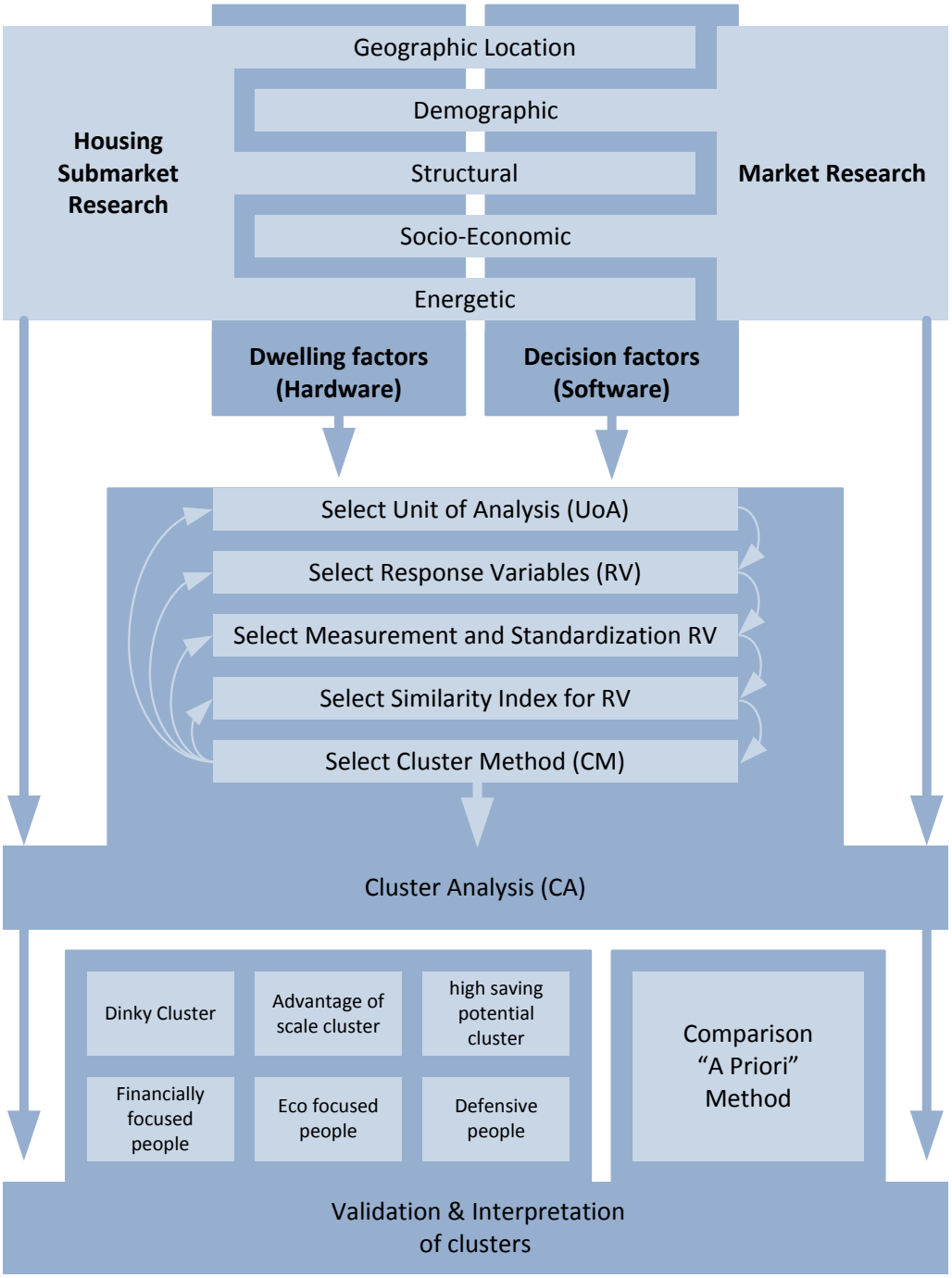


Figure 2.3 Research model

A further study on cluster analysis, PCA and GIS is conducted and reported in *Chapter 3 Theoretical Orientation.* Afterwards a case study on Eindhoven is conducted and reported on in *Chapter* 4.

3 Theoretical Orientation

3.1 Geographical analysis

In this case geographical analysis using Geographical Information Systems (GIS) will be used for the connection between geographical information in the database and a visualization of clusters on a map. This way the spatial proximity can be checked and spatially contiguous clusters are assured. At first the geographical information, which is still in postal code or street name format, needs to be converted into geographical coordinates. This process is called geocoding, and can be executed in both directions.

3.1.1 Geocoding

The conversion of an address into geographical coordinates can be done using different algorithms, with coordinates of different geographical coordinate systems as an output. The most used system for geographical referencing worldwide is WGS84. It has a very small deviation of the official system ITRS (International Terrestrial Reference System), but is much more accessible, and therefore used by e.g. most GPS systems. The coordinates in WGS84 are based on a metric three dimensional Cartesian coordinate system. With a reference ellipsoid these coordinates can be translated to a longitude and a latitude, often referred to as the X and Y coordinate of an object respectively.

In the Netherlands an alternative system is common used. The system referred to as the "Rijksdriehoeksmetingen" (RD coordinates) covers approximately 5600 carefully determined points in the Netherlands (Kadaster, 2003). The coordinates of these points are measured in centimeters. These points are used to locate all registered properties in the Netherlands, e.g. dwellings. This way they have their own RD-coordinates listed in the public register, which is accessible for everybody paying a fee.

In GIS it is often preferable to use a geocoded table with specific information and a map of the area based on the same geocoding system. Using online servers for geocoding is quite expensive; therefore it is advisable to obtain geocoded objects in a table. The municipality of Eindhoven has access to the all information in the public register of properties, this means they are able to geocode our data.

3.1.2 Displaying objects in clusters

With all commercial geographical information software, e.g. MapInfo, it is possible to point out our clusters on a map. All kind of maps are supported by Mapinfo. They can be topographical or satellite based, even a line drawing of areas can be used. Objects on a map can be highlighted in a color, cluster by cluster. The result will be a map of Eindhoven with the different clusters visualized.

For this research it is wise to use the RD coordinate system because it is widely used by governmental agencies and all properties are geocoded in the public register of properties. Usage of GIS-maps based on RD-coordinates is common in the Netherlands and MapInfo supports this coordinate system too.

3.2 Cluster analysis

Cluster analysis, sometimes referred to as classification, numerical taxonomy or typological analysis finds its applications, as mentioned in *Chapter 0* in both 1) market research in which the market is segmented into different groups of potential consumers and 2) housing

submarkets where different clusters of dwellings in a city are determined using a statistical algorithm. Let us first take a look at cluster analysis in general.

For evaluation of Cluster Analysis (CA) as a research method several books on statistical research methods are used to get an impression. In these books usually a complete chapter is dedicated to the different methods of cluster analysis, most of the chapters elaborate with examples to illustrate the methods used. The book sections of Burns & Burns (2008), Tryfos (1998), Huberty et al. (2005) and Norušis (2011) are used to get general insight to CA and statistics.

Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis.

Two different types of clustering can be distinguished, hard and soft clustering. Hard clustering divides every object into respectively a single hard cluster or more clusters for a certain degree. Soft clustering is also called fuzzy clustering. An example of hard clustering is partitioning clustering. In research for housing submarkets “k-means clustering for non hierarchical clustering” or “Ward’s method for hierarchical clustering” is mostly used. Both are methods that lead towards a division of hard clustered cases. Every object of the dataset belongs to a single cluster.

To determine the similarity of cases/objects measures of distance should be introduced. The next pages give a short introduction for a basic understanding of how cluster analysis work.

3.2.1 Measures of distance

In the additional chapter of the book written by Tryfos (1998), the basics for measures of distance for attributes and variables for the objects in a dataset are addressed. Measures of distance are needed to determine the similarity or closeness of objects and clusters.

Euclidean Distance

The Euclidean distance is the shortest distance between two points.

$$C(P0, P1) = \sqrt{\mathrm{A}^2 + \mathrm{B}^2} = \sqrt{(\mathrm{X_{P0}} - \mathrm{X_{P1}})^2 + (\mathrm{Y_{P0}} - \mathrm{Y_{P1}})^2} \qquad 3.1$$

In *equation 3.1* is X the horizontal axis and Y is the vertical axis. For everyone this is known as the most common used form of the Pythagoras theorem. If you describe this in Cartesian coordinates in Euclidean n-space *equation 3.2* can be used. Where *p* and *q* are the points (like *P*0 and *P*1) and *D* is the distance between them.

$$D(p, q) = D(q, p) = \sqrt{(\mathrm{q_1} - \mathrm{p_1})^2 + (\mathrm{q_2} - \mathrm{p_2})^2 + \cdots + (\mathrm{q_n} - \mathrm{p_n})^2} = \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{q_i} - \mathrm{p_i})^2} \qquad 3.2$$

This way the similarity of two objects with multi-variables can be calculated. If the distance is only calculated for comparative purposes we could suffice with the calculation of the squared Euclidean distance using *equation 3.3.*

$$D(p, q) = (\mathrm{q_1} - \mathrm{p_1})^2 + (\mathrm{q_2} - \mathrm{p_2})^2 + \cdots + (\mathrm{q_n} - \mathrm{p_n})^2 \qquad 3.3$$

In all cluster methods some kind of distance measures are used. The Euclidean distance is the most common technique used.

3.2.2 Connectivity based clustering

An example of connectivity based clustering is hierarchical clustering. Hierarchical clustering is the most common method for cluster analysis and ideal for explaining the basics of statistical clustering of objects. One can approach hierarchical clustering top-down or bottom up. This means 1) Divisive clustering for a top-down approach and 2) agglomerative clustering for a bottom-up approach. With divisive clustering you start with a single cluster containing all objects ending up with every object being a cluster for itself. With agglomerative clustering you start and end the other way around. The assessment of the right amount of clusters can be done looking at a dendogram. A dendogram is another word for a distance tree. How to construct a dendogram is most easily understood with the following example. A dendogram is shown in *Figure 3.4* at the end of this section.

Person	X_1	X_2
a	2	4
b	8	2
c	9	3
d	1	5
e	8.5	1

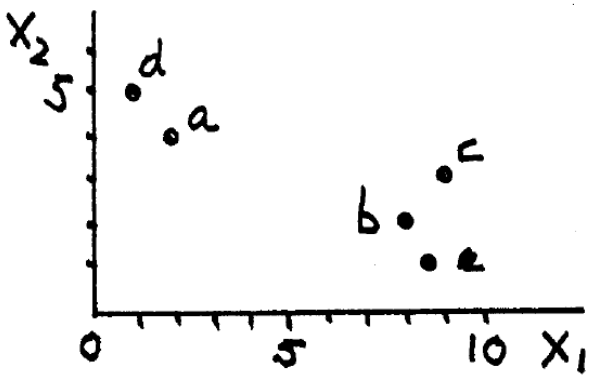


Figure 3.1 Illustrative data and grouping of objects plot by Tryfos (1998)

In *Figure 3.1* five persons are compared on two variables with the values mentioned in the table at the left. At the right the data is plotted. It is possible to calculate the Euclidean distance of all points, for example the distance between *a* and *b*:

$$D(a,b) = \sqrt{(2-8)^2 + (4-2)^2} = \sqrt{36+4} = 6.325$$

Resulting in the following cluster matrix displayed in *Figure 3.2.*

Cluster	a	b	c	d	e
a	0	6.325	7.071	1.414	7.159
b		0	1.414	7.616	1.118
c			0	8.246	2.062
d				0	8.500
e					0

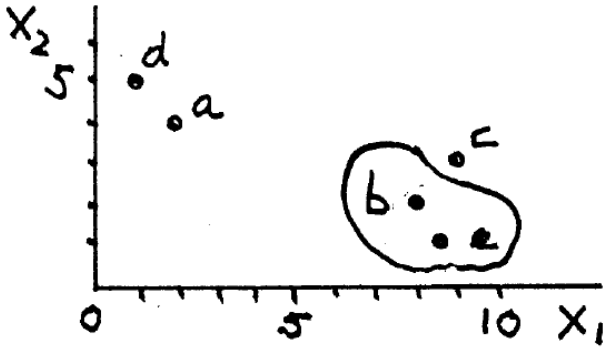


Figure 3.2 Illustrative data and clustering of objects plot by Tryfos (1998) first cluster

In this agglomerative cluster method the nearest neighbor method is used. Therefore the first cluster will be formed by the object with the smallest Euclidean distance, as shown in the plot above on the right, this is in this case object *b* and *e*. Using this nearest neighbor method from this step on *e* does not influence the further process because *b* is closer to all

other objects then e is. Using another cluster method the different steps could lead to other outcomes on the same dataset.

Following this method the last step will look as follows. The nearest neighbor in both clusters are still *a* and *b*. These objects determine the Euclidean distance between cluster *bce* and *ad*.

Cluster	(bce)	(ad)
(bce)	0	6.325
(ad)		0

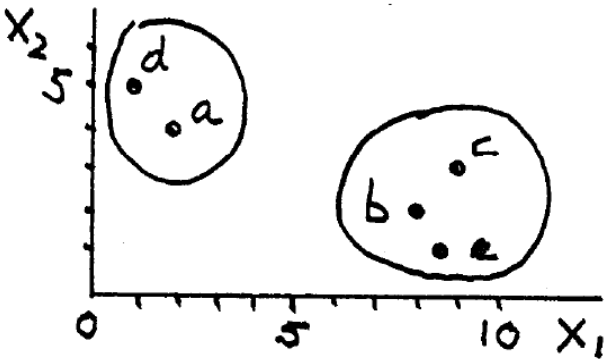


Figure 3.3 Illustrative data and clustering of objects plot by Tryfos (1998) last two clusters

Other cluster methods used for hierarchical clustering are the furthest neighbor method or the average linkage method. This agglomerative way of clustering assures that all objects are merged into one embracing cluster. However in cluster analysis you want to know when you have ended up with the right amount of clusters. On the one hand this is dependent on your research goal, on the other hand you can use the Euclidean distance between clusters to determine the stopping point of your agglomerative or divisive cluster analysis. To assess this, a dendogram is often helpful. A dendogram is a tree in which the distances are visualized. In the next section a more complex and comprehensive method for clustering is elaborated. In this example it advised to obtain two clusters.

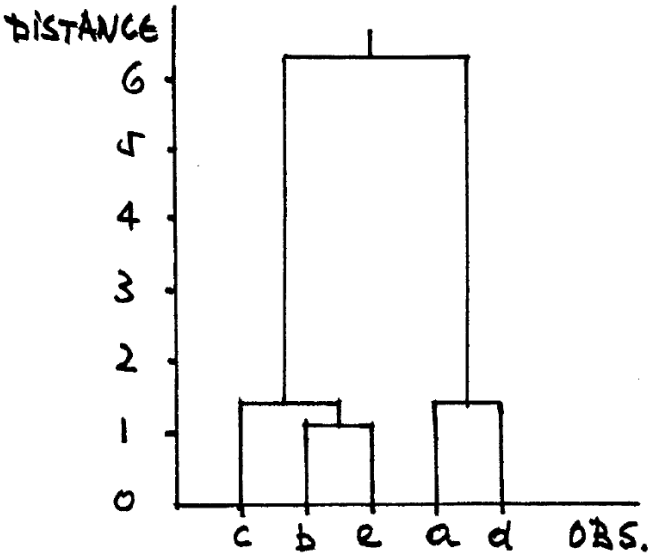


Figure 3.4 Dendogram of objects displaying their Euclidian distance by Tryfos (1998)

3.2.3 Centroid based clustering

As distance measure for *Centroid Based Clustering (CBC)* the Euclidean distance is most often used as well. Centroid based clustering does not use executive steps to come up with a division. CBC is non-hierarchical and based on an iterative process. The most used method for CBC is the *k-means method*, where regularly the k, the number of clusters, needs to be chosen upfront. There are different ways to determine the correct number of clusters for that specific research.

With the *k-means method* the initial centers can be randomly chosen, in every step the (Euclidean) distance towards all the cluster centers is measured. If an object is closer to another cluster center in comparison with the center of the cluster it is currently assigned to, the object will be reassignment. Before every step the mean, also called the central vector, of a cluster is calculated. Once again the process of reassigning begins. If every object is assigned according to this rule the clustering ends. This only works on small datasets with few variables. In many cases the number of iterative steps needs to be specified upfront. Otherwise the software will end in an infinite loop. Most of the time this problem is referred to as that CBC would be extremely vulnerable for outliers (Norušis, 2011). Outliers will be selected as initial cluster centers, resulting in clusters with only a few members. Therefore outliers should be considered to be left out of the analysis. How outliers are dealt with is described in *Chapter 0.*

As stated in the previous paragraph the Euclidean distance can be used as a distance measure for Centroid Based Clustering. However different from hierarchical clustering in Centroid Based Clustering we focus on the distance of an object towards the centre of the cluster it is initially assigned to. *Figure 3.5* illustrates what is mentioned above. The cluster centroids are simply calculated as the mean of the objects the cluster contains at that moment. The distance to the center is calculated using the Euclidean distance. This can be done using datasets considering thousands of objects or cases regarding a dozen of variables into dozens of clusters.

Cluster 1			Cluster 2		
Obs.	X_1	X_2	Obs.	X_1	X_2
a	2	4	c	9	3
d	1	5	e	8.5	1
			b	8	2
Ave.	1.5	4.5	Ave.	8.5	2

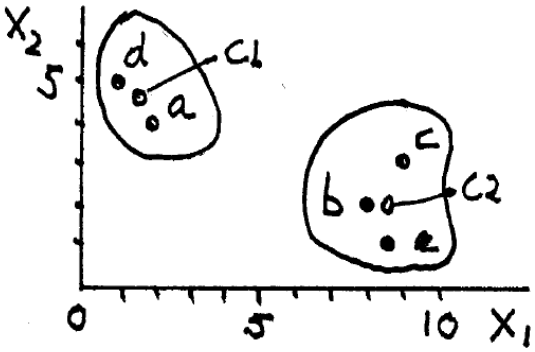


Figure 3.5 Illustrative data and clustering of objects plot by Tryfos (1998) using k-means method for CBC

If you want to divide your data set into five clusters, the borders will look like the well known Voronoi diagram in *Figure 3.6* at the left. The lines are the cluster borders and this way the division is made. Simple Voronoi diagrams can only be drawn when two variables of an object are considered. In *Figure 3.6* on the right a display in 3D is given. It already looks complicated. But remember only two clusters are displayed which were clustered on three variables/attributes. Cluster analysis is often done with 10 up to 20 variables. Displaying the

clusters and their objects in multi-variable and therefore multi-dimensional space is impossible.

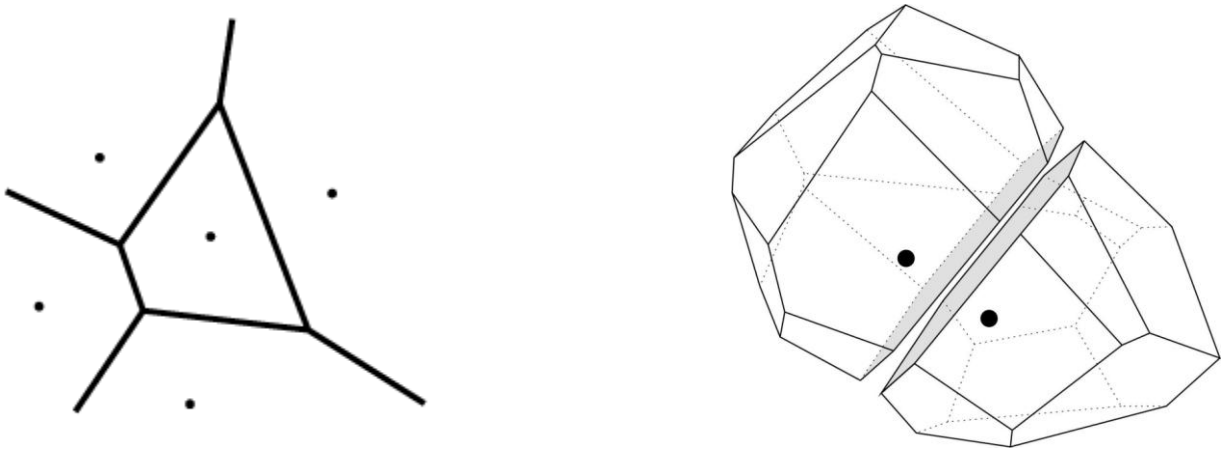


Figure 3.6 Visualizations of clusters considering them to be like Voronoi diagrams, illustrated by Brahmantia Iskandar Muda

3.3 Principal component analysis

Principal Component Analysis (PCA) is a quite complex statistical procedure to explore and in the end reduce the amount of variables in a dataset. PCA is one of many methods of factor analysis available where linear components are extracted out of the variables. The term factor and component are used interchangeably. PCA is widely used for scientific research on housing submarkets ((Bates, 2006) and (Wu & Rashi Sharma, 2011)), it reduces the complexity of the dataset used in the cluster analysis by indicating the contribution of variables to a component accounting for the most variance in the dataset. Therefore statistical terms used to express correlation (e.g. variance, the cross product deviations, covariance and the Pearson's correlation coefficient) are shortly introduced and wrapped up in the PCA method in the following section. For as far as co linearity is a problem in a dataset for cluster analysis, PCA will solve this problem. Much of what is written in the following section is inspired by a book of Field (2009).

3.3.1 Population, samples and cases

In statistical research it is rare to have access to information of the entire population. Therefore samples are used which are tested whether it is likely that they represent the total population. For cluster analysis it is important to realize that it is preferable to have access to the whole population your analysis is based on. Missing data, e.g. a case which is missing one or more value(s) for some variables is not allowed. For centroid based clustering it is already stated that outliers, which are cases with extreme values for a variable, should be left out the analysis. This is in contradiction with the statement that the whole population is used for the cluster analysis. This problem is addressed in the following section too.

3.3.2 Outliers

Outliers in a dataset are problematic for PCA because they influence the mean, and therefore the standard deviation of the variables. All the statistical procedures the data is used for are subject to and highly influenced by such extreme cases. Outliers are a specific problem for Centroid Based Clustering, which is the reason our data is closely checked for outliers. There are two ways to scan for and indicate outliers in your data. Firstly a histogram of the distribution of each can be used to scan the data for obvious problems. The second option is looking at the z-scores, this is preferred by (Field, 2009, p.153)

Calculating the z-scores is a way of standardizing the values in data set. This is done in such a way that a resulting distribution has scores with a mean of 0 ($\overline{X}$) and a standard deviation of 1 (s). The value of a case for a certain variable in *equation 3.4* is given by X.

$$z = \frac{X - \overline{X}}{s} \qquad 3.4$$

By using these z-scores, SPSS can come up with a frequency division of every variable too. An example of such a division on a dummy set is shown in Table 3.1.

outlier1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Absolute z-score less than 2	22259	96.0	96.0	96.0
	Absolute z-score greater than 1.96	391	1.7	1.7	97.7
	Absolute z-score greater than 2.58	201	.9	.9	98.6
	Absolute z-score greater than 3.29	333	1.4	1.4	100.0
	Total	23184	100.0	100.0	
Missing	System	1	.0		
Total		23185	100.0		

Table 3.1

Field suggests there are 3 ways to correct problems in the data. The first way is by removing the case. Because we know each case really exists, we need to be sure every case stays in the dataset so removing them is not an option. Secondly the data can be transformed, by doing this all values of a variable are changed. It is likely that transformation of data will fail and therefore the third method has been chosen. This is changing single values with a z-score of more than 3.29, with a score of 3.29 times the standard deviation above or below the mean for a variable.

3.3.3 Variance, Covariance and Correlation

After the dataset is completed and corrected for outliers, the correlation matrix is calculated. To understand the principal behind correlation first some basic statistics are addressed in the following section. Variance, covariance and correlation are the fundamental parts of a correlation matrix which is used for principal component analysis.

The dataset is considered to be the total population. The variance (s^2) in a dataset is calculated by dividing the squared sum of errors (SS) with the number of cases (N). This is shown by *equation 3.5.* The standard deviation is defined as the square root of the variance.

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{N} = \frac{\sum(x_i - \bar{x})(x_i - \bar{x})}{N} \qquad 3.5$$

The covariance (cov) is used to examine whether two variables are associated. The way covariance is measured is related to the equation for variance. The cross-product deviations between two variables (x and y) are calculated by multiplying the deviation of the mean of x with deviation of the mean of y. The covariance is defined as the cross-product deviation divided by the population size (N), see *equation 3.6.*

$$cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N} \qquad 3.6$$

A problem with covariance is that it does not account for the unit the different variables are measured in. For comparison of the covariance of all variables in a set the term correlation coefficient (r) is introduced. A way of calculating the correlation coefficient is by standardizing it as suggested by Pearson, resulting in the Pearson product-moment correlation coefficient, which is most often referred to as the Pearson correlation coefficient, Pearson's r or R. A way to standardize the covariance is by dividing it by the product of the standard deviations of the two variables, this leads to *equation 3.7* for R.

$$R = r = \frac{\mathrm{cov}(x,y)}{s_x s_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(N)s_x s_y} \qquad 3.7$$

3.3.4 Communality: Common and Unique variance

The variance in data is often split into 2 parts. The first part is the common variance. This is the segment of the variance that the variable has in common with other variables. The second part is the unique variance, which accounts for the variance that is owned solely by the variable itself. The common variance of a variable is called the communality. Communalities are used in decisions for the number of components to extract and in calculations concerning the KMO of variables and the complete dataset, see *paragraph 3.3.5*.

In PCA the communalities can be calculated after the components are extracted. This shows the multiple correlations between the variables and the factors extracted. This means that the communality is a measure for the amount of variance that can be explained by the extracted factors. A dummy example is shown in *Table 3.2*. All communalities of the variables should be 0.6 or greater. If this is not the case it is advised to extract more components or leave the variable out of the analysis.

Communalities

	Initial	Extraction
Variable 1	1.000	.672
Variable 2	1.000	.924
Variable 3	1.000	.547
Variable 4	1.000	.591
Variable 5	1.000	.836
Variable 6	1.000	.891

Table 3.2 Example of Communalities after extraction of components

3.3.5 Preliminary Analysis: Correlation Matrix

The correlation matrix is the standardized form of a covariance matrix. Such an R-matrix is very useful if variables are measured in different units. All variables are put into a matrix where the variables label the columns and rows. Therefore it contains a diagonal with values of 1 where the variables are compared with itself. Most often the values of cells above the diagonal, which is the upper right corner, are left empty because the same values are shown in the lower left corner.

Before conducting the PCA it is useful to look at the correlation matrix and run some tests that describe properties of the data and tell whether it is even allowed or meaningful to conduct a PCA. Two standard tests that indicate whether an analysis provides a reliable

solution are looking at the values for the KMO and Bartlett's test. These tests are described in the following section.

Most often the KMO is used to decide whether a sample size is big enough, for PCA it is used to determine whether it is smart to use the specific variables for linear components extraction. The Kaiser-Meyer-Olkin measure of sampling adequacy (Kaiser, 1970) is a ratio of the squared correlation between variables to the squared partial correlation between variables. The KMO of a set is always between 0 and 1, where 1 indicates that the sum of partial correlations is large in relation to the sum of correlations. Therefore PCA is probably not appropriate because the set contains diffusion in the pattern of correlation. The closer the value of KMO is to 1 the more compact the pattern is and the more components will be reliable and distinct. Kaiser (1974) states that the KMO of a set should be at least above 0.5, moreover below 0.7 they are mediocre and the values 0.8, 0.9, and 1.0 are good, great and superb respectively (Hutcheson & Sofronniou, 1999).

An anti-image matrix can be used to examine the KMO for the individual variables and can therefore be used to evaluate whether a variable meets the criterion of Kaiser (1974) to be at least 0.5. If a value for a variable in the diagonal cells is below 0.5 it is strongly advised to remove this variable from the analysis.

On similar grounds Bartlett's test of sphericity should be highly significant ($p<.001$) for PCA to be appropriate. Bartlett's test examines whether the correlation matrix is an identity matrix. This would indicate that all variables correlate very badly with each other and therefore factors cannot be extracted. According to Field (2009) it is very unlikely for a set to fail on this test. Drawing conclusions based on this test is not valid for a PCA. Therefore it is concluded that the KMO of the overall dataset and the anti-image matrix will be used to check the pattern of correlation for the individual variables.

3.3.6 Component Extraction: Component Loadings

To extract the components, the total of eigenvalues of a dataset regarding the different components is of crucial importance. Without looking at the mathematics concerning eigenvalues, the meaning of them is briefly discussed. The theory underlying eigenvalues can be visualized by looking at a scatterplot of two variables. A scatterplot contains the values for one variable on the x-axis and of the other values on the y-axis. The dimension of an ellipse around the plotted cases (points) can be specified with eigenvectors. The length of the eigenvector is a single value and defined as the eigenvalue. By looking at all the eigenvalues of a dataset the dimensions become clear. In other words they should show how the variance in the dataset is distributed.

In PCA the eigenvalues are determined for each linear component of the correlation matrix. There are as many components as there are variables. And the eigenvalues are a way to decide whether a component is statistically important. Remember, the dataset has to be reduced while keeping the most important components that explain a substantial amount of variance in the data. There are two ways to decide which components should be accepted, both are used in this research. The first is the Kaiser criterion Kaiser (1960) where components with eigenvalues greater than 1 are selected. For Kaiser's criterion to be valid the sample size must be above 250 and the average communalities after extraction should be greater than .6. The second alternative, a very useful option, is using a screeplot (Cattell, 1966). The eigenvalues of the components are plotted and the point of inflection is an

indication of how many components should be extracted. This is demonstrated in *Figure 3.7,* where 2 components are advised to extract. It is always advisable to compare the number of components both methods advocate to extract. It is expected both methods support the same number of components.

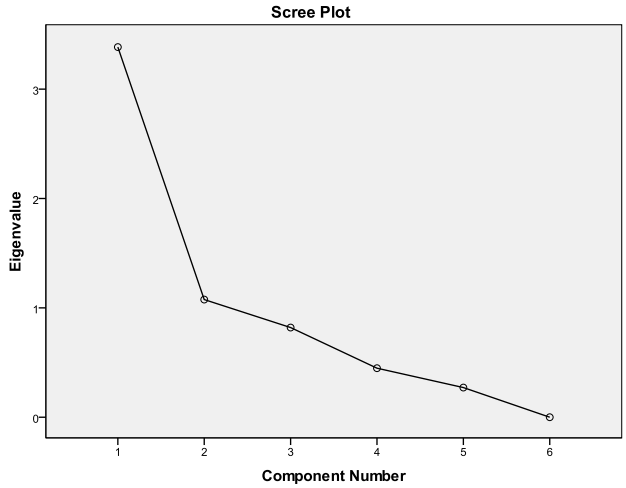


Figure 3.7 Example of a Screeplot advised to use by Cattell (1966)

The third test is looking at the residuals of the dataset. The residuals are calculated by subtraction of the reproduced correlations after component extraction from the original correlations. The amount of residuals above .05 should be lower than 50 percent (Field, 2009, p.664). The reproduced correlations of the variables with itself are called the communalities elaborated in *paragraph 3.3.4*.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.385	56.409	56.409	3.385	56.409	56.409
2	1.076	17.931	74.340	1.076	17.931	74.340
3	.820	13.666	88.006			
4	.448	7.473	95.479			
5	.271	4.521	100.000			
6	-4.761E-15	-7.935E-14	100.000			

Table 3.3 Example of Variance Explained table

In *Table 3.3* an example of a "variance explained" table is shown. Eigenvalues, as well as the percentage variance explained is shown for each component. The last column shows the cumulative value. Based on the criterion set for the analysis the extracted components are listed again.

The component loadings are the loadings of a variable on a component, which are the eigenvectors calculated from the eigenvalues of an R-matrix. This is often displayed in a component matrix; an example is displayed in *Table 3.4*. The cells represent the component loadings, how these are used is explained in *paragraph 3.3.8 Component Scores*.

Component Matrix

	Component	
	1	2
Variable 1	.775	.268
Variable 2	.001	.961
Variable 3	.737	.057
Variable 4	.762	.102
Variable 5	.891	-.206
Variable 6	.931	-.156

Table 3.4 Example of a component matrix

3.3.7 Component Rotation

Component rotation is introduced to maximize the loadings of variables onto the extracted components. This way the component loadings tend to polarize and thus being either high or low. There are two kinds of component rotation: orthogonal and oblique rotation. It is not in the scope of this chapter to elaborate further on the mathematics behind the techniques. Orthogonal rotation should be used if it is expected that the components are independent, if it is suspected that components might correlate oblique rotation could be used (Field, 2009, p.644). When the extracted components have a negligible correlation it is reasonable to use orthogonal rotation (Pedhazur & Schmelkin, 1991). This can be done by execution of an oblique rotation together with the PCA and examination of the component correlation matrix. If the components are not independent an oblique rotation could be used. The result is a rotated component matrix that looks the same as a normal component matrix.

3.3.8 Component Scores

It is important to understand the last step from component loadings towards component scores. It is stated that linear components (Y) are extracted and therefore it is a linear function (*equation 3.8)*. The component loadings for each variable (b) are multiplied with the value of a case for that specific variable. The summation of this product for all variables is the component score.

$$Y_i = b_1X_{1i} + b_2X_{2i} + \cdots + b_nX_{ni} + \varepsilon_i \qquad 3.8$$

If the variables have a different unit, this is a problem because the component score does not make sense anymore, especially when the scores are used in cluster analysis. This is the moment the standardization of our dataset takes place, which can be done using the regression method in SPSS.

The Pearson's correlation coefficient is already used as fundament for the KMO statistics of the data. Now the correlation matrix is used to standardize the linear principal components. To standardize the component loadings the inverse of the R-matrix is multiplied by the component matrix. The results are component score coefficients (B) which can replace the component loadings in *equation 3.8*. This results in *equation 3.9.*

$$Y_i = B_1X_{1i} + B_2X_{2i} + \cdots + B_nX_{ni} + \varepsilon_i \qquad 3.9$$

The components and their respective score for every case can be used for further analysis. Take note the scores are already standardized.

4 Case Study on Eindhoven

Although in theory the method discussed could easily have the total city of Eindhoven as its scope, a different approach is chosen. Beforehand it is hard to assess the quality of the data which is going to be used. It takes quite a lot of time to carefully scan the values of variables for all dwellings of Eindhoven. At the same moment it is time consuming to both interpret and validate (1) the components extracted in PCA and (2) the final cluster typology. For interpretation of the right amount of clusters and what they visualize, extensive knowledge of the neighborhoods you examine is needed. Therefore it is chosen to limit the case study to a single district of Eindhoven. The used district is introduced in the next section.

In paragraph *4.2 Principal Component Analysis* the amount of variables is reduced and transformed into fewer components and a characterization and interpretation of the extracted components is given. In *paragraph 4.3 Cluster Analysis* the cluster analysis is conducted and validated using the sketched criteria. The final interpretation is elaborated in paragraph *4.4 Cluster interpretation*.

For privacy reasons figures of clusters with less than 5 dwellings are not presented in the tables included.

4.1 Target Area

4.1.1 Eindhoven

Eindhoven consists of more than 96,000 dwellings. Almost half of them, about 40,000 dwellings, are located on property owned by a housing corporation. More than 56.000 dwellings have solely private owners. Certain types of dwellings have several private owners which are united in an association of owners. When a private dwelling is located in a multi-family building which is partly owned by a housing corporation it is wrongly assumed to be owned by a corporation. The municipality of Eindhoven has no record of whether a privately owned dwelling is inhabited by its owner or by tenants. (HetEnergiebureau BV, Q-Energy BV, Endinet BV and Gemeente Eindhoven, 2011, p. Appendix 1E)

Of the 56.000 privately owned dwellings in Eindhoven more than 45.000 are built before 1992 and still more than 30.000 are built before 1975. Before 1992 only few regulations on thermal insulation existed and based on this we expect a tremendous saving potential for Eindhoven is likely to exist.

In the upcoming years the population of Eindhoven is expected to grow, from 216,000 inhabitants now, to 226,000 in 2021 (Gemeente Eindhoven, 2010). Based on the prediction for development the dwelling market will most likely have a growth of about 5% in 10 years. How to tackle this need for growth is addressed by Van der Weerdt (2011). It is concluded that renovation is often not preferable for dwellings owned by corporations. However demolition and developing new estate on private property is far from likely to become mainstream in the upcoming decennia.

4.1.2 A Priori: Statistical Division of Eindhoven

The statistical division is introduced by the department of “Policy Information and Research” (BIO: afdeling “BeleidsInformatie & Onderzoek”) of the municipality of Eindhoven (Municipality of Eindhoven, 2011). This division is adopted by the municipality, the province and the Central Statistical Bureau (CBS: “Centraal Bureau voor de Statistiek”).

Eindhoven is divided into 7 quarters, i.e. Centrum, Tongelre, Gestel, Stratum, Strijp, Woensel and Gestel. These 7 quarters are split further into 20 districts and 116 neighborhoods, see *Figure 4.1*. Based on composition of the dwelling stock a neighborhood is further divided into clusters. The city of Eindhoven has more than 1,100 a priori clusters.

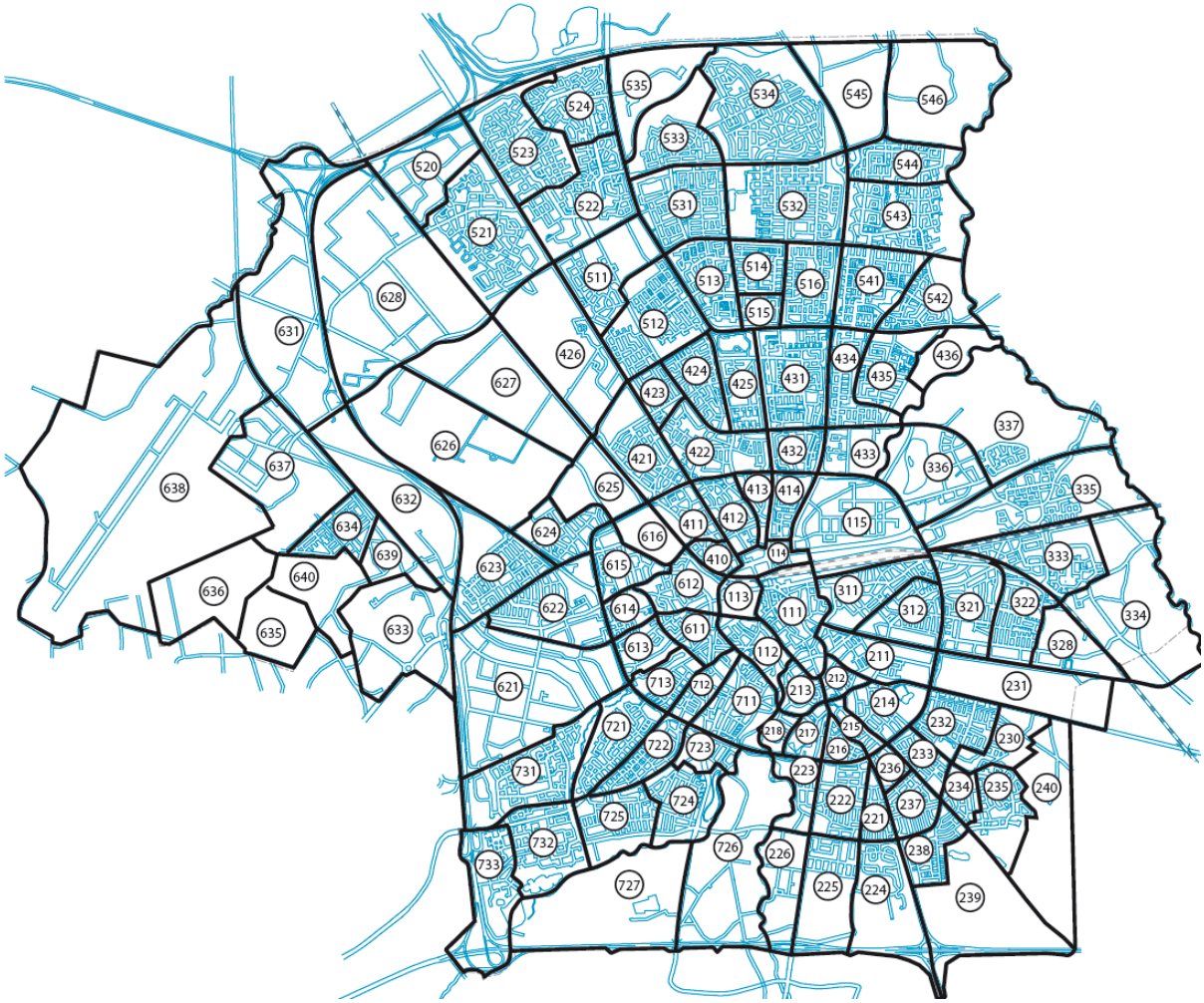


Figure 4.1 Eindhoven division in 116 statistical neighborhoods (Municipality of Eindhoven, 2011)

4.1.3 District: De Laak

The district “De Laak” is situated north of “Het Eindhovens Kanaal” and south of the railway tracks, between the city centre and the traffic access ring of Eindhoven. De Laak is divided into two neighborhoods, i.e. “Villapark” and “Lakerlopen”, see *Figure 4.2*. This district is chosen, because both neighborhoods diverge substantially in building periods and housing typology and therefore house prices and inhabitants. The two neighborhoods are subdivided into 5 sub-neighborhoods and 34 a priori clusters. The descriptives for the 34 clusters are listed in *Table 4.1*.

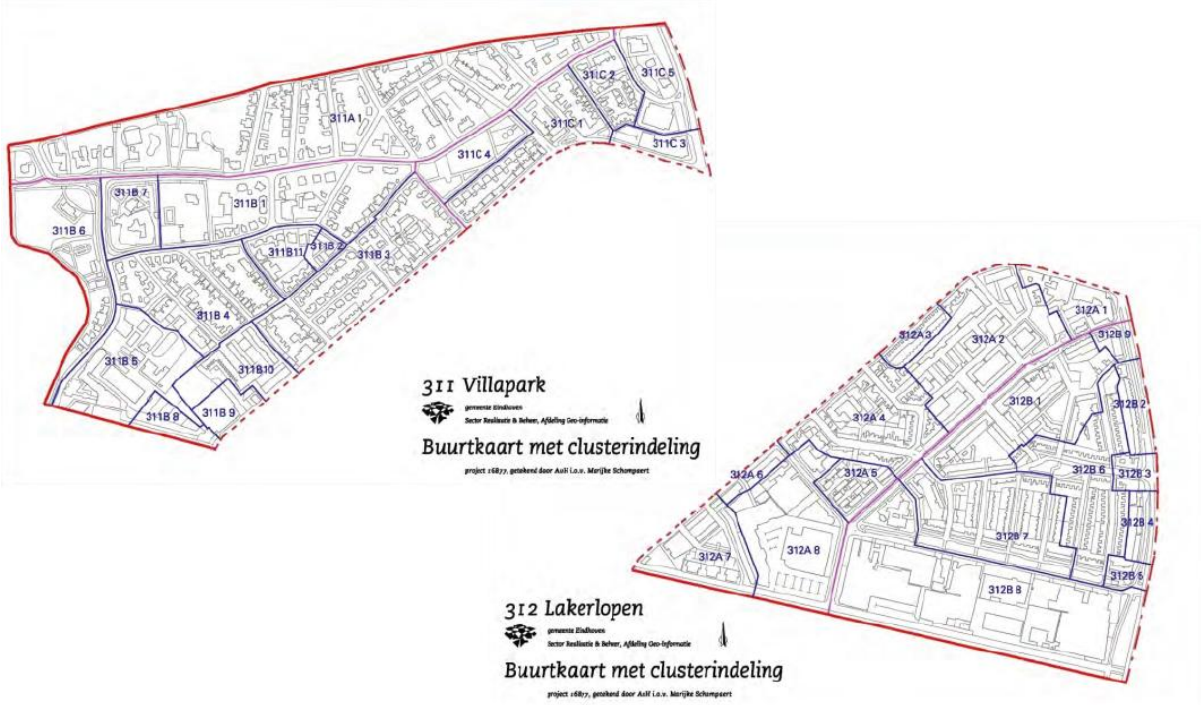


Figure 4.2 De Laak Neighborhood 311 and 312 in “Cluster distribution by neighborhood, división 2010, atlas” (Municipality of Eindhoven, 2010)

The visualization shown in Figure 4.3 is displayed in the format used for all maps. In the lower left corner a “variwide” plot of the different clusters is shown. On the y-axis the average saving potential is shown, the clusters are sorted on average saving potential. The surface of the bars/boxes is the total saving potential of a cluster. The total saving potential of the district is summation of the surface of the boxes. Knowing this, it is advocated to have as much of the surface in the first part of the distribution. This way we are able to select the neighborhoods that have the highest average saving potential as possible.

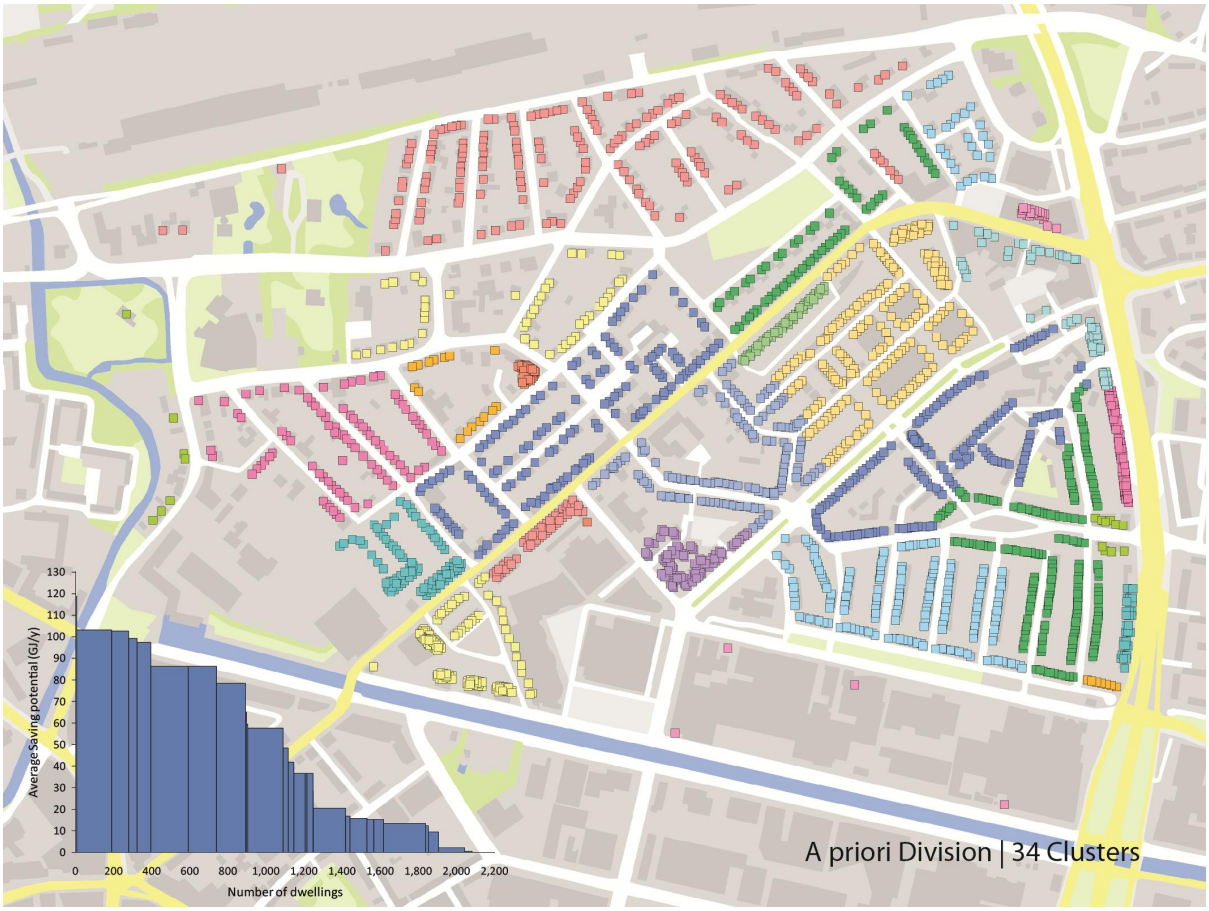


Figure 4.3 Visualization of District "De Laak" on most specific cluster level “a priori”

				Total	Average							
		Cluster code	Number of Dwellings	Saving potential	Saving potential	Dwelling age	WOZ value	Household size	Household Children	Household Oldest Age	Electricity SYU	Gas SYU
				GJ/y	GJ/y	y	k€			y	kWh	m^3
Villapark	A	311A 1	180	18,538	103	87	644	2.91	0.82	57	5,573	3,551
	Subtotal	**311A**	**180**	**18,538**	**103**	**87**	**644**	**2.91**	**0.82**	**57**	**5,573**	**3,551**
	B	311B 1	45	4,463	99	86	658	2.71	0.84	62	5,009	3,744
		311B 2	37	565	15	24	141	1.11	0.00	69	2,381	1,004
		311B 3	153	11,973	78	76	393	2.63	0.61	50	4,510	2,690
		311B 4	91	9,327	102	89	454	2.54	0.68	54	4,554	3,083
		311B 5	0									
		311B 6	7	830	119	99	1,109	2.14	0.43	47	11,486	8,814
		311B 7	0									
		311B 8	0									
		311B 9	0									
		311B10	108	0	0	1	239	1.78	0.19	41	2,539	918
		311B11	16	197	12	22	601	3.31	1.44	51	5,530	3,072
	Subtotal	**311B**	**457**	**27,354**	**60**	**57**	**393**	**2.31**	**0.52**	**51**	**4,072**	**2,424**
	C	311C 1	70	6,811	97	76	412	3.07	0.73	54	4,864	2,742
		311C 2	28	1,355	48	34	486	2.71	0.71	61	4,894	2,635
		311C 3	23	385	17	31	171	1.78	0.13	43	2,585	1,144
		311C 4	0									
		311C 5	0									
	Subtotal	**311C**	**121**	**8,552**	**71**	**58**	**383**	**2.74**	**0.61**	**53**	**4,438**	**2,413**
	Total	***311***	758	54,444	72	64	451	2.52	0.61	53	4,487	2,690
Lakerlopen	A	312A 1	27	1,126	42	67	201	2.22	0.15	38	5,993	2,415
		312A 2	222	2,968	13	19	234	2.30	0.59	55	3,196	1,163
		312A 3	48	733	15	37	175	1.48	0.04	44	2,296	1,116
		312A 4	148	12,763	86	79	196	2.55	0.45	47	3,890	2,138
		312A 5	90	1,421	16	24	140	1.68	0.03	36	2,889	1,046
		312A 6	53	500	9	21	156	1.77	0.06	28	2,653	785
		312A 7	138	286	2	11	255	1.78	0.17	54	2,830	936
		312A 8	1									
	Subtotal		**727**	**19,826**	**27**	**33**	**208**	**2.09**	**0.32**	**48**	**3,242**	**1,329**
	B	312B 1	169	3,445	20	22	222	2.64	0.83	41	2,898	1,278
		312B 2	60	2,196	37	54	157	1.33	0.07	53	2,822	1,188
		312B 3	11	403	37	54	211	7.00	0.36	40	6,435	3,595
		312B 4	32	1,171	37	56	154	2.41	0.13	31	2,630	1,174
		312B 5	7	416	59	56	207	5.00	1.43	40	4,875	2,445
		312B 6	198	17,094	86	66	161	2.10	0.56	45	2,787	1,523
		312B 7	185	10,643	58	61	158	2.40	0.78	47	2,600	1,614
		312B 8	4									
		312B 9	36	22	1	11	196	1.64	0.14	36	2,409	728
	Subtotal	**312B**	**702**	**35,650**	**51**	**50**	**178**	**2.33**	**0.60**	**44**	**2,834**	**1,450**
	Total	***312***	***1,429***	***55,476***	***39***	***41***	***193***	***2.21***	***0.46***	***46***	***3,042***	***1,389***
De Laak		***31***	***2,187***	***109,920***	***50***	***49***	***283***	***2.32***	***0.51***	***48***	***3,543***	***1,840***

Table 4.1 Descriptives District "De Laak" on several levels “a priori”

4.2 Principal Component Analysis

In the following paragraph the results of the PCA are addressed. For the theoretical underpinning of the method and the thresholds for selection of components used *paragraph 3.3 Principal component analysis* will provide. A summary of the PCA conducted is given below. A more comprehensive description is given in the next pages.

A principal component analysis (PCA) was conducted on the 9 variables with oblique rotation (Direct Oblimin). He KMO verified the sampling adequacy for the analysis, KMO - .726 ('Good" according to Hutcheson & Sofronniou (1999)), and all KMO values for individual items were > .618 which is well above the acceptable limit of .5 (Field, 2009). An initial analysis was run to obtain eigenvalues for each component in the data. Three components had eigenvalues over Kaiser's Criterion of 1 and in combination explained 78.40 percent of the variance. The screeplot was slightly ambigious and showed an inflexion that would justify retaining 2 or 3 components. Given the large dataset, and the convergence of the screeplot and Kaiser's criterion on three components , this is the number of components that were retained in the final analysis. *Table 4.9* Shows the component structure loadings after rotation. The items that cluster on the same components suggest that component 1 represents the "dwelling saving potential", component 2 the "Household characteristics" and component 3 that "Wealth comes with age".

4.2.1 Data preparation

The descriptives of the a priori clusters are based on the data after objects with missing values for some variables are deleted or supplemented based on identical dwellings in the direct proximity of the object. This was only the case for WOZ-values and electricity and gas standard year usage (SYU). The description of all steps in the data preparation is included in *appendix C Data preparation.*

Another step in the data preparation is correcting the data for outliers. Outliers tend to form a cluster consisting of only one object and therefore ruining the clustering process. It is already stated that the z-scores of the variables per object will be used to check for outliers. The outliers can only occur in the nine variables used in the PCA. Every variable is scanned and descriptive statistics are shown in *Table 4.2*. Afterwards which objects contain variable values with outliers is checked, the upper and lower boundary, i.e. the range per variable is listed in *Table 4.3*. In case a value is above the upper boundary, the value is replaced with the upper boundary itself.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
dwelling_age	2187	2	106	49.15	33.001
squared_dwelling_age	2187	4	11236	3504.07	3205.313
saving_potential	2187	0	129749	50260.82	43995.585
WOZ_value	2187	61000	4178000	282568.82	206520.007
household_size	2187	1	16	2.32	1.557
household_children	2187	0	7	.51	.946
age_oldest	2187	14	97	48.35	17.432
electricity_SYU	2187	2	58449	3542.78	3186.082
gas_SYU	2187	2	40977	1839.60	1607.620
Valid N (listwise)	2187				

Table 4.2 SPSS descriptive dataset statistics including range and standard deviation

The amount of values changed differs per variable. It is easily explained that the WOZ-value, the number of people and children in a household and the SYU's would contain some outliers. The most replacements are made in the WOZ-value of a dwelling, 27 replacements represent 1.2 percent of the objects in the data, and consequently all other variables have lower replacement rates. For all variables a table with descriptives focusing on outliers is provided in *appendix D Descriptives outliers*.

	lower boundary $(\bar{X} - 3.29*s)$	**Upper boundary** $(\bar{X} + 3.29*s)$	**Replacements made**
Dwelling age	0	157.72	0
Squared dwelling age	0	14,049.55	0
Saving potential	0	195,006.29	0
WOZ value	0	962,019.64	27
Household size	0	7.44	14
Household children	0	3.62	8
Household oldest age	0	105.70	0
Electricity standard year usage	0	14,024.99	25
Gas standard year usage	0	7,128.67	18

Table 4.3 Range after correction on outliers

4.2.2 Correlation

An important aspect of the PCA is the correlation, which can be presented in an R-matrix of Pearson's correlation coefficients of the data. The correlation matrix is used to calculate the component scores using the component loadings for each component. Therefore the matrix is shown in Table 4.4.

	Dwelling age	Squared dwelling age	Saving potential	WOZ value	Household size	Household children	Household oldest age	Electricity standard year usage	Gas standard year usage
Dwelling age	1.00								
Squared dwelling age	.973	1.00							
Saving potential	.908	.887	1.00						
WOZ value	.361	.447	.442	1.00					
Household size	.177	.19	.186	.2	1.00				
Household children	.125	.136	.147	.195	.697	1.00			
Household oldest age	.097	.113	.146	.279	-.156	-.115	1.00		
Electricity standard year usage	.264	.294	.275	.448	.456	.288	.043	1.00	
Gas standard year usage	.543	.565	.57	.656	.356	.2	.179	.582	1.00

Table 4.4 R-matrix of dataset "De Laak" (values above .3 are listed bold)

Execution of some tests should deliver supporting arguments for the suitability of the dataset containing 2187 dwellings for PCA. The KMO of the dataset and the individual KMO of the variables, which are the diagonal values in the anti-image correlation matrix, are shown in *Table 4.5 KMO measure of sampling adequacy of dataset "De Laak"*. The table is constructed using values adopted from the SSPS output in *appendix E Tables Output PCA*.

	Anti-image Correlation
Dwelling age	.654
Squared dwelling age	.709
Saving potential	.888
WOZ value	.652
Household size	.625
Household children	.593
Household oldest age	.742
Electricity standard year usage	.839
Gas standard year usage	.819
Total	**.726**

Table 4.5 KMO measure of sampling adequacy of dataset "De Laak"

The values of the KMO should be at least .5 for all variables in the dataset. The overall KMO of the dataset is .726, and according to Hutcheson & Sofronniou (1999) be considered as "good".

4.2.3 Component extraction

The number of components to be extracted is determined using the Kaiser criterion, the point of inflection in the screeplot and the percentage of number of residuals above .05. As you see in *Table 4.6* all communalities are above the mandatory boundary of .6. This supports the usage of the Kaiser Criterion, were components with a total eigenvalue above 1.0 are selected.

	Communalities after extraction
Dwelling age	.968
Squared dwelling age	.954
Saving potential	.908
WOZ value	.705
Household size	.819
Household children	.707
Household oldest age	.618
Electricity standard year usage	.627
Gas standard year usage	.751

Table 4.6 Communalities after extraction should be above .6 for Kaiser Criterion to be valid

In *Table 4.7* the first three components have an eigenvalue higher than 1.0 and therefore listed bold. These three components represent a substantial amount of variance in the data. In the dataset 33 percent (12 redundant cases) have residuals of greater than .05. This means that with extraction of 3 linear components out of the data the cumulative percentage for variance explained by the components, of about 78 percent, is high enough to have less than 50 percent of the residuals below .05.

	Initial eigenvalues		
Component	**Total eigenvalue**	**% of Variance**	**Cumulative % of Variance**
1	**4.072**	**45.25**	**45.25**
2	**1.78**	**19.783**	**65.032**
3	**1.203**	**13.37**	**78.402**
4	0.759	8.429	86.831
5	0.496	5.507	92.338
6	0.314	3.49	95.828
7	0.231	2.565	98.393
8	0.125	1.387	99.78
9	0.02	0.22	100

Table 4.7 Total Variance explained for components (3 extracted components are listed bold)

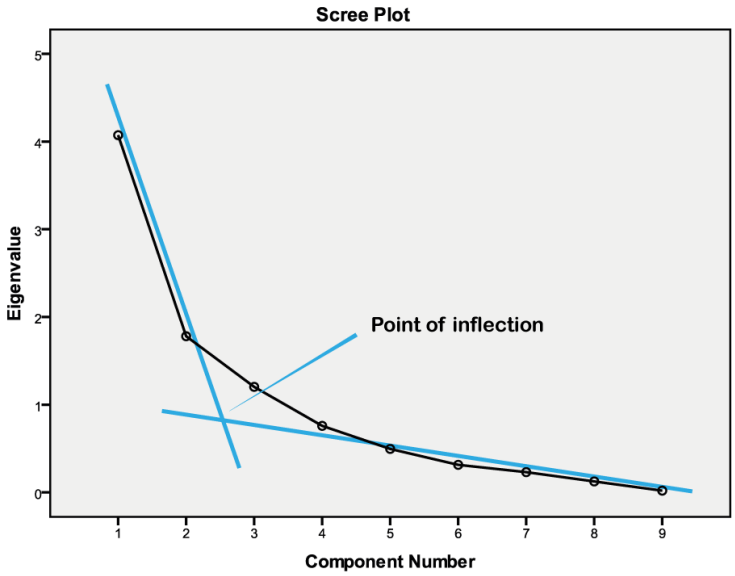


Figure 4.4 Screeplot with point of inflection of dataset "De Laak"

Moreover the point of inflection in the screeplot in *Figure 4.4* supports the selection of 2 or 3 components. It is decided to extract 3 linear components, which are evaluated in CA together with the X and Y coordinates of objects.

4.2.4 Component scores

Before the component scores per object are calculated an oblique rotation on component loadings is conducted. This way the loadings of variables on components are amplified, and components are discriminated. Overall, the variance in the component scores is maximized. It is decided to use oblique rotation (Direct Oblimin in SPSS) because it is suspected that the extracted components are not independent, this is supported by non zero values in the component correlation matrix in *Table 4.8*.

Component	1	2	3
1	1.000	.262	.315
2	.262	1.000	.102
3	.315	.102	1.000

Table 4.8 Component Correlation Matrix of dataset "De Laak"

	Pattern			Structure		
Component	**1**	**2**	**3**	**1**	**2**	**3**
Dwelling age	1.017			0.98		
Squared dwelling age	0.986			0.976		
Saving potential	0.953			0.952		
WOZ value			0.697	0.467		0.779
Household size		0.91			0.903	
Household children		0.851			0.83	
Household oldest age			0.763			0.705
Electricity standard year usage		0.582	0.474		0.633	0.537
Gas standard year usage			0.528	0.633	0.47	0.681

Table 4.9 Component Loadings displayed in Pattern and Structure Matrices after oblique rotation (values only shown when above .4)

It is important to give a characterization of the extracted components. This is done by interpretation of the loadings of variables on the different components. After oblique rotation two component matrices are presented. The pattern matrix contains the regression coefficients between each variable and a component. The structure matrix contains the correlation coefficients between each variable and a component. The structure matrix is multiplied by the inverse of the correlation matrix to get the component scores. The value of a component for an object is the summation of the component score of a variable multiplied with the Z-score of a variable for a certain object; this is called the regression technique, see *equation 3.8 and 3.9*.

The first component is strongly influenced by the presence of 2 identical variables representing the age of a dwelling. The saving potential of a dwelling is highly related with the age of a dwelling. Therefore the first component, which explains 45 percent of the variance in the data, is labeled “Dwelling saving potential”.

The second component represents the household characteristics of the objects in the study. As suspected the electricity usage of a dwelling and its occupants is related. This component, representing 20 percent of the variance in the data, is labeled “Household characteristics”.

The third component only represents 13 percent of the variance in the data. It is mainly influenced by the not earlier mentioned WOZ value of a dwelling and the age of the oldest inhabitant. This component is labeled “Wealth comes with age”.

4.3 Cluster Analysis

In the following paragraph several runs of a k-means cluster analysis are presented, with a different number of clusters as output. The objects are clustered based on the scores on three components extracted from the data, characterized as 1) dwelling saving potential, 2) household characteristics and 3) wealth comes with age, together with the weighted X and Y coordinates. The used distance measure is the Euclidean distance. The method used is described in *paragraph 4.3.1.* The output of the 2 different cluster analyses that were executed is presented in *paragraph 4.3.2*. In the same paragraph the clusters are validated using the evaluation criteria which are the total and the average saving potential and the spatial integrity of the clusters.

4.3.1 Method

As discussed in the previous chapter two main types of clustering exist. These are connectivity based, such as with hierarchical clustering, or centroid based as used in K-means clustering. In this study K-means clustering is used with an upfront specified number of clusters (K). This way it is possible to evaluate new spatial contiguous energy clusters and target group clusters as well. This specific method is chosen mainly because it is used by Wu & Rashi Sharma (2011) and the first explorative tests that were conducted pointed out K-means clustering is a well performing method on our data set.

4.3.2 Cluster output and validation

After the cluster procedure is executed the clusters are visualized on a map using GIS. The distribution of dwellings in one cluster over the district is inspected by looking at those maps. In most cases there is need for spatial contiguous boundaries of a cluster, this is inspected visually too (Wu & Rashi Sharma, 2011).

The homogeneity of the clusters, i.e. the similarity of the objects per cluster regarding the different dwelling and household characteristics, is checked using *equation 4.1*. The weighted average standard deviation (WASD) is a measure to evaluate and validate the cluster homogeneity (Wu & Rashi Sharma, 2011). The Standard deviation per cluster regarding a characteristic (s or SD) is the square root of the variance (s^2) per cluster regarding that characteristic. For easy comparison, one measuring unit per characteristic is adopted. The SD of all clusters in the district are multiplied with the number of dwellings in that specific cluster (N_i), these weighted SD's are summed. The WASD is calculated by division of the summed weighted SD's by the total number of dwellings in the district (N).

$$\mathrm{WASD}_{\text{per characteristic}} = \frac{\sum_{i=1}^{i=n}(N_i * SD_i)}{N} = \frac{\sum_{i=1}^{i=n}\left(N_i * \sqrt{\frac{\sum_{j=1}^{j=N_i}(x_j - \bar{x})^2}{N_i}}\right)}{N} \qquad 4.1$$

The WASD measures the intra-cluster homogeneity and is therefore a reasonable measure of the quality of the clustering regarding a specific characteristic. A low WASD indicates a high homogeneity within the cluster. This way the clustered district can be compared with the a priori classification method, i.e. the statistical neighborhood division of the municipality of Eindhoven. The WASD of the a priori division is shown in *Table 4.10*.

		Weighted Average Standard Deviation							
		Saving potential	Dwelling age	WOZ value	Household size	Household Children	Household Oldest Age	Electricity SYU	Gas SYU
		GJ/y	y	k€			y	kWh	m3
1 Cluster	A Priori	44.00	33.00	182	1.45	0.92	17.43	2,331	1,324
2 Clusters	A Priori	40.95	31.09	113	1.45	0.92	17.08	2,209	1,148
5 Clusters	A Priori	38.00	28.82	103	1.43	0.90	16.95	2,171	1,114
34 Clusters	A Priori	19.02	12.50	75	1.30	0.81	15.26	2,016	909

Table 4.10 Weighted Average Standard Deviation of the objects to the clusters mean per characteristic, A priori

The third measure used is the shape of the distribution in the "variwide" plot mentioned in *paragraph 4.1*. When the shape of the variwide plot of a CA is corresponding with the plot displayed in *Figure 4.5*, a high performance on the characteristic saving potential is achieved. As you can see in the chart approximately 1,000 dwellings in "De Laak" have a substantial saving potential. It is tried to achieve homogeneous clusters that contain those 1,000 dwellings, this means that the dwellings without saving potential should be clustered together in the remaining clusters. A well performing cluster division is shaped as much as possible like the chart. However this cannot be achieved completely because other variables and spatial contiguity are taken into account too.

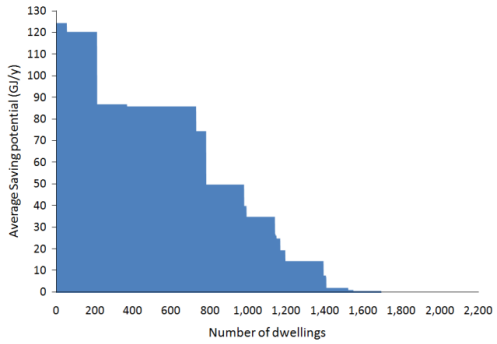


Figure 4.5 Saving potential of all individual dwellings in district "De Laak", sorted on saving potential.

New spatial contiguous energy clusters

The first variant that was executed focused on a new statistical division of the target area. To be able to make a comparison with the a priori division, the number of clusters of the a priori division was adopted up front. Therefore the number for k is 34. The second condition that should be met was that the clusters should have spatial contiguous boundaries. This means that nearly all the dwellings in one cluster should be in each other's direct proximity. To be able to reach this goal the dwelling's X and Y coordinate are weighted before 4 CA's were executed. The weights applied are the actual standardized units of the X and Y coordinate and the units to the 10nd power, i.e. $*10^1$, $*10^2$ and $*10^3$.

The number of objects in a cluster is listed in *Table 4.11*. The division of objects over the clusters fluctuates heavily when the weighting of the coordinates is increased. With every new CA the objects end up in other clusters and new divisions are made. Based on the development of these divisions it is hard to draw conclusions. The check for spatial contiguous boundaries can only be done after visualization of the clusters on a map. All the maps are shown in *Appendix F Visualizations* and the most important map is shown in *Figure 4.6.*

Cluster NUmber	3 Components and X Y coord. (unit)	3 Components and X Y coord. ($*10^1$)	3 Components and X Y coord. ($*10^2$)	3 Components and X Y coord. ($*10^3$)	Cluster NUmber	3 Components and X Y coord. (unit)	3 Components and X Y coord. ($*10^1$)	3 Components and X Y coord. ($*10^2$)	3 Components and X Y coord. ($*10^3$)
1	47	19	100	85	**18**	71	33	78	124
2	33	65	44	22	**19**	48	40	66	23
3	153	26	59	3	**20**	55	121	2	2
4	10	19	44	2	**21**	122	72	3	93
5	116	85	126	58	**22**	48	136	54	42
6	62	5	131	45	**23**	22	68	77	1
7	18	32	54	95	**24**	**208**	56	64	29
8	200	18	90	65	**25**	90	68	23	33
9	62	68	99	75	**26**	46	17	97	38
10	75	45	106	62	**27**	70	52	23	97
11	32	65	1	88	**28**	96	**238**	58	53
12	50	3	48	94	**29**	18	133	77	106
13	19	67	1	102	**30**	29	52	86	86
14	6	48	**127**	97	**31**	122	41	59	120
15	3	62	26	122	**32**	67	40	81	2
16	73	45	60	28	**33**	78	133	82	99
17	10	53	63	**126**	**34**	28	162	78	70

Table 4.11 Number of dwellings in each cluster after K means cluster analysis (K=34) with weighted X and Y coordinates

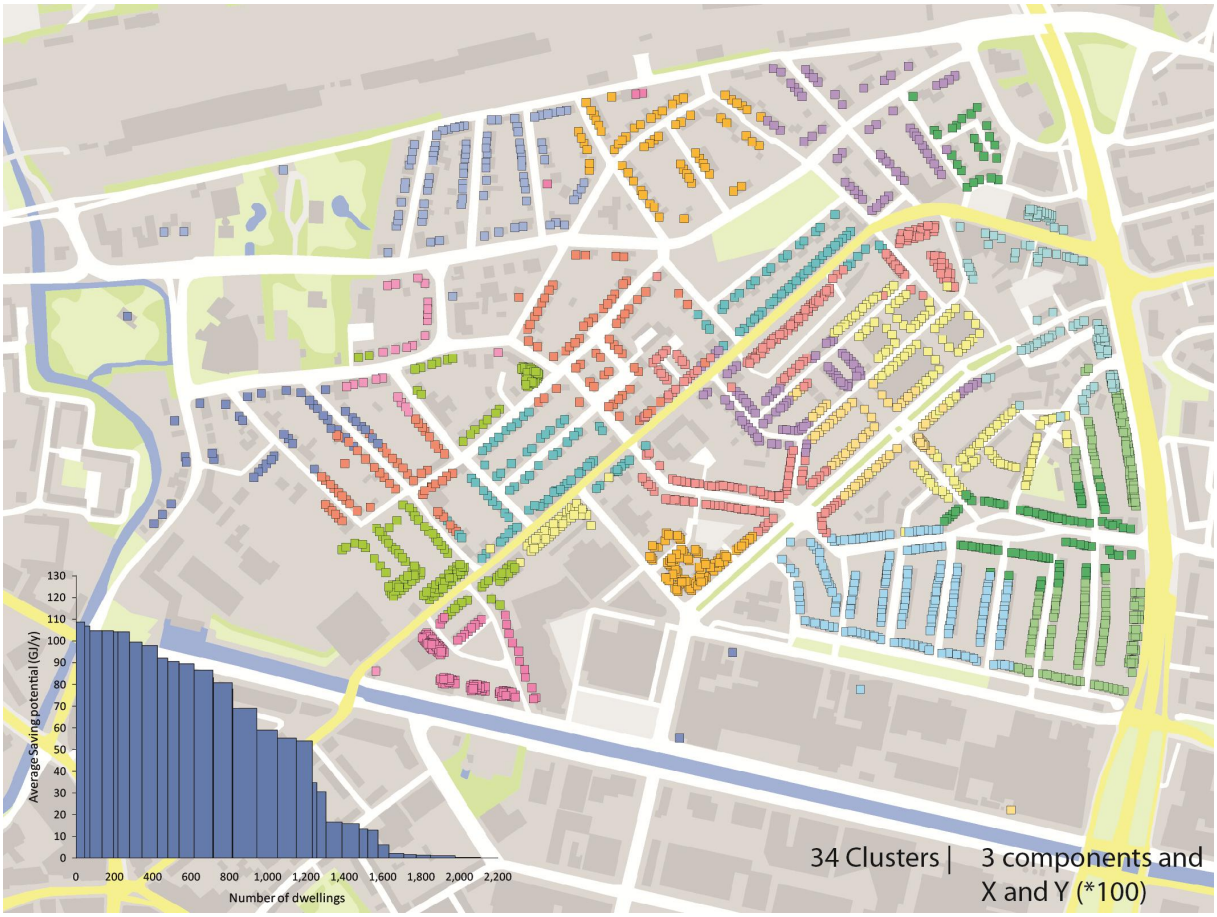


Figure 4.6 Visualization of K-means clustering (k=34) of the 3 components and X and Y coordinates weighted $*10^2$

The figure above shows 34 new clusters which are shown to have spatial contiguous boundaries. Some clusters have a few objects that are mixed on the edge of a cluster. However they are in such a minority that this is allowable. This is based on the figures found in the WASD calculations. The WASD calculations are shown in *Table 4.12*.

		Weighted Average Standard Deviation							
		Saving potential	Dwelling age	WOZ value	Household size	Household Children	Household Oldest Age	Electricity SYU	Gas SYU
		GJ/y	**y**	**k€**			**y**	**kWh**	**m3**
34 Clusters	A Priori	19.02	12.50	75	1.30	0.81	15.26	2,016	909
34 Clusters	3 components and XY (unit)	10.96	5.23	73	0.68	0.35	10.15	1,413	666
34 Clusters	3 components and XY ($*10^1$)	12.25	5.96	73	0.72	0.38	10.92	1,511	692
34 Clusters	3 components and XY ($*10^2$)	13.67	5.97	74	1.33	0.80	15.45	2,004	899
	Deviation A Priori 34	*-28%*	*-52%*	*-1%*	*2%*	*-1%*	*1%*	*-1%*	*-1%*
34 Clusters	3 components and XY ($*10^3$)	29.43	19.98	78	1.39	0.86	15.83	2,073	981

Table 4.12 Weighted Average Standard Deviation of the objects to the clusters mean per characteristic

The intra-cluster homogeneity of the 34 clusters that are found in the CA, with weighting $*10^2$ for XY coordinates, is much higher than for the existing a priori classification method for two of the characteristics. The WASD of the saving potential and the age of the dwellings

are respectively 28 and 52 percent lower. The objects in the new clustering have a more equal distribution regarding saving potential and dwelling age. To compare the descriptive of the k-means clustering with the a priori classification in *Table 4.1* the XY*10^2 case is shown in *Table 4.13*.

		Total	Average							
Cluster number	Number of dwellings	Saving potential	Saving potential	Dwelling age	WOZ value	Household size	Household Children	Household Oldest Age	Electricity SYU	Gas SYU
		GJ/y	GJ/y	y	k€			y	kWh	m^3
1	100	8,081	81	78	166	2.31	0.37	46	3,648	1,902
2	44	588	13	25	168	2.11	0.07	30	3,145	1,118
3	59	5,341	91	85	330	2.07	0.27	53	3,556	2,266
4	44	4,781	109	94	631	2.89	0.89	53	6,354	4,571
5	126	137	1	10	250	1.67	0.10	55	2,785	877
6	131	35	0	5	235	1.81	0.22	40	2,551	939
7	54	4,964	92	72	369	3.41	0.67	52	5,269	2,885
8	90	1,421	16	24	140	1.68	0.03	36	2,889	1,046
9	99	8,558	86	66	165	2.72	0.60	45	3,085	1,949
10	106	6,243	59	61	153	2.40	0.82	47	2,265	1,634
11	1									
12	48	1,459	30	51	186	2.02	0.15	41	4,543	1,869
13	1									
14	127	8,744	69	64	161	2.35	0.50	41	3,041	1,357
15	26	2,776	107	90	841	3.50	1.35	55	8,931	5,438
16	60	6,273	105	82	564	2.47	0.73	59	4,477	3,262
17	63	95	2	8	233	1.67	0.05	64	2,676	1,116
18	78	0	0	3	259	3.08	1.22	38	3,039	1,106
19	66	6,554	99	88	558	2.80	0.89	55	5,171	3,419
20	2									
21	3									
22	54	698	13	23	290	1.74	0.43	64	3,306	1,559
23	77	6,892	90	85	398	2.94	0.70	49	4,573	2,757
24	64	6,705	105	87	649	2.95	0.94	57	5,535	3,220
25	23	795	35	24	453	2.70	0.70	60	4,691	2,284
26	97	5,353	55	62	161	2.44	0.81	46	2,918	1,548
27	23	2,394	104	86	679	2.74	0.78	62	5,500	4,249
28	58	346	6	13	209	1.83	0.22	38	2,472	863
29	77	96	1	6	226	2.21	0.60	50	2,883	1,171
30	86	4,636	54	59	158	1.50	0.15	53	2,725	1,327
31	59	6,141	104	92	575	2.90	0.54	56	4,699	3,236
32	81	7,925	98	86	222	2.58	0.47	45	3,558	2,022
33	82	1,353	16	31	201	1.73	0.21	46	2,811	1,252
34	78	165	2	8	242	2.73	0.95	52	3,602	1,171
De Laak	***2,187***	***109,920***	***50***	***49***	***283***	***2.32***	***0.51***	***48***	***3,543***	***1,840***

Table 4.13 Descriptives District "De Laak" after K-means clustering (k=34) with weighting 10^2 for X and Y coordinates

Target Group clusters

Another part of the research focused on a target group clustering for energy effective renovation. In that case we focused on the clusters to represent target groups more than spatial contiguous clusters of dwellings. Of course GIS is used to produce a geographical representation of the target groups division in the district “De Laak”. However the maps are not checked for spatial contiguous boundaries.

To come up with a target group clustering the analysis is performed with the 3 components extracted in the PCA and the X and Y coordinate on unit level. The 2187 objects are clustered based on 5 variables with 3 different numbers of clusters as output. A 4-means, 6-means and 8 means clustering is conducted and the descriptives are shown in *Table 4.14, Table 4.15 and Table 4.16*. In *paragraph 4.4 Cluster interpretation* an effort is made to give an interpretation of the output of the CA with the selected number of clusters.

		Total	Average							
Cluster number	Number of dwellings	Saving potential	Saving potential	Dwelling age	WOZ value	Household size	Household Children	Household Oldest Age	Electricity SYU	Gas SYU
		GJ/y	GJ/y	y	k€			y	kWh	m^3
1	926	5,894	6	14	229	1.99	0.36	47	2,895	1,076
2	571	35,016	61	63	172	2.13	0.47	46	2,912	1,570
3	444	45,250	102	86	430	1.85	0.14	57	3,703	2,773
4	246	23,761	97	84	475	4.84	1.83	44	7,156	3,654
De Laak	***2,187***	***109,920***	***50***	***49***	***283***	***2.32***	***0.51***	***48***	***3,543***	***1,840***

Table 4.14 Descriptives District "De Laak" after K-means clustering (k=4) with no weighting for X and Y coordinates

		Total	Average							
Cluster number	Number of dwellings	Saving potential	Saving potential	Dwelling age	WOZ value	Household size	Household Children	Household Oldest Age	Electricity SYU	Gas SYU
		GJ/y	GJ/y	y	k€			y	kWh	m^3
1	210	14,214	68	65	194	4.30	1.59	41	4,489	2,113
2	753	4,927	7	14	214	1.54	0.04	48	2,455	959
3	431	43,626	101	86	430	1.85	0.13	57	3,739	2,764
4	190	19,627	103	88	558	4.62	1.82	44	7,388	3,990
5	175	1,024	6	11	295	3.94	1.74	43	4,980	1,639
6	428	26,503	62	63	171	1.51	0.09	49	2,500	1,452
De Laak	***2,187***	***109,920***	***50***	***49***	***283***	***2.32***	***0.51***	***48***	***3,543***	***1,840***

Table 4.15 Descriptives District "De Laak" after K-means clustering (k=6) with no weighting for X and Y coordinates

		Total	Average							
Cluster number	Number of dwellings	Saving potential	Saving potential	Dwelling age	WOZ value	Household size	Household Children	Household Oldest Age	Electricity SYU	Gas SYU
		GJ/y	GJ/y	y	k€			y	kWh	m^3
1	730	4,401	6	14	207	1.53	0.04	47	2,375	909
2	310	17,032	55	60	175	1.51	0.09	51	2,471	1,418
3	274	28,819	105	89	560	1.81	0.09	63	4,346	3,395
4	195	19,973	102	88	525	4.69	1.82	43	7,227	3,822
5	308	27,227	88	78	198	1.85	0.22	44	2,820	1,688
6	54	1,050	19	21	459	2.69	0.61	58	7,568	2,890
7	146	552	4	9	254	4.05	1.87	40	4,160	1,369
8	170	10,867	64	63	197	4.51	1.75	41	4,495	2,170
De Laak	***2,187***	***109,920***	***50***	***49***	***283***	***2.32***	***0.51***	***48***	***3,543***	***1,840***

Table 4.16 Descriptives District "De Laak" after K-means clustering (k=8) with no weighting for X and Y coordinates

It is typical for cluster analysis that a cluster division is always found. It is important to remember that cluster analysis will always produce a grouping, but these may or may not prove useful for classifying items (Burns & Burns, 2008). In *paragraph 4.4 Cluster interpretation* the implications of the CA are explained further.

		Weighted Average Standard Deviation							
		Saving potential	Dwelling age	WOZ value	Household size	Household Children	Household Oldest Age	Electricity SYU	Gas SYU
		GJ/y	y	k€			y	kWh	m3
5 Clusters	A Priori	38.00	28.82	103	1.43	0.90	16.95	2,171	1,114
4 Clusters	3 components and XY (unit)	16.18	9.29	129	1.16	0.75	16.74	1,978	961
6 Clusters	3 components and XY (unit)	16.27	8.89	121	0.82	0.47	16.44	1,838	930
	Deviation A Priori 5	*-57%*	*-69%*	*18%*	*-43%*	*-48%*	*-3%*	*-15%*	*-17%*
8 Clusters	3 components and XY (unit)	15.19	8.49	103	0.82	0.47	15.79	1,776	830

Table 4.17 Weighted Average Standard Deviation of the objects to the clusters mean per characteristic

4.4 Cluster interpretation

Of the three CA's conducted the 6-means analysis is selected as having the most meaningful output. The visualization of the K-means clustering (k=6) is shown in *Figure 4.7*. In contradiction with the earlier showed visualizations the colors in the legend, i.e. the color of the dwellings on the map, does have an intuitive meaning. In this research the goal was to find target group clusters for application in energy effective renovation. The clusters found in the previous paragraph need to have an interpretation. How can the clusters be characterized, or in others words what do they represent?

Figure 4.7 Visualization of K-means clustering (k=6) of the 3 components and X and Y coordinates no weighting (unit)

The 6 clusters found in the CA are easy to interpret. It was no surprise to find such a distinction between the 6 clusters. At first is noticed that there are three distinct levels in average saving potential. These values range from about 5 GJ/y per dwelling (indicated as a low saving potential, colored with red tints) up to approximately 100 GJ/y a dwelling saving potential (indicated as a high saving potential, colored with green tints). A mediocre average saving potential of about 65 GJ/y a dwelling is found for 638 of the dwellings (indicated as a medium saving potential, colored with yellow tints).

On every level of saving potential a distinct subdivision can be made. This is done with respect to one characteristic in particular, i.e. the size of a household. Every level of saving potential is divided into a light tint and heavy tint of the corresponding colors representing a cluster with mainly small and large households. To illustrate the cluster differences more accurate, the clusters' standard deviation regarding the saving potential of a dwelling and the size of a household are shown in *Table 4.18*.

Cluster description		Saving potential		Household size	
		Mean/ Avarage	Standard Deviation	Mean/ Avarage	Standard Deviation
		GJ/y	GJ/y		
Medium Saving potential	**Large Household**	67.68	24.01	4.3	1.46
Low Saving potential	**Small Household**	6.54	8.64	1.54	0.57
High Saving potential	**Small Household**	101.22	19.39	1.85	0.74
High Saving potential	**Large Household**	103.30	20.88	4.62	1.36
Low Saving potential	**Large Household**	5.85	9.64	3.94	1.13
Medium Saving potential	**Small Household**	61.92	23.43	1.51	0.65
De Laak		***50.26***	***43.00***	***2.32***	***1.46***

Table 4.18 Cluster interpretation with Average and Standard deviation of the 6 distinct target group clusters

5 Conclusion

In the conclusion and discussion section of this report the implications of this study for the research on energy performance in housing submarkets and the usability of target group clustering for the energy effective renovation program for private homeowners are discussed. KENWIB is in the early stages of exploring the field of cluster analysis for use in energy related submarket research. In this chapter the research questions are answered.

Available factors and variables

With the problem statement and research design in mind the goals of this study are evaluated and conclusions are formulated. It is important to realize that the data used for cluster analysis determines the quality of the output. Therefore it is investigated which variables or factors of all dwellings and households in Eindhoven are available for analysis.

This can mainly be answered by making use of the appendices *B Endinet Acquiring Data* and *C Data preparation*. The municipality of Eindhoven and network operator Endinet are the parties that grand access to their databases with data concerning energy use of dwellings and demographic information on households.

For this study the Standard Year Usage (SYU) for gas and electricity of a connection in a dwelling were used. Those figures are statistically compared with the figures used in the research of Brouwers et al. (2010). The current SYU's (figures of the last quarter of 2011) are available and they should represent the current electricity and gas use of a dwelling as accurately as possible. The introduction of smart meters in all dwellings will introduce a new interesting variable in dwelling related energy usage studies.

In the registers of the municipality a lot of information is availabe concerning the composition of the household and the age of the occupants, this information is deduced from the GBA. In the WOZ-database of the municipality the typology and year of constuction of each dwelling are stated. The information on variables describing the size of a dwelling, like surface or volume is not availabe for all dwellings. This is a real setback because the volume of surface combined with the known year of constuction and gas usage forms a reliable measure for energy performance of a dwelling.

Now the energy performance of a dwelling, and therefore the saving potential is determined based on dwelling typology and year of construction using the example dwellings of Agentschap NL. This method is presumed to be less accurate than using empirical data representing actual energy performance for analysis.

Decisions for participation

The goal to explore decisions for participation was not achieved to full satisfaction. Which variables or factors influence the decision for participation of private homeowners cannot be concluded based on this research. Of course the 8 variables used, characterize the dwelling and their occupants but it is not tested whether these variables determine the decision for participation in the EE-renovation program.

Interpretation of extracted components out of the data set

In the principal component analysis three components are extracted. These three components account for 78% of the variance in the data set used for this district. These

components are interpreted as 1) dwelling saving potential, 2) household characteristics and 3) “wealth comes with age”.

New spatial contiguous energy clusters

Based on our findings in *paragraph 4.3.2* where a new spatial contiguous cluster division for district “De Laak” is made. Whether the a priori classification is a well performing division regarding saving potential and homogeneity of clusters should be based on judgments for all different levels in the classification is answered next. For the most specific level (neighborhood clusters) the statistical cluster division of Eindhoven is quite a good representation. Moreover the homogeneity of the formed 34 clusters is significantly higher using the data-driven classification methodology, this conclusion is based on the found weighted average standard deviation for the characteristic saving potential which is 28 percent lower. Using the a priori division on any higher level of aggregation is risky, for the homogeneity is much lower in these divisions than in the division used in this study. There may be districts with highly coherent dwelling types and building periods but this may not be presumed.

The results of the first study support the belief that a data-driven classification method, such as cluster analysis, can lead to a better clustering of dwellings for energy effective renovation up to a certain level. The statistical a priori division of the municipality performed quite well at the lowest level. But a better spatial contiguous division is possible by using the 3 components deduced out of 8 variables and the geographical coordinates weighted by multiplying them with 10^2.

Target group clusters

The second study conducted focused on target groups in which the location of a dwelling was not used as a validation criterion upfront. To come up with a target group division for the district different amounts of clusters were generated as output. The cluster analysis where 6 distinct clusters were found was evaluated and an interpretation was formulated. Three levels of saving potential were split into two categories of household size, i.e. large and small. The insight that 6 target groups do characterize the district can be of use in the marketing campaign for BvB/e. 4 of the 6 clusters represent saving potential and a division into large and small households is made.

It does not come as a surprise that the output of the cluster analysis where target groups are distinguished seems to act upon the first two components characterizing the data set. Because the target groups are not spatially contiguous the homogeneity is much higher than the a priori division in 5 clusters.

6 Discussion

In the discussion section of this report limitations of the study are sketched, see *paragraph 6.1*. This will lead to some recommendations for further research. Moreover some first insights in using CA for energy related housing submarket research could lead to a broader usage of the method on several other specific energy related topics dealt with in KENWIB, focusing on the goal of an energy neutral Eindhoven in 2045. These recommendations are discussed in *paragraph 6.2*.

6.1 Limitations

The limitations of this research will be discussed in three parts. At first the availability of data is discussed. This is followed by remarks about the extent of which results of the district under study can account for the behavior of all districts of Eindhoven when analyzed. The last limitation that is addressed covers the introduction of tangible marketing aspects in the research design.

The source of the data of the different variables which are available really determines the quality and possibility to come up with a new target group division for neighborhoods. Most of the variables included consist of empirical data, i.e. data collected by (semi) direct observations. Only one variable cannot be described as empirical, this is the saving potential of a dwelling. To assess the saving potential of a dwelling the typology and building period are used. The adopted value is therefore an average value to sort identical dwellings build in a specific period of time. The way this is done, is included in *appendix C Data preparation*. It was attempted to come up with a more empirical measure for potential energy savings. But actual figures for all dwellings, e.g. the energy label, or volume or surface are not available. This means we still use the empirical SYU's for electricity and gas and a more raw method for saving potential.

The data collection for this experiment was a time consuming process. Once the data was collected the data had to be prepared for analysis, this process, especially enrichment of the data and searching reliable values for missing data took a lot of time. Therefore only one district is analyzed up until now. Even though the results are promising, no guarantees can be given that this approach will work for all other districts in Eindhoven. The conclusion that the maximum of 6 target group clusters will be enough to characterize all districts in Eindhoven should be seen as a hypothesis and should therefore be tested in further research.

Not all demographic data available is used for analysis. Culture and ethnicity are not included as factors. On the one hand because no direct relationship is suspected and more important because the municipality asked to leave this out of the variables because they had the presumption that it is a sensitive issue in society. Real research on buying behavior and sensitivity for marketing approaches is not included in the study, so the deduced target groups are only a further exploration of districts and characterize them regarding saving potential and household size. The communication agency could use this to adapt their strategy on district level. However the deduced target groups are not bearing target group strategies in mind themselves.

Some of the limitations can be tackled by conducting further research this is discussed in the next paragraph.

6.2 Recommendations

The recommendations are sketched in two directions. At first some recommendations are formulated based on issues indicated in the limitations section. Secondly some interesting further research based on the positive experiences with the use of CA for target group clustering is discussed.

Due to time constraints and the necessary practical output that had to be obtained, some research steps were executed rapidly. It is advisable to rerun the analysis with another measure for saving potential or at least leave it out of the analysis for once to see how clusters are formed then. It is known this does undermine the results. However it would be disputable to take the results of this study as a proof that the used factors are optimal. This may not be concluded before more experience is gained by actual using PCA and CA to divide object into target groups. As said this can be done analyzing different combinations of variables of the same district.

Another recommendation is that an even stronger belief in and further validation of the method can be gained by executing it on several other districts in Eindhoven. All districts have another division of dwelling typologies and construction periods. This will lead to further insights on the usability of housing submarket research for target group clustering. It is expected that conducting the study on other districts will take one-tenth of the time needed than when it was done for the first time. The hypothesis is that all districts can be divided into at most 6 target group clusters.

It is possible an even more specific target group division can be achieved by inclusion of more demographic variables like ethnicity, culture and lifestyle. A problem with this is that those aspects are intangible or of a too sensitive nature to be included. The only way intangible aspects can be measured is by using a survey. This would be a sort of a market survey. However it will not obtain those aspects for every object under study when conducting a CA. Therefore it is concluded that it will be rather difficult to execute this.

Looking at the literature on housing submarkets and marketing research it is advised to devote more research to efforts on these topics. In this study we looked at target group clustering of dwellings and its owners. But the practical potential of obtaining clusters of objects or people based on several characteristics could be versatile. The output which is visualized on a map is a very powerful tool to communicate your message or results. It is suspected that this can be used for all kind of other subjects, e.g. buildings in the commercial and industrial sector, a study on suitability of buildings for appliances of decentralized energy generation etcetera.

A research proposal for a study, which is currently in review, using CA can lead to a different optimal layout of a city. In this layout neighborhoods are not determined by historic development, housing types, functional clustering, or infrastructural developments, but by clustering functions according to energy use profiles, network usage, energy reduction, renewable energy generation or excess energy sharing (de Vries & Schaefer, 2012). The future will tell and it has been an honor to work on this graduation study which maybe is a useful contribution.

7 Acknowledgements

During this research collaboration of several parties took place. I really would like to thank all parties who supported me during all different phases of my graduation project. Many thanks to Jan Bekkering for the chance to graduate at "HetEnergieBureau". I really appreciate the way you gave me the opportunity to formulate my problem statement and supported me in the acquisition of the data needed.

Talking about the data, I will be the first to admit, a rare occasion of cooperation took place. The pilot project of "Blok voor Blok", in Eindhoven formulated as BvB/e, made it possible to bring parties together. Words of appreciation are for Maartje Essens of the municipality of Eindhoven for her introduction of Kees van der Hoeven of the department BIO. The greatest contribution to this project was done by him, by granting access to data in the WOZ-database and the GBA, his enthusiasm for my interest in this field of research and skepticism for the possible results of this research looking at the short period of time I could spent researching.

Thanks to Wiro Viergever for his early warning not to take the cooperation of other parties for granted and of course for his efforts to introduce me by Joost Toonen and Rick Donders. I would like to thank Rick for his extensive and careful planning of my short internship at Endinet. I really did not mind the repeated warnings to be really careful with the privacy-sensitive data all parties gave me access to. Bart Brouwers, I really liked it you took the time to explain to me how you analyzed your data at Endinet.

There was no place for all parties to be in the footnote of each page of this report. Nevertheless I would like to display them here once

endinet onderdeel van alliander

gemeente Eindhoven

Bringing me to my "colleagues" at "HetEnergieBureau". Jan, Serge, Jeroen, Thijs, Jan, Marjorie (for the use of her screen), Peter and Merian. Special thanks to Gilbert for the extensive checks of the English in some parts of my report, and to Stephan for calling me Pimmetje Panda and his support with imaging. Tanja, thanks for enlisting me on the Christmas card. Last but not least I would like to thank our fraternity-hero Maurice for the many cups of coffee I drank with him, or do I have to say mainly due to him. I would like to share with you one of the most motivating quotes I received during my research: "Als ik dit zo zie: Waarom heb je die 50.000 man niet opgebeld om te vragen of ze mee wilden doen, dat was een stuk minder werk geweest". Thanks Peter.

A word of thanks to prof. Schaefer for his enthusiasm and trust by signing of the confidentiality agreements. A bigger word of thanks to prof. De Vries for trying to put me in the right scientific direction. Dr. Blokhuis, thanks for your inspired mentoring over the last years. I really appreciated the, for me late night, and for you early morning, chats we had. The sound of the Australian birds via skype sounded great. Drs. Van der Waerden I appreciate your unprejudiced way of tutoring students in GIS.

I would like to mention Keep Rowing for giving me the feeling I was still a student. Mel I am already looking forward to see you in Africa in a few weeks. Tom thanks for supporting me when I was working late, which of course occurred more often than me working early in the

morning. Thanks to Dirk Jan for borrowing me the book Andy Field for a few months and Peter for drafting me a few study beers. Thanks to all friends and family, Van Speyk (Joris, Matthijs, Chris and Koops), Thêta in the broad meaning of the word. Of course I need to state all the people I forgot to list and would have appreciated it. Thanks to all readers of this report and word of thanks, I really appreciate the opportunity to write in the I form for more than one page in this thesis.

References

Bates, L.K., 2006. Does Neighborhood Really Matter? Comparing historically defined neighborhood boundaries with housing submarkets. *Journal of Planning Education and Research*, 26, pp.5-17.

Bourassa, S.C., Cantoni, E. & Hoesli, M., 2007. Spatial Dependence, housing submarkets, and house price prediction. *Journal of Real Estate Finance and Economics*, 35(2), pp.143-60.

Bourassa, S.C., Hamelink, F., Hoesli, M. & MacGregor, B.D., 1999. Defining housing submarkets. *Journal of Housing Economics*, (8), pp.160-83.

Bourassa, S.C., Hoesli, M. & Peng, V.S., 2003. Do housing submarkets really matter? *Journal of Housing Economics*, 12(1), pp.12-28.

Brouwers, B., Blokhuis, E.G.J., Putten, E.v.d. & Schaefer, W.F., 2010. Economic consequences of sustainable transition in energy supply systems. *Energy Economics*, in review.

Burns, R.B. & Burns, R.A., 2008. Chapter 23. In *Business Research Methods and Statistics Using SPSS.* London: Sage. pp.552-67.

Cattell, B.R., 1966. The scree test for the number of factors. *Multivariate Behavorial Research*, 1, pp.245-76.

CBS, 2011. *CBS statistics*. [Online] Available at: http://alturl.com/aukmo [Accessed 20 September 2011].

Clapp, J.M. & Wang, Y., 2006. Defining neighborhood boundaries: are census tracts obsolete? *Journal of Urban Economics*, 59, pp.259-84.

de Vries, B. & Schaefer, W.F., 2012. *Dynamic Energy Profiles, Energy production and consumption for different building types over the year (working title)*. Research proposal (in review). TU/e.

Field, A., 2009. *Discovering statistics using SPSS*. 3rd ed. London: Sage.

Gemeente Eindhoven, 2010. *Bevolkingsprognose 2010-2021*. BIO Beleidsinformatie en onderzoek.

HetEnergiebureau BV, Q-Energy BV, Endinet BV and Gemeente Eindhoven, 2011. *Subsidieaanvraag: Blok voor Blok: BvB/e*. Eindhoven.

Huberty, C.J., Jordan, E.M. & Brandt, W.C., 2005. Cluster Analysis in Higher Education Research. In Smart, J.C. *Higher Education: Handbook of Theory and Research*. Springer. pp.437-57.

Hutcheson, G. & Sofronniou, N., 1999. *The multivariate social scientist*. London: Sage.

Kadaster, 2003. *RD-brochure*. [Online] Available at: https://rdinfo.kadaster.nl/pdf/rd_brochure.pdf [Accessed 21 February 2012].

Kaiser, H.F., 1960. The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20, pp.141-51.

Kaiser, H.F., 1970. A second-generation little jiffy. *Psychometrika*, 35, pp.401-15.

Kaiser, H.F., 1974. An index of factorial simplicity. *Psychometrika*, 39, pp.31-36.

Kuo, R.J., Hob, L.M. & Huc, C.M., 2002. Cluster analysis in industrial market segmentation through artificial neural network. *Computer & Industrial Engineering*, (42), pp.391-99.

Marczinski, M.J.A.M., 2011. *Energy control in the dwelling market*. Graduation report. TU/e.

McDonough, W. & Braungart, M., 2002. *Cradle To Cradle / Remaking the Way We Make Things*. 1st ed. New York: North Point Press.

Ministry of Infrastructure and Environment, 2011. *Inventarisatie kennisvragen rondom Klimaatneutrale Steden*.

Motivaction, 2011. *Kansrijke aanpakken in gebouwgebonden energiebesparing, de particuliere eigenaar*. Agentschap NL.

Municipality of Eindhoven, 2008. *Uitvoeringsprogramma klimaatbeleid 2009-2012: Van succesvolle projecten naar structurele uitvoering*.

Municipality of Eindhoven, 2010. *Clusterindeling per buurt, indeling 2010, Kaartenboek*. [Online] Control BiO/Stadsontwikkeling Available at: http://alturl.com/248vz [Accessed 23 November 2011].

Municipality of Eindhoven, 2011. *Indeling in wijk en buurt*. [Online] Control BiO/Stadsontwikkeling Available at: http://alturl.com/xde3c [Accessed 22 Febraury 2012].

Nieuwenhuijsen, I., 2010. *Urging residents in Eindhoven to save energy*. Graduation report. TU/e.

Norušis, M.J., 2011. Chapter 17. In *IBM SPSS Statistics 19 Statistical Procedures Companion*. Pearson Higher Education. pp.375-404.

Pedhazur, E. & Schmelkin, L., 1991. *Measurement, design and analysis: an integrated approach*. Hillsdale, NJ: Erlbaum.

PRC Bouwcentrum & W/E Adviseurs, 2006. *Kern Publicatie WoON Energie 2006*. VROM/WWI.

PRC Bouwcentrum & W/E adviseurs, 2011. *Voorbeeldwoningen 2011, bestaande bouw*. VROM/WWI.

Punj, G. & Steward, D.W., 1983. Cluster analysis in marketing research: review and suggestions for applications. *Journal of Marketing Research*, (20), pp.134-48.

Tryfos, P., 1998. Chapter 15. In *Methods for Business Analysis and Forecasting: Text & Cases*. Wiley.

van der Weerdt, D., 2011. *Renovate or new estate?* Graduation report. TU/e.

van Duijn, C. & Stoeldraijer, L., 2011. *Huishoudensprognose 2011-2060: meer en kleinere huishoudens*. CBS.

Wesselink, L.G., Eerens, H. & Vis, J., 2008. EU 2020 climate target: 20% reduction requires five-fold increase in impact of CO2 policies., 2008.

Wu, C. & Rashi Sharma, R., 2011. Housing submarket classification: The role of spatial contiguity. *Applied geography*, 32, pp.746-56.

A Management Summary Project plan BvB/e

Slimme samenwerking voor energiezuinigere woningvoorraad

Energiebesparing, burgers aanspreken op hun energiegedrag en verlaging van woonlasten staan hoog op de agenda van het huidige kabinet. De ambitie van de gemeente Eindhoven om in 2040 een energieneutrale gemeente te zijn, sluit daar naadloos op aan. Essentieel daarbij is energiebesparing in de bestaande woningvoorraad. Deze besparing kan op drie fronten worden gerealiseerd, volgens de Trias Energetica-aanpak: door de vraag naar energie te verminderen, door in te zetten op duurzame energie en door zuiniger gebruik te maken van brandstoffen. Om op grote schaal particuliere woningbezitters ertoe te brengen energiebesparende maatregelen aan hun woning te treffen, heeft een aantal partijen zich aaneengesloten tot een consortium dat de samenwerking tussen burgers en bedrijfsleven organiseert: Buurt voor Buurt Eindhoven.

De kracht van Brainport

Eindhoven is de kern van Brainport regio Eindhoven, onlangs door het Intelligent Community Forum uitgeroepen tot slimste regio van de wereld. De regio ontwikkelt vernieuwende, hoogwaardige technologieën en samenwerkingsvormen op basis van open innovatiesamenwerking tussen kennisintensieve maakindustrie, onderzoeks- en onderwijsinstellingen en overheid. Deze triple helix biedt op vele fronten interessante kansen. Ook als het gaat om energiebesparing. Buurt voor Buurt Eindhoven zet het Brainport-netwerk in om bedrijfsleven en woningeigenaren bij elkaar te brengen.

Geïntegreerde inzet van gezamenlijke expertise

In Buurt voor Buurt Eindhoven werken lokale stakeholders op het gebied van energiebesparing en gebouwde omgeving samen: gemeente Eindhoven, Netwerkbedrijf Endinet, Q-Energy en HetEnergieBureau. Zij worden ondersteund door kennispartners met specifieke energetische, financiële en marketingcommunicatie deskundigheid: Rabobank Eindhoven-Veldhoven, Technische Universiteit Eindhoven - Kenniscluster energieneutraal Wonen in Brainport, Provincie Noord-Brabant, Samenwerkingsverband Regio Eindhoven, SVN en communicatiebureau Gleijm & Van der Waart. De geïntegreerde inzet van de gezamenlijke expertise gaat leiden tot verbetering van de bestaande woningvoorraad (minimaal 2.000 particuliere woningen in 2014). Dat resulteert niet alleen in reductie van het energieverbruik, maar ook in lagere woonlasten, meer wooncomfort en een positieve gedragsverandering ten aanzien van het omgaan met energie. Bovendien wordt hiermee een belangrijke bijdrage geleverd aan het waardebehoud van energiezuinigere woningen.

Aanbod per buurt of levensstijl

Particuliere woningeigenaren vormen een lastig te bereiken en te overtuigen doelgroep. Het is doorgaans erg moeilijk om hun woongedrag op individueel niveau in positieve zin te beïnvloeden. Daarom kiest Buurt voor Buurt Eindhoven voor een andere, collectieve aanpak. Een marketinggerichte aanpak waarbij woningeigenaren per buurt (met vergelijkbare woningen) en/of per leefstijlgroep (met vergelijkbare belangen en interesses) enthousiast worden. Ze krijgen overzichtelijke, hoogwaardige pakketten aangeboden met betrekking tot energetische verbetering van hun woning, inclusief financiële arrangementen daarvoor. Het gaat hierbij om drie energiemaatregelpakketten, die inzetten op vermindering van de energievraag, toepassing van duurzame energie en reductie van fossiel brandstofverbruik.

Advies, techniek en financiering

De pakketten bestaan uit een mix van maatwerk advies, isolerende maatregelen, slimme installaties en een intelligente energiemeter die de bewoner via gerichte feedback ondersteunt bij energiezuinig woongedrag. Ook voorstellen voor de financiering maken deel uit van elk pakket. Buurt voor Buurt Eindhoven biedt de deelnemers drie verschillende financieringsarrangementen: zelf financieren, veilig en aantrekkelijk lenen via de bestaande gemeentelijke energielening of het afsluiten van een contract met een nog te werven partij die (tegen servicekosten) de investering en de energiekosten gedurende enkele jaren overneemt. Alle arrangementen kennen een energielastengarantie. Bewoners maken zelf hun keuzes. Ze worden geheel ontzorgd, maar houden de regie volledig in eigen hand.

Projectorganisatie

Om de doelstelling van minimaal 2.000 energiezuinigere woningen in 2014 te bereiken, zullen de partners in Buurt voor Buurt Eindhoven een onafhankelijke logistieke eenheid oprichten die vraag en aanbod op het gebied van energiebesparing bij elkaar brengt. Deze eenheid opereert zelfstandig, als een juridische entiteit. De doelbuurten en doelgroepen in de stad worden volgens een gerichte marketingstrategie benaderd. De daadwerkelijke uitvoering van de energiemaatregelen en gedragsondersteuning zal in handen zijn van (combinaties van) bedrijven uit de regio Eindhoven: een stevige stimulans voor de locale bouwsector.

De consortiumleden van Buurt voor Buurt Eindhoven vaardigen een bestuurder af om plaats te nemen in een Raad van Toezicht. Daarnaast is er een Comité van Aanbeveling, waarin zowel de consortiumleden als kennispartners plaatsnemen. De kennispartners zullen ook actief worden betrokken bij kennisdeling en opschaling.

Transparant en onafhankelijk

Buurt voor Buurt Eindhoven voert voorbereiding, toezicht, marketingcommunicatie, inkoop, aanbesteding en kwaliteitscontrole op volstrekt transparante en onafhankelijke wijze uit. Het gehele traject wordt opgedeeld in drie fasen:

- Fase A: het opstellen en indienen van de projectaanvraag;
- Fase B: nadere uitwerking van de pilot (eind 2011 en Q1 2012);
- Fasen C en D: uitvoering van maatregelen aan minimaal 2.000 woningen (vanaf Q2 2012 tot 2014).

B Endinet Acquiring Data

Context

For this graduation project consumer data is investigated and finally used for analysis. The data was available and checked at the office of the network operator Endinet in Eindhoven. Endinet takes care of the electricity connections and network of Eindhoven and the gas connections and network area of south east Brabant. Since 2010 Endinet is a full affiliate of Alliander. For privacy concerns the used data was made anonymous or only presented on block level when.

In 2009 Bart Brouwers conducted an internship at Endinet. A part of his internship consisted of the calculation and extrapolation of billing information out of the SAP client information database. Out of this billing data the representative year usages for electricity and gas of every connection were determined. With these “Representative figures” the total electricity and gas usage of Eindhoven was calculated. Nowadays Endinet has another option to deliver figures of standard year usage of its connections. The client database contains a variable “Standard year usage” which is likely to represent the same information. To assess the reliability of the figures the methods of determination are compared at first. Secondly the correlation of the emerged figures for the connections was calculated.

Comparison of methods

In 2009 SAP already contained “Standard Year Usage” figures of every connection. At that time it was assumed to be inaccurate and not a good representation of every connection. Therefore a more substantial method was chosen. Without being too comprehensive in elaborating the method Bart Brouwers designed it is addressed briefly. For e-connections the method was checked on its reliability. Due to time concerns the values found for e-connections were not compared with the “Standard Year Usage” of every EAN code in SAP, it was assumed the calculated values were right because they were supported by electricity peak loads measured in transformer stations. On representative figures for gas connections such a check was not executed, the determined figures are not supported by a description of the method used.

Bart Brouwers electricity connections

For the 2004-2009 period billing data was used. For yearly invoiced connections four different situations were found to exist. For monthly invoiced connections only active an inactive are distinguished.

total connections E		107,180
subtotal connections E	annually invoiced	105,652
inactive		1,812
active	5 entire years 2004-2009	60,935
	1 entire year 2004-2009	36,802
	rest	6,103
subtotal connections E	monthly invoiced	1,528
inactive		60
active		1,468
total active connections E		**105,308**

For 60,935 connections the method is very accurate because it is the average of 5 years. For the category where only 1 year is known, 36,802 connections it is disputable whether it represents the current usage well. This is the case with the 6,103 rest connections too. For the monthly invoiced connections it is hard to assess the accuracy because there is nothing known on how many months the figures are based on.

Bart Brouwers gas connections

Nothing is known on the assessment of representative gas usage of the connections of Eindhoven. This is a real problem because gas is mainly used for heating of a dwelling. The figures Bart Brouwers found were not validated because he did not need to use them in his research.

Current "Standard Year Usage" SAP

The current SYU of a dwelling for electricity and gas is, according to sources inside Endinet, determined the same way Bart Brouwers did in his research. The SYU in SAP is not re-determined at the same moment. All connections have a date of revision on which the SYU is determined yet again.

Unique Key

The EAN code is a reliable and unique 18 digit code to identify a connection. Luckily the code is a variable in both files in the comparison. The files are merged using the EAN code as identifier. It was assumed that both files contained all the active connections of the Endinet region with the postal code 5600 up to and including 5658. Both files were imported in Excel 2007 and merged using the last 12 digits of the EAN code. This was done because Excel does not handle numbers with 18 digits the way it should.

Total and mean	2008	2011	% Deviance
Electricity (kWh)	1.042.374.845	1.042.998.846	0,06
number of connections	105.308	105.236	
Mean (kWh)	9.898	9.911	
Gas (m^3)	272.743.500	243.369.625	-12,07
number of connections	96.451	93.318	
Mean (m^3)	2.828	2.608	

Both files amount approximately 95.000 gas and 105.000 electricity connections. After confrontation 1530 electricity and 1703 gas records in the file of "Energie in Beeld" are not present in data Bart Brouwers used. It is assumed the amount of electricity connections should have been increased over time and for gas it is not likely the amount of gas connections dropped with more than 3.000. It seems some large gas connections are missing in the current files, because the overall mean of the connections in the file of Bart Brouwers after confrontation is significant lower.

Missing values in files

There are changes in the amount of connections of Eindhoven. Over time new connections are added and other existing connections disappear. To assess the variance and correlation of the variables electricity and gas usage of new and closed connections are left out of the analysis. The current "standard annual usage" file is assumed to contain all connections in Eindhoven. Therefore all connections which became inactive in recent years are not present in the comparison file. The total amount and mean per connection of electricity and gas use of Eindhoven in 2008 is calculated in the original files.

Standard variation and Correlation

In a quickscan it becomes obvious the EAN-code is the correct identifier for the data and values are somewhat alike. When values differ reasons can be twofold. Either the usage really changed in recent years, or it is caused by the different methods of calculation. The standard deviation and correlation is calculated with the use of IBM SPSS 19. To measure the correlation it is assumed the data is normally distributed, it is not tested whether this is true or not. Therefore standard deviations, variance and correlation may not be taken too strict and should be considered as a rough guide for comparison. In statistics a correlation coefficient of 1 corresponds with a perfect positive relationship. Values above .5 are considered as high effects when looking at the correlation of variables. However if the usage of connections did not change on individual level and both methods obtain the same pattern over all connections the correlation will be 1. It is hard to conclude on the correlation coefficient solely whether both methods are accurate.

Electricity

The correlation coefficient for electricity connections is .953 which is quite high.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
SJV_TOTAAL	104785	0	88209795	9953.70	342806.230	1.175E11
SJV_2008_BROUWERS	103705	.0000	53728006.204	9240.956807	262177.365244	6.874E10
Valid N (listwise)	103284					

Correlations

		SJV_TOTAAL	SJV_2008_BROUWERS
SJV_TOTAAL	Pearson Correlation	1	.953**
	Sig. (2-tailed)		.000
	N	104785	103284
SJV_2008_BROUWERS	Pearson Correlation	.953**	1
	Sig. (2-tailed)	.000	
	N	103284	103705

**. Correlation is significant at the 0.01 level (2-tailed).

Gas

The correlation coefficient for gas connections is .716 which indicates both methods have quite a high correlation. However for conclusions that implicate methods both measures the same standard year usage on gas it is for sure not high enough. Correlation between data is suspected, however nothing can be said about reliability.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
SJV_TOTAAL	92860	1	7959618	2620.82	42441.331	1.801E9
SJV_2008_BROUWERS	91616	.0000	580596.0000	2254.957815	11074.2412161	1.226E8
Valid N (listwise)	91201					

Correlations

		SJV_TOTAAL	SJV_2008_BROUWERS
SJV_TOTAAL	Pearson Correlation	1	.716**
	Sig. (2-tailed)		.000
	N	92860	91201
SJV_2008_BROUWERS	Pearson Correlation	.716**	1
	Sig. (2-tailed)	.000	
	N	91201	91616

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

		SJV_TOTAAL	SJV_2008_BROUWERS
SJV_TOTAAL	Pearson Correlation	1	.824**
	Sig. (2-tailed)		.000
	N	92389	90822
SJV_2008_BROUWERS	Pearson Correlation	.824**	1
	Sig. (2-tailed)	.000	
	N	90822	91228

**. Correlation is significant at the 0.01 level (2-tailed).

The second correlation table is filtered on private gas connections, indicated with a P in current file of “Energie in Beeld”. A substantial part of the lower correlation is hidden in the business gas connections, it is hard to point the source of this deviance.

Conclusion

No conclusions can be drawn based the statistical comparison of both methods of determination. For now we rely on the judgment of Endinet of how reliable the data regarding SYU’s in their SAP database really is. (1) Based on the statement that the method of determination of Bart Brouwers figures and the SYU’s in SAP are the same. And (2) a strict revision schedule is used by the customer department of Endinet to revise the figures. It is concluded that the current SYU’s of Endinet are most reliable to use in research were usage figures of connections are of importance.

C Data preparation

The variables in the files are presented in table 1. Every row represents one dwelling and the original file delivered by the municipality of Eindhoven contained 2212 records. Further on in this document the records with missing data are listed. 25 objects are deleted, because too much data was missing. For 15 records we were able to retrieve some information due to comparison with similar objects. Some dwellings situated in De Laak are not present in the file, those dwellings did not have occupants on the first of January 2011. With the scope of this research in mind this is allowed because dwellings without occupants cannot compete in a program for energy efficient renovation. This leaves us a file with 2187 records useful for PCA.

Raw-data set combined and with supplements	Data set with address information, typology and ownership	Enclosed in data set used for PCA
identifier_adress_formula		
identifier_adress	Identifier/Key	TRUE
Xco	X-coordinate	TRUE
Yco	Y-coordinate	TRUE
street	Street name	
hnr	House number	
hlt	House number character	
htv	House number supplement	
postalcode	Postal code	
cluster_code_apriori	Cluster code “a priori”	TRUE
year_of_construction		
dwelling_age	Dwelling age	TRUE (variable)
squared_dwelling_age	Squared dwelling age	TRUE (variable)
WOZ_type		
type		
combined_type	Typology	
combined_type_code		
combined_year_type_code		
savingpotential_formula		
savingpotential	Saving potential	TRUE (variable)
ev	Ownership	TRUE
WOZ_ev		
WOZ_corp1		
WOZ_corp	Ownership corporations	TRUE
WOZ_value	WOZ value	TRUE (variable)
household_size	Number of persons in household	TRUE (variable)
household_children	Number of children in household	TRUE (variable)
age_oldest	Age of oldest person in household	TRUE (variable)
elek_Bart_Brouwers		
gas_Bart_Brouwers		
elek_2011_aangevuld_formule		
gas_2011_aangevuld_formule		
elek_2011_aangevuld	Electricity standard year usage	TRUE (variable)
gas_2011_aangevuld	Gas standard year usage	TRUE (variable)

The variable year of construction contains categorical variables. About 50 cells have values with building periods. The dwellings are mainly of a somewhat older age and it is likely the exact year of construction is not known. Therefore is chosen to select the upper boundary of the category as year of construction. For use in PCA the variable should be at least of interval level and if possible of ratio level. This means the ratio of the data on the scale must make sense, e.g. 20 years is twice as old as 10 years.

Eh	normaal, 2^1 kap, standaard/algemeen
Et	normaal, 2^1 kap, standaard/algemeen
Eh	praktijkwoning, 2^1 kap, standaard/alg
Eh	bedrijfswoning, 2^1 kap, standaard/alg

Table 2

Ev	Detached single family dwelling	"Vrijstaande woning"	1,600
Eh	Terraced corner single family dwelling	"Rijwoning eind"	1,000
Et	Terraced in between single family dwelling	"Rijwoning tussen"	9,000
Eg	Attached single family dwelling	"Rijwoning eind"	1,000
Bh	Ground floor attached single family dwelling	"Maisonnette hoekwoning onder het dak op begane grond"	8,00
2o1	Semi-detached single family dwelling	"2 onder 1 kap woning"	1,200
M	Multi-family dwelling	"Flatwoning gemiddeld"	600

Table 3

In table 3 the different type of dwellings are listed. In the third column the chosen corresponding sub-type of the "Example dwellings 2011"of AgentschapNL is added. These sub-types are used to determine the saving potential of a dwelling knowing the year of construction. For every type en building period the average current energy use and potential energy use after energy efficient renovation is known. This way it is possible to converse a typology and year of construction of a dwelling into the interval variable "saving potential". In table 4 the figures for every used subtype are given.

	Construction period	Saving potential (MJ/y)	Current standard gas usage (m^3/y)
"Vrijstaande woning"	<1964	129,749	4,731
	1965-1974	100,745	4,110
	1975-1991	41,824	2,616
	1992-2005	7,476	1,882
	2006-now	0	1,600
"Rijwoning eind"	<1945	125,546	4,274
	1945-1964	77,755	2,948
	1965-1974	64,405	2,707
	1975-1991	26,134	1,740
	1992-2005	2,408	1,186
	2006-now	0	1,000
"Rijwoning tussen"	<1945	89,662	3,337

	1945-1964	52,127	2,246
	1965-1974	39,606	2,030
	1975-1991	20,510	1,542
	1992-2005	2,406	1,135
	2006-now	0	900
“Maisonette”	<1964	98,754	3,396
	1965-1974	45,361	2,044
	1975-1991	16,825	1,324
	1992-2005	1,415	896
	2006-now	0	800
“2 onder 1 kap woning”	<1964	90,699	3,453
	1965-1974	72,891	3,046
	1975-1991	27,584	1,915
	1992-2005	8,319	1,497
	2006-now	0	1,200
“Flatwoning gemiddeld”	<1964	36,602	1,512
	1965-1974	28,322	1,570
	1975-1991	15,261	1,004
	1992-2005	922	724
	2006-now	0	600

Table 4

Of 122 dwellings the representative gas usage is not known. For those records is decided to replace the blanks with a value using the representative current gas usage of dwellings acoording to the “Example dwellings 2011” of AgentschapNL. For dwellings with values mentioned in column 4 of table 4 are used. For the values in the 2005-now period raw extrapolations of the representative year usage of the same type dwelling in the 1991-2005 period.

Of 56 dwellings the representative year electricity usage is not known. For those records is decided to replace the blanks with a value using the following equation. This way an average household has a representative electricity usage of 3.300 kWh a year.

$$E_{repr.elec.} = 1100 + 1000 * n_{persons\ in\ household} \quad \text{(kWh)}$$

The different categories of ownership are listed below in table 5. The code with value 9 and 99 is used after confrontation of two databases. The value of 9 represents houses in the private rental sector. The value of 99 represent real private dwelling occupied by its owner. 0, 1, 2, 3 represent houses in possession of a corporation.

0	Woningbouwcorporatie, onzelfstandig (student)	Vestide	Rental Corporation
1	Woningbouwcorporatie	Woonbedrijf	Rental Corporation
2	Woningbouwcorporatie	Trudo	Rental Corporation
3	Woningbouwcorporatie, (meergezins)?	Wooninc.	Rental Corporation
9	Particulier		Private Rental
99	Particulier		Private

Table 6

In table 6 the key/identifier of deleted objects are listed. The missing values are given as reason for deletion.

Key/Identifier	Missing Value	Comment
5613BC 97	WOZ-value, typology	Elderly care house
5613DP 150	WOZ-value, typology	
5613EL 1 F	WOZ-value, typology	Assisted living house
5613DA 29	WOZ-value, typology	
5613DA 31 A	WOZ-value, electricity usage	Combined with 5613DA 31
5613SC 24 A	WOZ-value, typology, energy usage	
5613SC 38 A	WOZ-value, typology, energy usage	
5613SC 48 A	WOZ-value, typology, energy usage	
5613SC 60 A	WOZ-value, typology, energy usage	
5613SC 72 A	WOZ-value, typology, energy usage	
5613SC 82 A	WOZ-value, typology, energy usage	
5613SC 102 A	WOZ-value, typology, energy usage	
5613SC 104 A	WOZ-value, typology, energy usage	
5613DA 47	WOZ-value, typology, energy usage	
5613 DB 81	WOZ-value, owner type	“Apollo huis”
5613 DT 226	WOZ-value, typology, energy usage	
5613 EM 10	WOZ-value, typology	Assisted living house
5613HL 22	WOZ-value, typology, energy usage	
5613BC 93	WOZ-value, typology	
5613 KB 2A	WOZ-value, typology	Combined with 5613KB 2
5613EB 26	WOZ-value, typology, electricity usage	
5613EA 27 A	WOZ-value, typology, energy usage	
5613GP 4	WOZ-value, typology, energy usage	
5613 DW 5	WOZ-value, typology	
5613JA 15 F	Coordinates, energy usage	

Table 6

In table 7 the key/identifier of deleted objects are listed. The missing values are given and in the third column the entered values are listed.

Key/Identifier	Missing Value	Values entered
5613 DC 87	WOZ-value, owner type	159,000; 1
5613 DC 115	WOZ-value, owner type	150,000; 1
5613 DC 117	WOZ-value, owner type	150,000; 1
5613 DD 149	WOZ-value, owner type	273,000; 2
5613 DD 144	WOZ-value, owner type	366,000; 1
5612 DT 230	WOZ-value, owner type	225,000; 2
5613 DV 258	WOZ-value, owner type	150,000; 1
5613 DV 262	WOZ-value, owner type	150,000; 1
5613 DV 270	WOZ-value, owner type	150,000; 1
5613 DV 282	WOZ-value, owner type	150,000; 1
5613AA 51	WOZ-value, owner type	200,000; 1
5613HK 17	WOZ-value, owner type	150,000; 1
5613JB 57	WOZ-value, owner type	264,000; 1
5613SG 244	WOZ-value, owner type	295,000; 2
5613GJ 6	WOZ-value, owner type	174,000; 2

Table 7

D Descriptives outliers

Dwelling_age outlier descriptives

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Absolute z-score less than 2	2187	100.0	100.0	100.0
	Absolute z-score greater than 3.29	0	0	0	100.0
Total		2187	100.0		

Saving_potential outlier descriptives

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Absolute z-score less than 2	2187	99.9	100.0	100.0
	Absolute z-score greater than 3.29	0	0	0	100.0
Total		2187	100.0		

WOZ_value outlier descriptives

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Absolute z-score less than 2	2083	95.2	95.2	95.2
	Absolute z-score greater than 1.96	54	2.5	2.5	97.7
	Absolute z-score greater than 2.58	23	1.1	1.1	98.8
	Absolute z-score greater than 3.29	27	1.2	1.2	100.0
	Total	2187	100.0	100.0	

Household_size outlier descriptives

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Absolute z-score less than 2	2102	96.1	96.1	96.1
	Absolute z-score greater than 1.96	46	2.1	2.1	98.2
	Absolute z-score greater than 2.58	16	.7	.7	98.9
	Absolute z-score greater than 3.29	23	1.1	1.1	100.0
	Total	2187	100.0	100.0	

Household_children outlier descriptives

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Absolute z-score less than 2	2071	94.7	94.7	94.7
	Absolute z-score greater than 2.58	90	4.1	4.1	98.8
	Absolute z-score greater than 3.29	26	1.2	1.2	100.0
	Absolute z-score greater than 3.29	0	0	0	100.0
	Total	2187	100.0	100.0	

Age_oldest outlier descriptives

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Absolute z-score less than 2	2134	97.6	97.6	97.6
	Absolute z-score greater than 1.96	49	2.2	2.2	99.8
	Absolute z-score greater than 2.58	4	.2	.2	100.0
	Absolute z-score greater than 3.29	0	0	0	100.0
	Total	2187	100.0	100.0	

Electricity_SYU outlier descriptives

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Absolute z-score less than 2	2127	97.3	97.3	97.3
	Absolute z-score greater than 1.96	21	1.0	1.0	98.2
	Absolute z-score greater than 2.58	14	.6	.6	98.9
	Absolute z-score greater than 3.29	25	1.1	1.1	100.0
	Total	2187	100.0	100.0	

Gas_SYU outlier descriptives

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Absolute z-score less than 2	2105	96.3	96.3	96.3
	Absolute z-score greater than 1.96	38	1.7	1.7	98.0
	Absolute z-score greater than 2.58	26	1.2	1.2	99.2
	Absolute z-score greater than 3.29	18	.8	.8	100.0
	Total	2187	100.0	100.0	

E Tables Output PCA

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
dwelling_age	49.15	33.001	2187
squared_dwelling_age	3504.07	3205.313	2187
saving_potential	50260.82	43995.585	2187
WOZ_value	279357.36	181862.509	2187
household_size	2.30	1.455	2187
household_children	.50	.922	2187
age_oldest	48.35	17.432	2187
electricity_SYU	3428.41	2330.893	2187
gas_SYU	1813.18	1324.192	2187

Correlation Matrix[a]

		dwelling_age	squared_dwelling_age	saving_potential	WOZ_value	household_size	household_children	age_oldest	electricity_SYU	gas_SYU
Correlation	dwelling_age	1.000	.973	.908	.361	.177	.125	.097	.264	.543
	squared_dwelling_age	.973	1.000	.887	.447	.190	.136	.113	.294	.565
	saving_potential	.908	.887	1.000	.442	.186	.147	.146	.275	.570
	WOZ_value	.361	.447	.442	1.000	.200	.195	.279	.448	.656
	household_size	.177	.190	.186	.200	1.000	.697	-.156	.456	.356
	household_children	.125	.136	.147	.195	.697	1.000	-.115	.288	.200
	age_oldest	.097	.113	.146	.279	-.156	-.115	1.000	.043	.179
	electricity_SYU	.264	.294	.275	.448	.456	.288	.043	1.000	.582
	gas_SYU	.543	.565	.570	.656	.356	.200	.179	.582	1.000
Sig. (1-tailed)	dwelling_age		.000	.000	.000	.000	.000	.000	.000	.000
	squared_dwelling_age	.000		.000	.000	.000	.000	.000	.000	.000
	saving_potential	.000	.000		.000	.000	.000	.000	.000	.000
	WOZ_value	.000	.000	.000		.000	.000	.000	.000	.000
	household_size	.000	.000	.000	.000		.000	.000	.000	.000
	household_children	.000	.000	.000	.000	.000		.000	.000	.000
	age_oldest	.000	.000	.000	.000	.000	.000		.021	.000
	electricity_SYU	.000	.000	.000	.000	.000	.000	.021		.000
	gas_SYU	.000	.000	.000	.000	.000	.000	.000	.000	

a. Determinant = .001

Inverse of Correlation Matrix

	dwelling_age	squared_dwelling_age	saving_potential	WOZ_value	household_size	household_children	age_oldest	electricity_SYU	gas_SYU
dwelling_age	28.624	-23.097	-6.375	3.610	.334	-.154	.004	.050	-1.341
squared_dwelling_age	-23.097	23.509	1.062	-3.124	-.302	.222	.164	-.112	.798
saving_potential	-6.375	1.062	6.360	-.805	-.001	-.152	-.177	.202	-.292
WOZ_value	3.610	-3.124	-.805	2.415	.338	-.341	-.324	-.260	-1.162
household_size	.334	-.302	-.001	.338	2.427	-1.472	.236	-.532	-.535
household_children	-.154	.222	-.152	-.341	-1.472	2.024	.045	.053	.349
age_oldest	.004	.164	-.177	-.324	.236	.045	1.157	.032	-.101
electricity_SYU	.050	-.112	.202	-.260	-.532	.053	.032	1.742	-.749
gas_SYU	-1.341	.798	-.292	-1.162	-.535	.349	-.101	-.749	2.780

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.726
Bartlett's Test of Sphericity	Approx. Chi-Square	16234.212
	df	36
	Sig.	.000

Anti-image Matrices

		dwelling_age	squared_dwelling_age	saving_potential	WOZ_value	household_size	household_children	age_oldest	electricity_SYU	gas_SYU
Anti-image Covariance	dwelling_age	.035	-.034	-.035	.052	.005	-.003	.000	.001	-.017
	squared_dwelling_age	-.034	.043	.007	-.055	-.005	.005	.006	-.003	.012
	saving_potential	-.035	.007	.157	-.052	-7.429E-5	-.012	-.024	.018	-.016
	WOZ_value	.052	-.055	-.052	.414	.058	-.070	-.116	-.062	-.173
	household_size	.005	-.005	-7.429E-5	.058	.412	-.300	.084	-.126	-.079
	household_children	-.003	.005	-.012	-.070	-.300	.494	.019	.015	.062
	age_oldest	.000	.006	-.024	-.116	.084	.019	.865	.016	-.031
	electricity_SYU	.001	-.003	.018	-.062	-.126	.015	.016	.574	-.155
	gas_SYU	-.017	.012	-.016	-.173	-.079	.062	-.031	-.155	.360
Anti-image Correlation	dwelling_age	.654[a]	-.890	-.473	.434	.040	-.020	.001	.007	-.150
	squared_dwelling_age	-.890	.709[a]	.087	-.415	-.040	.032	.032	-.017	.099
	saving_potential	-.473	.087	.888[a]	-.205	.000	-.042	-.065	.061	-.069
	WOZ_value	.434	-.415	-.205	.652[a]	.140	-.154	-.194	-.127	-.449
	household_size	.040	-.040	.000	.140	.625[a]	-.664	.141	-.259	-.206
	household_children	-.020	.032	-.042	-.154	-.664	.593[a]	.030	.028	.147
	age_oldest	.001	.032	-.065	-.194	.141	.030	.742[a]	.023	-.056
	electricity_SYU	.007	-.017	.061	-.127	-.259	.028	.023	.839[a]	-.340
	gas_SYU	-.150	.099	-.069	-.449	-.206	.147	-.056	-.340	.819[a]

a. Measures of Sampling Adequacy(MSA)

Communalities

	Initial	Extraction
dwelling_age	1.000	.968
squared_dwelling_age	1.000	.954
saving_potential	1.000	.908
WOZ_value	1.000	.705
household_size	1.000	.819
household_children	1.000	.707
age_oldest	1.000	.618
electricity_SYU	1.000	.627
gas_SYU	1.000	.751

Anti-image Matrices

		dwelling_age	squared_dwelling_age	saving_potential	WOZ_value	household_size	household_children	age_oldest	electricity_SYU	gas_SYU
Anti-image Covariance	dwelling_age	.035	-.034	-.035	.052	.005	-.003	.000	.001	-.017
	squared_dwelling_age	-.034	.043	.007	-.055	-.005	.005	.006	-.003	.012
	saving_potential	-.035	.007	.157	-.052	-7.429E-5	-.012	-.024	.018	-.016
	WOZ_value	.052	-.055	-.052	.414	.058	-.070	-.116	-.062	-.173
	household_size	.005	-.005	-7.429E-5	.058	.412	-.300	.084	-.126	-.079
	household_children	-.003	.005	-.012	-.070	-.300	.494	.019	.015	.062
	age_oldest	.000	.006	-.024	-.116	.084	.019	.865	.016	-.031
	electricity_SYU	.001	-.003	.018	-.062	-.126	.015	.016	.574	-.155
	gas_SYU	-.017	.012	-.016	-.173	-.079	.062	-.031	-.155	.360
Anti-image Correlation	dwelling_age	.654[a]	-.890	-.473	.434	.040	-.020	.001	.007	-.150
	squared_dwelling_age	-.890	.709[a]	.087	-.415	-.040	.032	.032	-.017	.099
	saving_potential	-.473	.087	.888[a]	-.205	.000	-.042	-.065	.061	-.069
	WOZ_value	.434	-.415	-.205	.652[a]	.140	-.154	-.194	-.127	-.449
	household_size	.040	-.040	.000	.140	.625[a]	-.664	.141	-.259	-.206
	household_children	-.020	.032	-.042	-.154	-.664	.593[a]	.030	.028	.147
	age_oldest	.001	.032	-.065	-.194	.141	.030	.742[a]	.023	-.056
	electricity_SYU	.007	-.017	.061	-.127	-.259	.028	.023	.839[a]	-.340
	gas_SYU	-.150	.099	-.069	-.449	-.206	.147	-.056	-.340	.819[a]

Extraction Method: Principal Component Analysis.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	4.072	45.250	45.250	4.072	45.250	45.250	3.612
2	1.780	19.783	65.032	1.780	19.783	65.032	2.449
3	1.203	13.370	78.402	1.203	13.370	78.402	2.115
4	.759	8.429	86.831				
5	.496	5.507	92.338				
6	.314	3.490	95.828				
7	.231	2.565	98.393				
8	.125	1.387	99.780				
9	.020	.220	100.000				

Extraction Method: Principal Component Analysis.

a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

Component Matrix[a]

	Component		
	1	2	3
dwelling_age	.852	-.343	-.353
squared_dwelling_age	.874	-.324	-.293
saving_potential	.859	-.319	-.259
WOZ_value	.680	.010	.493
household_size	.454	.773	-.120
household_children	.362	.742	-.156
age_oldest	.181	-.374	.668
electricity_SYU	.591	.405	.336
gas_SYU	.814	.062	.291

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

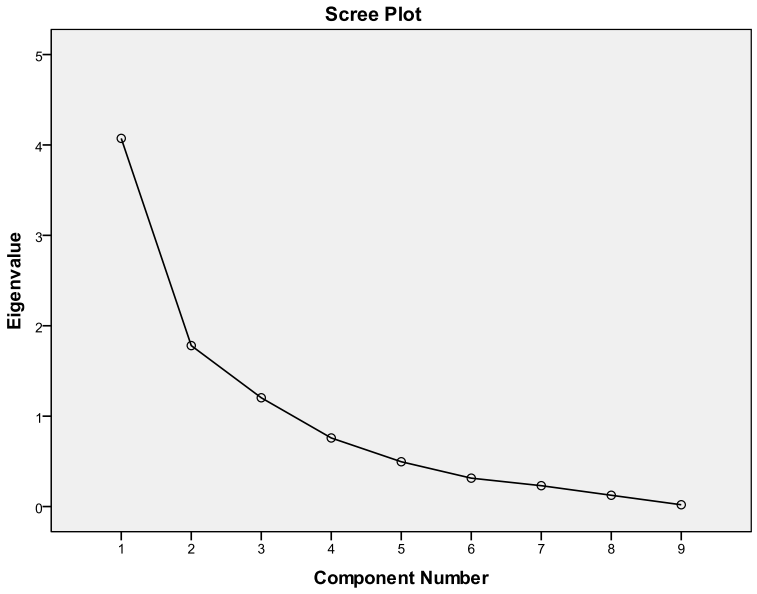
Reproduced Correlations

		dwelling_age	squared_dwelling_age	saving_potential	WOZ_value	household_size	household_children	age_oldest	electricity_SYU	gas_SYU
Reproduced Correlation	dwelling_age	.968[a]	.959	.933	.402	.164	.109	.046	.246	.569
	squared_dwelling_age	.959	.954[a]	.930	.446	.182	.122	.083	.287	.606
	saving_potential	.933	.930	.908[a]	.453	.175	.115	.102	.291	.604
	WOZ_value	.402	.446	.453	.705[a]	.257	.177	.448	.571	.697
	household_size	.164	.182	.175	.257	.819[a]	.757	-.287	.542	.383
	household_children	.109	.122	.115	.177	.757	.707[a]	-.316	.463	.295
	age_oldest	.046	.083	.102	.448	-.287	-.316	.618[a]	.180	.318
	electricity_SYU	.246	.287	.291	.571	.542	.463	.180	.627[a]	.604
	gas_SYU	.569	.606	.604	.697	.383	.295	.318	.604	.751[a]
Residual[b]	dwelling_age		.014	-.025	-.040	.013	.015	.051	.018	-.027
	squared_dwelling_age	.014		-.043	.000	.008	.014	.030	.008	-.041
	saving_potential	-.025	-.043		-.011	.011	.032	.045	-.016	-.035
	WOZ_value	-.040	.000	-.011		-.057	.018	-.169	-.123	-.042
	household_size	.013	.008	.011	-.057		-.060	.130	-.086	-.027
	household_children	.015	.014	.032	.018	-.060		.201	-.175	-.096
	age_oldest	.051	.030	.045	-.169	.130	.201		-.137	-.139
	electricity_SYU	.018	.008	-.016	-.123	-.086	-.175	-.137		-.022
	gas_SYU	-.027	-.041	-.035	-.042	-.027	-.096	-.139	-.022	

Extraction Method: Principal Component Analysis.

a. Reproduced communalities

b. Residuals are computed between observed and reproduced correlations. There are 12 (33.0%) nonredundant residuals with absolute values greater than 0.05.



Pattern Matrix[a]

	Component		
	1	2	3
dwelling_age	1.017	-.057	-.069
squared_dwelling_age	.986	-.032	-.004
saving_potential	.953	-.034	.025
WOZ_value	.191	.217	.697
household_size	-.004	.910	-.062
household_children	-.033	.851	-.123
age_oldest	-.077	-.324	.763
electricity_SYU	.012	.582	.474
gas_SYU	.384	.316	.528

Extraction Method: Principal Component Analysis.

Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 6 iterations.

Structure Matrix

	Component		
	1	2	3
dwelling_age	.980	.202	.245
squared_dwelling_age	.976	.225	.303
saving_potential	.952	.218	.322
WOZ_value	.467	.338	.779
household_size	.214	.903	.030
household_children	.151	.830	-.046
age_oldest	.078	-.267	.705
electricity_SYU	.314	.633	.537
gas_SYU	.633	.470	.681

Pattern Matrix[a]

	Component		
	1	2	3
dwelling_age	1.017	-.057	-.069
squared_dwelling_age	.986	-.032	-.004
saving_potential	.953	-.034	.025
WOZ_value	.191	.217	.697
household_size	-.004	.910	-.062
household_children	-.033	.851	-.123
age_oldest	-.077	-.324	.763
electricity_SYU	.012	.582	.474
gas_SYU	.384	.316	.528

Extraction Method: Principal Component Analysis.

Rotation Method: Oblimin with Kaiser Normalization.

Extraction Method: Principal Component Analysis.

Rotation Method: Oblimin with Kaiser Normalization.

Component Correlation Matrix

Component	1	2	3
1	1.000	.262	.315
2	.262	1.000	.102
3	.315	.102	1.000

Extraction Method: Principal Component Analysis.

Rotation Method: Oblimin with Kaiser Normalization.

Component Score Coefficient Matrix

	Component		
	1	2	3
dwelling_age	.335	-.025	-.090
squared_dwelling_age	.322	-.017	-.048
saving_potential	.309	-.019	-.028
WOZ_value	.028	.063	.425
household_size	.000	.432	-.090
household_children	-.006	.407	-.124
age_oldest	-.061	-.195	.510
electricity_SYU	-.020	.248	.271
gas_SYU	.099	.119	.303

Extraction Method: Principal Component Analysis.

Rotation Method: Oblimin with Kaiser Normalization.

Component Scores.

Component Score Covariance Matrix

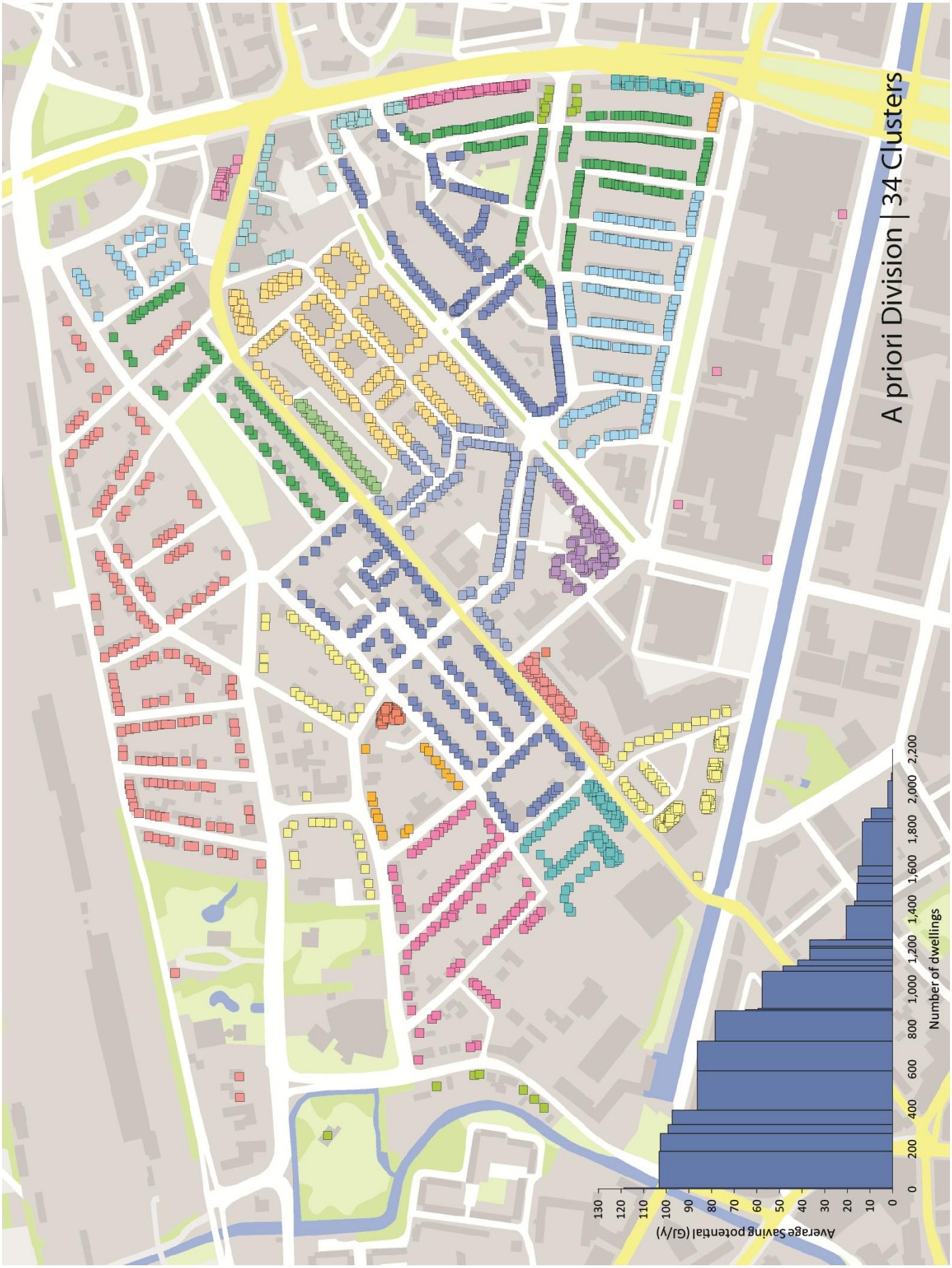
Component	1	2	3
1	1.483	.658	2.397
2	.658	1.106	.940
3	2.397	.940	3.414

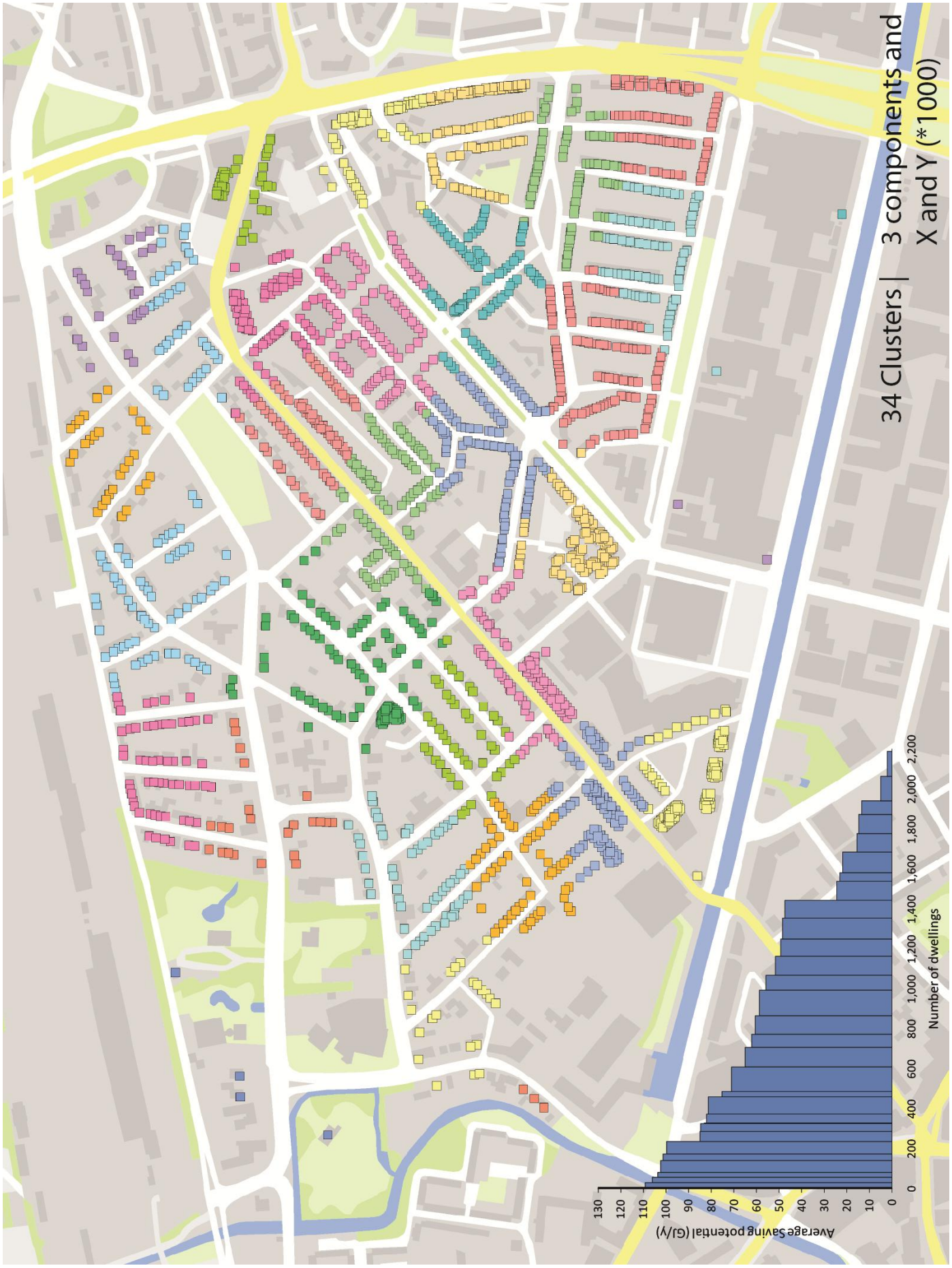
Extraction Method: Principal Component Analysis.

Rotation Method: Oblimin with Kaiser Normalization.

Component Scores.

F Visualizations

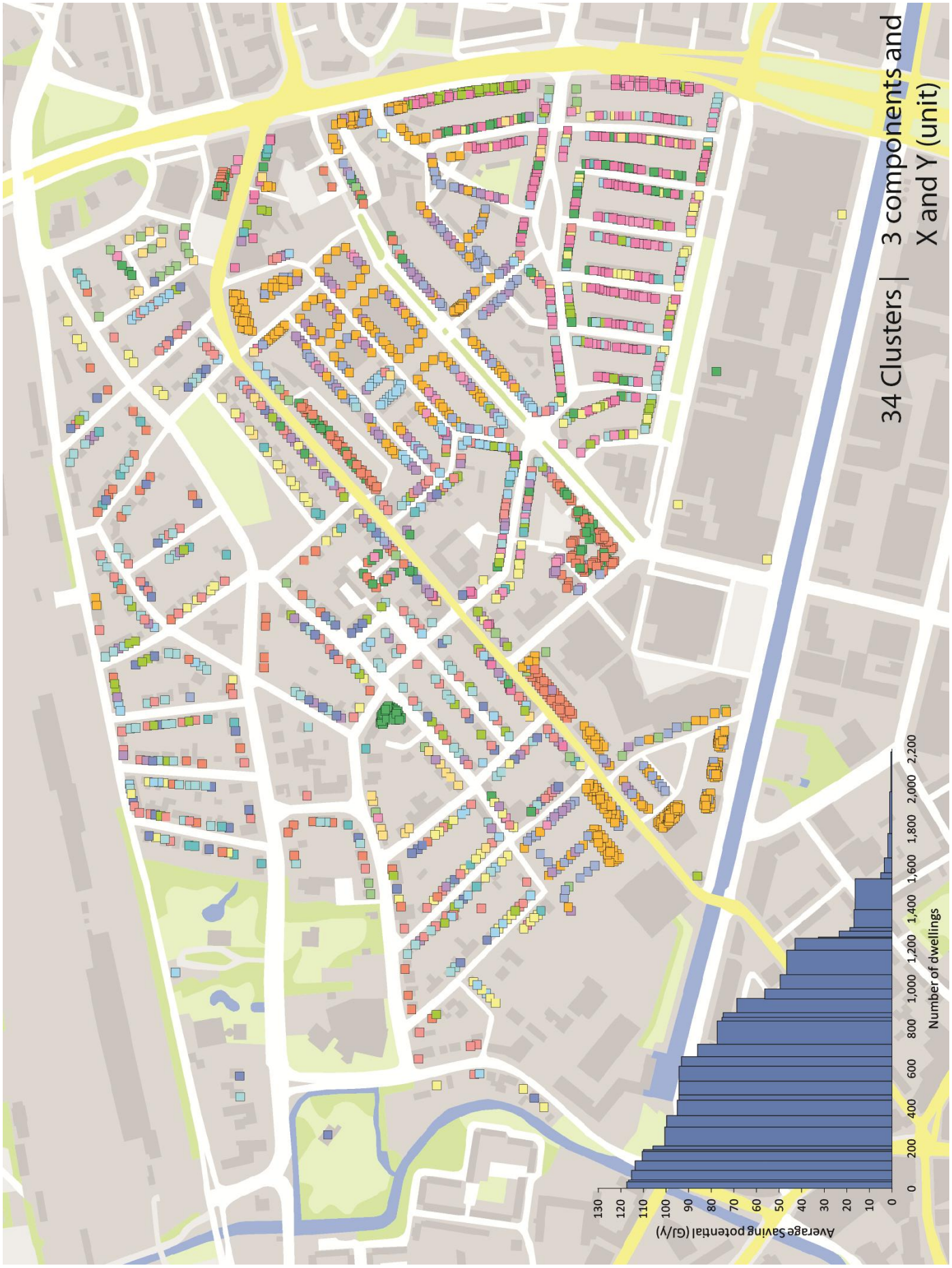

A priori Division | 34 Clusters
Average Saving potential (GJ/y)
130
120
110
100
90
80
70
60
50
40
30
20
10
0
Number of dwellings
0
200
400
600
800
1,000
1,200
1,400
1,600
1,800
2,000
2,200

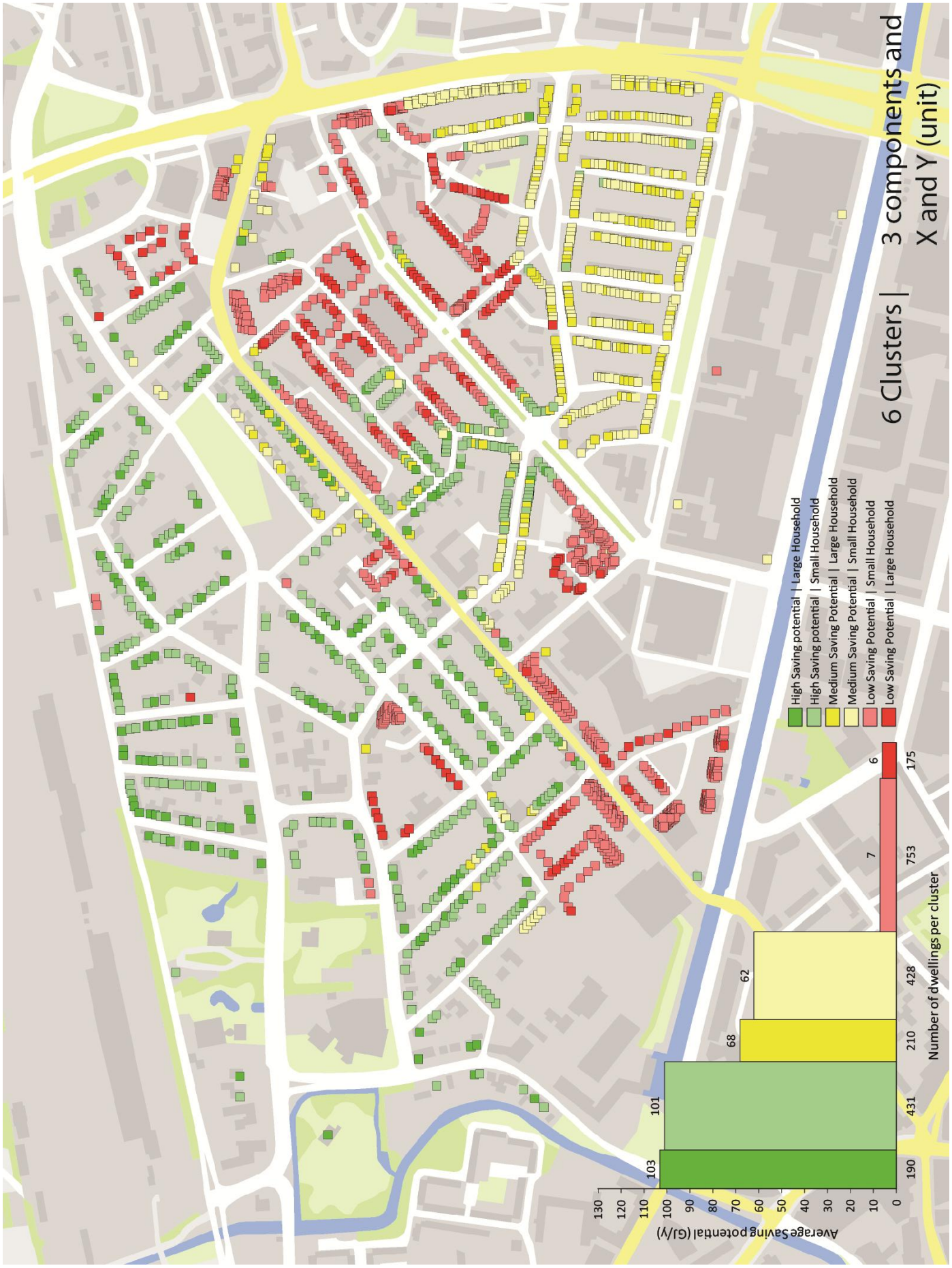

34 Clusters | 3 components and X and Y (*1000)
Average Saving potential (GJ/y)
Number of dwellings
0
200
400
600
800
1,000
1,200
1,400
1,600
1,800
2,000
2,200
10
20
30
40
50
60
70
80
90
100
110
120
130

		Total	Average							
Cluster number	Number of dwellings	Saving potential	Saving potential	Dwelling age	WOZ value	Household size	Household Children	Household Oldest Age	Electricity SYU	Gas SYU
		GJ/y	GJ/y	y	k€			y	kWh	m^3
1	100	8,081	9	24	39	1.46	0.74	16	2,307	981
2	44	588	10	8	62	1.10	0.33	12	2,304	1,448
3	59	5,341	15	32	110	1.17	0.64	14	1,998	880
4	44	4,781	9	22	190	1.51	1.13	12	2,984	1,691
5	126	137	1	0	55	0.62	0.31	20	1,333	517
6	131	35	2	1	77	0.96	0.60	16	1,455	740
7	54	4,964	11	16	169	2.14	1.06	18	2,919	1,236
8	90	1,421	0	2	30	1.31	0.18	12	2,614	928
9	99	8,558	5	24	33	1.77	0.94	16	2,141	1,159
10	106	6,243	3	12	21	1.40	1.04	16	1,224	696
11	1		0	0	0	0.00	0.00	0	0	0
12	48	1,459	26	23	40	1.31	0.41	17	3,522	1,441
13	1		0	0	0	0.00	0.00	0	0	0
14	127	8,744	5	26	15	1.48	0.86	14	1,989	732
15	26	2,776	8	19	197	1.58	1.38	12	3,902	1,738
16	60	6,273	10	20	233	1.42	1.13	17	2,303	1,528
17	63	95	7	5	58	0.84	0.21	17	1,783	994
18	78	0	0	0	17	1.40	1.12	9	1,434	512
19	66	6,554	7	20	178	1.50	1.23	16	2,270	1,002
20	2									
21	3									
22	54	698	2	4	222	1.08	0.79	11	1,959	917
23	77	6,892	7	27	146	1.74	0.99	15	2,844	982
24	64	6,705	7	18	130	1.45	1.18	15	2,685	1,097
25	23	795	2	8	144	1.29	0.93	10	2,135	777
26	97	5,353	4	10	27	1.62	1.16	15	1,948	532
27	23	2,394	9	19	139	1.57	1.31	17	2,591	1,285
28	58	346	8	10	26	0.88	0.56	13	1,015	404
29	77	96	3	1	50	1.58	1.16	18	1,718	864
30	86	4,636	10	27	4	0.84	0.47	21	1,623	686
31	59	6,141	5	18	133	1.82	0.84	19	2,322	1,323
32	81	7,925	5	24	49	1.58	0.76	16	1,875	1,215
33	82	1,353	11	7	58	1.05	0.62	14	1,491	704
34	78	165	0	0	21	1.38	1.14	16	1,453	306
De Laak	***2,187***	***109,920***	***50***	***49***	***283***	***2.32***	***0.51***	***48***	***3,543***	***1,840***

34 Clusters | 3 components and X and Y (*100)
Average Saving potential (GJ/y)
0
10
20
30
40
50
60
70
80
90
100
110
120
130
Number of dwellings
0
200
400
600
800
1,000
1,200
1,400
1,600
1,800
2,000
2,200

34 Clusters | 3 components and X and Y (*10)
Average Saving potential (GJ/y)
0
10
20
30
40
50
60
70
80
90
100
110
120
130
Number of dwellings
0
200
400
600
800
1,000
1,200
1,400
1,600
1,800
2,000
2,200


34 Clusters | 3 components and X and Y (unit)
Average Saving potential (GJ/y)
Number of dwellings
0
200
400
600
800
1,000
1,200
1,400
1,600
1,800
2,000
2,200
10
20
30
40
50
60
70
80
90
100
110
120
130


6 Clusters | 3 components and X and Y (unit)
High Saving potential | Large Household
High Saving potential | Small Household
Medium Saving Potential | Large Household
Medium Saving Potential | Small Household
Low Saving Potential | Small Household
Low Saving Potential | Large Household
Average Saving potential (GJ/y)
0
10
20
30
40
50
60
70
80
90
100
110
120
130
103
101
68
62
7
6
190
431
210
428
753
175
Number of dwellings per cluster

G Summary KENWIB

TARGET GROUP CLUSTERING FOR APPLICATIONS OF ENERGY EFFECTIVE RENOVATION CONCERNING PRIVATELY OWNED DWELLINGS

Author: P.M.T. van Loon

Graduation program:
Construction Management and Urban Development 2011-2012

Graduation committee:
prof. dr. ir. B. de Vries (TU/e)
dr. ir. E.G.J. Blokhuis (TU/e)
J. Bekkering (HetEnergiebureau BV)

Date of graduation:
21-03-2012

ABSTRACT
In programs for energy effective renovation of dwellings it is hard and still not clear how to select the right target group regarding the dwellings saving potential (hardware) and decision making private homeowner (software). In this research linear components of different factors and variables of dwellings and their occupants are extracted in a principal component analysis (PCA). With the components and the actual geographical coordinates of the dwellings different cluster analyses are conducted searching for new spatial contiguous energy clusters and target group clusters. The target area of the research is only one district in Eindhoven. Therefore the promising results obtained should be tested in further research.

Keywords: Energy Efficient Renovation, Target Group Clustering, Housing Submarkets, Marketing, Cluster Analysis

INTRODUCTION

Context

All the things we do in and for life on earth consumes energy. Currently fossil fuels are used for the majority of our energy production. Looking at future energy scenarios it becomes clear that fossil fuels need to be replaced by other (renewable) energy sources. The Trias Energetica is a simple and logical concept that stimulates to achieve energy savings, reduce our dependence on fossil fuels, and for the remaining part use fossil fuels as efficient as possible.

Governments in Europe decided there should be common objectives for all countries in the European Union. This leaded to European energy and sustainability objectives for the Netherlands in 2020: 1) 20 percent reduction of greenhouse gasses compared to 1990 2) 14 percent share of renewable and 3) an annual energy consumption reduction of 2 percent. The municipality of Eindhoven has even more ambitious goals, she aims to be energy neutral between 2035 and 2045 (Municipality of Eindhoven, 2008). Energy neutral in this case means that the (remaining) energy demand for the own organization, dwellings, industry and remaining connections is generated with renewables inside the borders of Eindhoven.

Campaigns of the government to reduce energy use in existing dwellings were not often successful in the past. The participation rate was in most cases not exceeding 5%, of which 3% was not attracted by the campaign but already was intrinsically motivated to compete in a program. Idea owners of such programs at municipal level are in a real need for ways to increase this participation rate. The government just finished a report with best practices for building related energy savings for private owners (Motivaction, 2011), this report has been made as part of the "more with less" program (meer met minder). In the report Do's and Don'ts are formulated. One of the Do's only raises more questions:

Choose the target group and their homes with care. There must be a potential saving in the houses and it is important to focus on a target group. A group is characterized by shared values, needs and ages. All residents of a neighborhood are rarely a target. Make sure your approach fits the target audience. Are they sensitive to comfort, money savings or unburdening? Adjust your approach to it.

Easier said than done, but how do you do such a thing if you have more than 50,000 potential dwellings in, for example, Eindhoven? Some of the questions that rose are listed below; these questions will transform into the research questions:

Problem Statement

How can we select the dwellings with the biggest saving potential bearing the characteristics of the household in mind? Is an evaluation method available which can integrate characteristics of hardware and software?

In programs for energy effective renovation of dwellings it is hard and still not clear how to select
the right target group
regarding the dwellings saving potential (hardware) and decision making private homeowner (software).

In the next paragraph research questions are formulated that will lead to the design of a study.

Research questions

- Which variables or factors of all dwellings and households in Eindhoven are available for analysis?
- Which variables or factors influence the decision for participation of private homeowners?
- Is the statistical cluster division of Eindhoven a reliable target group division?

Is an optimization of target group size and geographical distribution possible, when focusing on maximization of (1) participation and (2) reduced energy demand in a program for energy effective renovation for private owners of dwellings?

Relevance of research

In a graduation research including an internship both practical and theoretical relevance are important. In the transition towards a more sustainable Eindhoven, i.e. total energy neutrality, an important factor is upgrading the energy performance of the existing stock. Programs to achieve this are rarely successful, since too little is done to target the right dwellings and owners. The practical relevance of this research lies in its usability for BvB/e. The foundation tries to recruit up to 2,000 participants in a program for EE-renovation of privately owned dwellings. To reach this goal 20,000 potential participants and their dwellings are selected. The designed method and output will be used to select the 20,000 dwellings used by BvB/e.

Expected results

- A set of decision variables that tries to reflect the actual behavior and characteristics of private owners and their dwellings considering to participate in an energy effective renovation program;
- A cluster analysis (clustering). Resulting in a certain number of clusters (geographical constrained typological target groups) in Eindhoven;
- A dataset of all the dwellings and their owners to use in the communication strategy for BvB/e;
- A map of Eindhoven visualizing the target group clusters for energy saving.

Research design

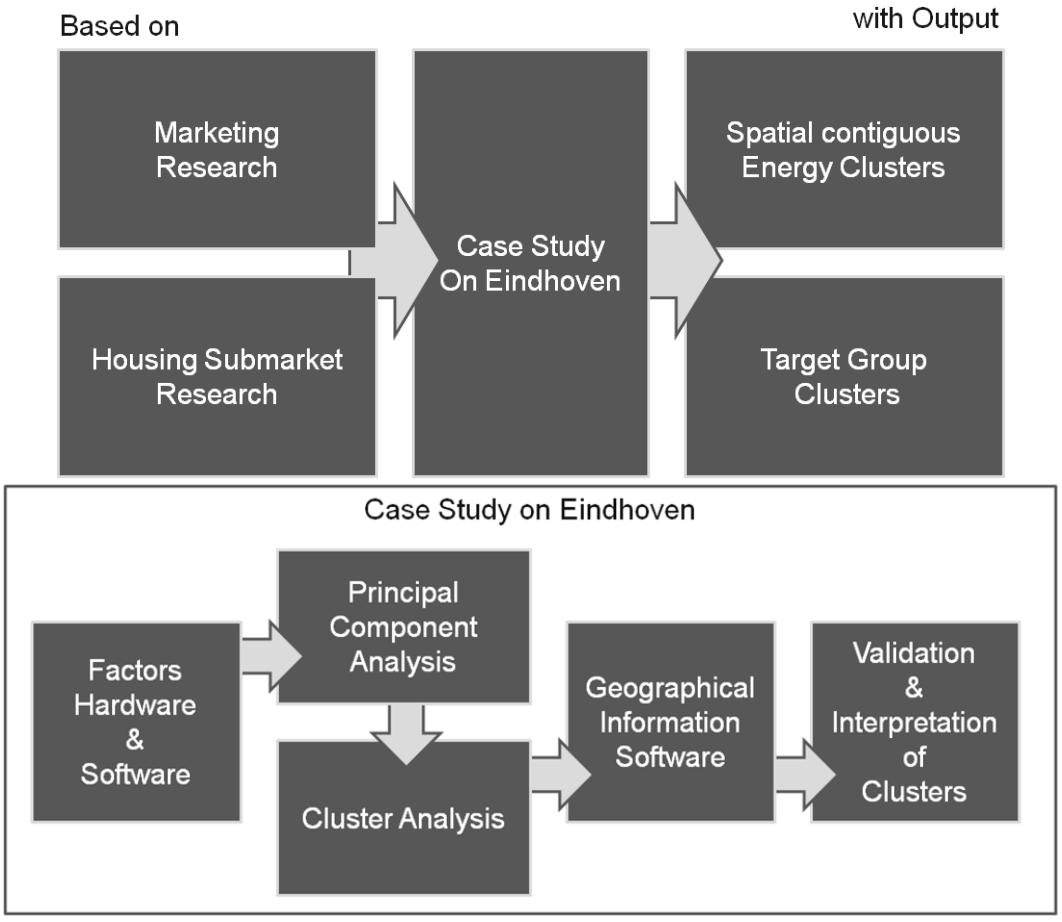


Figure 1: research model

In figure 1 the research model is visualized. With the characteristics and methods of marketing and housing submarket research in mind a case study on a district in Eindhoven is conducted. 2 studies are designed which should lead to spatial contiguous energy clusters and target group clusters. The case study consists of a study on available hardware & software factors. With these factors a PCA and CA is conducted the results are presented with maps using GIS. With the last step the found clusters are validated and an interpretation is given.

THEORETICAL ORIENTATION

Housing submarket research

Neighborhoods are a historically grown, physical presentation of groups of buildings. In which neighborhood a property is located is influenced by administrative decisions of planners and therefore historically determined. In the 1960's research on housing markets started, it was based on a belief that the prices of property are not only defined by its physical location but also structural, demographic and socio-economic characteristics have influence. The most often used definition is given by Bourassa et al. (1999) were a submarket is defined as a set of dwellings that are reasonably close substitutes of one another, but relatively poor substitutes for dwellings in other submarkets. Transferring this towards the challenge for target group clustering in Eindhoven this research indicates we should look further then an "a priori" division of the municipality of Eindhoven for neighborhoods and small clusters.

It is disputable that with the use of housing submarkets, factors related to energy usage are left out of the response variables. In this research it could be wise to use a statistical submarket housing model, where technical, structural and energetic characteristics of a dwelling are taken in account too. Wu & Rashi Sharma (2011) deals with the topic of housing submarket classification and the role of spatial contiguity, sometimes called nearness or proximity. A spatially constrained data-driven classification methodology is used to deduce spatially integrated housing market segments. This research is extremely useful because it links geographical constrained data into a model where different variables are statistically considered in their coherence. Submarkets are formed with houses more similar to each other based on location and their physical, typological and demographic properties.

It is concluded that the method advocates the utility of spatial submarkets where public and private organizations can identify specific geographic zones of potential growth or with special needs. Is it possible to identify these regions and use them as target groups for energetic effective renovation programs?

Marketing research

In a program for energy effective renovation there is a need for a division of target groups. People are attracted to different aspects of the results of a renovation program and therefore have different grounds for participation. Cluster analysis is used in market research for selection of possible or preferred consumers, the so called market segmentation. With market segmentation the market is divided into target groups or sub-markets. An older overview for possible application is given by Punj & Steward (1983) and an implementation is conducted by Kuo et al. (2002).

CASE STUDY ON EINDHOVEN

Eindhoven is divided into 7 quarters, i.e. Centrum, Tongelre, Gestel, Stratum, Strijp, Woensel and Gestel. These 7 quarters are split further into 20 districts and 116 neighborhoods. Based on composition of the dwelling stock a neighborhood is further divided into clusters. The city of Eindhoven has more than 1,100 a priori clusters. In figure 2 the target area is visualized, district "De Laak" consists of 2 neighborhoods and 34 a priori cluster representing 2187 occupied dwellings.

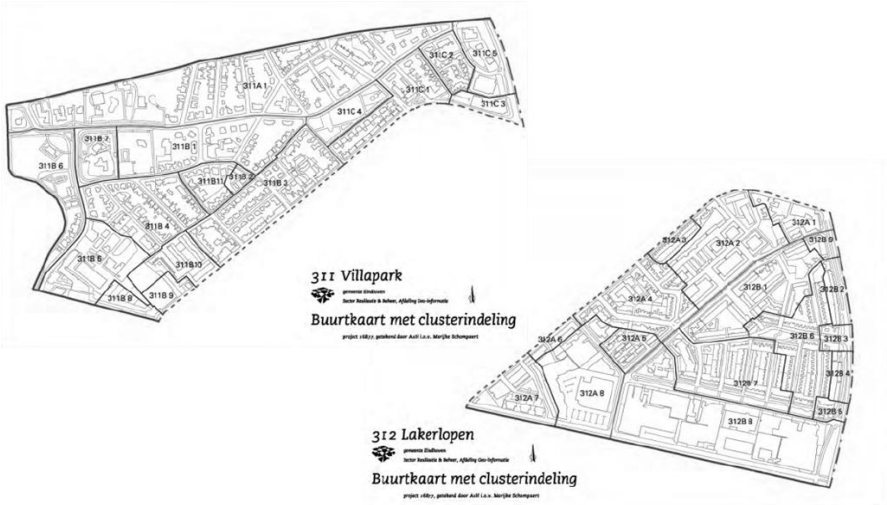


Figure 2: Target Area district "De Laak" Eindhoven, a priori division

In figure 3 the following available and useful factors of hard & software are enlisted. The municipality delivers information about hardware and software. The hardware factor "typology" is converted to a interval level using the example dwelling saving potential of Agentschap NL. Endinet delivers standard year usage figures of electricity and gas connections of every dwelling in the area under study.

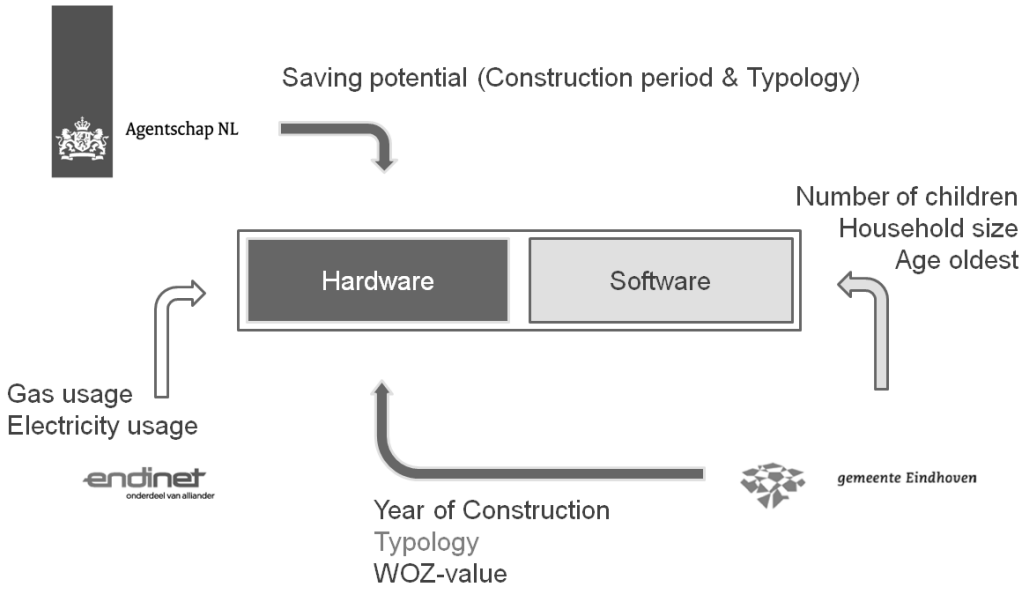


Figure 3: Factors of hardware & software available for analysis

On the dataset of 2187 dwelling a principal component analysis (PCA) was conducted on the 9 variables with oblique rotation (Direct Oblimin). The KMO verified the sampling adequacy for the analysis, KMO - .726 ('Good" according to Hutcheson & Sofronniou (1999)), and all KMO values for individual items were > .618 which is well above the acceptable limit of .5 (Field, 2009). An initial analysis was run to obtain eigenvalues for each component in the data. Three components had eigenvalues over Kaiser's Criterion of 1 and in combination explained 78.40 percent of the variance. The screeplot was slightly ambigious and showed an inflexion that would justify retaining 2 or 3 components. Given the large dataset, and the convergence of the screeplot and Kaiser's criterion on three components , this is the number of components that were retained in the final analysis. The items that cluster on the same components suggest that component 1 represents the "dwelling saving potential", component 2 the "Household characteristics" and component 3 that "Wealth comes with age".

The two studies are validated using the weighted average standard deviation (WASD) (Wu & Rashi Sharma, 2011).

$$\mathrm{WASD}_{\text{per characteristic}} = \frac{\sum_{i=1}^{i=n}(N_i * SD_i)}{N} = \frac{\sum_{i=1}^{i=n}\left(N_i * \sqrt{\frac{\sum_{j=1}^{j=N_i}(x_j - \bar{x})^2}{N_i}}\right)}{N}$$

In most cases there is need for spatial contiguous boundaries of a cluster, this is inspected visually too (Wu & Rashi Sharma, 2011). The third measure used is the shape of the distribution in the "variwide" plot. This plot is shown in de lower left corner of figure 4. The results of a CA are better when it is shaped as much as possible like the plot of the individual saving potential of dwellings. Two different studies are executed, both are discussed in the next subparagraphs.

New spatial contiguous energy clusters

At first a new spatial contiguous cluster division for district "De Laak" is made. Whether the a priori classification is a well performing division regarding saving potential and homogeneity of clusters should be based on judgments for all different levels in the classification is answered next. For the most specific level (neighborhood clusters) the statistical cluster division of Eindhoven is quite a good representation. Moreover the homogeneity of the formed 34 clusters is significantly higher using the data-driven classification methodology, this conclusion is based on the found weighted average standard deviation for the characteristic saving potential which is 28 percent lower. Using the a priori division on any higher level of aggregation is risky, for the homogeneity is much lower in these divisions than in the division used in this study. There may be districts with highly coherent dwelling types and building periods but this may not be presumed.

The results of the first study support the belief that a data-driven classification method, such as cluster analysis, can lead to a better clustering of dwellings for energy effective renovation up to a certain level. The statistical a priori division of the municipality performed quite well at the lowest level. But a better spatial contiguous division is possible

by using the 3 components deduced out of 8 variables and the geographical coordinates weighted by multiplying them with 10^2.

Target group clusters

The second study conducted focused on target groups in which the location of a dwelling was not used as a validation criterion upfront. To come up with a target group division for the district different amounts of clusters were generated as output. The cluster analysis where 6 distinct clusters were found was evaluated and an interpretation was formulated. Three levels of saving potential were split into two categories of household size, i.e. large and small. The insight that 6 target groups do characterize the district can be of use in the marketing campaign for BvB/e. 4 of the 6 clusters represent saving potential and a division into large and small households is made.

It does not come as a surprise that the output of the cluster analysis where target groups are distinguished seems to act upon the first two components characterizing the data set. Because the target groups are not spatially contiguous the homogeneity is much higher than the a priori division in 5 clusters.

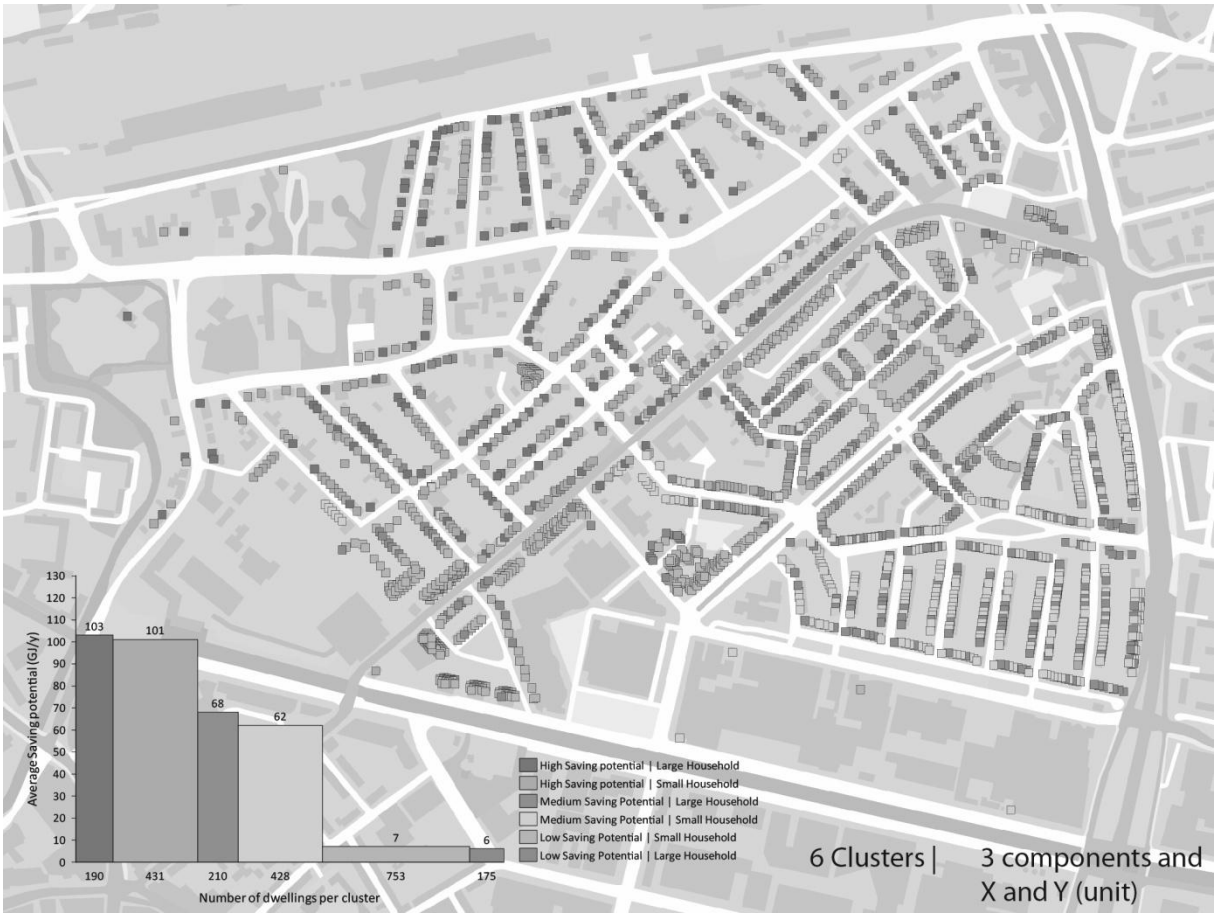


Figure 4: Visualisation of "De Laak" divided into 6 target group clusters

CONCLUSION AND DISCUSSION

In the conclusion and discussion section the implications of this study for the research on energy performance in housing submarkets and the usability of target group clustering for the energy effective renovation program for private homeowners are discussed. KENWIB is in the early stages of exploring the field of cluster analysis for use in energy related submarket research. In this section the research questions are answered.

Conclusion

Available factors and variables

With the problem statement and research design in mind the goals of this study are evaluated and conclusions are formulated. It is important to realize that the data used for cluster analysis determines the quality of the output. Therefore it is investigated which variables or factors of all dwellings and households in Eindhoven are available for analysis.

For this study the Standard Year Usage (SYU) for gas and electricity of a connection in a dwelling were used. Those figures are statistically compared with the figures used in the research of Brouwers et al. (2010). The current SYU's (figures of the last quarter of 2011) are available and they should represent the current electricity and gas use of a dwelling as accurately as possible. The introduction of smart meters in all dwellings will introduce a new interesting variable in dwelling related energy usage studies.

In the registers of the municipality a lot of information is availabe concerning the composition of the household and the age of the occupants, this information is deduced from the GBA. In the WOZ-database of the municipality the typology and year of constuction of each dwelling are stated. The information on variables describing the size of a dwelling, like surface or volume is not availabe for all dwellings. This is a real setback because the volume of surface combined with the known year of constuction and gas usage forms a reliable measure for energy performance of a dwelling.

Now the energy performance of a dwelling, and therefore the saving potential is determined based on dwelling typology and year of construction using the example dwellings of Agentschap NL. This method is presumed to be less accurate than using empirical data representing actual energy performance for analysis.

Decisions for participation

The goal to explore decisions for participation was not achieved to full satisfaction. Which variables or factors influence the decision for participation of private homeowners cannot be concluded based on this research. Of course the 8 variables used, characterize the dwelling and their occupants but it is not tested whether these variables determine the decision for participation in the EE-renovation program.

Interpretation of extracted components out of the data set

In the principal component analysis three components are extracted. These three components account for 78% of the variance in the data set used for this district. These components are interpreted as 1) dwelling saving potential, 2) household characteristics and 3) "wealth comes with age".

Discussion

Limitations

The source of the data of the different variables which are available really determines the quality and possibility to come up with a new target group division for neighborhoods. Most of the variables included consist of empirical data, i.e. data collected by (semi) direct observations. Only one variable cannot be described as empirical, this is the saving potential of a dwelling. To assess the saving potential of a dwelling the typology and building period are used. The adopted value is therefore an average value to sort identical dwellings build in a specific period of time. It was attempted to come up with a more empirical measure for potential energy savings. But actual figures for all dwellings, e.g. the energy label, or volume or surface are not available.

Because the data collection for this experiment was a time consuming process, only one district is analyzed up until now. Even though the results are promising, no guarantees can be given that this approach will work for all other districts in Eindhoven. The conclusion that the maximum of 6 target group clusters will be enough to characterize all districts in Eindhoven should be seen as a hypothesis and should therefore be tested in further research.

Not all demographic data available is used for analysis. Culture and ethnicity are not included as factors. Real research on buying behavior and sensitivity for marketing approaches is not included in the study, so the deduced target groups are only a further exploration of districts and characterize them regarding saving potential and household size. The communication agency could use this to adapt their strategy on district level.

Recommendations

Due to time constraints and the necessary practical output that had to be obtained, some research steps were executed rapidly. It is advisable to rerun the analysis with another measure for saving potential or at least leave it out of the analysis for once to see how clusters are formed then. It is known this does undermine the results. However it would be disputable to take the results of this study as a proof that the used factors are optimal. This may not be concluded before more experience is gained by actual using PCA and CA to divide object into target groups. As said this can be done analyzing different combinations of variables of the same district.

Another recommendation is that an even stronger belief in and further validation of the method can be gained by executing it on several other districts in Eindhoven. All districts have another division of dwelling typologies and construction periods. This will lead to further insights on the usability of housing submarket research for target group clustering. It is expected that conducting the study on other districts will take one-tenth of the time needed than when it was done for the first time. The hypothesis is that all districts can be divided into at most 6 target group clusters.

ACKNOWLEDGEMENTS

During this research collaboration of several parties took place. Many thanks to Jan Bekkering for the chance to graduate at “HetEnergieBureau” on the pilot project of “Blok voor Blok”, in Eindhoven formulated as BvB/e. All participating parties in BvB/e: HetEnergiebureau BV, network operator Endinet, Q-energy and the municipality of Eindhoven. Special thanks to ir. Rick Donders, Kees van der Hoeven, Maartje Essens, drs. Van der Waerden, prof. De Vries and dr. Blokhuis.

REFERENCES

Bourassa, S.C., Hamelink, F., Hoesli, M. & MacGregor, B.D., 1999. Defining housing submarkets. *Journal of Housing Economics*, (8), pp.160-83.

Field, A., 2009. *Discovering statistics using SPSS*. 3rd ed. London: Sage.

Hutcheson, G. & Sofronniou, N., 1999. *The multivariate social scientist*. London: Sage.

Kuo, R.J., Hob, L.M. & Huc, C.M., 2002. Cluster analysis in industrial market segmentation through artificial neural network. *Computer & Industrial Engineering*, (42), pp.391-99.

Motivaction, 2011. *Kansrijke aanpakken in gebouwgebonden energiebesparing, de particuliere eigenaar*. Agentschap NL.

Municipality of Eindhoven, 2008. *Uitvoeringsprogramma klimaatbeleid 2009-2012: Van succesvolle projecten naar structurele uitvoering*.

Punj, G. & Steward, D.W., 1983. Cluster analysis in marketing research: review and suggestions for applications. *Journal of Marketing Research*, (20), pp.134-48.

Wu, C. & Rashi Sharma, R., 2011. Housing submarket classification: The role of spatial contiguity. *Applied geography*, 32, pp.746-56.

PIM MATHIEU THEODORUS VAN LOON

p.m.t.v.loon@student.tue.nl

After finishing a broad bachelor ABP, with extensive extracurricular activities at the student boat club with the associated development of social and management skills, he started his master CME at the TU/e. In his graduation thesis on the topic of target group clustering for energy effective renovation many stakeholders participated believing in his enthusiastic and inspiring attitude. Basically on his own he managed to set project boundaries and implemented research methods such as cluster analysis for implementation on housing submarket and marketing research.

2004 – 2009 Bachelor Architecture Building and Planning
2007 – 2008 Vice-president of executive board E.S.R. Thêta
2009 – 2010 TU/e Certificate in “Technology management”
2009 – 2012 Master Construction Management and Engineering