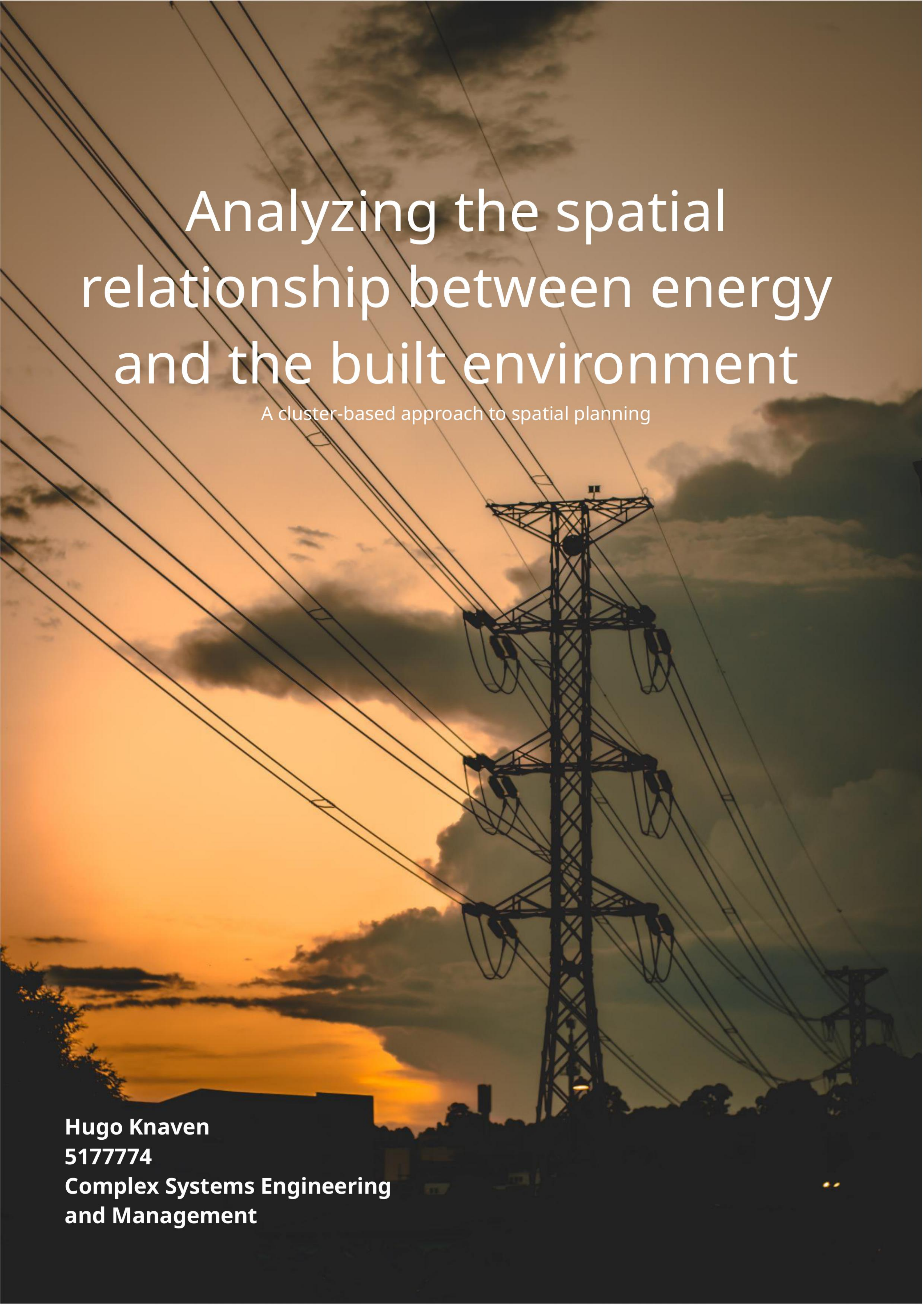


Analyzing the spatial relationship between energy and the built environment

A cluster-based approach to spatial planning

Hugo Knaven
5177774
Complex Systems Engineering and Management

Analysing the spatial relationship between energy and the built environment

A cluster-based approach to spatial planning

By

Hugo Knaven

in partial fulfillment of the requirements for the degree of

Master of Science
in Complex Systems Engineering and Management

at the
Delft University of Technology,
to be defended publicly on Friday, November 7, 2023, at 15:30.

1st supervisor: Nazli Aydin
2nd supervisor: Trivik Verma
Chair: Igor Nikolic
Advisor: Özge Okur

An electronic version of this thesis is available at
http://repository.tudelft.nl/.

Acknowledgements

Dear reader,

The thesis you are currently reading is part of the graduation process for the Master Complex Systems Engineering and Management. In total, the writing of this thesis took eight months of hard work. I would like to express my gratitude to some individuals who supported and helped me during this process.

I would like to express my deepest appreciation to my graduation committee. Their input and feedback during these months have profited numerous insights and helped shape the thesis to its current form. I am also extremely grateful to my supervisor from the province of South Holland who provided the case for this thesis. His insights into the topic and feedback on the results have helped me enormously during the making of this thesis. I also want to express my sincere thanks to all the interviewees who participated in my research. Without their input, it would not have been possible to provide an analysis of the system on the same level of detail. Besides this, I also want to thank my fellow students and supervisors from the Center of Urban Science and Policy. Our weekly discussions have served as a source of inspiration and feedback during the creation of this thesis. Lastly, I want to thank my girlfriend, my family, and my friends for their support during the writing of this thesis.

Executive Summary

As a result of climate change, there is a push to reduce greenhouse emissions in the Netherlands. This requires large investments in energy infrastructure, which can both be costly and might not always be practical in areas where the limited availability of space hinders its placement. For this reason, it is important that decision-makers have a clear picture of the current state of the energy system. One field where these insights into the energy system are required is spatial planning. This is because energy characteristics like supply and demand are directly linked to the built environment. This means that any changes to the built environment will affect the energy system and vice versa. These changes to the built environment are mostly being made by the province and municipalities through spatial policies. It is therefore important that these policymakers are able to combine aspects of both the energy system and the built environment within their spatial policies. However, spatial planners currently lack the insight into their relationship to use this combined approach effectively.

Though research into these relationships has been performed in the past, these were often limited to a small set of characteristics. This research also often focused on a small area like a single town or neighborhood. For this reason, there is currently a gap in knowledge on how the relationship between energy and built environment characteristics can be mapped. In this thesis, a spatial analysis of the province of South Holland was performed. The goal of this analysis was to answer the following question: ***"How can different sets of characteristics be used in spatial clustering to identify the relationship between the energy system and the built environment for use in spatial planning?"***.

To reach this goal, three subquestions were answered. The first of these questions was: *"What characteristics of the energy system and build environment are relevant for defining clusters of the province of South Holland?"*. To answer this question the relevant aspects of the energy system and built environment were defined. This was done through a combination of literature research and interviews with experts in the field of energy and spatial planning. Based on this, characteristics were identified for the energy system, built environment, and social environment. The main characteristics of the energy system were defined as demand, supply, network capacity, and storage for both heat and electricity. For the built environment, These were defined as land use, building typology, renewable potential, and population density. Lastly, the social characteristics were defined as attitudes toward the energy transition and energy poverty.

Using these characteristics the second subquestion was answered. This question was: *"What clustering methodology can be used to identify the relationships between energy and spatial characteristics?"*. To answer this question first suitable datasets were found for each of the characteristics. For some characteristics, no suitable dataset was available therefore these characteristics were not included. In total 9 energy characteristics, 5 built environment characteristics, and 4 social characteristics were included. Besides missing datasets, not all datasets included data for the full province of South Holland. In particular, data was missing from rural and industrial areas, for this reason the performed analysis mostly focused on the urban areas.

To analyze the characteristics, a k-prototype cluster analysis was performed. Here, different sets of characteristics were combined to analyze how they were spatially related to each other. In total four analyses were performed with one focusing on only the energy characteristics, one focusing on the energy and built environment characteristics, one focusing on energy and social characteristics, and one focusing on a combination of all characteristics.

In general, the clusters that were created using this method were relatively weak. This meant that there was an overlap between the different clusters. It also meant that the same characteristics often had a wide variety of values for all areas in one cluster. Although this limited the possibility of performing a detailed analysis of aspects of the clusters, this still showed that it was possible to identify differences between areas in the province of South Holland based on their characteristics. It also showed that relationships between different characteristics can be identified based on their occurrence in different clusters. An example of this was the energy label and average energy gas demand of households which showed that all clusters with a high energy label also had a low gas demand and the other way around.

The analysis also showed that the characteristics of each cluster vary depending on the cluster set that was used. Here each cluster set showed unique features that could not be found in the other cluster analyses. The combined analysis of all characteristics showed that although some of these unique features are still visible when combining all characteristics many also get lost. Not all characteristics seemed to be equally important in deciding the cluster of an area, with all clusters mostly looking at the heat demand of an area. This however did not mean that the remaining characteristics were not important. This is because most of the variation between the clusters and between the cluster sets of each analysis could be explained by the differences between the other characteristics. Combining these findings showed that it was possible to find spatial relationships between energy and built environment characteristics using the methodology used in this thesis. However, finding these insights does require a balance between the number of characteristics included in the analysis and the detail level of the analysis.

Based on these findings and a discussion with experts in energy and spatial planning the last research question was answered. This question was as follows: *"What kind of policy insights can be identified based on the proposed method of identifying and analyzing energy clusters?"*. Here the results showed that this type of analysis is most suited for providing insights into the relationship between areas and their characteristics. Based on this it was determined that clustering is mostly suitable for an exploratory analysis of the energy system and the built environment. The findings of this analysis can then be used to inform decision-making for spatial planning or serve as the basis for a more in-depth analysis of the energy system and built environment.

By combining these results it could be concluded that the method described in this thesis could be used to analyse the relationship between energy and landscape for use in spatial planning. However, more research into alternative datasets, alternative clustering methods, and a more result-based set of characteristics would be required to further optimize this methodology.

Table of Contents

Acknowledgements **i**

Executive Summary **ii**

Table of content **v**

List of figures **vii**

List of tabels **viii**

1 Introduction **1**
1.1 Background . . . 1
1.2 South Holland . . . 2
1.3 Existing research . . . 4
1.3.1 Current use of energy characteristics in spatial analysis . . . 5
1.3.2 Knowledge gap . . . 7
1.4 Research approach . . . 8
1.5 Research questions . . . 10
1.6 Research outline . . . 12

2 Defining relevant energy and build environment characteristics **13**
2.1 Literature . . . 13
2.1.1 Review process . . . 13
2.1.2 Review results . . . 14
2.2 Interviews . . . 18
2.2.1 Interview process . . . 18
2.2.2 Interview results . . . 19
2.3 Cluster characteristics . . . 23

3 Methodology for spatial clustering of energy and spatial characteristics **25**
3.1 Data collection . . . 25
3.1.1 Dataset selection . . . 25
3.1.2 Data cleaning . . . 27
3.1.3 Dataset analysis . . . 29
3.2 Clustering method . . . 31
3.2.1 Cluster algorithm selection . . . 32
3.2.2 Cluster model design . . . 34
3.2.3 Experimental design . . . 35
3.3 Model iteration . . . 39

4 Identifying relation between energy and spatial characteristics **41**
4.1 Energy clusters . . . 41
4.1.1 Spatial distribution of energy clusters . . . 41
4.1.2 Cluster characteristics of energy clusters . . . 45
4.1.3 Defining characteristics of energy clusters . . . 49
4.2 Energy and built environment cluster . . . 51
4.2.1 Spatial distribution of energy and built environment clusters . . . 51

4.2.2 Cluster characteristics of energy and built environment clusters . . . 53
4.2.3 Defining characteristics of energy and built environment clusters . . 57
4.3 Energy and social cluster 59
4.3.1 Spatial distribution of energy and social clusters 59
4.3.2 cluster characteristics of energy and social clusters 61
4.3.3 Defining characteristics of energy and social clusters 64
4.4 Combined cluster 66
4.4.1 Spatial distribution of the combined clusters 66
4.4.2 cluster characteristics of the combined clusters 68
4.4.3 Defining characteristics of combined clusters 71

5 Spatial policy insights for clustering 73

6 Discussion of methodology and results 76
6.1 Result discussion 76
6.2 Limitations 78

7 Conclusion 80

Appendix 81

List of Figures

1.1 A overview of different areas of interest in the province of South Holland . 3
1.2 Selection process for papers used in the literature review 4
1.3 Modeling cycle Brooks and Robinson (2000) 9
1.4 Process for finding suitable characteristics for clustering 11
1.5 Research flow diagram of thesis 12

2.1 Selection process for papers concerning energy and spatial characteristics 14
2.2 A overview of the interviewees chosen to represent each topic 18

3.1 Transformation methods used in creation of ESDL dataset 29
3.2 Comparison completeness of land use and demand datasets 30
3.3 Value distribution for mean electricity demand of households and industry 31
3.4 Overview of the K-prototype algorithm 33
3.5 Pseudocode of script used for clustering algorithm 34
3.6 An overview of the silhouette scores of the energy and built environment clusters 35
3.7 An overview of the silhouette scores of social and total clusters 36
3.8 An overview of the silhouette scores of Huang and Coa 37
3.9 An overview of the silhouette scores 25 and 100 iterations 38
3.10 Difference in the clusters when including wind and conventional electricity supply 40

4.1 Spatial distribution of energy clusters 42
4.2 Number of areas per energy cluster 43
4.3 Spatial distribution of energy clusters in The Hague 44
4.4 Energy demand distribution per energy cluster 45
4.5 Required temperature for heating per energy cluster 46
4.6 A overview of the characteristics of the first three energy clusters 46
4.7 A overview of the characteristics clusters 3 to 6 47
4.8 A cumulative plot of the match between areas and clusters based on their energy and built environment characteristics 49
4.9 Primairy match between areas and the energy clusters 50
4.10 Locations of build environment clusters 52
4.11 Number of areas per built environment cluster 52
4.12 Locations of build environment clusters 53
4.13 Mean electricity demand and number of houses per built environment cluster 54
4.14 Build year of average house per built environment cluster 54
4.15 Overview of the first, second and third cluster of the built environment . . . 55
4.16 Overview of the last cluster of the built environment 56
4.17 A cumulative plot of the match between areas and clusters based on their built environment characteristics 57
4.18 Primairy match between areas and the built environment clusters 58
4.19 location of social clusters 59
4.20 Number of areas per social cluster 60
4.21 Spatial distribution of social clusters in The Hague 60
4.22 Percentage of rental housing and average income per cluster 61

4.23 Average yearly household energy demand social clusters 61
4.24 Overview of social clusters 62
4.25 A cumulative plot of the match between areas and clusters based on their social characteristics 64
4.26 Primairy match between areas and social clusters 65
4.27 Location of combined clusters 66
4.28 Number of areas per combined cluster 67
4.29 Spatial distribution of social clusters in The Hague 68
4.30 Overview of combined clusters 69
4.31 Build year of average house per combined cluster 70
4.32 A cumulative plot of the match between areas and clusters based on their combined characteristics 71
4.33 Primairy match between areas and combined clusters 72

5.1 Steps for use of cluster analysis within spatial planning 75

List of Tables

1.1 Overview of the papers used in the literature analysis and for which of the findings they were used. 6

2.1 An overview of the reviewed papers and their corresponding findings . . . 16
2.2 Overview of characteristics for use during interviews. 17
2.3 Overview of the use of characteristics during the interviews. 20
2.4 Characteristics defining energy and spatial evironments 24

3.1 Overview of characteristics excluded from the cluster analysis 26
3.2 Overview of characteristics included in the cluster analysis 27

1 | Introduction

1.1 Background

To limit the effects of climate change, the Dutch government is striving to become carbon-neutral by 2050 (Ministerie van Economische Zaken en Klimaat, 2022). To reach this goal, changes to the current energy system that focus on a reduction in greenhouse gas emissions are required. Performing these changes comes with a set of challenges. Altering and creating new energy infrastructure requires significant investments both in capital and workforce. It is expected that this will require an investment of about 102 billion euros in the energy grid alone (PCW, 2021). Besides the cost, it is also not always easy to make these changes in urban areas where competing interests exist for the use of limited space. Combined with other challenges like limitations in the current network capacity and possible mismatches between demand and supply, this creates a complex problem that requires a good understanding of the energy system. For this reason, it is essential that decision-makers understand the current state of the energy system and how changes can affect how it functions.

One of the parties responsible for reaching the goal of net zero emissions is the province of South Holland. As a provincial government, the province of South Holland is tasked with planning the provincial energy transition and assisting municipalities in their energy transition planning. As a part of their provincial energy transition strategy, the province of South Holland wants to increase renewable energy production and reduce the overall energy demand (Provincie Zuid Holland, 2022b). This is particularly challenging as South Holland has the largest overall energy consumption of any province in the Netherlands (Rijksoverheid, 2021). This is due to the large urban centers and industrial areas that are present within this province. Besides that, there are also regional differences in energy consumption, which further complicate this process.

Reaching the energy transition goals requires the province to be able to make changes to the current energy system. In the Netherlands, the energy distribution networks are the responsibility of distribution system operators (DSO) and transport system operators (TSO) (Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, 2016). Moreover, energy supply and demand are managed by private entities. This means that the province is not directly involved in the management of the energy system. However, they can influence the energy system through policy decisions.

An important policy tool that the province can use to influence the energy system is spatial planning. Depending on spatial policies, the energy landscape of an area can change. This is because not all types of built environments require and/or provide identical amounts of energy (S.-H. Wang et al., 2018). An example of this would be industrial energy consumption, which is often higher than residential energy consumption for an area of the same size. This means that changes to the built environment can influence the energy characteristics of an area and vice versa. The provincial government can use this to reach the energy transition goals by including sustainable energy within its spatial policies.

Including these energy goals in their spatial policies requires the province to identify which policies should be enacted and what locations are suitable for each of these policies. This requires a clear understanding of the current energy landscape, the built environment and their relationship. Though information is available on individual aspects, like which areas are suitable for district heating, limited information is available on the relationships between the energy landscape and the built environment. This limits the spatial planners in their ability to include all relevant aspects of the energy system and the built environment in their decision-making. Decision-making based on limited information might result in ineffective usage of the existing and planned energy infrastructure (Asarpota & Nadin, 2020). It also makes it harder for the province to identify areas with similar characteristics. This information could be used to create a joined energy transition strategy instead of making an individual plan for each neighborhood. This would reduce the overall complexity of the energy transition and allow for more effective spatial planning. As both of these limitations affect the province's ability to reach its energy transition goals efficiently, it is essential to further analyze the relationship between energy and the built environment.

This thesis looks at which relationships exist between the spatial and energy characteristics of an area and how this information can be used to inform choices regarding spatial planning. Although this thesis looks specifically at the case of the province of South Holland, the methodology used within this thesis is universal. This means that it can also be used to provide similar insights for other provinces, municipalities, or researchers in the fields of energy and spatial planning.

1.2 South Holland

South Holland is particularly interesting for this type of analysis because of the wide variety in both energy and built environment characteristics from one area to the other. This allows for a broad analysis of the similarities and differences between different types of built environments. In Figure 1.1, a map of the province of South Holland can be seen. This shows that the northwest and middle of the province of South Holland has an urban area consisting of multiple cities in close proximity to each other. This area includes the cities of Rotterdam and The Hague. These are the second and third largest cities in the Netherlands, with an approximate population of 650.000 and 550.000 residents, respectively (Centraal Bureau voor de Statistiek, 2023b). Around this urban center are more rural and less populated areas in the (north)east and south of the province. This contrast between urban and rural areas shows that there is a variation in the type of built environment within the province. Besides the distinction between urban and rural, the province of South Holland also has other areas that are of interest. One of these is the port of Rotterdam and its surrounding industry, which are situated along the Maas River. This area is the biggest port in Europe and the second-largest industrial area in the Netherlands (Centraal Bureau voor de Statistiek, 2018). Because of this, both the type of land use and energy landscape differ from most of the rest of the province.

The last areas that are of interest are the regions below The Hague and between The Hague and Rotterdam. These areas are Westland and Oostland and have a high concentration of greenhouses. This means that although they can be defined as rural areas, their gas and electricity usage is significantly higher compared to other rural areas within the province.

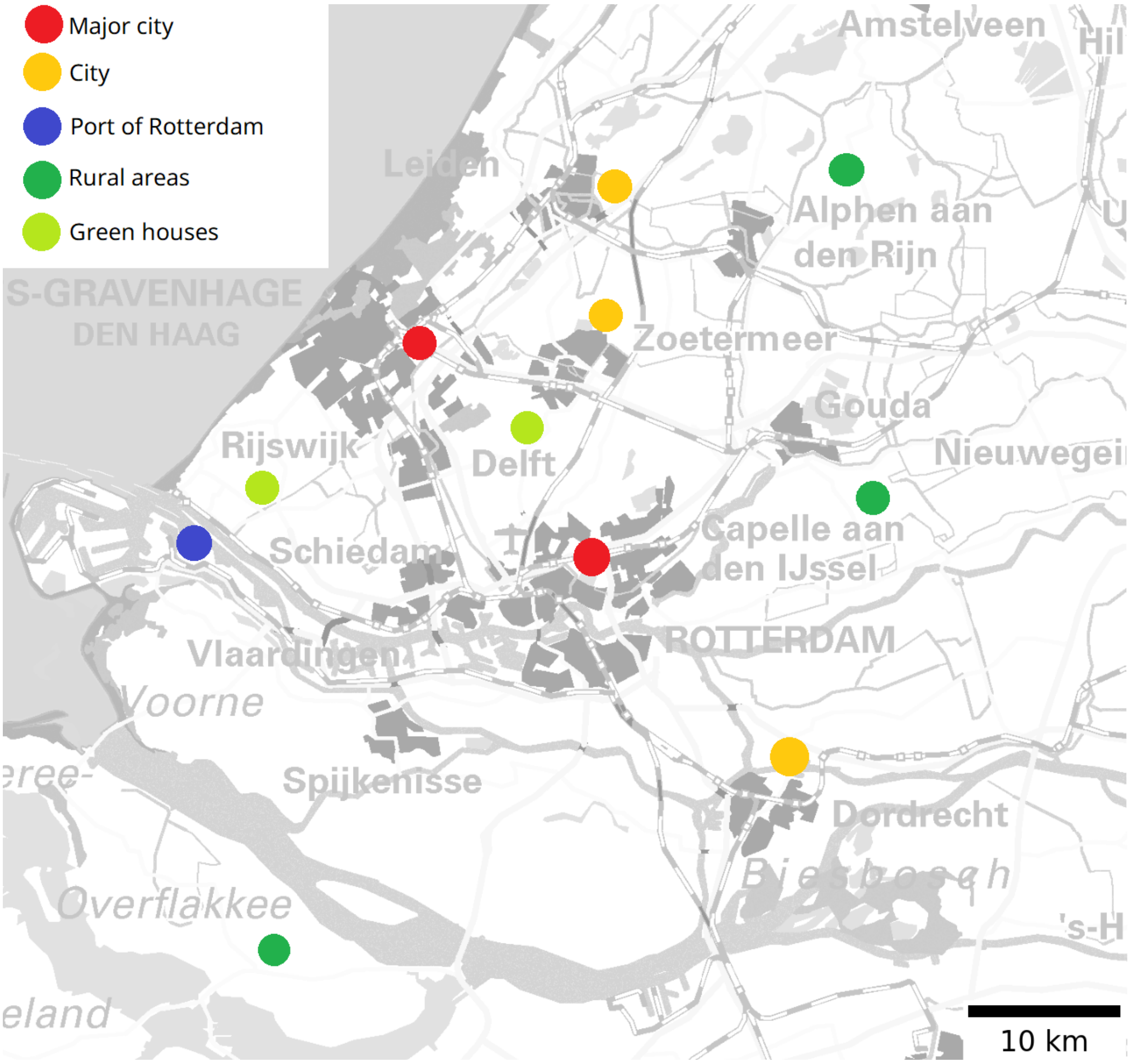


Figure 1.1: A overview of different areas of interest in the province of South Holland

1.3 Existing research

A literature review was used to gain insight into existing research regarding the combination of spatial characteristics and energy characteristics. Papers on spatial analysis in energy systems were used as these show the current methods and knowledge about this topic. In total, twelve papers were reviewed. A systematic approach was used to filter the papers. First, a broad search term was used in the literature search engine Scopus. This resulted in 877 papers that referenced a combination of landscape and energy. As this set of papers was too big for a detailed analysis, only a small subset was reviewed that looked promising based on their titles and abstracts. From these, a set of relevant papers was identified, and their keywords were used to create a search term that was more tailored to the subject. In total, 87 papers were identified using the search term *"TITLE-ABS-KEY((Spatial-planning OR Spatial-analysis) AND (Energy-model OR Energy-system OR Energy-network OR Energy-planning OR energy-cluster) AND (Demand OR Consumption OR Production) AND (Urban or Regional))"*. Based on the title and abstracts, these papers were further filtered. All papers that did not make use of spatial analysis combined with energy demand or supply were removed. This resulted in a set of 32 papers. From these papers, eleven were chosen for the literature review. This was done based on the level of detail in their description of the spatial analysis method. After this, one more paper was added by Kraaijvanger et al. (2023). This paper did not meet the search requirements but was found during the literature review and did contain information relevant to the topic. For this reason, it was added after the search.

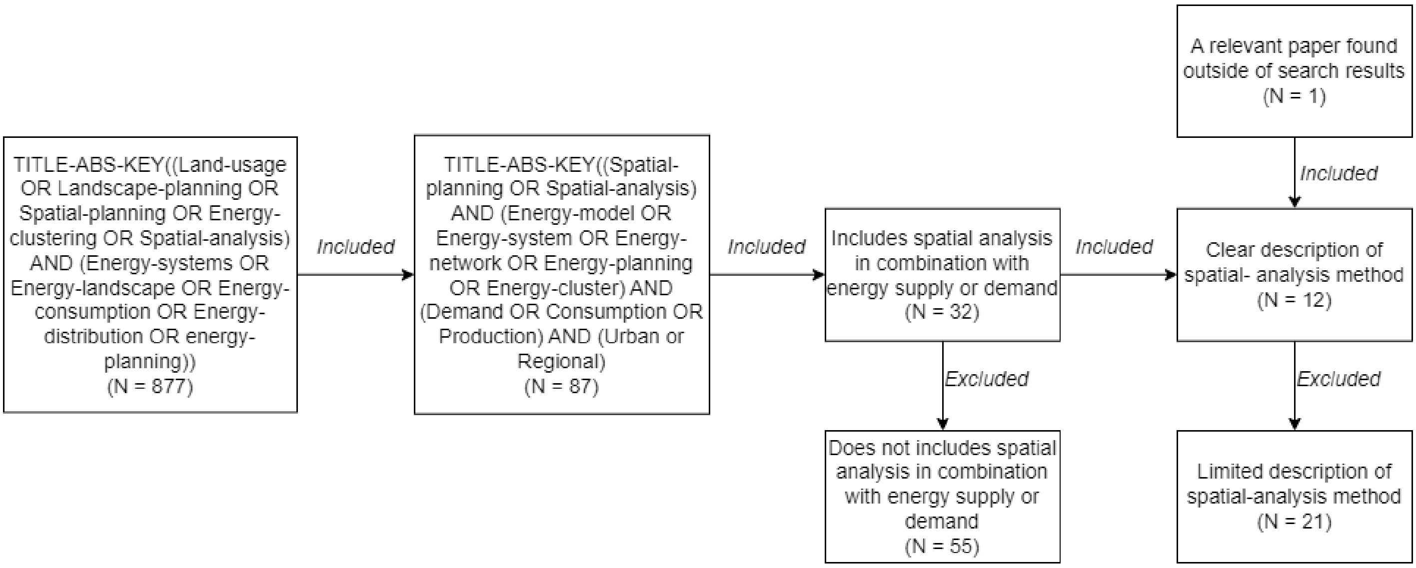


Figure 1.2: Selection process for papers used in the literature review

1.3.1 Current use of energy characteristics in spatial analysis

Most of the findings that will be discussed in this section are a combination of details from two or more papers. For this reason, Table 1.1 shows an overview of each of the papers and their findings. Each of these findings is numbered in the text.

The analysis showed that spatial analysis is currently already being used to analyze the relationship between land and energy both in and outside of the Netherlands. The scale of these analyses ranges from individual neighborhoods to entire countries. To identify the current use of spatial and energy data, an overview was made of the types of characteristics that were used in the papers. To perform this analysis, a separation was made between energy characteristics and spatial characteristics. As this literature review focuses on combinations between energy and spatial characteristics, all papers included at least one characteristic of both. For energy, the main characteristics used in the papers were mean energy demand and mean energy supply. This focused on electricity, gas, or heating, depending on the paper(1). An example of this can be seen in Moya et al. (2022), which looks at the energy consumption per region in Ecuador.

For spatial characteristics, two types of characteristics can be defined. One type of spatial data that was often used was built environment characteristics. This includes but is not limited to housing type, land use and housing density. Besides this, it also includes characteristics of the natural environment, like elevation, average solar radiation, and average wind speed(2). Although solar radiation and windspeed can both be identified as energy and spatial characteristics, for this thesis, a spatial definition will be used. This is because a high potential for energy production does not always match with high actual production, as other elements like land use and accessibility also play a role.

The second type of spatial characteristic that can be identified is social characteristics. This includes information regarding people's attitudes, income, and the inclusivity of the energy systems(3). Although examples of these types of characteristics can be found, their use is limited. Only Kraaijvanger et al. (2023) focuses on this type of characteristic within their analysis, with other papers like Moya et al. (2022) using but not focusing on similar characteristics. This does not have to mean that this topic is irrelevant however, as this might signal that more research is required or that the current search query might not be optimized to find papers regarding this topic. One thing to note is that the definitions and categorizations of these characteristics are not universal, as no unified framework for spatial analysis in energy systems exists. Depending on the research goal and what stakeholders are involved, some characteristics might be relevant in one context but not in another. They can also be interpreted differently depending on context. For this reason, it will be important to identify what characteristics will have to be included.

Currently, research into spatial planning and energy systems is used for several goals. The most frequent type of research was the identification of the potential for renewable energy sources within an area(4). An example of this was the paper by Q. Wang et al. (2014). In this paper, the renewable potential of the Fukushima area in Japan was mapped using spatial data like wind and solar potential per year. Most of these papers only use geographical location as a spatial characteristic. However, some papers like Kraaijvanger et al. (2023) also include other spatial characteristics like attitude toward renewable energy and the ability of an area to invest in renewable energy sources.

Another area where spatial analysis is already being used is in the prediction of current and future energy demand(5). In Fichera et al. (2016), an example of this can be seen with an analysis of energy demand within a neighborhood based on housing characteristics. Here, energy demand in an area is measured by combining the average energy demand per housing type with the distribution of housing types in an area.

The last way that spatial planning is currently used in an energy setting is to determine the relationship between energy and spatial characteristics(6). Examples of this can be found in Chen et al. (2020) and Moya et al. (2022). Within these papers, the energy demand of an area is compared to other factors like temperature and population density to see how these characteristics are related. Interestingly, almost all papers approach their analysis within a static build environment. This means that research into the effect of changes to the built environment on the energy supply and demand is currently lacking.

In general, there is no consistent analysis method within the papers, with every paper providing a different method. However, some similarities can be seen between their methods of data analysis. Based on these similarities, two different types of data analysis were identified. The first type combines characteristics and constraints based on spatial data to identify a certain area's supply and/or demand(7). In Oudes and Stremke (2018), this is done by combining the energy potential with stakeholder constraints to show the potential of energy sources in an area. The second spatial analysis method makes use of clustering(8). Here, different areas are clustered based on their energy and spatial characteristics. Within this method, areas are grouped based on similarities in characteristics. This analysis method allows for identifying differences and similarities between areas based on their characteristics. Using this method, it is also possible to see the relationship between characteristics by comparing their values within different clusters. From the reviewed papers, only Fonseca and Schlueter (2015), Unternährer et al. (2017) and Kraaijvanger et al. (2023) made use of this method.

Table 1.1: Overview of the papers used in the literature analysis and for which of the findings they were used.

Author	**1 Energy**	**2 Built**	**3 Social**	**4 Renewable**	**5 Demand & Supply**	**6 Relations**	**7 prediction**	**8 Clustering**
Chen et al. (2020)	x	x				x	x	
Fichera et al. (2018)	x	x			x		x	
Fonseca and Schlueter (2015)	x	x			x			x
Leduc and Van Kann (2013)	x	x			x		x	
Moya et al. (2022)	x	x	x		x	x	x	
Oudes and Stremke (2018)	x	x		x			x	
Ramachandra and Shruthi (2007)	x	x		x			x	
Unternährer et al. (2017)	x	x		x	x			x
N. Wang, Verzijlbergh, et al. (2020)	x	x		x			x	
Q. Wang et al. (2014)	x	x		x			x	
Yao and Zang (2021)	x	x		x	x		x	
Kraaijvanger et al. (2023)	x	x	x	x			x	x

1.3.2 Knowledge gap

The literature review shows that there is already research with regard to the relationship between energy and the built environment. However, none of these are directly focused on the province of South Holland. This means that an analysis focusing on this region is required. Although the literature does provide some insight into this type of analysis, there are some important aspects currently missing. In this chapter, an overview of these aspects is given. Based on this, a knowledge gap is identified.

The analysis showed that there are a multitude of characteristics that can be used to define energy systems, the built environment, and the social environment. However, these characteristics are currently being chosen based on the scopes that are used within the papers. As each of these papers looks at the energy system from a different perspective this means that the importance and presence of each characteristic may vary between papers. This in itself is not a problem as focusing on a small set of characteristics can lead to a more detailed analysis of the energy system. However, this does mean that there currently is no definition of what characteristics are relevant for spatial planning. A universal list of characteristics would not be possible as this would either include too many or too few characteristics depending on the research goal for which they are used. However, a list tailored specifically to the case of the province of South Holland could help to show what characteristics are relevant for this context. Furthermore, this could also help to provide a base framework of what types of characteristics can be used to analyze energy and built environments in other contexts.

This use of a limited set of characteristics can also be seen in the different methods for performing spatial analysis within energy systems with most of these methods being used within a narrow scope. This means that though individual relationships like electricity demand and housing type are already mapped, there is currently no way to see these relations within the broader context of the entire energy system. Because of this, there is no insight into how different sets of characteristics might influence the outcome of a spatial analysis. For example, how focusing on social characteristics might lead to different insights than focusing on built environment characteristics.

Combined this shows that although there is information about the energy and built environment this is mostly focused on specific narrow contexts. This limits decision-makers like the province of South Holland in their ability to get an overview of the current energy system and the spatial relationship that affects this current system. This has the potential to lead to decision-making based on trends in a subset of characteristics, which might unintentionally negatively influence other aspects of the system. To adequately describe the energy system of the province of South Holland thus requires a broader analysis of both energy and built environment characteristics. It also requires insights into how choosing a certain set of characteristics influences the insights that can be gathered and thus the spatial policies.

Combined, these limitations show a lack of knowledge of how the energy system and built environment are influenced by each other. Although none of the models that were used within the papers could directly be used to provide this knowledge, clustering does show potential. However, there are some aspects missing from the current use of clustering that are required for this thesis. Current uses of clustering in the reviewed papers focus on a neighborhood or city. This means that there is currently no research available about the use of clustering on a provincial level. As a result, insights into some aspects, such as the difference between rural and urban environments, are currently not mapped. Besides that, there is also a lack of research into clustering using a broad set of energy and built environment characteristics. Current papers mostly provide a method suited for one specific question. For example, Fonseca and Schlueter (2015) only focuses on clustering based on housing conditions. However, for this topic, insight into a broad set of characteristics is necessary as this allows for a more complete picture of the energy- and built environment of the province of South Holland. Lastly, though these clustering models do give insight into relations between energy characteristics and spatial characteristics, they only give limited information about how this could be used within spatial planning. This means there is still a gap between the theoretical results and the practical implementation within spatial planning. Based on this analysis, it can be seen that there is a lack of knowledge on how the relationships between energy-, and built environment can be mapped in different areas based on a broad set of characteristics from both environments.

1.4 Research approach

Looking at the problem defined in the previous section, it is clear that the goal is to identify the relationship between energy and built environment characteristics in different areas. To further the research into this topic, a cluster analysis of the province of South Holland was performed. As these relations are dependent on the focus of the research it will be important to see whether these findings change when looking at different sets of characteristics. To reach this goal, the following research question has been answered: ***"How can different sets of characteristics be used in spatial clustering to identify the relationship between the energy system and the built environment for use in spatial planning?"***

The literature review showed that there is currently limited information available regarding this topic within the context of a province. For this reason, it was necessary to use an approach that is able to show these relationships. Because of the complex nature of the built environment and the energy systems, these clusters cannot be obtained purely by quantitative analysis. An in-depth analysis of the system and its characteristics is required to be able to cluster areas. Therefore, a modeling approach was used. A drawback of this approach is that as not every aspect of a real-world scenario can be modeled, the results might not always be a one-on-one match with how the system reacts in the real world. Still, the modeling approach is the most suitable method to reach the research goal as it is the only method that can be used to accurately analyze and show spatial clusters within a complex environment such as South Holland.

To perform the modeling process, the cycle of Brooks and Robinson (2000) was used. This divides the modeling process into four steps, as shown in Figure 1.3. The first step focuses on the conceptualization of the model. Conceptualization shows how the real world will be represented within the model and what considerations and constraints come with the representation. This is then followed by the creation of a model based on the conceptualization that can be used to answer the research question. The third step is to analyze the results of the model. This can then be used to create an understanding of the modeled system. Lastly, the gathered understanding can be used to implement changes in the real world. It is important to note that this process is not linear, as results from the next step can influence the previous steps, meaning multiple iterations are required.

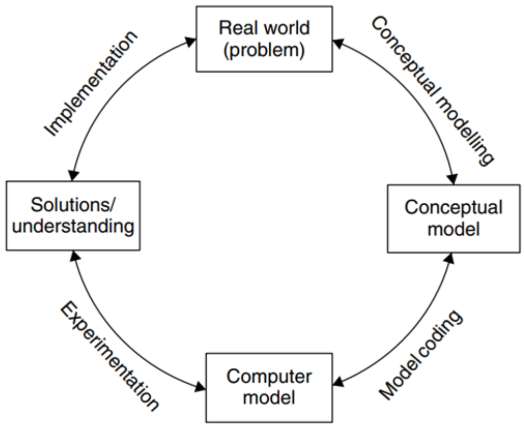


Figure 1.3: Modeling cycle Brooks and Robinson (2000)

Based on the method above, a research approach for this thesis was defined. As each of the steps taken within the approach was dependent on the previous steps, only a general overview will be given within this section. A more in-depth description of the methods used will be provided at the start of each chapter. In the approach, the conceptualization step was used to find out what characteristics should be used to define the energy- and the built environment. As noted before, this was necessary because no single definition of what characteristics define each of these systems exists. Here, a combination of literature and interviews was used. After this, a clustering model was created that could be used to create clusters based on these characteristics. A clustering model in this context refers to a model that can compare the characteristics of predefined areas, like the province of South Holland. The model can then group these areas based on their similarities and dissimilarities. After the model and results where created, the third step was used to analyze the clusters to get an understanding of the spatial relations between clusters and the relationship between different characteristics within clusters. The last step was to see how these insights could be used to guide decision-making for spatial planning. Though this thesis has applied the approach specifically to the case of the province of South Holland, it should also be applicable in other contexts with a similar goal. To further define how this method has been used, the following section will discuss the sub-questions that where used to answer the main research question. In total, three sub-questions were created, which link to different parts of the modeling cycle. Here, the first question focuses on conceptualization, the second on modeling and analysis, and the third on implementation.

1.5 Research questions

For the first subquestion the goal was to conceptualize what is meant by a cluster. As this definition is scenario-dependent, the goal was not to determine an overall definition of a cluster but to find one that fits within the context of the research. As this subject looks into the relationship between energy systems and the built environment, a list of characteristics defining both was formed. For this reason, the sub-question was as follows: ***"What characteristics of the energy system and build environment are relevant for defining clusters of the province of South Holland?"***.

Answering this question required data about two topics. First, what energy characteristics are relevant for defining the energy system within the province of South Holland? Secondly, what spatial characteristics are appropriate for defining the spatial environment within the province of South Holland? To answer these questions, information was used that has not yet been provided by the original literature review. For this reason, further literature research was performed to analyze what characteristics are currently used to define energy systems and the built environment. This was used to create a preliminary list of characteristics. To identify the relevance of each characteristic and to validate if all characteristics were identified, interviews have been used. For these interviews, a protocol was used that can be found in Appendix 1. During the interviews, the interviewees were first asked what characteristics they used themselves. After this, they were presented with a list of characteristics collected from the literature review. Using this list, they were asked to identify which of these characteristics they find relevant, which of the characteristics they might find irrelevant, and what characteristics might be missing. This way, an overview of relevant characteristics based on the interviews can be created.

In total, seven interviews were held. As the energy transition is a complex problem with multiple different stakeholders involved, it was important to involve as many relevant views as possible. For this reason, the interviews were divided as follows. Four interviews were held within the province of South Holland with experts in spatial planning, provincial energy policy, and heating infrastructure. These were combined with three interviews with experts outside the province about the local electricity infrastructure, national energy infrastructure, and heat transition. A combined list of characteristics was created using the insights provided by the literature review and the interviews. This list was used as a first draft of relevant characteristics for the creation of clusters. This draft was then discussed with the province of South Holland to see if there were any characteristics that the province did not agree with or needed to be included. These changes were then compared to the literature and interviews to assess their validity. Based on this, some changes were made, which led to the final list of characteristics. This process of co-design was used to create a shared definition of what characteristics are relevant for energy and spatial clusters. These characteristics could then be used within the data collection process. They could also be used to create an overview of general groups of characteristics that can be used to define energy clusters in future research. An overview of this process can be seen in Figure 1.4.

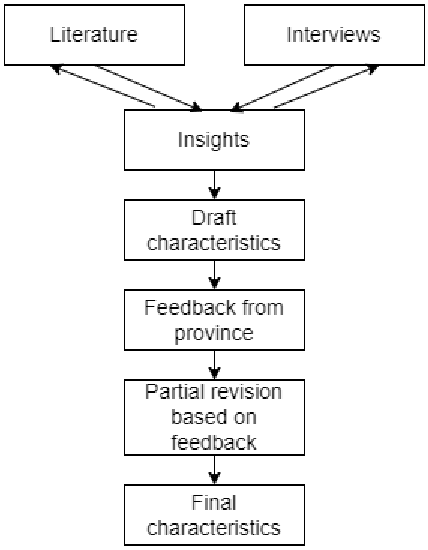


Figure 1.4: Process for finding suitable characteristics for clustering

After creating the definitions for the energy system and built environment, a model was designed to show the relations between energy and spatial characteristics. This was done using the following sub-question: ***"What clustering methodology can be used to identify the relationships between energy and spatial characteristics?"***. To create this model, two things were needed. First, a clustering algorithm had to be selected. This was done using the literature combined with the results of the first sub-question. Besides this, a data collection process was performed. This was done based on the characteristics defined in the first sub-question. Only open-source data was used, which was provided by open-source databases like the province of South Holland's open data portalProvincie Zuid Holland, 2023. This was done to make sure all data used in the thesis can be published. Based on the clustering method and the input data, the model was created. This model was produced within Python and uses the Energy system description language(ESDL). This is a description language that can be used to create energy models in a uniform matter. This way, the dataset from this thesis can easily be used in other research on similar topics. Using this model, a clustering analysis was performed. This looked at the current state of the energy system and built environment by identifying and comparing different clusters within the province of South Holland.

In the last part of the research, the following question was answered: ***"What kind of policy insights can be identified based on the proposed method of identifying and analyzing energy clusters?"*** To answer this question, data on how energy cluster data can be used in spatial planning was required. For this reason, a group discussion was held with spatial planners from the province of South Holland. Within this meeting, the focus was on discussing the results of the analysis. The outcomes of these interviews were used for two purposes. Firstly, they were used as a validation step for the model. This was done to improve the reliability of the results. Secondly, as the interviewees were experts in spatial planning and/or energy, they could give insights into how the results and research methods can be used within policy-making for spatial planning. This, combined with the insides from the previous chapters, has been used to formulate ways to use energy and spatial clustering results in spatial planning policies.

1.6 Research outline

Combined, the three sub-questions were used to answer the main research question. This report will show the results of each of these questions and how they were obtained. In Figure 1.5, an overview of the inputs, outputs, and methods for each sub-question is shown. This also shows the relationship between the questions. Each of these sub-questions will be answered within its own chapter. In chapter two, an overview of the literature search and interviews will be given. It will also show the characteristics selection process, which has led to the eventual list of characteristics. After this, chapter three will discuss the data collection process and the model creation. Chapter four will then show the results and analysis of this cluster model. In chapter 5 the discussion can be found. This is before the last research question as this question requires findings from the discussion. These findings will then be used in Chapter 6 in combination with the expert feedback to define a methodology for using this type of analysis in spatial planning. Lastly, chapter 7 will discuss the scientific implications of this thesis as proving a conclusion regarding the entire thesis.

In the thesis, various visualizations will be included that would require a full page or even a3 page size to be fully readable. This would either affect the readability of the document or require many of the pictures and plots to be moved to the appendix. To prevent this the choice was made to make this document digital readable only. This means that some pictures will need to be zoomed in to be fully readable.

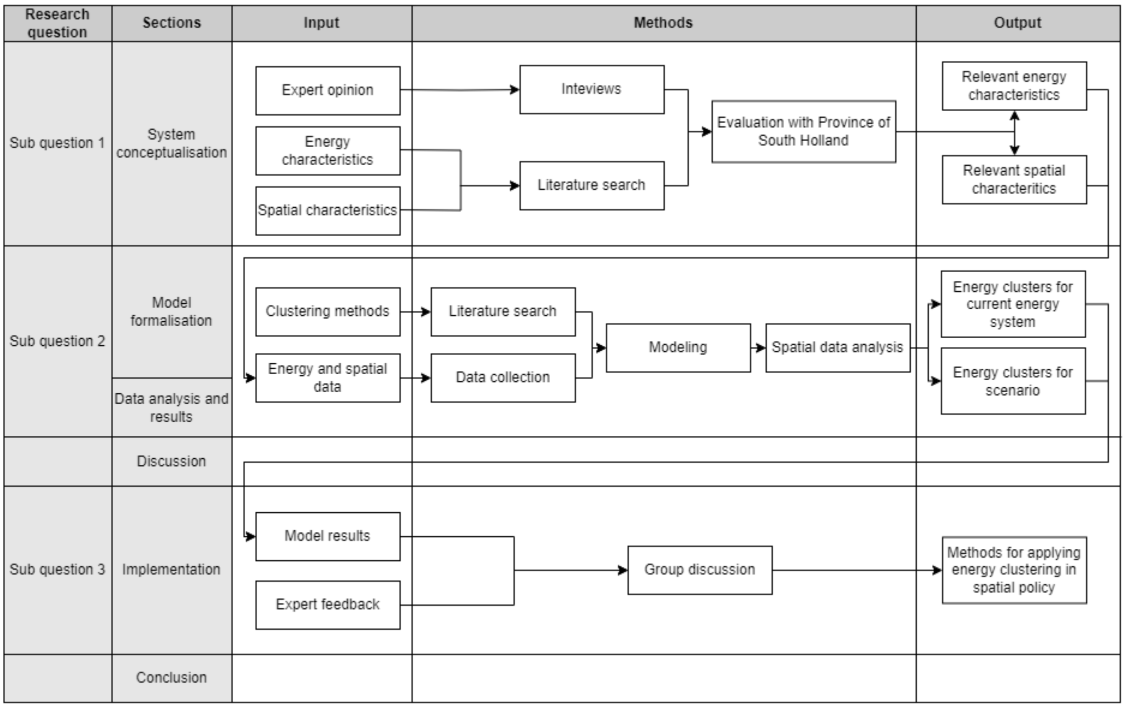


Figure 1.5: Research flow diagram of thesis

2 | Defining relevant energy and build environment characteristics

Before the creation of a clustering model was possible, it was important to map what characteristics should be included for both the energy system and the built environment. In this chapter, the results of the literature review and interviews will be discussed. This will then be used to create a list of characteristics that define energy systems and the built environment.

2.1 Literature

To define the relevant characteristics of the energy and built environment, a literature review was performed. As there was no literature available specifically focusing on defining what characterizes the energy and built environment, a broad literature review was performed. In this literature review, multiple different papers looking at a combination of energy and landscape were compared to see what characteristics were often used. This way, often-used characteristics were defined, and a preliminary list of characteristics was made.

2.1.1 Review process

This literature review focused on finding papers that modeled parts of the energy system and spatial environment to see what characteristics were used. The review looked for characteristics that were both relevant in a spatial and energy setting. For this reason, a combined search term was used to find both sets of characteristics at once. This limited the risk of finding characteristics that were inrelevant in the scope of this research. The collection of the papers was done using Scopus. The topic of finding spatial and energy characteristics is quite broad, resulting in more results than could be analyzed. For this reason, a narrower search term was created based on keywords used in papers that fit within the goal of the literature review. The results were also filtered on a geographic size as characteristics like the number of rooms within a house might be relevant for determining energy demand in a household but less so for energy use within a province. This led to the following search term: *"TITLE-ABS-KEY ((energy-system OR energy-landscape or energy-network) AND (model OR simulation) AND (built-environment OR spatial-planning OR land-use) AND (regional OR provincial OR urban OR rural))"*. In total, this resulted in 135 results. As this search term was narrower, some characteristics might not have been included in the resulting literature. However, as this list was validated and expanded on using interviews, this should have had a low impact on the eventual list of characteristics. After further filtering the results based on their applicability, 23 papers were selected.

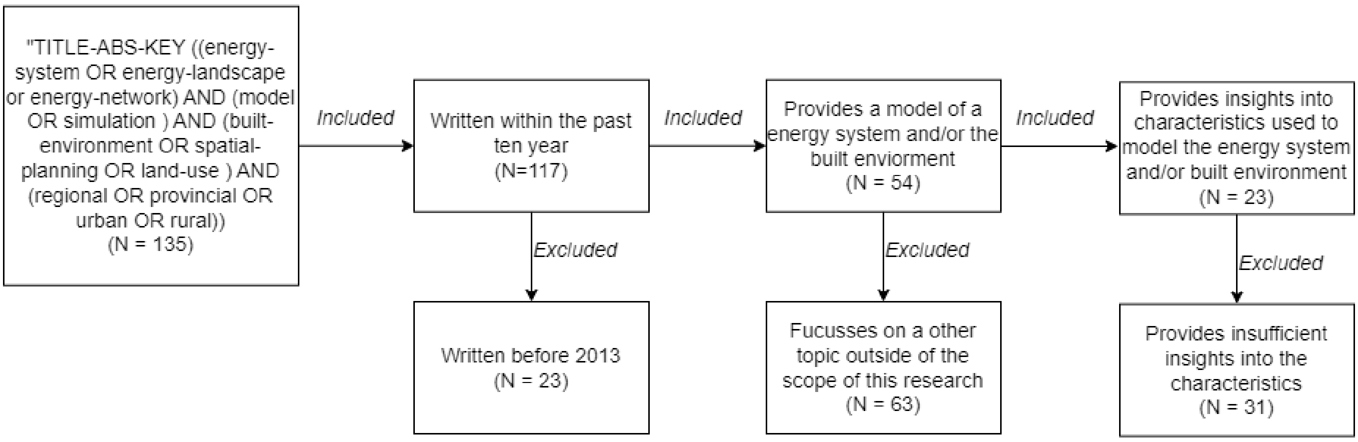


Figure 2.1: Selection process for papers concerning energy and spatial characteristics

2.1.2 Review results

To analyze the literature the three types of characteristics defined in Chapter 1 were used. These are energy characteristics, built environment characteristics, and social characteristics. In total, 28 different characteristics were identified that were used to characterize the energy and spatial landscape. Of these, fourteen focused on the energy sector, eleven on the built environment, and three on social aspects. An overview of all characteristics can be found in Appendix 2. This also shows in which of the papers the characteristics are present. Within this list, some characteristics are amalgamations or combinations of different characteristics used within the papers. This was done to prevent a long list with similar but slightly different characteristics, as some papers use the same characteristics but have different ways of defining them. Noteworthy is that the demarcations between different characteristic types in this list can sometimes be fuzzy. For example, the potential of renewable energy sources can be seen as a factor of the built environment and a factor of the energy system. The same can also be said for population density and population size, as these are characteristics of the built environment but are influenced by the social environment. This means that this distribution had to be further validated and refined using interviews.

Similar to the previous literature analysis most findings are derived from a combination of multiple pieces of literature. For this reason, Table 2.1 provides an overview of the findings for each paper. Looking at the literature, it was clear that the characteristics used to model energy systems are highly dependent on the goal of the model. Four different types of energy characteristics could be defined. The most prevalent of these are the demand characteristics(1) and the supply characteristics(2). These characteristics were used in 17 and 16 papers respectively. Most papers used the total energy demand with a region as the main characteristic for the demand characteristics, with some subdividing this based on energy carrier or type of user(3). For the supply characteristics, a distinction is often made between conventional and renewable energy. An example of this can be seen in N. Wang, Verzijlbergh, et al. (2020) where both renewable energy sources like wind and solar and conventional energy sources like coal and gas are included.

In this thesis, a similar distribution will be used to Sahoo et al. (2023), where renewables include solar, wind, geothermal, residual heat, and biomass. This means that gas, oil, coal, and nuclear are defined as conventional energy sources. Energy sources like tidal energy and hydro-power, are not included as these are not prevalent in the province of South Holland. With the current trend towards renewable energy systems, most of the papers including energy supply were focused on renewable energy supply(4), with only a small subset looking at the supply of conventional energy sources(5). Both supply and demand characteristics were often used as the main indicator of the energy landscape within an area.

Besides the demand and supply, the energy system can also be characterized by its distribution network. The main characteristic linked to this topic is network capacity which determines the amount of energy that can be transported within the energy network. The use of this topic was far less prevalent than supply and demand, as it was only included in seven papers(6). However, these papers showed the importance of this topic, as limitations in network capacity can have an effect on the ability of an energy network to match the supply to the demand within different areas. An example of this can be seen in Sahoo et al. (2022) where network capacity is used in the calculation of the regional energy mix. Lastly, the papers showed that storage availability also plays a role in defining an energy system. This storage can be in the form of different mediums, with the most prevalent mentioned in the papers being hydrogen-based storage and batteries. Only Jäger et al. (2016), N. Wang, Heijnen, et al. (2020), and N. Wang, Verzijlbergh, et al. (2020) included storage within their research. However, these papers do show the importance of storage technology in future energy systems where intermittent energy sources will become a large part of the energy supply.

The energy characteristics can also be grouped based on the type of demand. In most papers, a subdivision was made between energy used for heating(7) and energy used for electricity(8). In these cases, heating was often defined based on gas usage, district heating, or heat pumps. One thing to note is that this distinction only works for some characteristics. An example of this is gas supply which can be used for electricity generation and heating. Similarly, electricity can also be used for heating using electric heaters or heat pumps. For this reason, this definition is mostly applicable to energy demand.

Looking at the spatial aspect, most characteristics focus on the built environment. Within these, some form of land use was the main characteristic of most of the papers. Examples of land use used within the papers are agriculture, industry, and residential buildings. These characteristics were then compared with energy characteristics or used to predict the energy usage of an area(9). Besides this, a broad set of characteristics describing the building within an area was used, like the energy label, building sizes, and building type(10). These characteristics were mostly used to explain the relation between the energy demand of individual houses and the characteristics of these houses. However, in some cases, they were also included in the analysis of broader areas like neighborhoods or cities. An example of this type of analysis can be seen in Buckley et al. (2021). Here multiple building characteristics like building age are used to identify the energy resilience of the housing. As many different characteristics within a house might influence its energy usage, these characteristics together where named building characteristics.

Another characteristic often used to define the built environment is the potential for various renewable energy sources(11). Though this can also be seen as an energy characteristic, it highly depends on the physical environment. For example, solar potential depends on multiple factors from the built environment, like the number of roofs suitable for solar panels. These characteristics were often combined with conventional demand and supply to predict the future energy landscape. Besides built environment characteristics, there were some characteristics that focused on social aspects. similar to the literature provided in the previous chapter, these characteristics were far less prevalent than the other two types of characteristics. In total, three characteristics were identified within this category. These were income, employment rate, and acceptance of renewable energy sources. Of these, only income and acceptance of renewable were used as main characteristics in the models (Jäger et al., 2016; Yamagata & Seya, 2013), with employment rate only getting a mention(Yamagata & Seya, 2013). Though this might show the lower importance of social characteristics, it might also be caused by the search query needing to be more optimized to find this characteristic type. For this reason, it was important to validate the importance of this topic for use in spatial planning using interviews.

Table 2.1: An overview of the reviewed papers and their corresponding findings

Author	**1 Demand**	**2 Supply**	**3 Main & variable**	**4 Renewable**	**5 Conventional**	**6 Capacity**	**7 Heating**	**8 Electricity**	**9 Predict & demand**	**10 Find & relations**	**11 Map & potential**
Liu et al. (2023)		x		x				x		x	x
Sahoo et al. (2022)	x	x	x	x		x	x	x		x	x
Bao et al. (2022)		x		x				x			x
de Vries and Schrey (2022)	x	x	x	x	x	x		x			x
Sahoo et al. (2023)	x	x	x	x	x		x	x		x	x
Zhou et al. (2021)		x			x	x	x	x			
Buckley et al. (2021)	x	x	x		x	x	x	x	x		
Then et al. (2021)	x	x	x		x	x	x	x	x		
Bao et al. (2020)		x		x						x	x
Chen et al. (2020)	x		x				x	x		x	
N. Wang, Heijnen, et al. (2020)	x	x		x			x				
N. Wang, Verzijlbergh, et al. (2020)	x	x	x	x	x		x	x			x
Drechsler et al. (2011)		x		x							x
Bosch et al. (2020)											x
Eicker et al. (2020)	x		x			x	x		x	x	x
Fichera et al. (2018)	x	x	x		x		x	x			
Szarka et al. (2018)	x	x		x	x		x	x			
Afshari and Friedrich (2017)	x		x				x	x			x
Mutani et al. (2016)	x		x			x	x	x	x		
Jäger et al. (2016)	x	x		x			x		x		
Nolde et al. (2016)	x						x				x
Bustos-Turu et al. (2016)	x		x		x	x	x	x			
Yamagata and Seya (2013)	x	x	x	x	x		x	x	x		

Based on the insights gathered from the literature, a generalized list of characteristics was created. This shows what characteristics can be used to identify energy systems and the built environment. As not all characteristics were used the same amount of times further insights are required to define what characteristics are important. Furthermore, as the importance of these characteristics is scenario-dependent, their relevance in the case of the province of South Holland also had to be validated. To collect these insights interviews were held. The following list of characteristics was used as the basis for the interviews.

Table 2.2: Overview of characteristics for use during interviews.

Characteristic	**Unit of measurement**
Energy	
Peak demand electricity	GW
Demand electricity	GWh/year
Supply electricity conventional	GWh/year
Supply electricity renewable (wind, solar, biomass)	GWh/ year
Electricity network capacity	GW
Peak heat demand	GW
Heat demand	GWh/year
Heat supply gas	GWh/year
Heat supply geothermal	GWh/year
Capacity district heating	GW/year
Capacity gas network	GW/year
Energy storage	GW
Built environment	
Land use	Type (agriculture, industry, etc)
Building type	Type (flat, terraced house, etc)
Building density	#/km^2
Solar potential	GW
Wind potential	GW
Geothermal potential	GW
Population size	#/km^2
Population density	#/km^2
Social	
Average income	€
Employment rate	%
Acceptance of renewables	Scale(0-10)

2.2 Interviews

To further define what characteristics are relevant for characterizing the energy and spatial landscape, interviews were held. These interviews aimed to verify the current list of characteristics defined by the literature and to alter this list if any characteristics needed to be added or removed. Interviews were chosen over other ways of collecting these characteristics, like surveys, because this would allow for a more in-depth discussion about why certain characteristics should or should not be included.

2.2.1 Interview process

As discussed in the research method, a total of seven interviews were held. These interviewees were chosen based on two criteria. The first criterium was that experts from inside and outside the province of South Holland should be involved. This was to ensure that the final analysis did not only reflect the view of the province but also the view of other stakeholders within the energy system. The second criterium was that the interviewees should be able to provide insights into one of the topics that were prevalent in the literature review. Based on these criteria, a list of possible parties to interview was created. These parties where filtered based on their suitability to provide information about these prevalent topics. This list was then discussed with the province of South Holland, which provided the contact information of all interviewees. In Figure 3.2, an overview can be seen of all topics. It also shows which stakeholders were chosen to interview regarding each of these topics. The numbers given to these interviewees will be used as an identifier during the discussion of the interview results. Regarding the energy characteristics, a split was made between the heating and electricity systems. Within the province of South Holland, two people were found with knowledge on these topics, with a third one having both knowledge of energy and spatial characteristics. For the external expert, it was decided to interview the network operators as these parties are directly involved with the network and indirectly with the supply, demand, and storage. This meant that knowledge about a broad set of subjects regarding the energy system could be received with a limited number of interviews. Only interviews within the South Holland province were held for the built environment and social environment as both the decision-making and the experience regarding these topics lie with the province, limiting the need for external insights. In total, two interviews were performed with experts within this field. Both of these interviewees also had knowledge of the energy system.

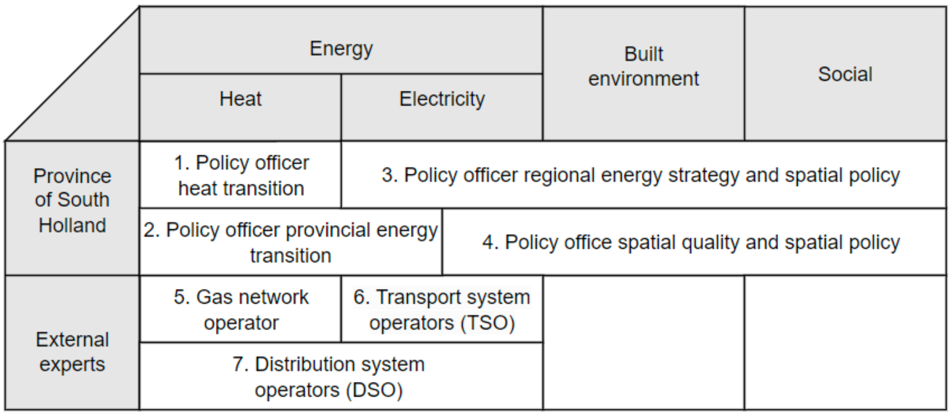


Figure 2.2: A overview of the interviewees chosen to represent each topic

For the interviews, a protocol was created which was used to ensure that every interview followed the same structure and used the same questions. This interview protocol can be found in Appendix 2. The interviews were separated into two sections focusing on current and future scenarios. The future scenario questions were originally meant to be used during the modeling to create a model that resembles a future energy situation. However, the future scenario was later removed from the scope of the thesis. For this reason, the future scenario will not be discussed in this section.

During the interview, the interviewees were asked a set of identical questions concerning energy, the built environment, and social aspects. The interviewees were asked to identify which characteristics they currently use for each of these topics within their work field. The goal of these questions whas to see what characteristics they would mention from the top of their head. This way, important characteristics could be identified without influencing the participants by showing them the characteristics overview from the literature. After this, the participants were shown the characteristics overview and asked to identify any characteristics they thought were missing. They were also asked to identify if there were any characteristics that they would not define as important for the definition of spatial and energy systems. The results of these two types of questions were combined to get a better understanding of what characteristics are relevant.

2.2.2 Interview results

Based on the interviews, a few observations can be made concerning what characteristics are relevant. An overview of the main characteristics provided within the interviews was made to visualize some of these observations. This overview can be seen in Table 2.3. This table shows an overview of how often characteristics were mentioned in the interviews. It also has an additional list of characteristics that were mentioned in the interviews but were outside the original set of characteristics. Though this already gave some insights into what characteristics were important, it did not show the full picture. This is because not everyone had the same expertise, leaving some subjects under-exposed in some interviews. To further explore these under-exposed subjects, a detailed analysis of the interviews was performed in the folowing section.

Table 2.3: Overview of the use of characteristics during the interviews.

Energy	#	Built environment	#	Social	#
Peak demand electricity	4	Land use	6	Average income	2
Demand electricity	5	Building type	4	Employment rate	0
Supply electricity conventional	4	Building density	1	Acceptance	5
Supply electricity renewable	4	Solar potential	4		
Electricity network capacity	4	Wind potential	4		
Peak heat demand	2	Geothermal potential	2		
Heat demand	3	Population size	0		
Supply Geothermal	2	Population density	1		
Supply gas	2				
Capacity district heating	3				
Capacity gas network	3				
Energy storage	4				
Aditional characteristics					
Availablity of renewable energy sources	2	Industry types	2	Energy poverty	3
Demand profile	4	Zoning	3		
Supply residual heat	1	Available space	5		
Supply green gas	2	Ownership	1		
Energy efficiency	1	Ground type	1		
Ramp up / down times	1				

Energy characteristics

Regarding energy characteristics, almost all characteristics from the literature review were mentioned during the interviews. After showing the characteristics list to the participants, there was a consensus that all characteristics were in some way relevant for identifying energy systems. However, there were some recurring remarks about the way they were written down. One recurring remark was to split up conventional and renewable energy sources into their subtypes. This was mentioned some energy sources, like nuclear, can be fuzzy to identify in the current overview. Interviewee 1 (2023), Interviewee 2 (2023), and Interviewee 5 (2023) also mentioned that heat demand is not only differentiated by peak and yearly demand but also based on temperature ranges. For this reason, heat demand should be further split into high-temperature, medium-temperature, and low-temperature demand. The supplied list of characteristics also defined the supply of gas as one characteristic. However, as pointed out by Interviewee 4 (2023) and Interviewee 5 (2023), this should be further split into the supply of natural gas and the supply of green gasses such as biogas and hydrogen. The same was the case with heat supply which only included geothermal in the original list but should also include residual heat. In general, the most discussed topics were the supply of renewable energy sources, energy demand, and the capacity of the energy networks. The capacity of the electricity grid was particularly often seen as an important factor as this is a limiting factor in many current energy infrastructure projects. This was not the case for gas infrastructure, as Interviewee 5 (2023) pointed out that due to the energy transition, only limited changes to the capacity of the gas network would be needed. However, it was noted that changes to this network would be required if a hydrogen-based gas network would be implemented in the future. This shows a clear difference in the importance of capacity depending on the energy source.

Besides the characteristics present in the list, some characteristics that were not part of the original list of characteristics were defined as relevant. The first of these characteristics was the availability of renewable energy sources. This characteristic was meant to signal how often a windmill or solar panel can create energy, as both are intermittent energy sources. The second characteristic was demand profiles. This characteristic was mentioned as defining energy demand by base and peak load is not always enough. An example was given by interviewee one, who noted that the demand for heat differs between seasons. As this difference can not be seen with the base demand or peak energy demand, it was suggested to add an extra characteristic to show these differences. The third characteristic mentioned was energy efficiency, which was mentioned by Interviewee 2 (2023) and was meant to signal how efficiently the consumers were using energy. The same interviewee also mentioned ramp-up time and ramp-down time. These characteristics describe the amount of time an energy source takes to change its energy output. This was mentioned as this can have an effect on what energy sources are suitable to supply energy depending on how often and how abrupt changes in energy needs occur within an area.

A shared feature of three of these characteristics is that they were meant to signify the changes in the energy system over time. This was a recurring theme during the interviews, as matching supply and demand at different time scales was seen as one of the challenges of the current and future energy system. As most types of cluster models are static in nature and lack a time aspect, it will be impractical to include this type of characteristic within the model. The above-mentioned characteristics were not included for this reason. Only one suggestion was made for the removal of energy characteristics. This suggestion was made by Interviewee 6 (2023) and was about the removal of the electricity network capacity. The interviewee noted that this was not because of a lack of importance but because of the complexity of implementing network capacity in a valid way. Here, the interviewee noted that ”Network capacity is very hard. Network capacity fluctuates about every second within the network”(Interviewee 6, 2023). This is because the available capacity on a cable can change from moment to moment depending on demand. Although this does not automatically mean that capacity should not be included in the list of characteristics, it is something to keep in mind for the design of the cluster model.

Built environment characteristics

Almost all built environment characteristics from the literature review were mentioned at least once during the interviews. The main exception is population density, with building density and population size being two other characteristics that were sparsely mentioned. After showing the characteristics to the interviewees, the general consensus was that most of the characteristics that were mentioned should be included in the definition of the built environment. The one characteristic that was deemed unimportant by Interviewee 7 (2023) was building density. One reason for this was the partial overlap with population density. It was also said that building density only has a relationship with energy characteristics for residential housing, as building density for commercial building and factories are not directly tied to energy consumption.

Though most characteristics were deemed important, some adjustments were proposed. Concerning renewable energy potential, an important distinction was made by Interviewee 7 (2023) between technical and feasible potential. This was because there can be a large difference between the technical and feasible potential as factors like building type and acceptance of energy sources also influence the feasible potential of renewable energy sources. For this reason, it was suggested to split this up into separate characteristics or clearly state which characteristics would be included in the research. As factors like the acceptance of renewable energy and regulation are hard to map, it was decided only to include technical potential within this paper.

Another characteristic was the availability of space above and below the ground to place energy infrastructure (Interviewee 1, 2023; Interviewee 2, 2023; Interviewee 4, 2023; Interviewee 6, 2023; Interviewee 7, 2023). This was mentioned as, in some mostly urban areas, there is limited space left for the placement of these types of infrastructure, making it a limiting factor. Regarding building types, it was also mentioned that the current characteristics focused on residential buildings. Here, the suggestion was made that the industry should also be subdivided into different types. This is because the energy characteristics might vary drastically based on the type of industry. Another aspect that was mentioned was the distinction between current and possible future land uses in the form of zoning. This was mentioned because some areas might currently have only one type of land use but can be suitable for multiple others. By analyzing these zoning policies, possible changes to land use in the future can be mapped. The last characteristic that was mentioned was ground type within the context of geothermal potential. However, as geothermal potential is already based on the ground type, this characteristic was combined with geothermal potential. One thing that was mentioned by Interviewee 4 (2023) was that the built environment can be characterized in many different ways depending on the type of analysis that is performed. This was visible throughout the interviews, where the interviewees analyzed the subject through the lens of a specific spatial planning project or energy project. For this reason, many characteristics were mentioned which were too small for the scope of this thesis. Examples of this were whether a building had a lift or not, regulations surrounding the size of windmills, and the possible locations for the placement of distribution stations. Though these characteristics would be useful on other levels of detail, they were impractical to include within this thesis as the level of detail required to analyze these characteristics would be too in-depth for an analysis of the entire province. These characteristics were either combined with an overarching characteristic like building type or not included.

Social characteristics

Regarding social characteristics, most participants stated that they did not regularly use them in their work. However, some interviewees did state that this type of characteristic could be useful for this type of analysis to show the underlying relationship between the social environment and the energy system. Because of the lack of use of social characteristics, only a limited amount of characteristics were mentioned during the interviews. The main characteristic that was mentioned was the degree of acceptance of proposed energy plans by the community. However, it was also mentioned this is mostly used during infrastructure projects to see whether the local population would support these projects. This means that this type of characteristic currently sees limited use within similar research as this thesis.

Another characteristic that was mentioned by multiple interviewees was the concept of energy poverty. The importance of this topic was stated due to rising energy prices leading to more people who are barely able to pay their energy bills. One characteristic that was mentioned regarding this topic was the percentage of rental housing to privately owned housing. However, Interviewee 7 (2023) also stated that this might not be a direct measurement of energy poverty as low-income homeowners can also struggle to pay their energy bills due to a relatively high mortgage compared to their income. Besides this, education level was also mentioned as a possible indicator. Looking at the list of characteristics that were provided to the interviewees, a general consensus was that all characteristics would be useful to define the social environment. However, Interviewee 1 (2023), Interviewee 2 (2023), Interviewee 3 (2023), and Interviewee 5 (2023) did not specifically mention the topic of acceptance of renewable energy sources instead focusing on the more generalized acceptance of energy projects. For this reason, this broader definition will be used.

2.3 Cluster characteristics

By combining the characteristics and insights from the literature and the interviews, a revised overview of the relevant characteristics was made. This list was then verified with the province of South Holland to check if they agreed with these definitions of energy systems and the spatial environment. Though the province agreed with most of the characteristics, the availability of space for energy infrastructure was added. Although this characteristic was mentioned within the interviews it was originally removed from the list as it would be hard to identify on the level of a province. However, the province noted that it was still a characteristic that was deemed important to define the built environment which meant that it was re-added to the list. Based on this feedback, the final version of the cluster characteristics was made. This overview can be found in Table 2.4. For the energy characteristics, most remarks from the interviewees were included. However, all characteristics that would require a time sequence like demand profiles, and ramp-up and ramp-down times were not added as these would be hard to include in any meaningful manner. This is because most cluster analysis methods do not include data from more than one time-step. This means yearly demand can be included as this gives one value to represent the entire year but demand profiles can not as this would require multiple values describing the energy demand at different times of the day, month, or year. For the built environment characteristics, almost all characteristics that were regularly mentioned were added. Besides this, some changes were made to clarify some parameters, like potential. For the social parameters, the concept of energy poverty was further defined by adding possible indicators for this in the form of education level, Average housing value (WOZ value), and percentage rental housing.

Table 2.4: Characteristics defining energy and spatial evironments

Energy	**Built environment**	**Social**
Peak demand electricity	Land use type	Average income
Total yearly demand electricity	Zoning policy	Employment rate
Peak heat demand	Building type	Education level
Total yearly heat demand:	Industry type	Percentage rental housing
- High temperature	Available space for energy infrastructure	Average attitude towards energy transition
- Medium temperature		
- Low temperature	Technical potentials:	Average WOZ value
Total yearly supply conventional electricity:	- wind	
- Coal	- Solar	
- Gas	- Geothermal	
- Nuclear	Population size	
Total yearly supply renewable electricity:	Population density	
- Solar		
- Wind		
- Biomass		
Total yearly supply conventional heat:		
- Gas		
Total yearly supply renewable heat:		
- Hydrogen		
- Biogas		
- District heating		
Network capacity electricity		
Network capacity district heating		
Network capacity gas network		
Energy storage capacity:		
- Hydrogen		
- Battery		
- Heat		

The list above gives a general overview of what characteristics are important to define the energy and built environment in the case of the province of South Holland. However, this list can not be directly translated to clusters as spatial data of the province of South Holland is required for each of these characteristics. The following chapter will describe how the list of characteristics was used to find datasets and how this was used to create the clustering model.

3 | Methodology for spatial clustering of energy and spatial characteristics

After the relevant characteristics for each of the topics of interest were defined, a cluster model was designed. This process was subdivided into two sections: data collection and model operationalization.

3.1 Data collection

3.1.1 Dataset selection

Based on the list of characteristics that was defined in the previous chapter, a search was performed for matching datasets. As the idea was to perform a spatial analysis, the goal was to find datasets that included a spatial factor and had the area of South Holland in their scope. The last criterium was that the datasets should be open source as this would allow the results to be directly presented within this thesis without any changes due to confidentiality. Based on these criteria, datasets containing a total of 24 characteristics were collected. Most of these datasets were collected through the open data portal of the Provincie Zuid Holland (2023) or through the database of the Centraal Bureau voor de Statistiek (2023a). Though the original goal was to find datasets for all characteristics, this was not possible for a variety of reasons. In Table 3.1, an overview of each characteristic that could not be included in the final dataset was given. For each of these, a reason is also provided for why they were excluded. Lastly, a list of stakeholders is given who are the most likely to have access to this data. This can possibly be used in future research to improve the clustering model by providing a more complete dataset.

Three reasons for the exclusion of characteristics were identified. The first reason was a lack of available datasets. In total eleven characteristics were excluded for a lack of publicly available data. This was either because the data was unavailable, the dataset included too little information to be useful within the analysis, or because the resolution was too big for the required analysis. Other characteristics, like the total yearly supply of gas, were excluded because there was no clear way to implement them into a clustering model. A total of four characteristics were excluded for this reason with Table 3.1 giving a more in-depth description for each.

The last reason why some characteristics were excluded had nothing to do with a lack of datasets. Instead, these characteristics were removed later in the modeling process during testing. Although there are datasets available for these characteristics they are still excluded, and therefore mentioned in Table 3.1. This is because this provides a full overview of all excluded characteristics and because these characteristics can not be found in the dataset used for the final cluster analyses. A full explanation of behind their removal will be given in the Model iteration section in section 3.3.

Table 3.1: Overview of characteristics excluded from the cluster analysis

Characteristic	**Data**	**Reason for Exclusion**	**(Possible) providers of data**
Peak demand electricity	No	No publicly available dataset: *No geospatial dataset found*	TSO or DSO
Peak heat demand	No	No publicly available dataset: *No geospatial dataset found*	DSO or gas network operator(GTSO)
Total yearly supply district heating	No	No publicly available dataset: *No dataset found including a detailed overview of district and residual heat supply*	DSO or GTSO
Energy storage capacity	No	No publicly available dataset: *No geospatial dataset found*	DSO, TSO or energy storage operators
Network capacity district heating	No	No publicly available dataset: *No dataset found including both a detailed network and the capacity of each branch of the network*	DSO
Network capacity gas network	No	No publicly available dataset: *No dataset found including both a detailed network and the capacity of each branch of the network*	DSO or GTSO
Industry type	No	No publicly available dataset: *No geospatial dataset found*	CBS or land registry
Education level	No	No publicly available dataset: *No geospatial dataset found including enough data point for the entire province*	CBS
Attitude towards energy transition	No	No publicly available dataset: *No geospatial dataset found*	CBS
Total yearly supply from biomass	No	No publicly available dataset: *No geospatial dataset found*	Energy producers, DSO or TSO
Available space for energy infrastructure	No	No publicly available dataset and no method of implementation: *As available space can vary from street to street it is not possible to implement this type of characteristic on the level a province in a meaningful way*	Province of South Holland, CBS or land registry
Zoning policy	No	No method of implementation: *Though multiple datasets exist with spatial policies there is no one combined policy. Instead, each subject has its own dataset. For example, the dataset for suitable locations for windmills is separate from the one defining nature areas. For this reason further analysis into what zoning policies are usefull would be required before they can be used*	Province of South Holland
Total yearly supply of gas	No	No method of implementation: *Gas can be consumed in other regions than its produced. This means that there is not a clear way to implement this in a model of the province of South Holland without an dataset of the gas grid and its network capacity*	DSO or GTSO
Total yearly supply biogas	No	No method of implementation: *Bio gas can be consumed in other regions than its produced. This means that there is not a clear way to implement this in a model of the province of South Holland a dataset of the gas grid and its network capacity*	DSO or GTSO
Total Yearly Electricity Demand (Household)	Yes	Not compatible with the cluster model: *The way this characteristic was implemented in the dataset led to unwanted cluster results therefore the characteristic was removed*	Centraal Bureau voor de statistiek (2023)
Total Yearly Gas Demand (Household)	Yes	Not compatible with the cluster model: *The way this characteristic was implemented in the dataset led to unwanted cluster results therefore the characteristic was removed*	Centraal Bureau voor de statistiek (2023)
Installed Wind Supply	Yes	Not compatible with the cluster model: *The way this characteristic was implemented in the dataset led to unwanted cluster results therefore the characteristic was removed*	Rijksinstituut voor Volksgezondheid en Milieu (2023)
Installed Conventional Supply	Yes	Not compatible with the cluster model: *The way this characteristic was implemented in the dataset led to unwanted cluster results therefore the characteristic was removed*	CE Delft et al. (2021)
Total Solar Potential	Yes	Not compatible with the cluster model: *Potential was excluded during modeling as this did not show the current state of the energy system*	Provincie Zuid Holland (2017)
Geothermal Potential (Return at 35°C)	Yes	Not compatible with the cluster model:*Potential was excluded during modeling as this did not show the current state of the energy system*	Provincie Zuid Holland (2016a)
Geothermal Potential (Return at 25°C)	Yes	Not compatible with the cluster model: *Potential was excluded during modeling as this did not show the current state of the energy system*	Provincie Zuid Holland (2016b)

For the remaining 17 characteristics one or more datasets could be found. In table 3.2 an overview of each of these characteristics and their unit of measurement is given. A complete overview of the sources, geographical scope, resolution, and unit of measurement of the datasets can be found in Appendix 3. As data was found for the last seven characteristics of table 3.1 these are also included within this appendix. One thing to note is that not all data from the industrial energy demand was included. Specifically, the energy demand of industrial areas was only included if the area also included a household energy demand. This was because the outliers within these datasets led to unwanted clustering results during testing. The reasoning behind this decision will be further discussed in chapter 4.5.

Table 3.2: Overview of characteristics included in the cluster analysis

Characteristic	**Unit of Measurement**
Mean Yearly Electricity Demand (Household)	KWh/year
Mean Yearly Gas Demand (Household)	KWh/year
Mean Yearly Electricity Demand (Industry)	KWh/year
Mean Yearly Gas Demand (Industry)	KWh/year
Total Yearly Heat Demand	KWh/year
Required Temperature for Heating	High/medium/low
Installed Conventional Supply	N/A
Network Capacity for Electricity Users	Available, limited, not available in the near future or not available
Network Capacity for Electricity Producers	Available, limited, not available in the near future or not available
Landuse Type	Residential, commercial, greenhouse, recreational, public facilities, agricultural, transport, nature and forest, water or miscellaneous
Number of Houses	#
Number of Inhabitants	#
Mode Build Year of Building	<1945, 1945-1965, 1965-1975, 1975-1985, 1985-1995, 1995-2005, 2005-2015 or >2015
Energy Label of Building	A+++,A++,A+,A,B,C,D,E or F
Average Income of Household	€
Average WOZ Value of Household	€
Percentage Rental Housing	%

3.1.2 Data cleaning

To be able to analyze the collected datasets, a combined dataset was required. For this, a raster-based approach was chosen as rasters could easily be used in the spatial analysis as every pixel is equal to an area of the same size. From the resolutions of the datasets, a raster with pixels from 100 x 100 meters was chosen as this was the minimum size possible that would still provide valid data for all of the used datasets. As not all files were stored in a similar format, some needed to be converted. Seeing that the data provided by the CBS was already in a 100 x 100 grid format, this was chosen as the basis for the raster (Centraal Bureau voor de Statistiek, 2023c). To store the data, the Energy System Description Language was chosen (TNO, 2021). ESDL is a data format that can be used to map the characteristics of an energy system in a standardized way. What was relevant for this case is that these characteristics can be mapped to user-defined areas. This way, an area could be created for each pixel in the grid that contained all information about the area. The reason why ESDL was chosen over other data storage methods is that it allows the dataset to easily be converted for use in other energy models that make use of the ESDL language. For the combined dataset, an ESDL file was made with one overarching area the size of the province of South Holland and a grid of smaller areas, all with sides of 100 meters by 100 meters.

Because not all characteristics were compatible with this ESDL structure, a series of data transformations had to be performed. In general, three types of data conversion were necessary. The first type was a conversion from shape data to grid data. This type of conversion was applicable to all datasets that provided data either at the level of a postal code or neighborhood. An example of shape shape-based dataset was total solar potential, for which only data was available on a neighborhood level. To convert this type of data to a usable format, a grid that matches the coordinates of the ESDL areas was put over the shapes. For each pixel in the grid, the value of the shape at that location was taken. If a pixel fell within two or more shapes, the mean or mode value was taken depending on the dataset. An overview of the resolutions of each dataset can be found in Appendix 3. It is important to note that some datasets included shapes that were bigger than 100 meters. An example of this can be seen in the income distribution dataset, which used a grid structure but with a 500-meter resolution instead of 100. For these types of datasets, a uniform distribution within each shape was assumed. In the case of the income dataset, this meant splitting up the 500 by 500 meter pixels into 25 pixels of 100 by 100, all with the same value as the original pixel. The limitation of this is that internal differences within the area of the shape are not accounted for. This was seen as a necessary tradeoff as the alternative would be to change all datasets to a lower resolution, which would lead to less accurate results.

The second type of transformation that was performed had to do with already rasterized datasets that had a pixel resolution that were smaller than 100 meters. Similar to the previous type of transformation, a grid was placed over these datasets. Subsequently, the mean or mode value of all data within the 100-meter squares was calculated. In the case of the land use dataset, these values were subsequently grouped based on their similarities. For example, all agricultural land use categories where combined into one land use type called agriculture. This way, a limited number of land use types were included in the analysis.

The last type of transformation that was performed was for datasets using point data. Only the wind supply dataset and the conventional supply dataset were converted using this type of transformation. These characteristics, however, were not included in the final cluster analysis as they where removed during the testing portion of the modeling process. Their exclusion will be discussed will be discussed in chapter 4.5 of the thesis. For this reason, their methodology will be discussed in this section as this gives context for the future section and offers a possible method to work with this type of data. If untouched, these datasets would only provide data at the precise location of the energy supply itself. However, as energy supply can be distributed over a large area, this would have given limited insights into the supply side of the energy landscape. For this reason, it was decided to calculate the supply suitability of each point within the datasets. Calculating the distance from each area to the energy supplier using the existing energy grid was not possible as no datasets with this information were available. For this reason, the assumption was that points closer to a source of energy supply are more suitable to receive energy from the energy source. The reasoning behind this assumption was that less infrastructure would be required, and the chance of a capacity bottleneck would be lowered the closer a point is to the energy source.

Based on this assumption, the following calculation method was used. For each pixel, the Euclidean distance from the pixel to all energy suppliers was calculated. Euclidean distance calculates the straight line distance between two points in a grid. For example, point 2.1 would be 2 meters from point 4.1 in an X,Y grid with a resolution of one meter. The Euclidean distance was then combined with the power output of each energy supplier to calculate the suitability using the formula below. These results were then normalized over the entire dataset, making 0 the least suitable location and 1 the most suitable.

$$\mathsf{V}_n = \sum_{i=1}^{n} \left(\mathsf{P}_i \cdot \left(1 - \frac{\mathsf{D}i}{\mathsf{Dmax}} \right) \right)$$

In this formula, Vn is the normalized value, p is the power output of the energy supplier in Megawatt, D is the Euclidean distance between the area and the energy supplier, and D max is the maximum distance possible. In the case of wind, this process was performed using all on- and offshore windmills in the Netherlands. As no reliable spatial dataset could be found regarding the conventional energy generators in the Netherlands, only the generators within the province of South Holland were included in this transformation.

In Figure 3.1, a simplified representation of the three methods of transformation is shown. Using these methods, all collected datasets could be combined into one ESDL.

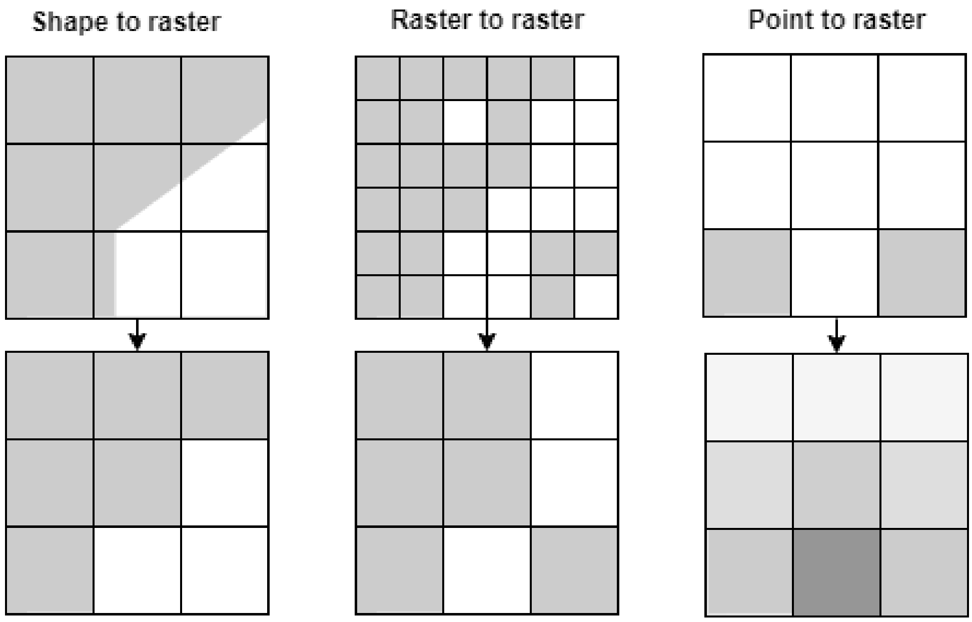


Figure 3.1: Transformation methods used in creation of ESDL dataset

3.1.3 Dataset analysis

Before the dataset could be used within a clustering algorithm, a better understanding of the dataset was required. To reach this goal, plots of all characteristics within the dataset were made. A complete overview of these plots can be found in Appendix 4. Based on these plots, two important observations were made.

The first observation that was made was that not all datasets are equally complete. In some cases, large areas of the province of South Holland are not included in the dataset. An example can be seen when comparing the land use dataset to the mean yearly electricity demand per household. As can be seen in Figure 4.2 the land use dataset has an almost complete coverage of the province, while the energy demand data is missing data for many areas. This missing data can mostly be explained by demographics. Most datasets only include data from areas where large groups of people live. For example, datasets from the CBS only included areas where more than 5 people live. For this reason, much of the rural and industrial areas were not included. As most clustering methods can only include areas that have a complete set of characteristics, some areas of the province can not be analyzed. One way to solve this would have been to interpolate missing data. Interpolation would mean that areas that were missing characteristics would get a value appointed to them. The value would be based on the average value of the same characteristics of surrounding areas. This way the value would be similar to known values in the area. The decision was made not to use interpolation as the dataset is not suitable. This has to do with the fact that the missing data is not randomly distributed through the map. Instead, it is clustered in distinct areas. This means that there are no surrounding values that could be used to approximate the missing values in a reliable way. There is also the risk of introducing artificial relationships between characteristics that can influence the cluster results. For this reason, it was decided only to use areas that included all characteristics that were necessary to run the cluster analysis. One important result of not interpolating the data is that the data analysis only includes data-rich environments. As these areas are mostly centered around urban areas, the performed analysis almost exclusively focuses on energy and spatial characteristics within urban residential areas. This has to be kept in mind while reading the results, as rural and industrial areas will be underrepresented within the results compared to the real world.

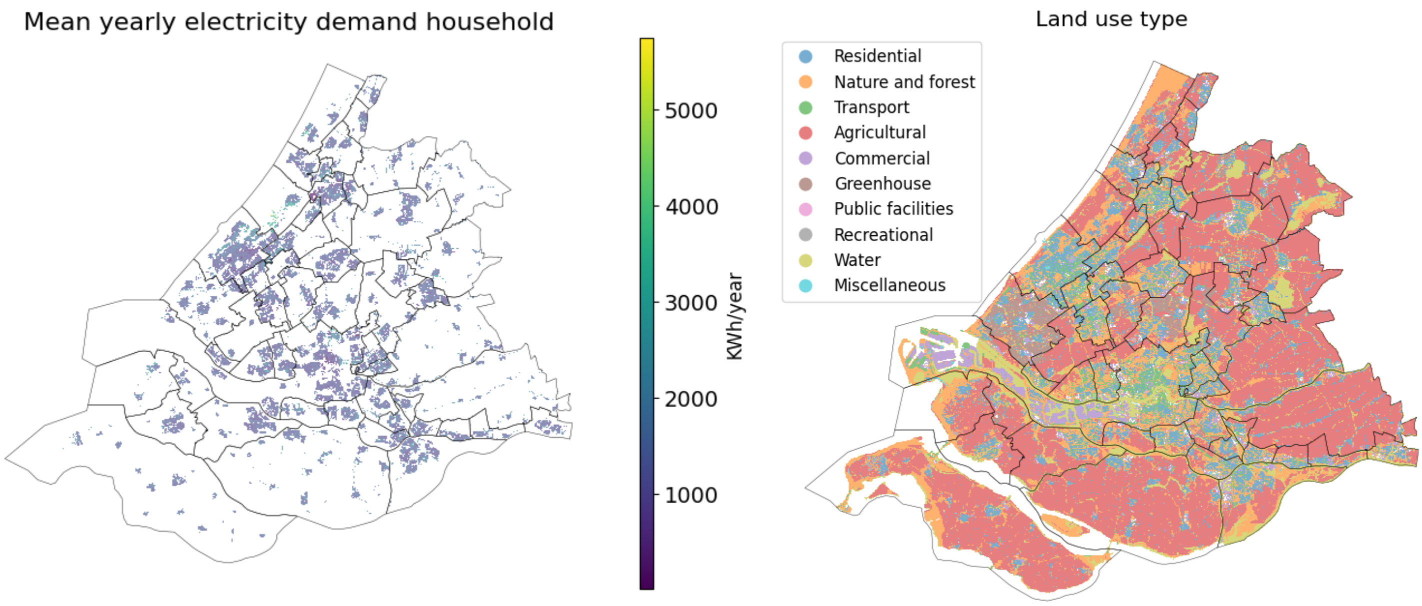


Figure 3.2: Comparison completeness of land use and demand datasets

The second observation that was made regarding the source dataset was that most of the datasets were heavily skewed. A good example of this can be seen in Figure 3.3, which shows the distribution of the total yearly electricity demand of an area. Most of the values fall within a narrow band. However, there is a significant set of outliers that fall far outside of this band. Another dataset where this was very noticeable was the gas and electricity use by industry. Here, a big gap can be observed between the mode value, which is close to 10^4, and the maximum value which is closer to 10^6. It is important to note that this plot is logarithmic as a normal boxplot would show almost exclusively the outliers. Though these outliers can be explained for most of the characteristics, it could lead to problems as some types of cluster algorithms, like K-mean, are susceptible to outliers (Gupta et al., 2017). For this reason, some precautions were taken to limit the impact of these outliers. The implementation of these precautions will be further discussed during the experimental setup.

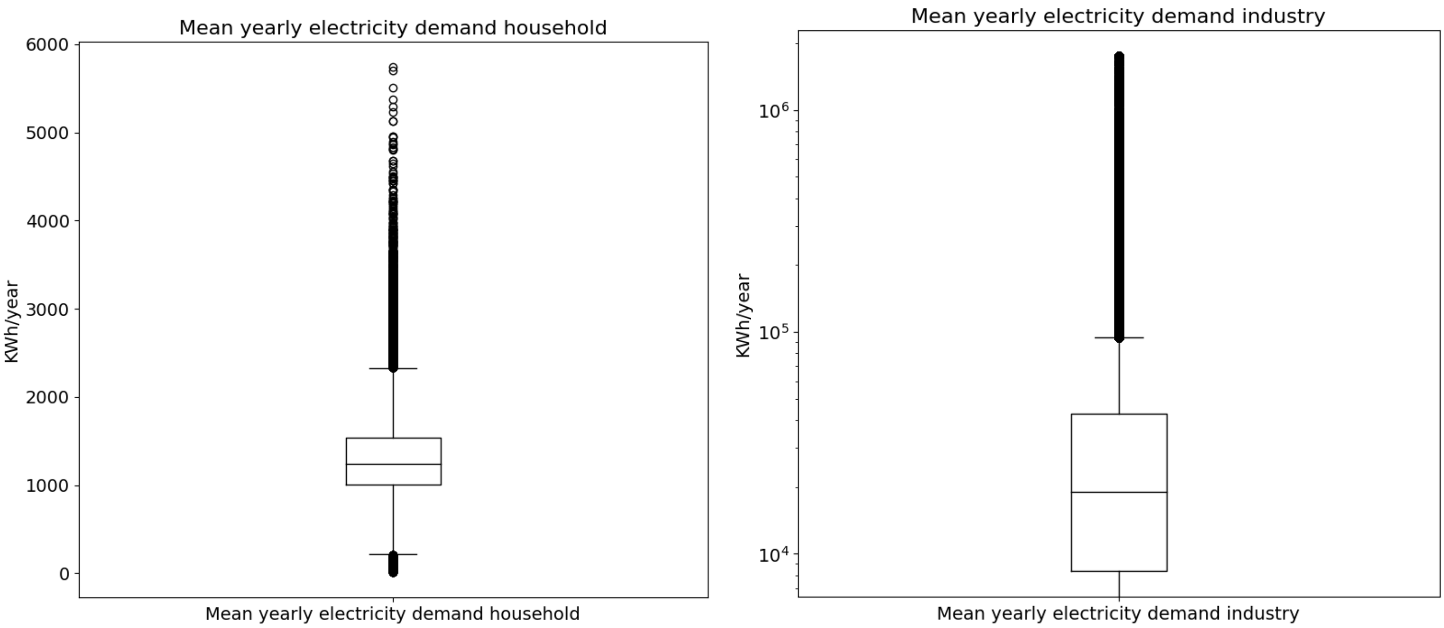


Figure 3.3: Value distribution for mean electricity demand of households and industry

3.2 Clustering method

After the dataset was created and analyzed, a clustering algorithm was chosen. The goal of this algorithm was to analyze all areas within the province of South Holland and make groupings of areas with similar characteristics. By making these groupings, two things could be analyzed. Firstly, the relationship between the characteristics of an area and its geographical location. Secondly, the relationships between different characteristics based on their values within different clusters.

3.2.1 Cluster algorithm selection

For the selection process of a clustering algorithm, multiple features of the dataset were taken into account. An important feature of the dataset is that it includes both numeric and categorical parameters. The reason why this feature is important is that not all clustering algorithms are inherently built to deal with this type of mixed data. For example, a commonly used algorithm for clustering based on numeric data is K-means. To calculate in which cluster a data point fits, the K-means algorithm makes use of distance calculations. This means that it looks at how each of the data points differs on each variable from the centroids that define the clusters. For these calculations to work, there has to be a measurable distance between the two variables. However, categorical data, like land use types, do not have a defined distance between their different categories. This means that the distance-based algorithms would not be able to calculate clusters for categorical datasets without any additional changes to the algorithm (Costa et al., 2022).

Another aspect that was included in the decision-making process regarding the clustering algorithms was the scope of the dataset. As the dataset has around 300,000 different areas, with each area having up to 17 different characteristics, it was important to find an algorithm that could efficiently find clusters within a large dataset. The last characteristic that was considered was the high amount of outliers that could be found in the dataset. The reason why this is important is because outliers can have a significant impact on the clustering results. For example, if the difference between the outlier and the rest of the dataset is big enough, the outlier can pull the centroid of the cluster towards itself, making the cluster less effective in describing the overall dataset. Although the effect of these outliers on the results can not be completely removed, some algorithms are more resilient than others, making them more suitable for this type of analysis.

Based on these three criteria, multiple algorithms were analyzed to asses there suitability. For this assessment, a set of benchmark studies was used to compare their performance in datasets with mixed categorical and numeric data. The result of the analysis was that both the K-Means clustering algorithm for mixed large datasets (KAMILA) (Foss & Markatou, 2018) and K-prototype (Huang, 1998) performed well in clustering a wide variety of datasets with mixed datatypes. Although KAMILA scored on average comparable to or slightly better than K-prototype in the benchmarks, the decision was made to use K-prototype (Costa et al., 2022; Jimeno et al., 2021). The reasoning behind this choice was that an implementation of the K-prototype already existed for Python. By utilizing this existing model, more time could be spent on testing out different parameterizations and analyzing the results. Seeing that this thesis is a preliminary look at the use of clustering for the purpose of analyzing a combination of the energy system and the built environment, a choice was made to prioritize these aspects of the analysis over finding the most optimal clustering algorithm.

The K-prototype is an algorithm that combines aspects of the K-means and the K-modes algorithms. In this algorithm, numerical and categorical data are analyzed separately and later combined. Similar to the normal K-means algorithm, K-prototypes uses an iterative process to calculate in which cluster each of the data points fits. In figure 4.4 an overview of the steps used by K-prototype is given.

For the combined distance calculation, the following formula is used (Huang, 1998).

$$d(X,Y) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2 + \gamma \sum_{j=p+1}^{m} \delta(x_j, y_j)}$$

In this formula, the numeric distance is calculated with Euclidian distance using the following formula:

$$\mathrm{D}_e = \sum_{j=1}^{p}(x_j - y_j)^2$$

In this calculation D_e is the Euclidean distance, p represents the set of numerical characteristics and $(x_j - y_j)$ the distance between the value of the characteristic and the value assigned to the centroid. Because of the outliers in the data, experiments with Manhattan distance were also done as a replacement of Euclidean distance. This is because the Manhatten distance is more robust with outliers. The results of these experiments can be found in the experimental setup paragraph.

The categorical distance is calculated using the following formula.D =

$$\mathrm{D}_b = \sum_{j=p+1}^{m} \delta(x_j, y_j)$$

Here,D_b is the distance measure, m is the set of all categorical characteristics and $\delta(x_j, y_j)$ a boolean that returns 1 if the centroid and characteristic have different categorical values and 0 if they are the same. To calculate the distance to each cluster centroid, the numeric and categorical distance are then combined using γ. This value can be set between 0 and 1 and regulates the relative importance of the categorical and numeric data. Here, a higher γ results in a higher importance of categorical data during the clustering. As this value can have a big impact on the creation of the clusters, multiple experiments were run to decide the most optimal value for this parameter.

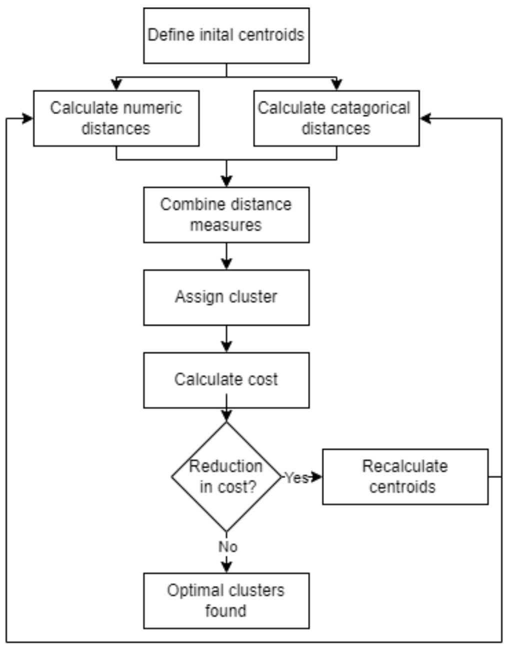


Figure 3.4: Overview of the K-prototype algorithm

3.2.2 Cluster model design

After the algorithm was chosen, a model was made that could combine the dataset with the K-prototype algorithm. For this, the K-prototype package from the K-modes Python library was used (de Vos, 2015). To connect this package to the dataset, a translation layer had to be written between ESDL and a file format that could be read by the K-prototype package. This was done by converting every area in the ESDL file into a row of a geopanda dataframe. Here, the values were normalized between 0 and 1 to make sure that all characteristics were treated equally by the clustering algorithm. After this, a script was made that could run cluster analysis and generate the resulting clusters. Below, a simplified representation of this code can be seen. The complete code and datasets can be found within the TU data repository. As some of the variables, like the minimum and maximum number of clusters, can have multiple values the following section will discuss the parameterization used within this analysis.

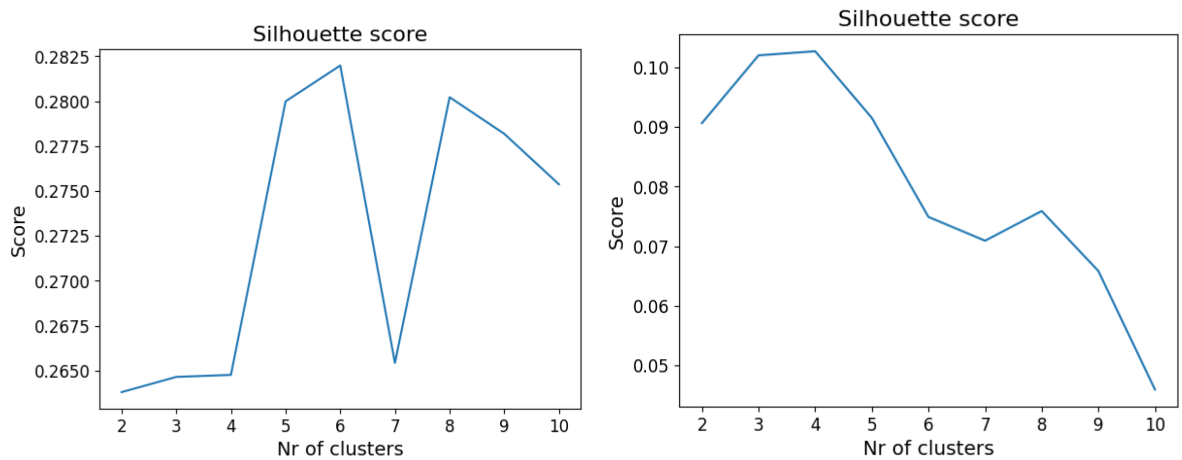
Figure 3.5: Pseudocode of script used for clustering algorithm

```
Import characteristic dataset
Set the number of cluster analyses to be performed
foreach analysis in number of cluster analyses do
    Set list of characteristics to be clustered
    Set list of Categorical columns
    Set clustering parameters
Set minimum number of clusters to be calculated
Set maximum number of clusters to be calculated
foreach Analysis in number of cluster analyses do
    Set cluster dataset based on list of characteristics to be clustered
    Filter out rows with missing values
    Remove outliers
    foreach Cluster amount in range(min number of clusters, max number of clusters) do
        Initialize cluster analysis based on characteristics of cluster
        Define the first set of centroids
        Calculate the distance between data points and centroids
        Calculate cost
        while Cost < cost of previous iteration do
            Reassign new centroids based on new clusters
            Calculate the distance between data points and centroids
            Calculate new cost
            Compare the previous cost to new cost
        Save optimal cluster for each datapoint
        Save list of cluster centroids
        Save cost optimal clustering solution
```

3.2.3 Experimental design

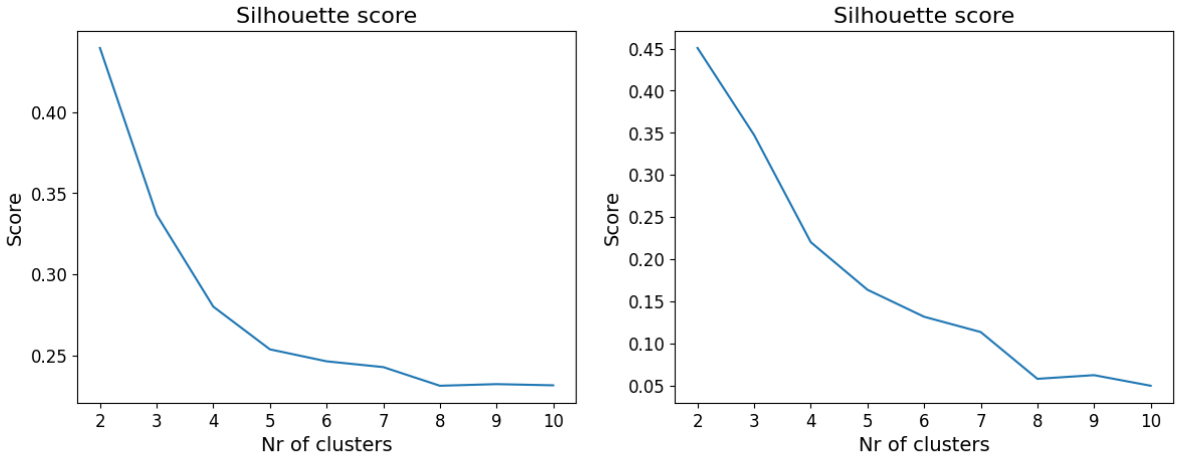
Using this code, multiple cluster analyses are performed under different parameterizations. Based on the results of these experiments, a final set of experiments was created which can be found in Appendix 5. In total four different analyses were performed that looked at different combinations of characteristics. The first cluster analysis looks solely at the energy characteristics. This will be used to see how the cluster forms if the built and social environment are not considered. The second analysis included both energy and built environment characteristics. Similarly, the third analysis focused on a combination of energy and social aspects. The last of the analysis looked at all characteristics combined.

The first aspect that had to be defined for each of the cluster analyses was the number of clusters that would be created. As the K-prototype algorithm can not find the optimal amount of clusters automatically, separate solutions were calculated for each number of clusters in a range from 2 up to 10 clusters. Though better solutions may exist upwards of 10 clusters, these were not analyzed as this many clusters would be difficult to visualize and analyze. To analyze the optimal number of clusters, the silhouette score was employed. This metric looks at the similarity between areas within the same clusters and compares this with the similarity between areas within different clusters. This metric is given in a range from -1 to 1. Here, a value close to -1 means that there either are limited similarities between areas within the same cluster or too many similarities between areas in different clusters. +1, on the other hand, means that all characteristics are within a well-defined cluster with many similarities between areas in the same clusters and only limited overlap between values in different clusters. Using this method, the most suitable number of characteristics was defined by taking the number of clusters with the highest silhouette score. In Figure 3.6 and 3.7, an overview of silhouette scores for each cluster analysis is shown.



(a) Energy (b) Energy and built environment

Figure 3.6: An overview of the silhouette scores of the energy and built environment clusters



(a) Energy and social

(b) Total

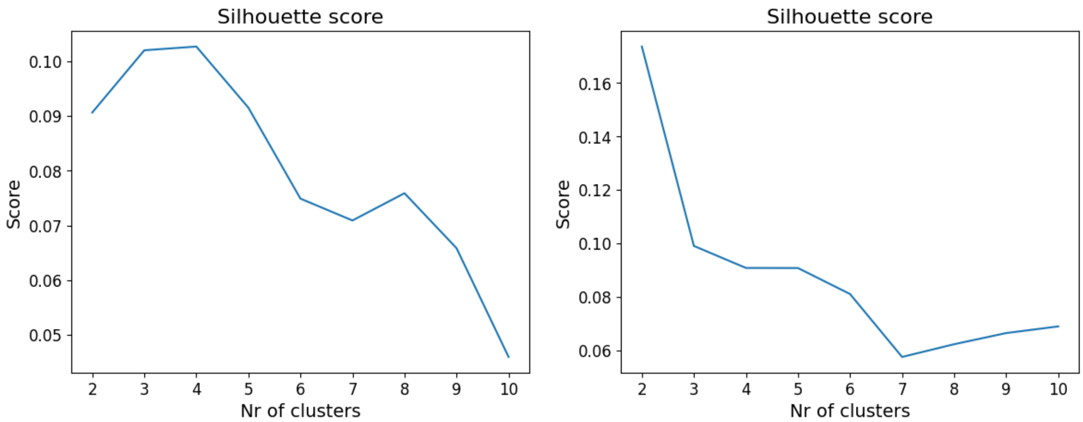
Figure 3.7: An overview of the silhouette scores of social and total clusters

Based on these scores, it is clear that six and four clusters would be the ideal number of clusters for the energy and built environment analysis. However, for the remaining datasets, the right amount of clusters is less clear. This is because the silhouette scores decline with the addition of more clusters. As only limited insights can be held from clustering with two clusters, the different numbers of clusters were instead visually inspected and compared. This was done by plotting the cluster for each number of clusters individually and comparing these plots to each other. From this, it was clear that the most usefull amount of clusters for both datasets was four. This number was chosen because adding more than four clusters seemed to only result in small clusters that only included a few areas. These clusters were similar to one of the other clusters except for one characteristic. This meant that adding more clusters did not result in any more big clusters being formed. This number of clusters also seems to match the silhouette plots as, in both cases, the silhouette score is still relatively high when clustering with four clusters. Based on this, it was decided to analyze the energy, built environment, social, and total datasets with a total of 6, 4, 4, and 4 clusters, respectively.

Besides the number of clusters, this silhouette score also shows that the clusters are relatively weak. This is especially the case for the built environment, which shows a silhouette score of around 0.1. This shows that the boundaries between the clusters are not well-defined. Because of this, some areas that are at the boundary of two clusters might exhibit features of both clusters. This results in less well-defined clusters. One of the factors influencing this low value seems to be the high amount of outliers still present in the data. This is because these outliers are harder to group together with the other areas, which makes the range of values within each cluster bigger and thus less distinct. Another important aspect seems to be the wide variety in values for categorical characteristics, leading to many mismatches. As a categorical characteristic can only be a match or a mismatch with the value of the cluster the amount of mismatches is significantly higher than with numeric data which can have a partial match. As some of the categorical characteristics in the current dataset have more than 10 different options this means that many mismatches between categorical characteristics within clusters exist, lowering the strength of the cluster.

Because of these relatively poorly defined boundaries between the clusters, only limited results could be drawn from the analysis. The main limitation is that the characteristics of individual areas can not be assumed based on the cluster it is assigned to. This meant that no analysis of individual data points could be performed. Broader trends within the cluster can be analyzed, however. An example of this would be an above-average value within one type of characteristic. Because of this limitation, the cluster analysis will exclusively look at this type of result.

After the number of clusters was defined the initialization method was defined. This parameter decides how the initial cluster centroids are to be determined. This is important as the K-prototype algorithm is susceptible to finding local optimums instead of global optimums. For this reason, the starting point of the cluster analysis can have an impact on the results. K-prototype allows for two initialization methods. Both of these methods can be used to automatically find the most optimal starting points for the cluster analysis. These are Huang and Coa. The main distinction between these is that Huang requires multiple runs to analyze the most optimal results, while Coa uses an algorithm to determine the most optimal starting positions beforehand (Cao, 2009; Huang, 1997). This means that this algorithm only requires one run to find the optimal cluster composition. Below a side-by-side comparison can be seen between Coa and Huang. Here it can be seen that excluding the first data point, they create similar results with only minor differences in their overall silhouette score. The same was also the case when comparing their clustering results. In the end, Huang was chosen as these cluster results were more consistent when running the cluster analysis multiple times in a row.

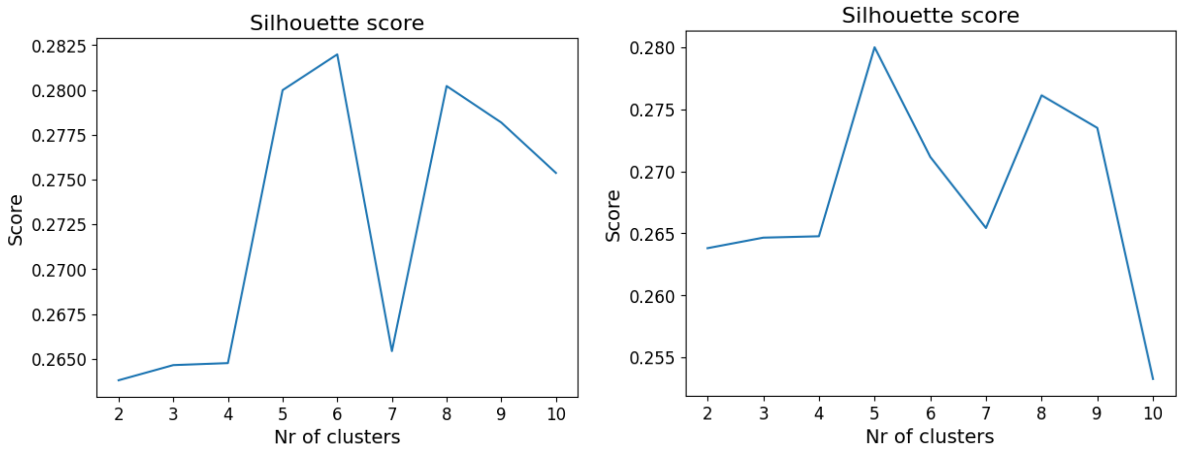


(a) Huang

(b) Coa

Figure 3.8: An overview of the silhouette scores of Huang and Coa

Because Huang was chosen as the initialization method, multiple iterations with different starting points had to be performed. The goal was to find a balance between computational time and result accuracy, This was done by comparing the same dataset with 25 and 100 different starting points. In Figure 3.9 a side-by-side comparison of the silhouette score for 25 and 100 runs can be seen. This shows that there is no significant difference between these two values. A lack of distinction was also observed when visually expecting the cluster of the two analyses with only a few differences in cluster placement. For this reason, a value of 25 runs was chosen for the final analyses.



(a) 25 runs

(b) 100 runs

Figure 3.9: An overview of the silhouette scores 25 and 100 iterations

During the testing of the model, it also became apparent that the impact of the outliers on the clustering results was significant. This resulted in clusters containing single data points and limited distinctions between the clusters. To limit this impact, it was decided to remove all outliers that fell outside a certain number of standard deviations from the rest of the dataset. Based on experimentation, it was decided to set a limit at 3 standard deviations as an analysis of the results showed that this provides a good balance between the removal of outliers and limiting the loss of information. An unfortunate result of this change is that the dataset became less representative of the entire environment. However, this step proved to be necessary as no meaningful cluster could be created using the original dataset containing the outliers.

Related to this, some experimentation was also done with regard to the use of Manhattan distance calculations instead of Euclidean distance. As this resulted in no significant improvement in the clustering results, Euclidian distance was therefore chosen for the final runs as this was the one that was originally used in the K-prototype model. Lastly, some different values were tried for gamma. This way, a possible difference in clustering results based on the importance of categorical data could be observed. The original paper by Huang uses the following calculation to define gamma as $0.5 \times \text{mean}(\text{std}(\text{p}))$ where p is all numerical columns (Huang, 1997). To see the impact of lowering and raising the relevance of categorical columns, 0.5 was replaced with values between 0.1 and 1.5. As this can not be directly measured by looking at the cost function a more subjective method was used whereby clusters were analyzed to see what the balance was between categorical and numerical influences. The resulting clusters showed that for this dataset a value around 0.5 led to both numeric and categorical data being used to define the clusters. For this reason, this number was used for the final run.

Based on these findings the final parameterization was defined. In Appendix 5, an overview of the parameterization of each cluster analysis is given. In the next section, the results of each of these clustering algorithms will be shown to see whether there is any difference depending on the characteristics that were used.

3.3 Model iteration

The previous sections have shown the setup for cluster analysis. Based on this multiple cluster analyses were performed. The procurement of these results was not a linear process. As stated in the research methodology a co-design process was used for to form these clusters. This means that results were shown to and checked by the province of South Holland during multiple meetings. The goal of these checks was twofold. Firstly the goal was to check if there were any changes that had to be made in the characteristics used within the analysis. Secondly, the goal was to compare the results of the model to the real world to check whether this matches the expectations of the experts. The reason why this was important is because no known good cluster analysis exists for the province of South Holland. This means that there is no set of previously established results to which the results can be compared. For this reason, the province's view of the real-world energy and built environment based on their expertise was used.

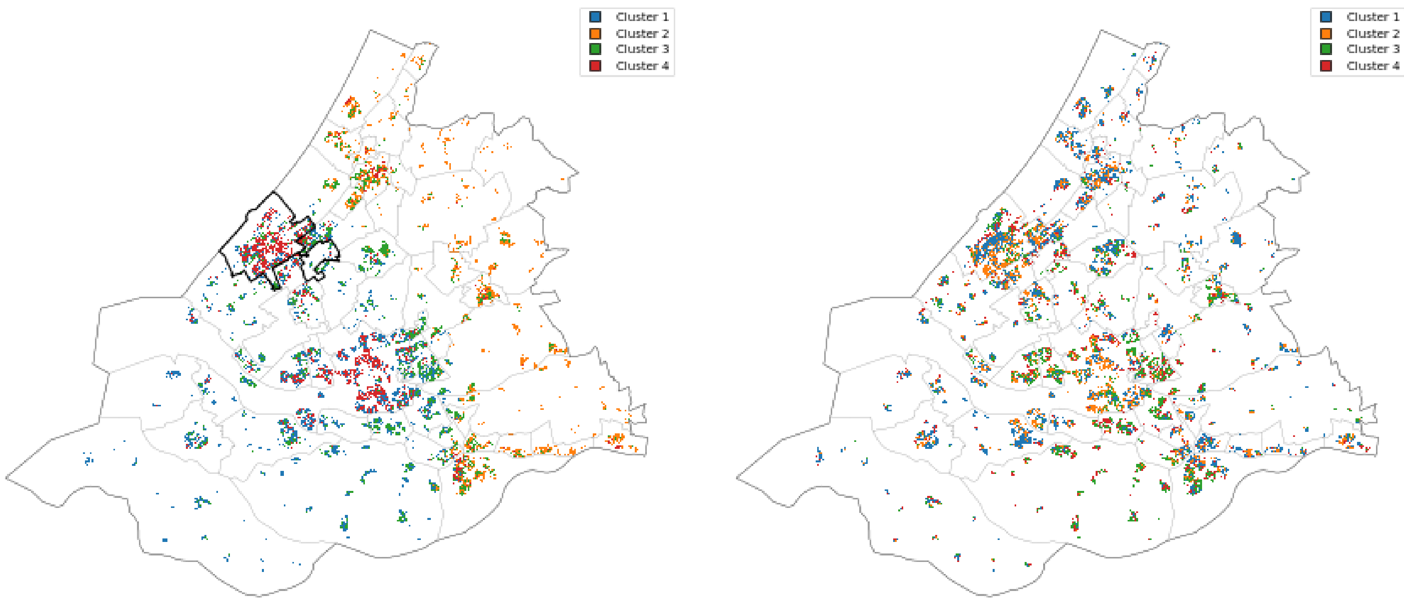
Most meetings with the province were held with the client for the thesis who is an expert in spatial planning and the energy transition in the province of South Holland. Besides this one group discussion was held with six employees of the province of South Holland who were either experts in energy or spatial planning. Unfortunately, the final version of the model was not checked as the last meeting resulted in some changes that required a new cluster analysis to be performed. However, as this model was relatively similar to the final model with the exception of two characteristics the findings of this meeting will be used.

During the meetings, a few observations were made. The first observation made during one of the early meetings was regarding the datasets of the average yearly industrial electricity and gas demand. From experimentation, it became clear that the high outliers in these datasets limited the ability of the algorithm to cluster on any other characteristic than industrial energy demand. It was determined that this was an unwanted result as this would limit the further insights that could be gathered. It was therefore proposed to the province that this dataset should be removed from the dataset. However, during one of the earlier meetings, areas with a combined residential and industrial energy demand were flagged as an interesting finding by multiple experts. For this reason, it was decided not to remove the entire dataset but to limit it to areas that have both industrial and residential energy demand. This way large industrial complexes like the port of Rotterdam would be excluded without removing smaller more urban industrial centers.

Another type of characteristic that was flagged during the meetings was renewable potential. For these characteristics, the consensus was that they provided little relevant insights into the existing relations between energy and landscape as they look to the future situation of the energy system. For this reason, they were excluded from the final cluster analyses.

Other characteristics that were removed from the final dataset were total electricity demand of households, total gas demand of households, supply of wind energy, and supply of conventional energy. For the first two, this decision was made based on a set of preliminary results. These results mostly showed clusters based on the total energy demand with some minor changes in the remaining characteristics. Here the household electricity and gas demand were removed in favour of a more in-depth analysis of the other characteristics. This was motivated by the fact that the province was more interested in the differences in mean household energy consumption than the total consumption as more data is already available on this topic. Besides this, it was also noted that these characteristics were already indirectly included in the data set in the form of a combination of mean electricity and gas demand and the number of houses within an area.

In the case of the wind supply dataset and conventional supply dataset, the removal was done because the clusters were overfitting based on these characteristics. Both of these datasets made use of the point-to-raster transformation method described in chapter 3 and were interpolated based on the distance between the area and the energy supply. Because of this interpolation, the transition between the values was relatively smooth. This resulted in the clusters strongly favoring these characteristics over the other characteristics when clustering. In Figure 4.34 an example of a dataset with wind and conventional supply can be seen compared to one without these factors. Here, a clear orange and blue area can be seen within the clusters. This shows a distinction based on the distance to most of the energy suppliers, which are in the southwest of the province. This was seen as an interesting find by the province however similar to the previous characteristics they were removed in favor of a more in-depth analysis of the remaining characteristics. One important factor in this decision was that the province noted that this methodology currently is of little value as the Dutch energy grids assume that all energy supplied to the network can move freely throughout the entire energy grid. However, it was noted that this type of analysis could be more interesting in the future when the plan is to have a more localized match between demand and supply.



(a) Clusters with supply (b) Clusters without supply

Figure 3.10: Difference in the clusters when including wind and conventional electricity supply

4 | Identifying relation between energy and spatial characteristics

Based on the results of the previous chapter, the four cluster analyses were performed. In this chapter, the results of these analyses will be shown and discussed. Although this chapter will give some remarks on the implications of these results this will mostly be discussed in chapter 5. After and during the cluster analyses multiple feedback meetings were held with the province of South Holland. The feedback from these meetings and their impact on the final clusters will also be discussed. During the discussion of each section references to the map of South Holland will be made. These references will be numbered in the text and on the map. Besides the analysis of the entire province, each cluster analysis also looked specifically at the city of the Hague. This was done as this allows for a more in-depth analysis of local differences in clusters. The reason The Hague was chosen is because the dataset of this area is relatively complete. Besides this, the energy and built environment landscape of this area is also relatively diverse meaning that in most of the analysis, almost all of the clusters were included within this area. A side-by-side map of all cluster sets can be seen in Appendix 6.

4.1 Energy clusters

The first cluster analysis that was performed only looked at the energy characteristics. The goal of this analysis was twofold. Firstly, it was used to analyze the current state of the energy system in the province of South Holland. Secondly, it served as a baseline as it shows the clusters without the inclusion of other types of characteristics.

4.1.1 Spatial distribution of energy clusters

In total, six clusters were identified. In Figure 4.1, a map of the location of the different clusters can be seen. Looking only at the spatial distribution of the clusters without their characteristics some observations can be made. One observation is that the clusters are not confined to one region of the province. Instead, the clusters are spread in patches throughout different urban areas. Notably, these patches do not align with city or municipal borders. Instead, the clusters show different areas with similar energy characteristics. This shows the similarities in energy between areas within a city are smaller than the similarities with other areas in the province. Another thing to note is that the separation between the clusters is relatively strict. Meaning that within a patch of one cluster, there are only a few areas that are a part of a different cluster. This shows that although the clusters do differ on a provincial scale they are relatively uniform on a local scale. Although most clusters are present in multiple areas of the map there are some trends. On the map, it can be observed that clusters 4 and 5 can mostly be found within major cities like Rotterdam and The Hague (1), while clusters 1, 3, and 6 can both be found in larger and smaller urban areas (2). In the case of cluster 2, a clear patch can be seen in the South of the province, with the only other areas with cluster 2 being at the top of the map near Leiden (3). Besides this, there are also notable differences between similar areas. For example, comparing the clustering of cities like the Hague and Rotterdam shows that The Hague mainly contains clusters 1 and 4 with some areas 6. Meanwhile, Rotterdam mainly has 6 with almost no cluster 1. This again points to other factors influencing these energy clusters than the physical location. This can also be seen in the large number of cities that are in the same clusters as smaller towns (2).

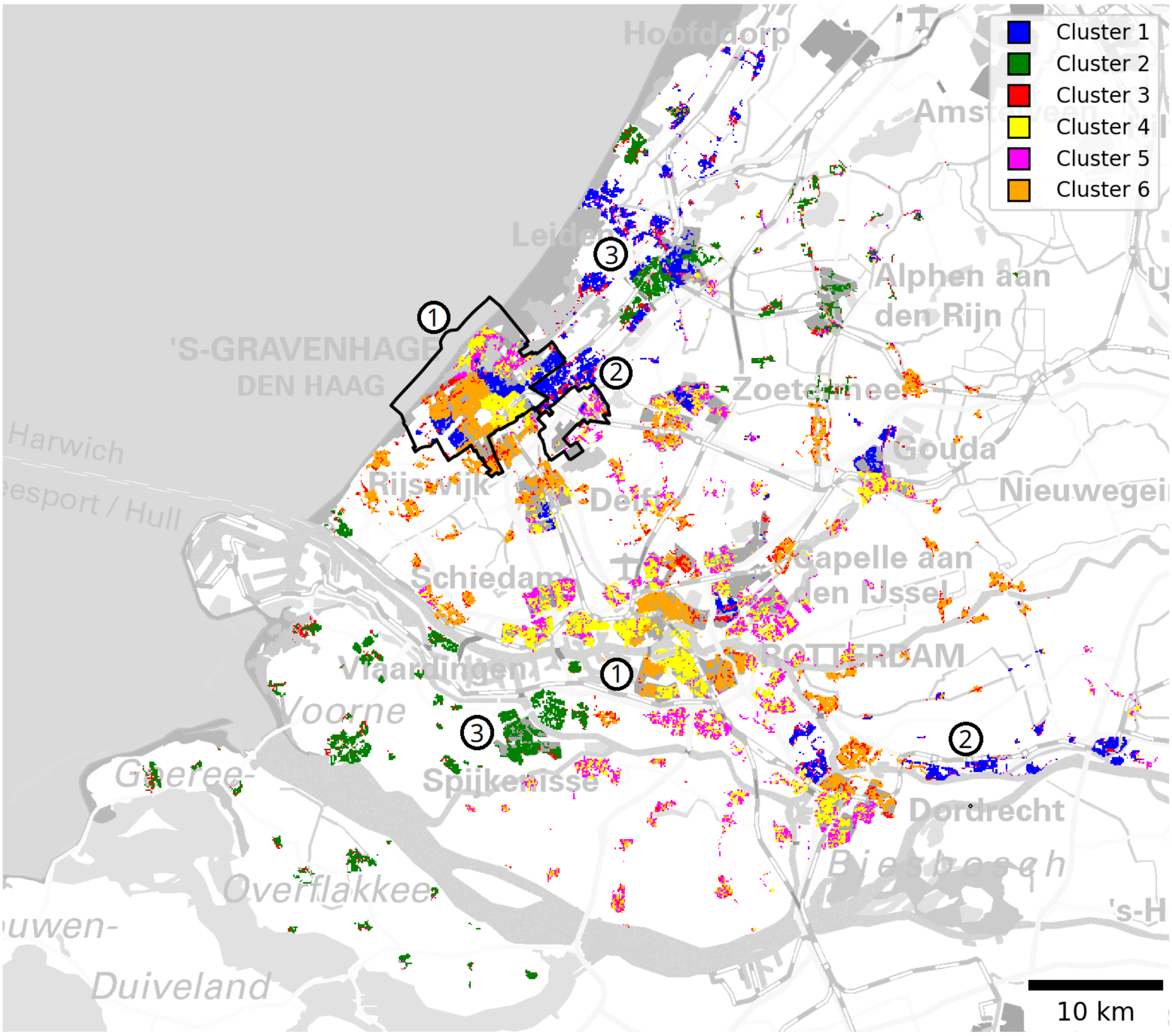


Figure 4.1: Spatial distribution of energy clusters

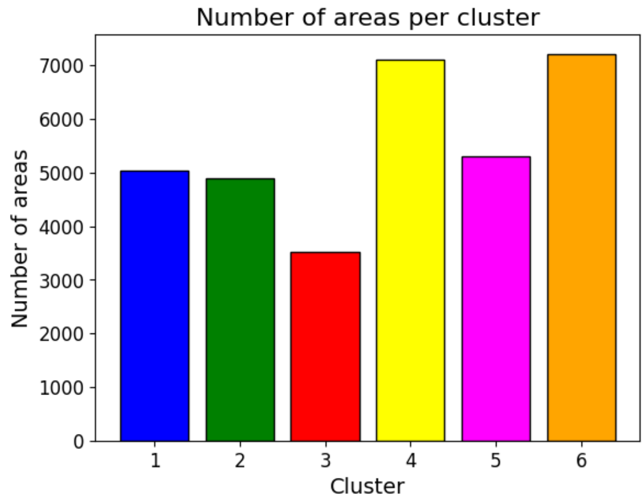


Figure 4.2: Number of areas per energy cluster

To get a more detailed idea of the placement of these clusters, a map focusing on the city of The Hague was made. This map, which is shown in Figure 4.3, shows that clusters 1,4 and 6 are the most common clusters within this area. The clusters do not strictly follow neighborhood borders with cluster borders often being within the neighborhoods themself. That being said, the majority of neighborhoods do seem to have only one or two clusters with them. It can also be observed that the clusters are not continuous. Instead, they show multiple areas that are not connected but have similar characteristics. Though some borders between the clusters are relatively well defined some areas are a mix of multiple clusters. A good example of this can be seen at the top of the map where cluster 4 and cluster 5 are mixed through each other (1). Looking at this in more detail, the maps show the center of the Hague (2), which has been marked with a red outline, to be mostly occupied by clusters 1 and 4. Besides this, a large area of cluster 6 can be seen in the (south)west of the map (3). Besides this, clusters 3 and 5 are also present. However, they are confined to far smaller areas than the other clusters. Cluster 2 is not included within this area at all so will only be discussed in the context of the entire province.

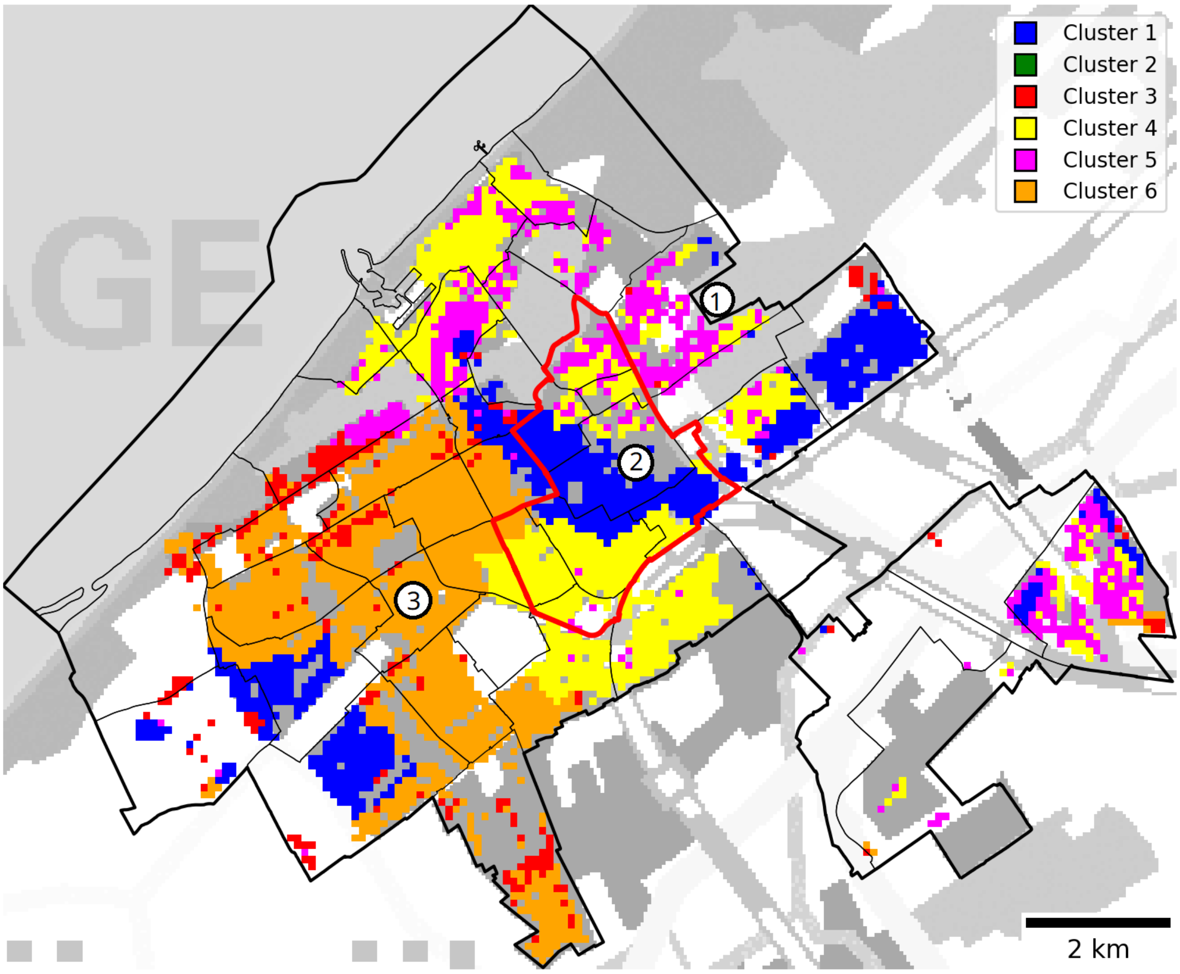


Figure 4.3: Spatial distribution of energy clusters in The Hague

4.1.2 Cluster characteristics of energy clusters

Although this spatial distribution already gives an idea about which areas in the province of South Holland have similar energy characteristics, it does not yet show the defining features of each of these clusters. To this end, an analysis of the differences and similarities between the clusters was performed. One thing that can be seen when comparing these clusters is that all clusters contain a broad range of values, resulting in a lot of overlap. An example of this can be seen in Figure 4.4, which shows the distribution of values within each cluster. Here the plot shows that the highest value of each cluster often overlaps with the mode values of other clusters. Because of this, the clusters can only be analyzed in terms of differences in medium, minimum, and maximum value instead of describing the complete cluster set. These images show that there is only limited variation in the industrial electricity demand between clusters. Though some differences can be seen in the maximum values, the mode value of all clusters is at 0. A similar distribution can also be seen for the gas demand of industry. This shows that the algorithm was unable to incorporate these characteristics effectively within the clusters. One possible reason for this is the removal of all large industrial demands as this analysis only looked at areas that also have household energy demands. Another aspect that might cause this behavior is not correlated to other energy characteristics like the mean heat demand of households. This means that these industrial areas are placed in clusters based on other characteristics of the area instead of their industrial electricity and gas demand. This does not automatically mean that this factor is unimportant for defining the energy system. However, it does mean that the current dataset is not suited for finding relationships of the industrial energy demand characteristics.

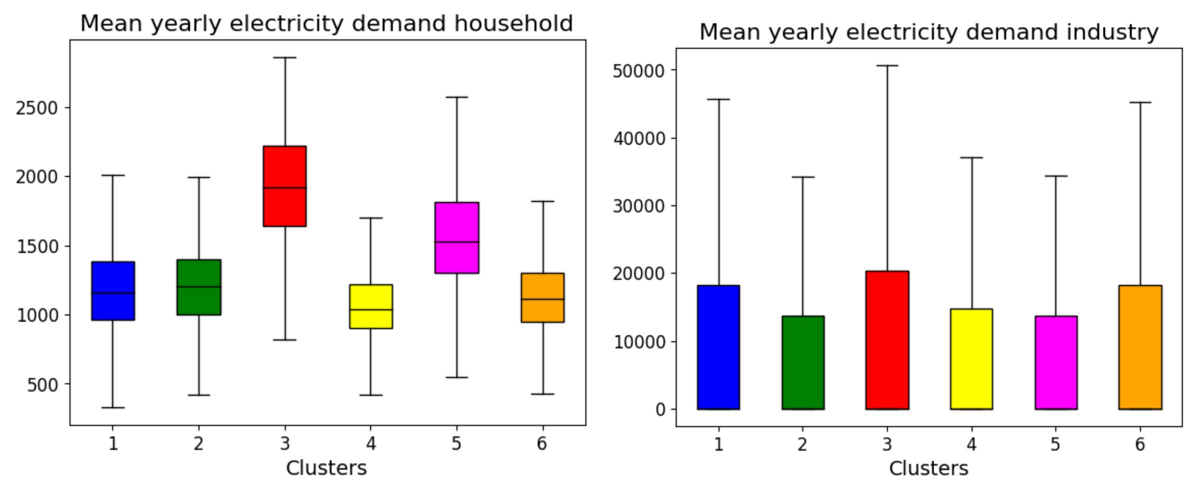


Figure 4.4: Energy demand distribution per energy cluster

A similar trend can be seen with the temperature requirements of buildings. Here, all clusters show a similar distribution, with most of the areas needing high-temperature heat. This shows that under the current parametrization, other characteristics like total yearly heat demand have a bigger influence on the overall differences in the energy system.

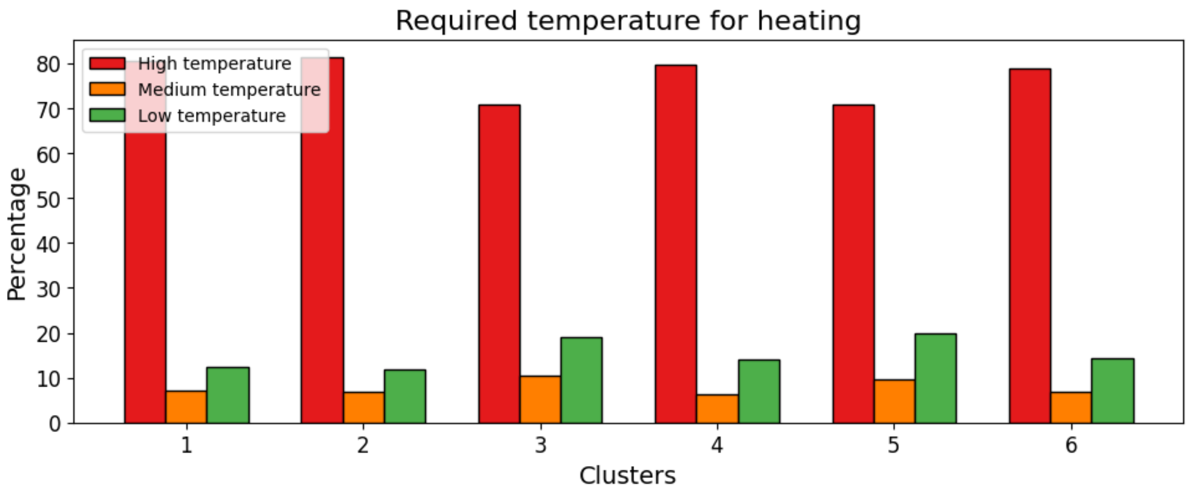
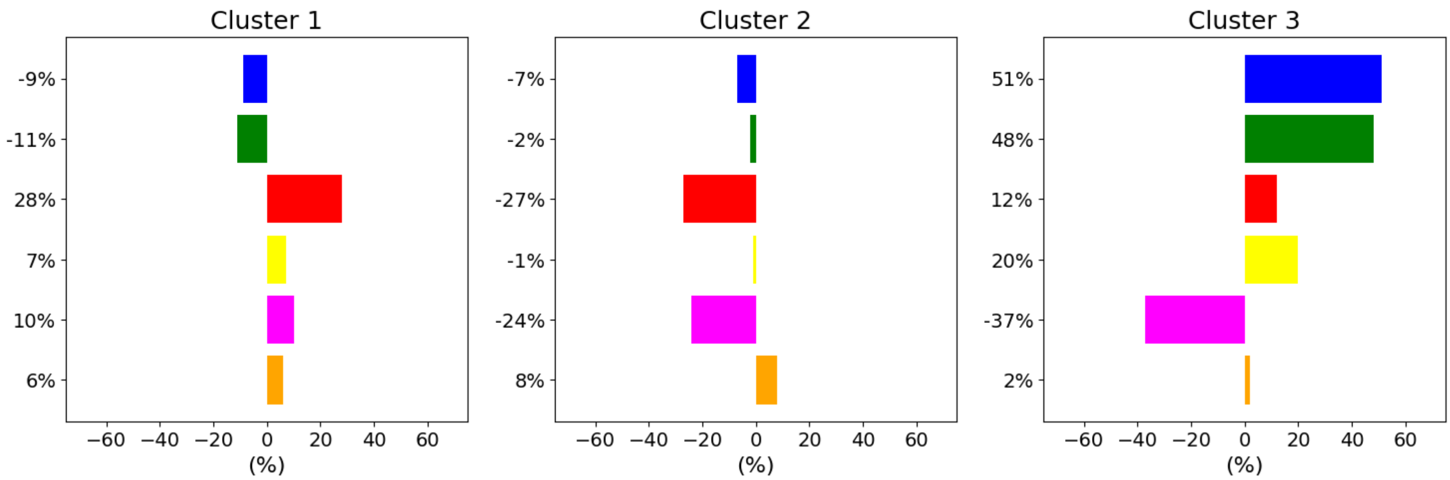


Figure 4.5: Required temperature for heating per energy cluster

To visualize the features of each cluster, the percentile difference between the mean value of the entire dataset and the centroid of each cluster's characteristics is shown in Figures 4.6 and 4.7.

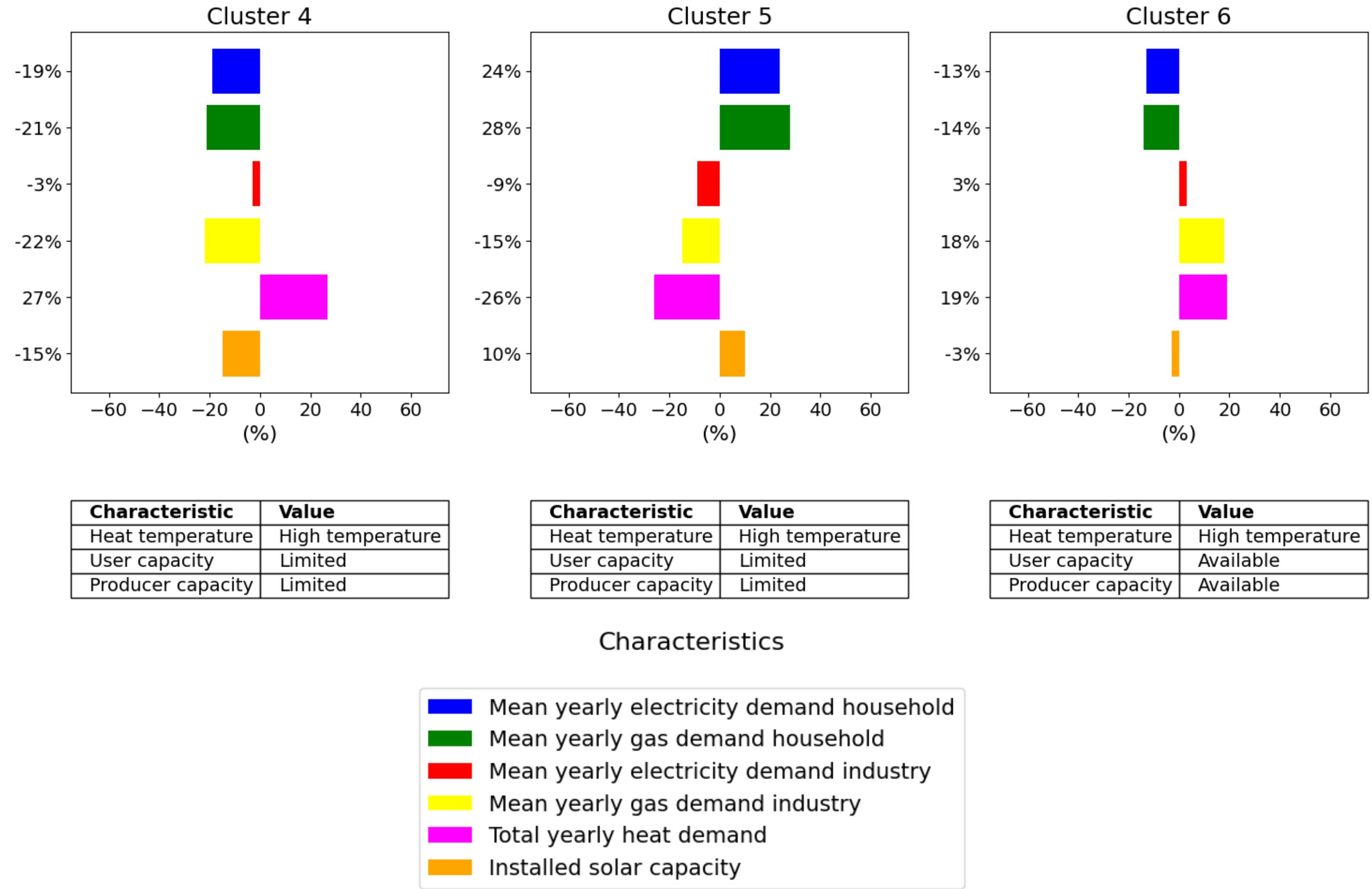


Characteristic	Value
Heat temperature	High temperature
User capacity	Limited
Producer capacity	Available

Characteristic	Value
Heat temperature	High temperature
User capacity	Not in near future
Producer capacity	Available

Characteristic	Value
Heat temperature	High temperature
User capacity	Available
Producer capacity	Available

Figure 4.6: A overview of the characteristics of the first three energy clusters



Characteristic	Value
Heat temperature	High temperature
User capacity	Limited
Producer capacity	Limited

Characteristic	Value
Heat temperature	High temperature
User capacity	Limited
Producer capacity	Limited

Characteristic	Value
Heat temperature	High temperature
User capacity	Available
Producer capacity	Available

Figure 4.7: A overview of the characteristics clusters 3 to 6

These plots show, that cluster 1 contains areas with lower-than-average household energy demand. In contrast, the yearly heat demand is slightly above average. This points to an above-average amount of houses, with each house using below-average amounts of energy. Notably, the electricity demand from industry and commercial buildings is relatively high within these areas. However, as pointed out before, this dataset only gives limited insights into the differences in industrial energy use. One interesting factor of this cluster is its similarity to the mean characteristics of the entire dataset. Because of this, it can be seen as a baseline cluster.

For cluster 2, the characteristics with the highest deviation from the mean value are reduced electricity demand for industrial use and reduced yearly heat demand. Another defining feature of this cluster is that it is the only cluster that mainly exists in areas with a significant limitation in the available network capacity of energy consumers. This may point to a causal relationship between these factors. However, further research would be required to confirm whether this is the case.

For cluster 3, the highest deviation can be seen in its relatively high energy usage per household. Notably, this cluster also has a reduced heat demand. This again possibly points to a reduced housing density. however, without the inclusion of built environment characteristics, these types of results can not be directly verified. For this reason, the built environment analysis will further look into the relationship between average and total demand energy demand. Cluster 4 shows exactly the opposite of cluster 3, with a reduced average energy demand over the board and a high total heat demand. This is also the only cluster with a significant reduction in the installed solar capacity compared to the average. This points to these areas either being underdeveloped with regard to solar energy or having a reduced potential.

Cluster 5 has similar characteristics to Cluster 3, with the notable exception of the energy demand of the industry and the availability of network capacity. This shows that although the differences between some areas are only limited a change to one characteristic can be enough to warrant a new its own cluster. Lastly, Cluster 6 follows a similar trend to Cluster 1. However, one major difference is the network capacity as this is the only cluster that almost exclusively contains areas with no network capacity restrictions.

By combining these energy characteristics with their geographical location, some trends could be observed. One notable thing is that areas with opposing energy demand characteristics can often be found close to each other. For example, clusters four and five are often found next to each other. In this case, four has a low average electricity demand, low average gas demand, and high total heat demand while five has a high average electricity demand, High average gas demand, and low total heat demand. Looking at the remaining characteristics of these clusters it can be seen that they are relatively similar. This shows that energy demand characteristics play a role in distinguishing in which of these clusters the areas are placed. The same can also be seen with characteristics 3 and 6.

Another factor that seems to impact the differences between clusters is the available network capacity. For example, when looking at The Hague, it can be seen that most of the areas are in clusters 1,4,6 which have many similarities in their characteristics. However, available network capacity does show a significant change from one cluster to another. This points to this being one of the factors influencing the differences between these areas. One thing to note however is that capacity does not explain all variations. For example, cluster two is mostly defined by the limited capacity in the South of the province. However, these limitations do not apply to the areas of cluster two surrounding Leiden.

4.1.3 Defining characteristics of energy clusters

To get a better understanding of the energy system, not only insights into the clusters were required but also into why certain areas were assigned to certain clusters. To visualize this, a plot was made that shows the characteristics that influence the cluster in which an area is placed. This was done by calculating the distance from each characteristic to the centroid using the same methodology as the k-prototype algorithm. These distances were then ordered to see which characteristics match the best with the cluster that the areas are appointed to. One problem with this approach is that the distance of categorical data can only be a one or a zero, depending on whether their values match or not. This would mean that the shortest distance would always be one of the categorical characteristics, as an area will always have at least one categorical characteristic in common with its cluster. Because of this, it was decided to treat the categorical data as one characteristic. This was done by taking the mean distance over all categorical characteristics. A drawback to this approach is that it limits the insights into the individual categorical characteristics. However, it does give better insights into what type of characteristics cause an area to be placed within a cluster. In Figure 4.8, a plot of all characteristics is shown with each step of the x-axis representing a position in the distance ranking and y representing the cumulative number of areas. This means that place one shows the amount of area for which the characteristic has the lowest distance value and place two is the sum of the areas where the characteristic has either the lowest or the second to lowest distance value. This plot shows that most of the areas have the categorical characteristics and the total yearly heat demand as their primary link to their cluster. Notably, the link between mean yearly electricity and gas demand for households seems to be relatively weak. This is interesting as the previous section showed the clusters had visible differences in household energy demand. This shows that although some clusters as a whole have a higher energy demand this does not have to be the case for individual areas as they are only loosely connected to their cluster based on this characteristic.

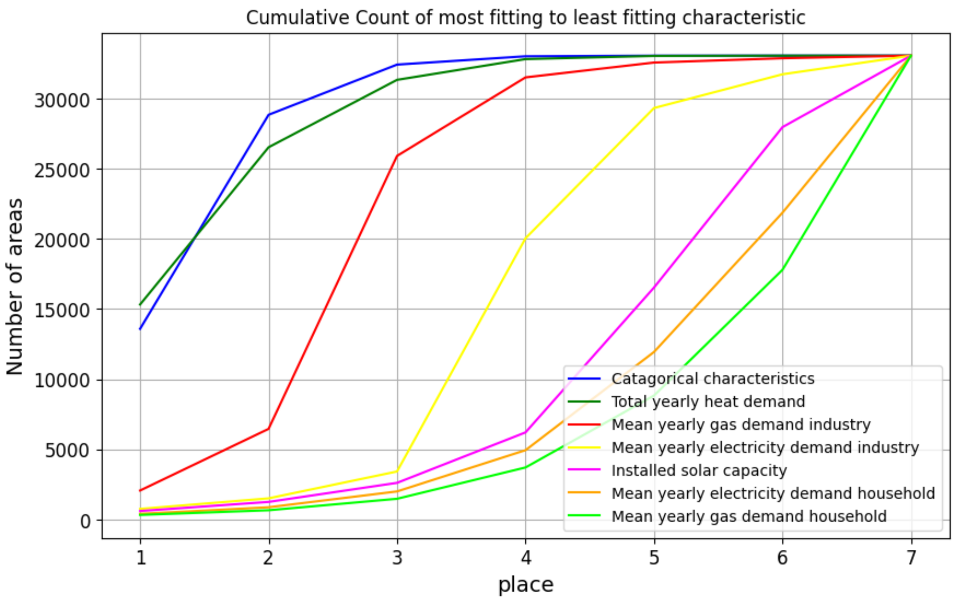


Figure 4.8: A cumulative plot of the match between areas and clusters based on their energy and built environment characteristics

Though this plot already shows the differences between characteristics it does not yet show the spatial aspect. In Figure 4.9 a plot is made of the characteristic with the closest match to the cluster for each area within the province. As expected this plot also shows that the total yearly heat demand and the categorical characteristics are the primary factors in the placement of areas within certain clusters. Based on these results, it can be seen that the categorical characteristics often are the main link between the areas and their clusters in urban areas like The Hague. On the contrary, the clustering in almost all of the rest of the province mostly focused on the total energy demand, with some small areas being defined by other characteristics. Interestingly, this map shows a similar disparity between The Hague and the other cities in the province as can be seen in the cluster map. this means that this can possibly be explained by a difference in the importance of a categorical characteristic, like the network capacity. Also noteworthy is that some areas that share the same cluster seem to have different primary characteristics linking them to this cluster. This shows that although the clusters are built on a set of similar characteristics, areas within these clusters should not be seen as uniform. Based on this analysis, the first insights into the energy system were collected. However, to better understand these insights, the relationship to social and built environment characteristics was analyzed within the following sections.

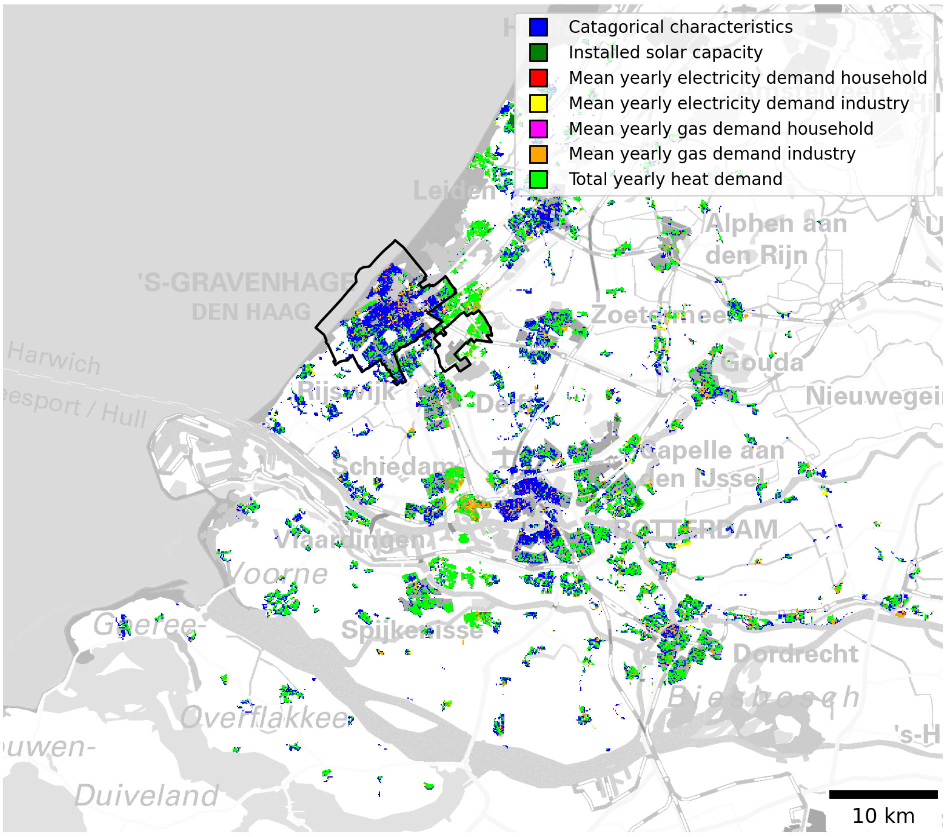


Figure 4.9: Primairy match between areas and the energy clusters

4.2 Energy and built environment cluster

After the baseline analysis, a cluster analysis involving both energy and build environment characteristics was performed. The goal of this analysis was to see how the clusters would change with the inclusion of other factors than energy characteristics.

4.2.1 Spatial distribution of energy and built environment clusters

In Figure 4.10, a map can be seen showing the location of each of the clusters. As noted before, this cluster analysis only made use of four clusters. This in itself already shows that by adding the built environment characteristics, changes to the clusters are made. Similar to the energy cluster analysis, most clusters are distributed over the entire area. However, the boundaries between different areas are less clear than with the energy clusters. This means that there are more areas within this analysis that show a mix between two clusters. One possible cause for this is lower boundaries between clusters compared to the energy clusters. This is supported by the silhouette score which is significantly lower for this set of clusters. Comparing this map to the energy clusters, some similarities in cluster location are visible. For example, cluster 3 fills a similar role to cluster 6 of the energy analysis. This is most apparent when looking at the map of The Hague (1). Here an almost identical layout of these clusters can be seen. However, it is also apparent that adding build environment characteristics does cause some of the clusters to change. An example of this can be seen in the portion of the map below the river Maas (2). In the energy clusters, this area was almost completely defined by one cluster, while the current clusters split these areas up into many different sections. Another noticeable difference between these results is that this analysis contains significantly fewer data points. This was because of two reasons. Firstly, less information about the built environment was available limiting the amount of data points in the dataset. Secondly, as this analysis includes characteristics the chance of a single missing data point is larger leading to the exclusion of more areas.

Looking at the spatial distribution of the points without the comparison to the energy clusters, it was observed that most major cities primarily were a combination of clusters 3 and 4 (3) with some smaller patches of cluster 1. The other areas of the province however had a more even spread with all clusters being present within these areas. In general, it can be seen that clusters 1 and 2 are more spread around the province with the exception of a few major hotspots around still a few major hot spots (4).

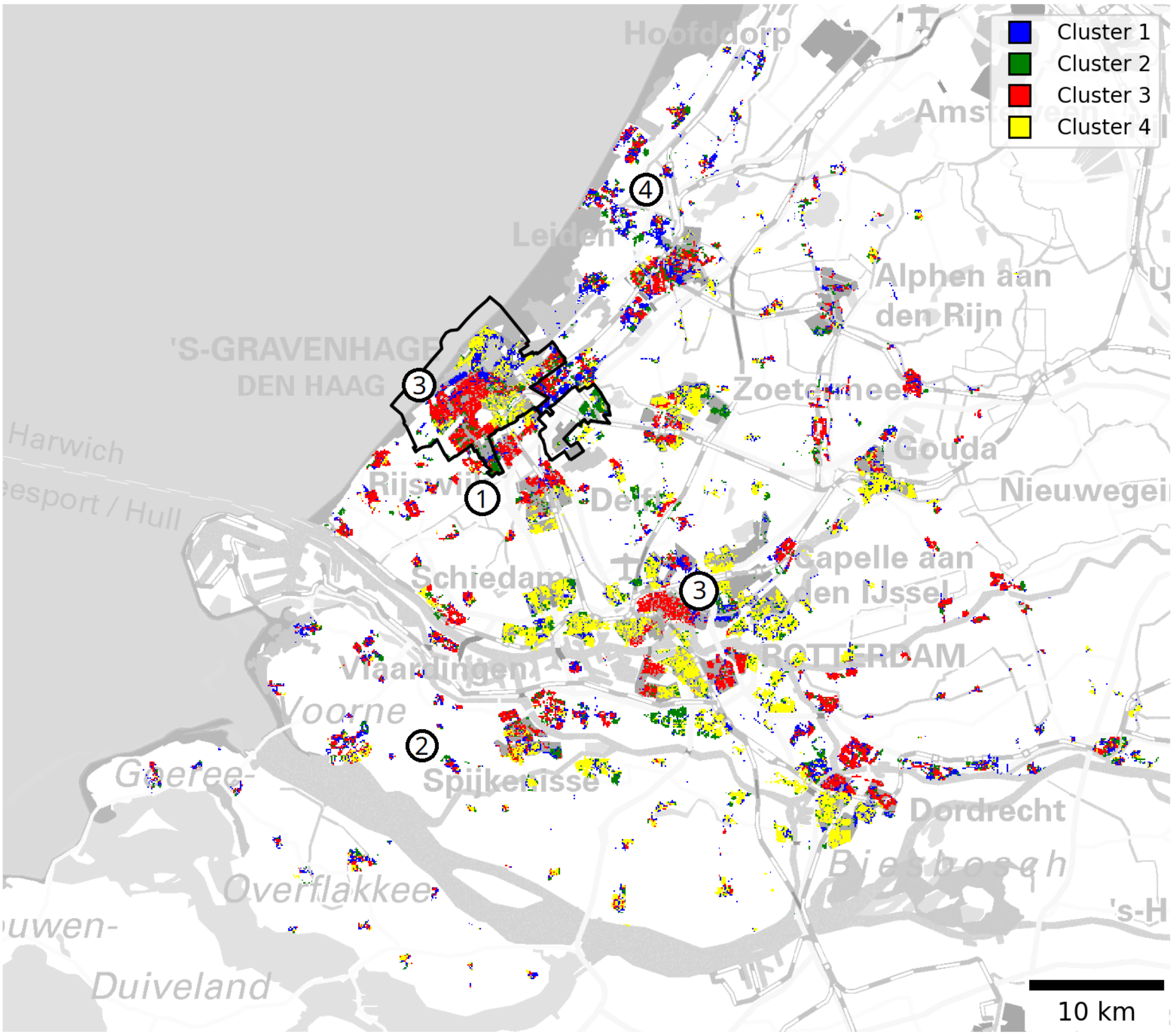


Figure 4.10: Locations of build environment clusters

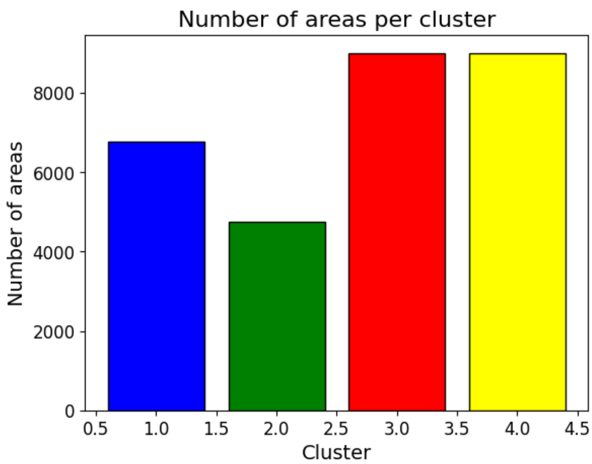


Figure 4.11: Number of areas per built environment cluster

Looking at the city of The Hague, many similarities can be seen to the energy clusters. similar to the energy clusters, one cluster is dominant in the (south)west of the city (1). In this case, this is cluster 3. Meanwhile, clusters 1 and 4 are dominant in most of the rest of the city (2). Cluster 2 can mostly be found within two groups on the edge of the city with some smaller areas mixed through the city (4). Though some separation between the clusters is still visible there is also more overlap within this analysis. For example, the city center is now a mix of all of the clusters.

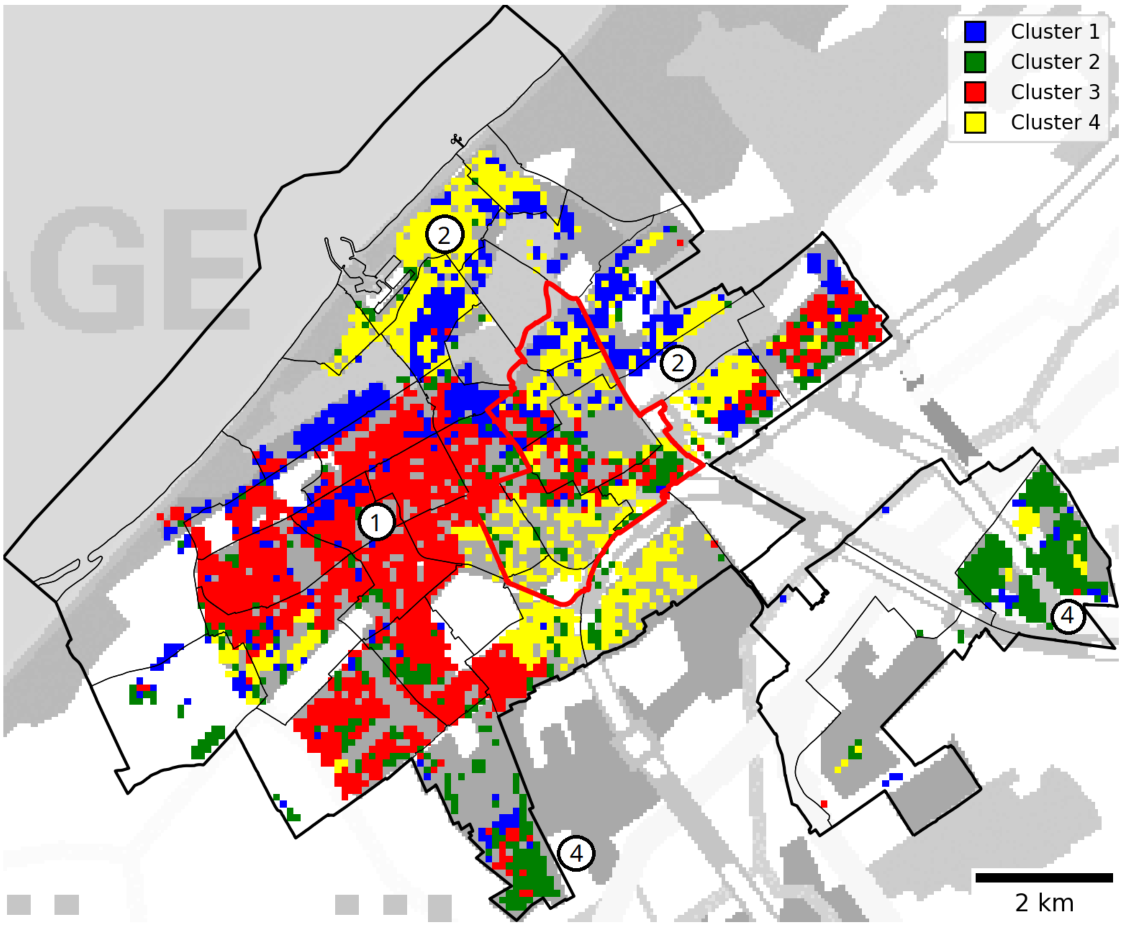


Figure 4.12: Locations of build environment clusters

4.2.2 Cluster characteristics of energy and built environment clusters

To get a better insight into what this cluster distribution means for the energy and built environment, the clusters were compared. Here, the main focus was on analyzing how the inclusion of the built environment would affect the clusters. Looking purely at the average electricity demand suggests that cluster 1 has the highest energy demand. However, the total heat demand shows that the clusters with a high total heat demand have a relatively low energy demand and the other way around. One of the main factors behind this seems to be the number of inhabitants as can be seen in Figure 4.13. This is in line with the expectations stated in the energy cluster analysis and shows that energy and the built environment are indeed connected.

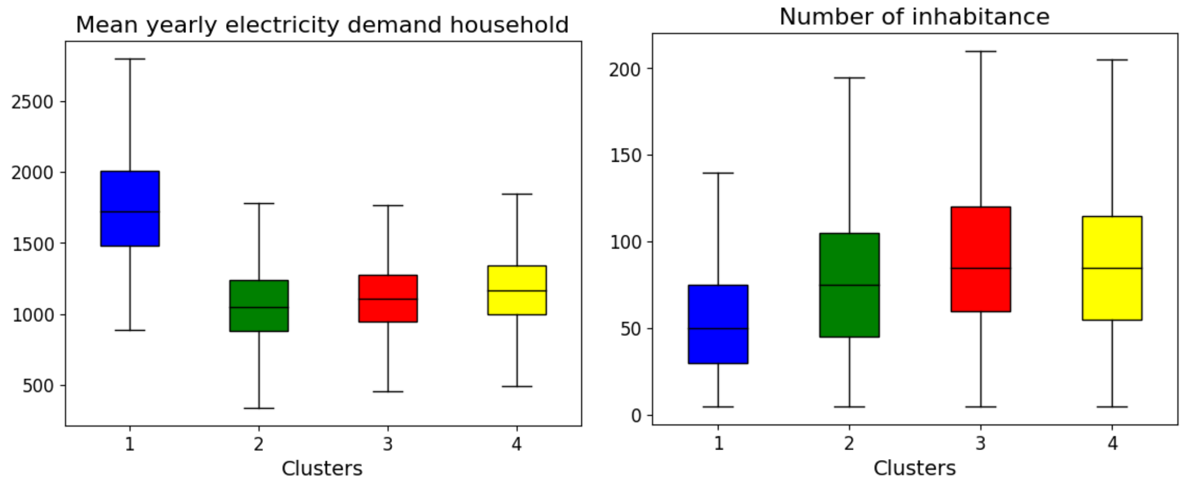


Figure 4.13: Mean electricity demand and number of houses per built environment cluster

Another of the main distinctions that can be made is in the built year of housing. Here, a clear distinction can be made between cluster 1, which mostly has houses from before 1945, and cluster 2, which mostly contains housing built after the year 2000. This shows a link between energy and the built environment as the clusters with older housing also contain higher average electricity and gas demand for households. Notably all clusters include at least some houses from every category of built year. This again shows that not all areas with a cluster are the same and that some differences between the area within a cluster will exist.

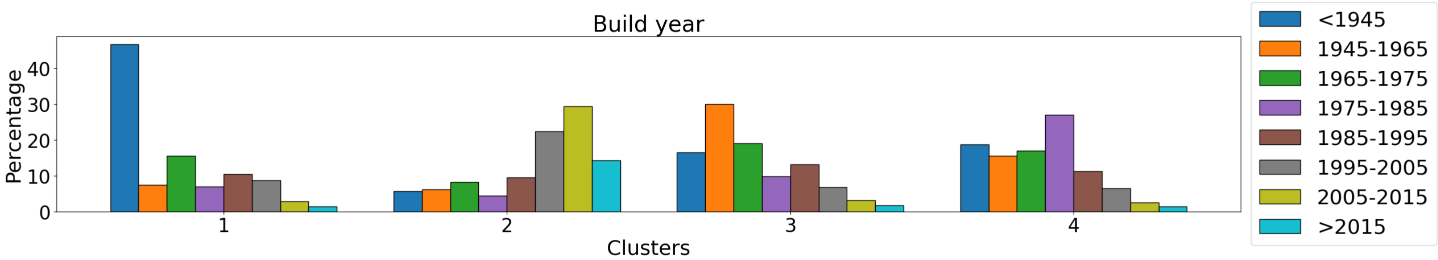
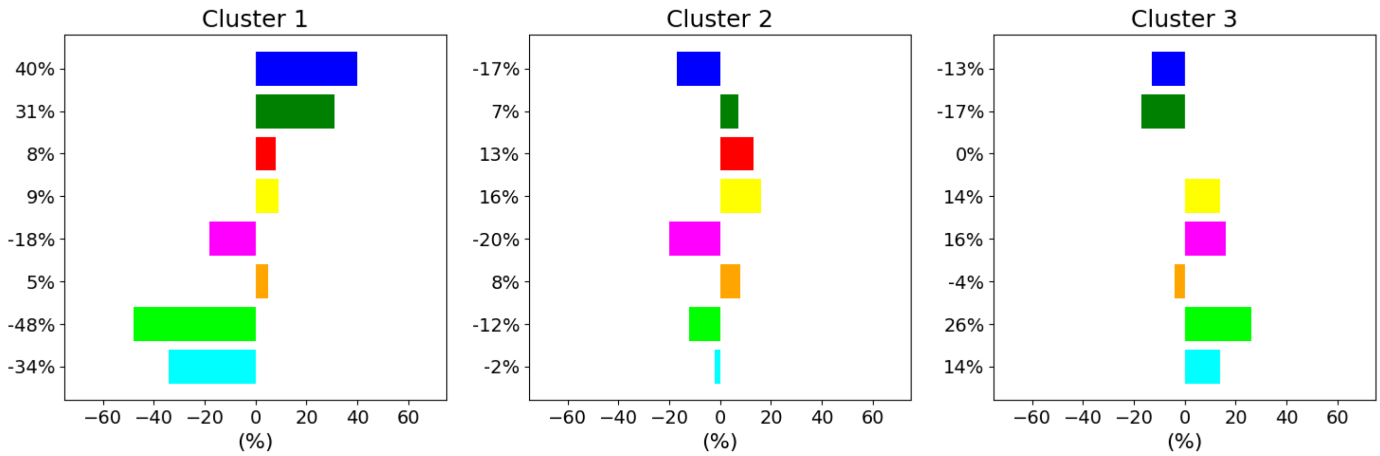


Figure 4.14: Build year of average house per built environment cluster

To get a better insight into how these building characteristics affect the distribution of the clusters, a plot of each cluster was made. These plots show that cluster 1 is characterized by older buildings with a relatively low energy label. Because of this, gas and electricity usage is relatively high per household. However, because of the low housing density, the overall heat demand is relatively low. Just like most of the clusters, the main type of land use of this cluster was residential.

The second cluster, on the other hand, mostly consists of newer houses with a higher energy label. Because the building density within these areas is also relatively lower, this means that the total heat demand is also below the average. One noteworthy feature of this cluster is that it is the only cluster where a majority of the housing can be provided with low-temperature heating. It is also the only one with an alternative land use in the form of transport. This however does not have to mean that this area mostly includes highways or strain tracks as small roads were also included in this category. This most likely means that these areas have a relatively high amount of roads compared to the amount of housing.

Cluster 3 shows similarities with the first cluster in the build year and energy label. However, its energy characteristics are different. This can mostly be seen based on its below-average energy usage and relatively high heat demand. Besides this, it also has a higher-than-average housing density. Lastly, cluster 4 mostly has a combination of features of the previous clusters with a relatively low energy label, a slightly raised housing density, and slightly lower household energy demand. Two features that do distinguish the cluster are its relatively low gas demand for industrial uses and the fact that it is the only cluster where almost all of the areas have a limited network capacity both for energy suppliers and for users.

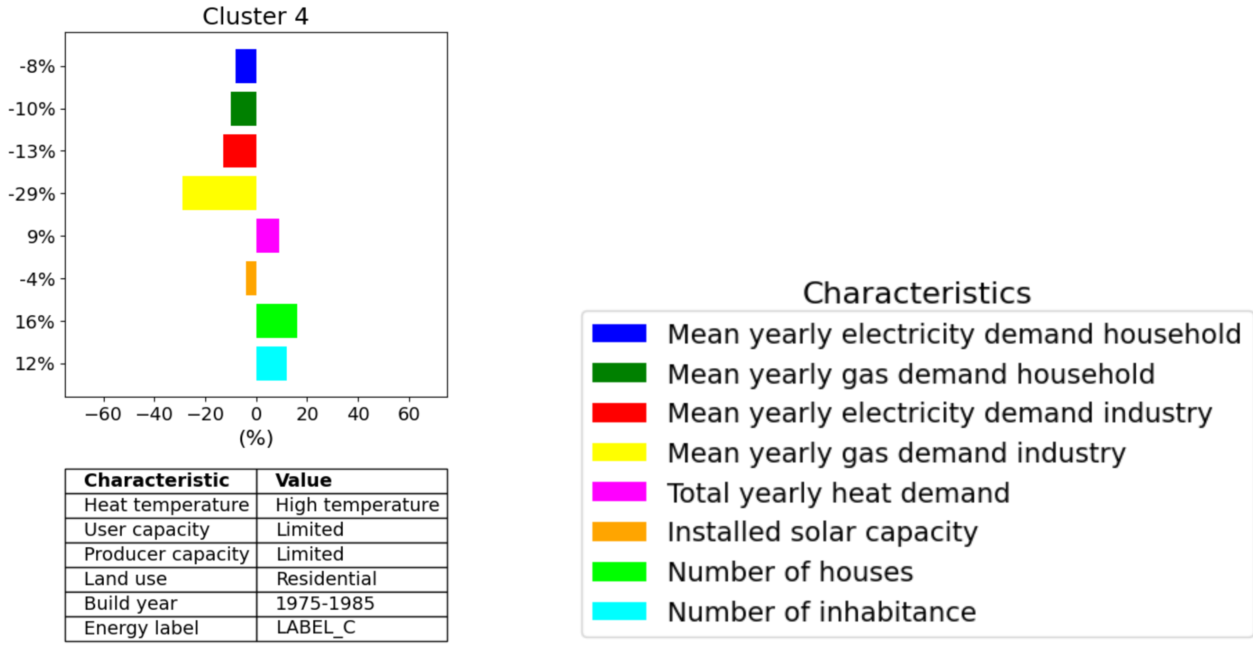


Characteristic	Value
Heat temperature	High temperature
User capacity	Limited
Producer capacity	Available
Land use	Residential
Build year	<1945
Energy label	LABEL_C

Characteristic	Value
Heat temperature	Low temperature
User capacity	Limited
Producer capacity	Available
Land use	Transport
Build year	2005-2015
Energy label	LABEL_A

Characteristic	Value
Heat temperature	High temperature
User capacity	Available
Producer capacity	Available
Land use	Residential
Build year	1945-1965
Energy label	LABEL_C

Figure 4.15: Overview of the first, second and third cluster of the built environment



Characteristic	**Value**
Heat temperature	High temperature
User capacity	Limited
Producer capacity	Limited
Land use	Residential
Build year	1975-1985
Energy label	LABEL_C

Figure 4.16: Overview of the last cluster of the built environment

A common trend between the clusters is that housing density and total heat demand seem to be linked to each other. Furthermore, a clear trend can be seen where a high total heat demand often means a low per-average energy demand and vice versa. As to be expected, there also seems to be a relationship between housing density and the number of inhabitants. In the case of the land use type all but one cluster where marked residential with the last being marked as transport. Transport in this context includes all transport infrastructure like roads or train tracks. Although this does not mean that these areas only include roads it does mean that on average more surface area is assigned to roads and other forms of transport than housing. Notably, none of the clusters include agriculture while the original dataset was predominantly agriculture as can be seen in Appendix 4. This is because most of the rural areas were removed because of a lack of data leaving predominately residential areas.

By combining these findings with the spatial layout, some further observations were made. The city of the Hague shows similar characteristics to the energy cluster. However, two differences can be seen. The first of these is a difference in industrial energy demand. Here areas that previously were part of areas with a low demand are now in a cluster with a low demand. The second more prevalent difference is the availability of network capacity. This characteristic is far more prevalent in the energy clusters than in the built environment clusters. However, there are still areas where this characteristic can be observed like in cluster 3.

Another observation is that housing with a relatively higher average energy demand is spread through the province with only a few major clusters. This can be linked to the built environment characteristics as these are also the areas with on average the older buildings. The same can be seen with cluster 2, which can mostly be found in places with relatively new housing.

4.2.3 Defining characteristics of energy and built environment clusters

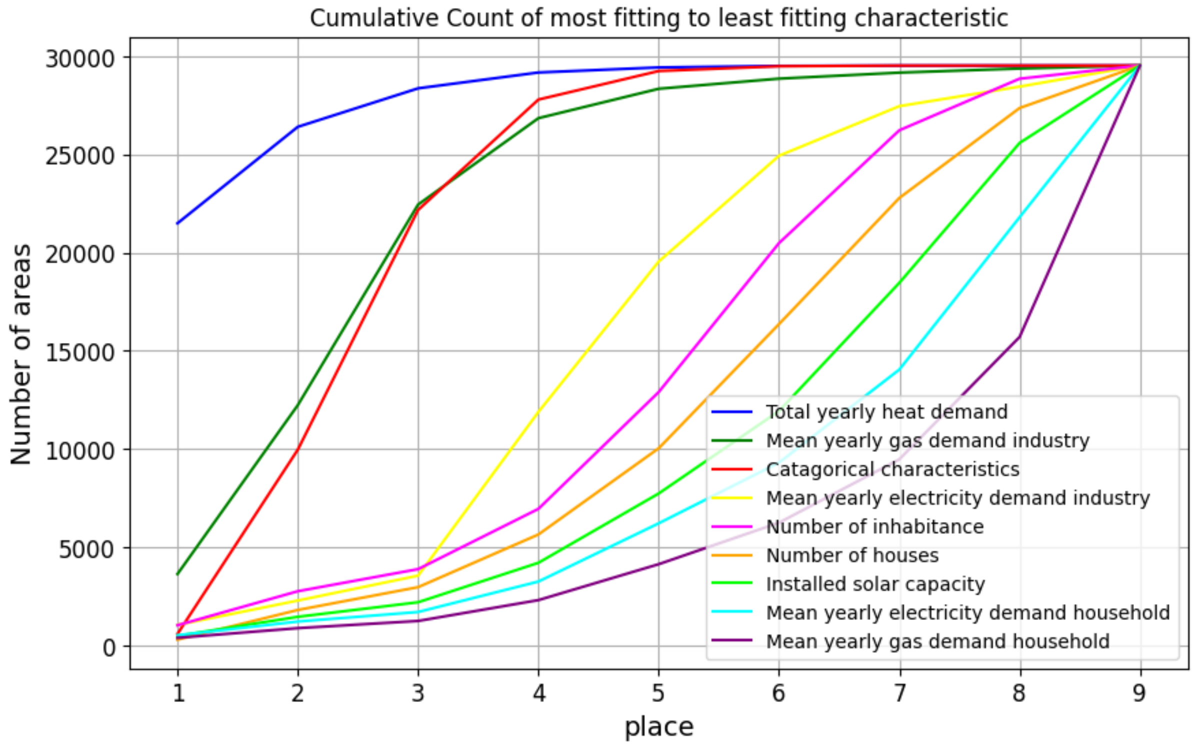


Figure 4.17: A cumulative plot of the match between areas and clusters based on their built environment characteristics

Similar to the energy characteristics the main factors for cluster selection were mapped. In Figure 4.17 an overview of these clusters can be shown. Just as in the previous analysis most clusters either have a close match with their total yearly heat demand or their categorical characteristics. This shows that these characteristics often have an influence on the cluster designation of areas. Notably industrial characteristics also often seem to have a close match even though the clusters only show a small difference between these characteristics. This seems to be caused by the fact most areas do not have any industry which means that their value for these characteristics is 0. As the average value of most clusters is also close to 0, this means that the distance between these values is relatively small. Interestingly, all of the newly added characteristics seem to score better on this metric than the mean yearly electricity and gas demand of households. This points to these characteristics having more influence on the clusters than these energy characteristics.

To further analyze the relationship between the clusters and their areas, the primary characteristic linking each area to its cluster was plotted in Figure 4.18. Similar to the previous plot, many areas are primarily defined by their total heat demand. What is different is the placement of areas based on their categorical characteristics. This characteristic is way less apparent in these clusters than in the energy clusters. Although the relationship between the categorical data and the cluster is still relatively strong, as can be seen in Figure 4.17, this does mean that these characteristics play a smaller role compared to the previous cluster analysis. Besides this, there is also a smaller subgroup of characteristics that are defined by the energy demands of the industry. These mostly match up with areas in cluster four, which is a cluster with a relatively low industrial energy demand.

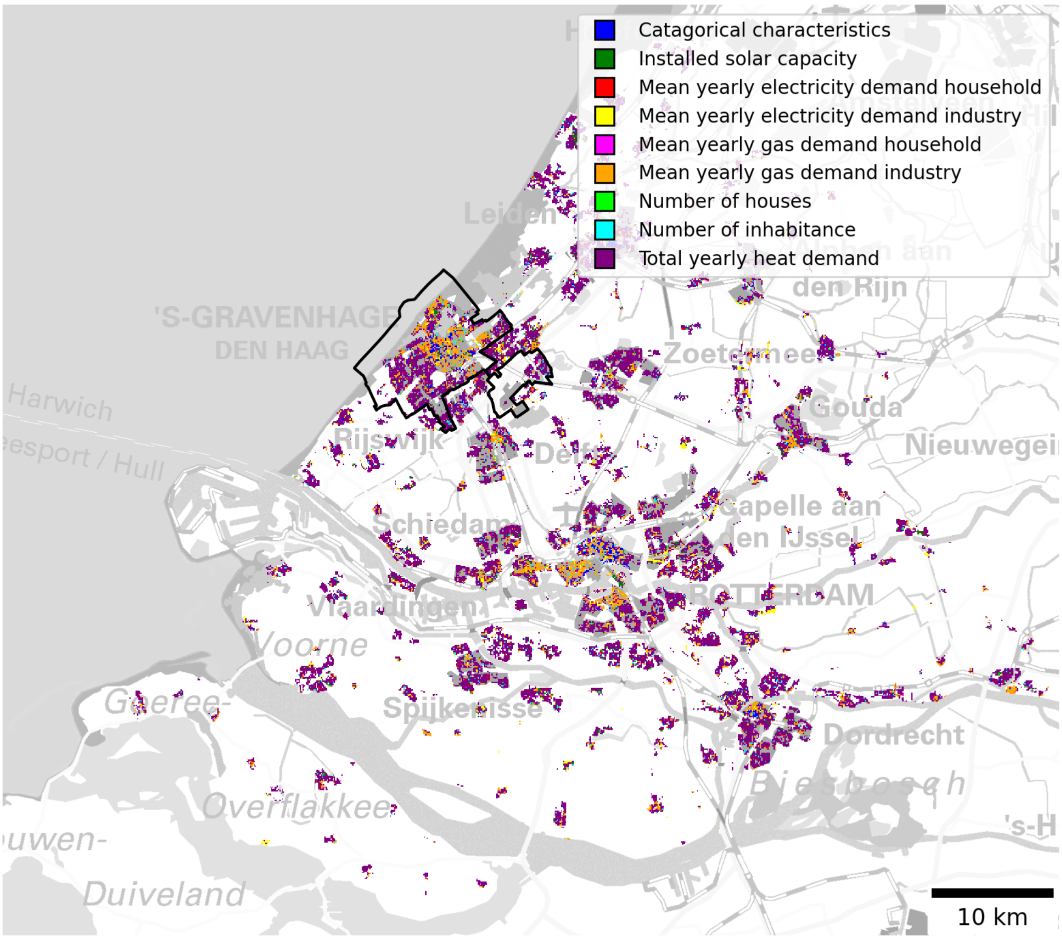


Figure 4.18: Primairy match between areas and the built environment clusters

4.3 Energy and social cluster

The third analysis that was performed was a combination of energy and social characteristics. Just like the previous analysis, the goal was to see the impact of a combination of energy characteristics and another set of characteristics. Here the intention was not only to see whether any changes would occur compared to the case with only energy characteristics but also compared to the clusters from the combined energy and built environment analysis.

4.3.1 Spatial distribution of energy and social clusters

Similar to the built environment clusters, a clear reduction is visible in the amount of areas that are included in this analysis. This resulted in only limited data for some areas, like Goeree Overflakke that were present in the previous analysis(1). Although enough information was present to perform this type of analysis, this does mean that this analysis was even more focused on large urban areas than the previous clusters. In Figure 4.19 an overview of all social clusters is shown. Comparing this to the previous two sets of clusters it can be seen that this analysis leads to different clusters than either analysis. One major difference is that there are fewer areas with a high concentration of the same cluster. An example of this can be seen in Leiden which is a mix of all clusters(2). Besides this, there is also no clear difference between different urban areas with almost all containing a mix of the four clusters(3).

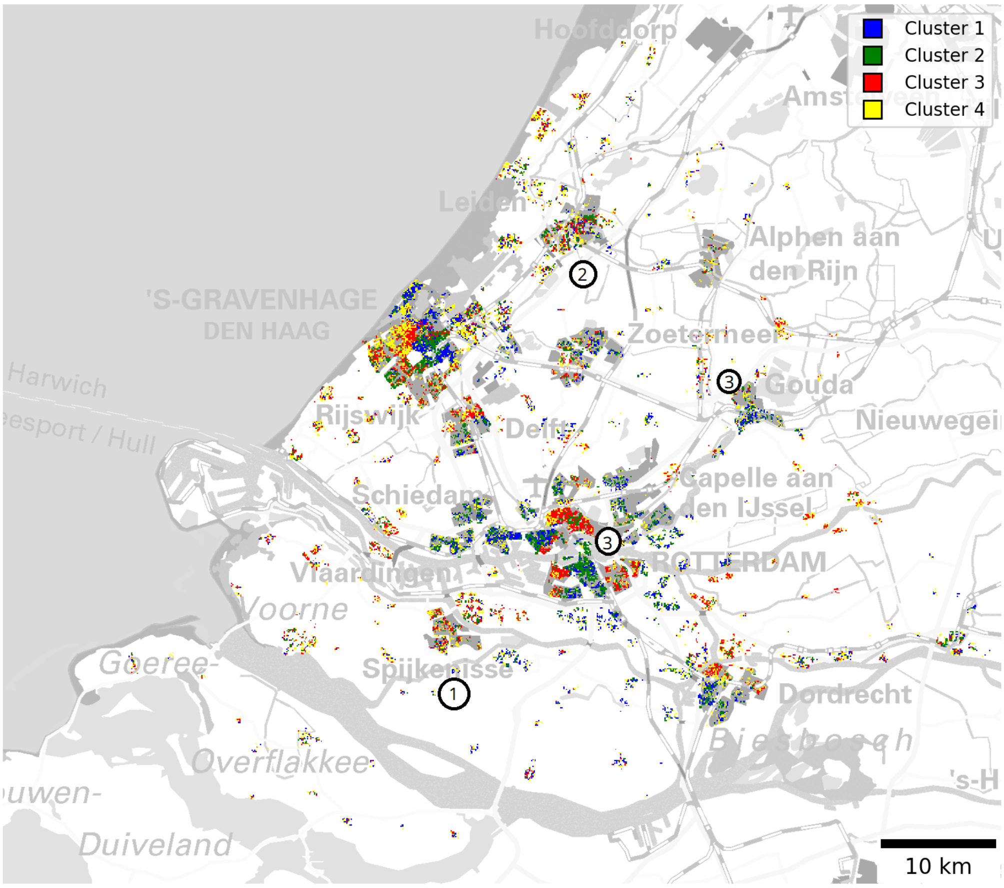


Figure 4.19: location of social clusters

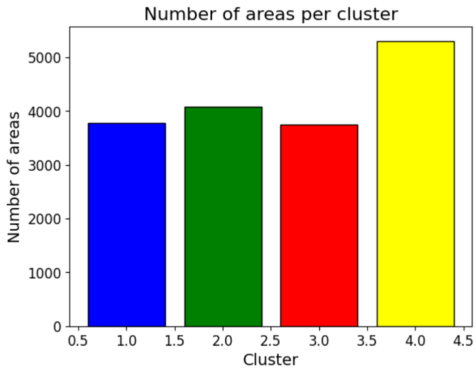


Figure 4.20: Number of areas per social cluster

One place where the differences compared to the other cluster sets are especially apparent is The Hague. Here, clearly grouped clusters could be found in previous cluster analyses which are less apparent in the social clusters. Although some areas have similarities to the previous analyses like one cluster being prevalent in the (South)West of the city(1), these clusters are more mixed with another cluster. They often include areas that were previously part of other clusters. Because the borders of the cluster are less defined, no clear insights could be found solely based on the cluster distribution. For this reason, a further analysis of the clusters themself was performed.

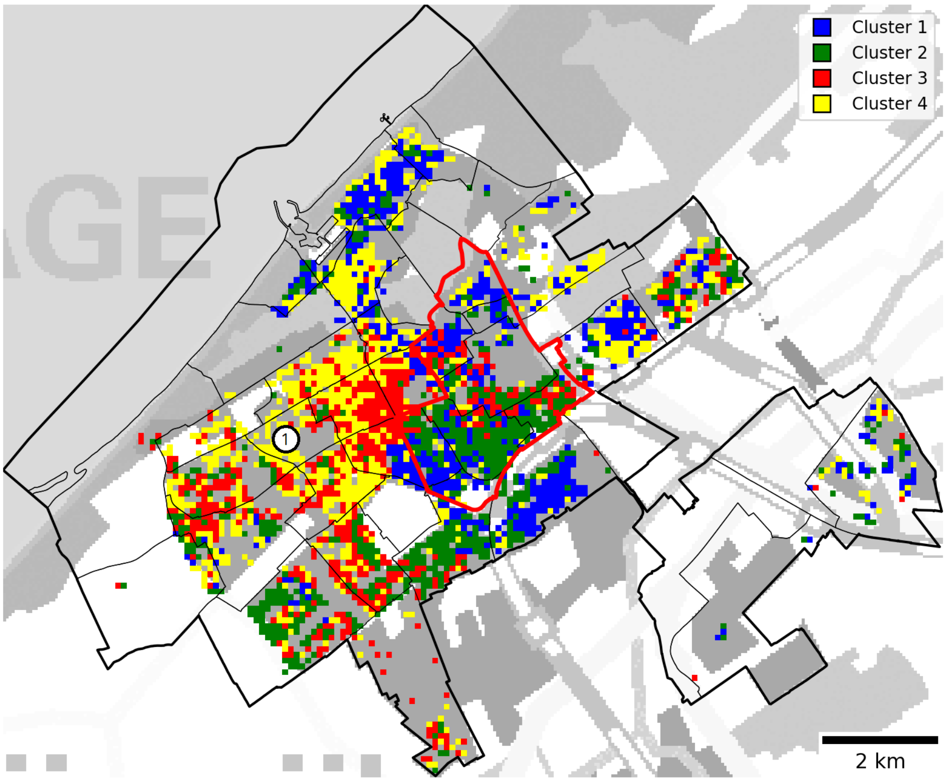


Figure 4.21: Spatial distribution of social clusters in The Hague

4.3.2 cluster characteristics of energy and social clusters

Looking at the newly included spatial characteristics it can be seen that these characteristics are different between clusters. In Figure 4.22 an overview of the percentage of rental housing and the average income per household can be seen. This shows that there are clear differences between clusters. In particular, the percentage of rental housing shows well-defined differences with only limited overlap between the clusters. As expected there are also similarities between these two characteristics with a high percentage of rental housing matching with a higher average income.

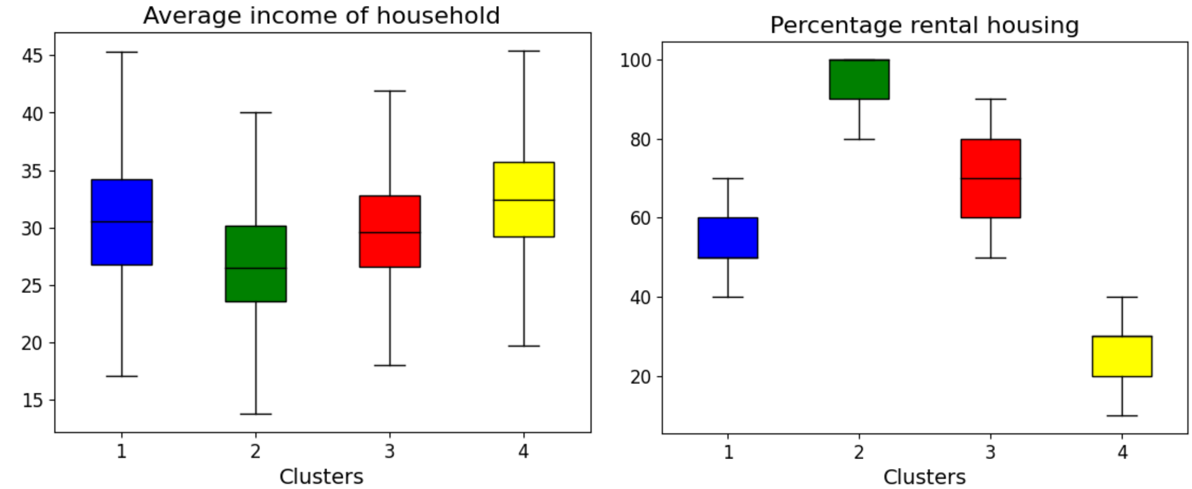


Figure 4.22: Percentage of rental housing and average income per cluster

Interestingly this same relationship can also be seen between the energy and social characteristics of clusters. Comparing the mean yearly electricity demand of households with their percentage of rental housing shows that clusters with a high value on characteristics on average have a lower value on the other and vice versa. Although this information is not enough to define a correlation between these characteristics this does show that the social aspect influence the clusters.

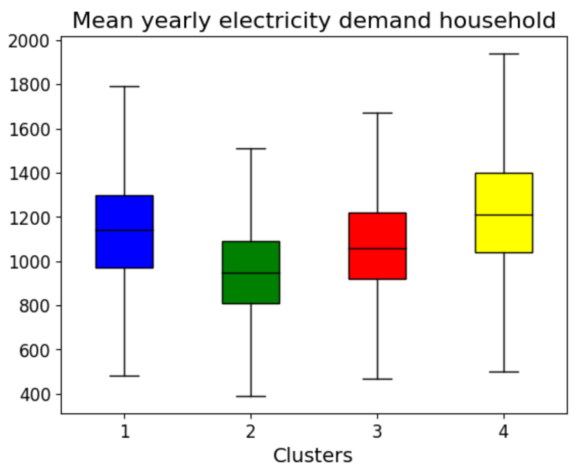
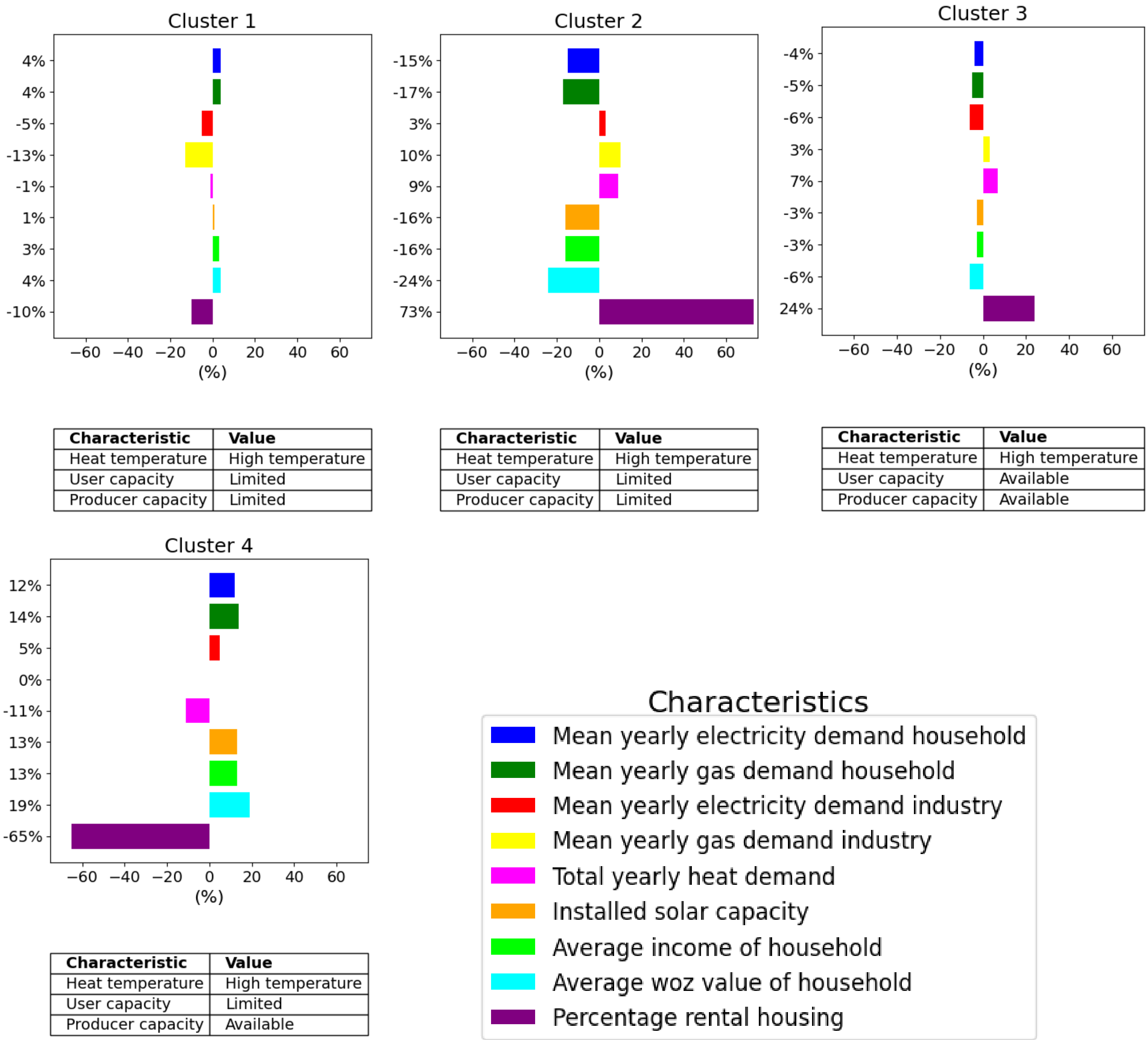


Figure 4.23: Average yearly household energy demand social clusters

Similar to the previous analysis the clusters were compared based on their deviation from the mean value for the characteristics. In Figure 4.24 an overview of each of the clusters can be seen.



Characteristic	**Value**
Heat temperature	High temperature
User capacity	Limited
Producer capacity	Limited

Characteristic	**Value**
Heat temperature	High temperature
User capacity	Limited
Producer capacity	Limited

Characteristic	**Value**
Heat temperature	High temperature
User capacity	Available
Producer capacity	Available

Characteristic	**Value**
Heat temperature	High temperature
User capacity	Limited
Producer capacity	Available

Figure 4.24: Overview of social clusters

Looking at the characteristics of each cluster, a clear difference can be seen between cluster 2 and cluster 4. Here, cluster 2 is characterized by a relatively low average energy demand combined with a high percentage of rental housing, low average woz value, and low average income. Cluster 4, on the other hand, is defined by a relatively high average energy demand, low percentage rental housing, high average woz value, and high income. This relation shows that the algorithm was able to find areas within the province with matching social and energy characteristics.

Cluster 1 mostly follows similar trends to Cluster 4. However, the deviation compared to the mean value is smaller. What is also different is the network capacity. Here, Cluster 1 has a higher amount of available network capacity compared to Cluster 4 where this is mostly limited. The same can also be said for cluster 3 compared to cluster 2. Noteworthy is the relatively big distance between the installed solar capacity between clusters 2 and 4 as this only saw limited differences in the other cluster analysis. This shows that relatively more solar capacity is installed in areas with fewer rental homes and a higher income. This can possibly be explained by the difference in effort it takes to install solar panels on privately owned homes compared to rental housing. Besides this, the relative income difference might also play a role in this difference.

Although these clusters are different from the energy and build environment clusters some similarities can be seen. For example, the heat demand seems to have the same relationship with the average energy demand in the social cluster as in the other cluster analysis. This shows that although this analysis mainly focuses on social characteristics, there seems to be some overlap between the energy, built environment, and social characteristics that lead to similar behaviors. Based on these findings, it can be seen that the clusters from this analysis show a relationship between the energy characteristics and the social characteristics. However, other factors like network capacity and energy demand of industry also still play a role within these clusters.

4.3.3 Defining characteristics of energy and social clusters

Similar to the other analysis figure 4.25 shows that most areas have the closest match to their clusters with the total heat demand characteristic. Interestingly the relationship between the social characteristics and the clusters is less apparent even though these relationships are visible in the clustering results themselves. This is especially apparent for the characteristic percentage of rental housing which has the lowest match even though the clusters do show a significant variation in this characteristic. This shows that there is not only variation in the average percentage of rental housing between clusters but also between areas within the clusters. It also shows that although all sets of clusters primarily follow heat demand and the categorical characteristics the other characteristics do have an influence on the clustering.

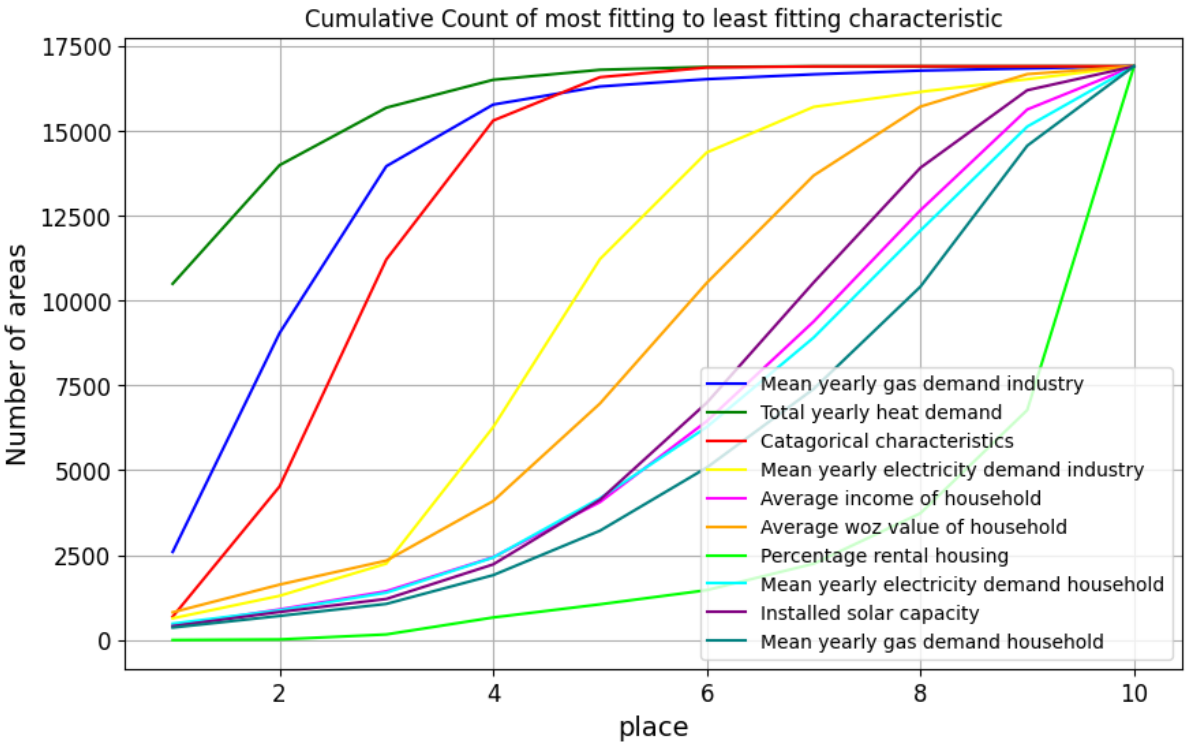


Figure 4.25: A cumulative plot of the match between areas and clusters based on their social characteristics

In Figure 4.26 a map of the primary factors can be seen. Similar to the built environment clusters there are no clear areas where another characteristic than heat demand is dominant. This is mostly because the different characteristics are spread throughout all areas. This shows that the placement of a cluster in a certain cluster can differ between neighboring areas.

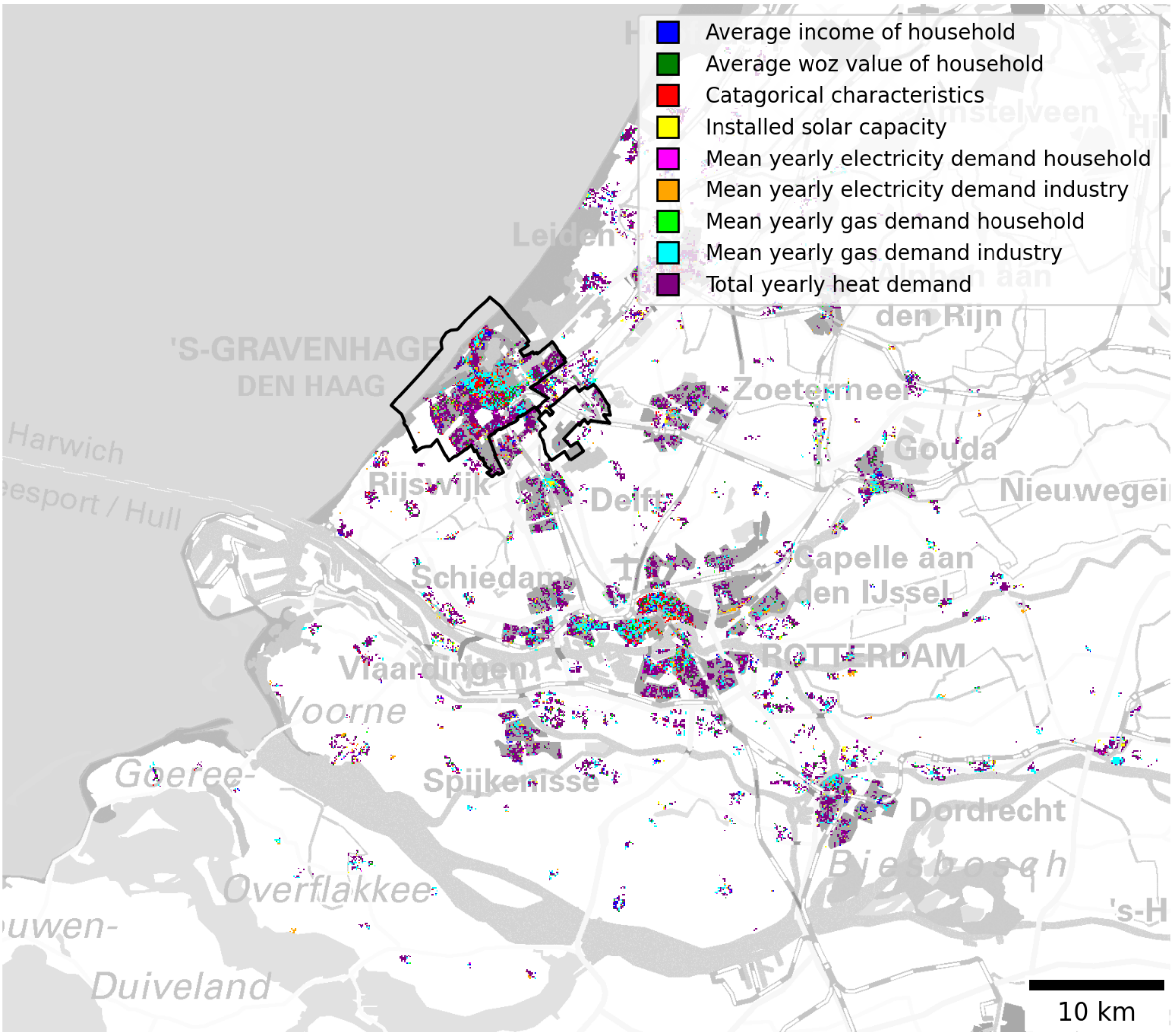


Figure 4.26: Primairy match between areas and social clusters

4.4 Combined cluster

The last analysis that was performed was a combination of all other cluster analyses. The main goal of this analysis was to see whether the same features that are present in the separate analysis can also be seen in a combined analysis or where these features get lost.

4.4.1 Spatial distribution of the combined clusters

As this analysis includes all characteristics the amount of areas with one or more clusters missing is higher than the previous three analyses. This means that fewer areas are included in this analysis. Similar to the previous analysis the removed areas were mostly rural areas and small towns. This means that the remaining areas are mostly the larger urban areas. Looking at Figure 4.27 it can be seen that the major groupings of clusters that were present in the energy and built environment clusters are less prevalent in this analysis. The exception to this area is a few areas in The Hague and Rotterdam which still have areas relatively separated clusters.

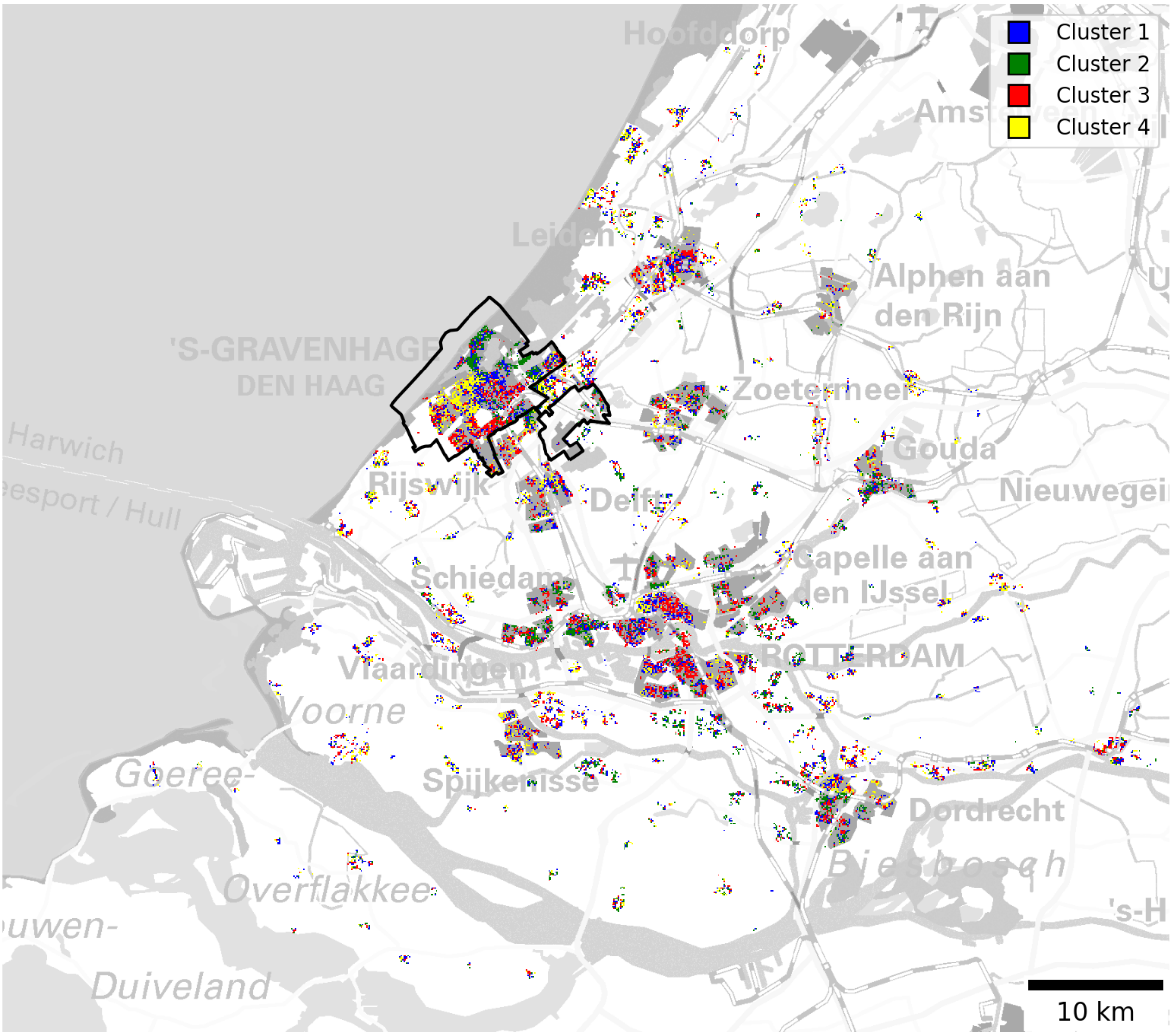


Figure 4.27: Location of combined clusters

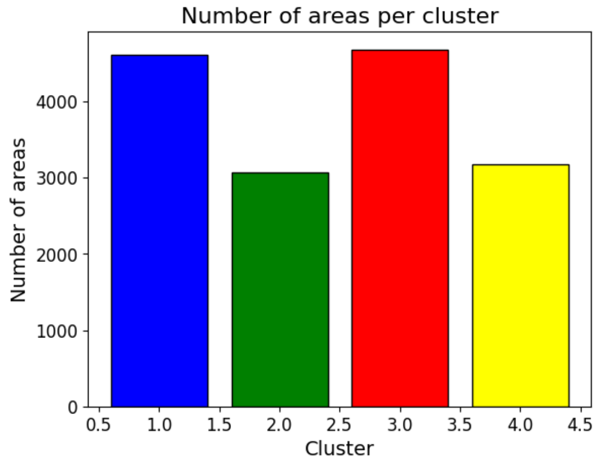


Figure 4.28: Number of areas per combined cluster

Looking only at The Hague as plotted in Figure 4.29 some similarities between these clusters and some of the clusters from other analyses can be seen. For example, the coastal areas fall within two clusters similar to both the energy clusters and the built environment clusters (1). Though some similarities can be seen many of the previous features are also missing. One major feature that is missing is the big cluster in the (South)west of The Hague that was present within the energy and built environment cluster (2). This already shows that these combined clusters contain features from all different cluster analyses leading to different clusters. Though this does show that all clusters are probably involved within these clusters this also already shows a drawback of using one combined cluster. This is because specific cluster characteristics of some of the clusters get lost during the combination.

Besides this, the separation between the clusters is also reduced compared to the other analysis with more neighborhoods having multiple clusters within them. Notably, the number of areas that are assigned to each cluster seems to be more evenly divided within this analysis compared to the other analysis with none of the clusters having a large majority within the city.

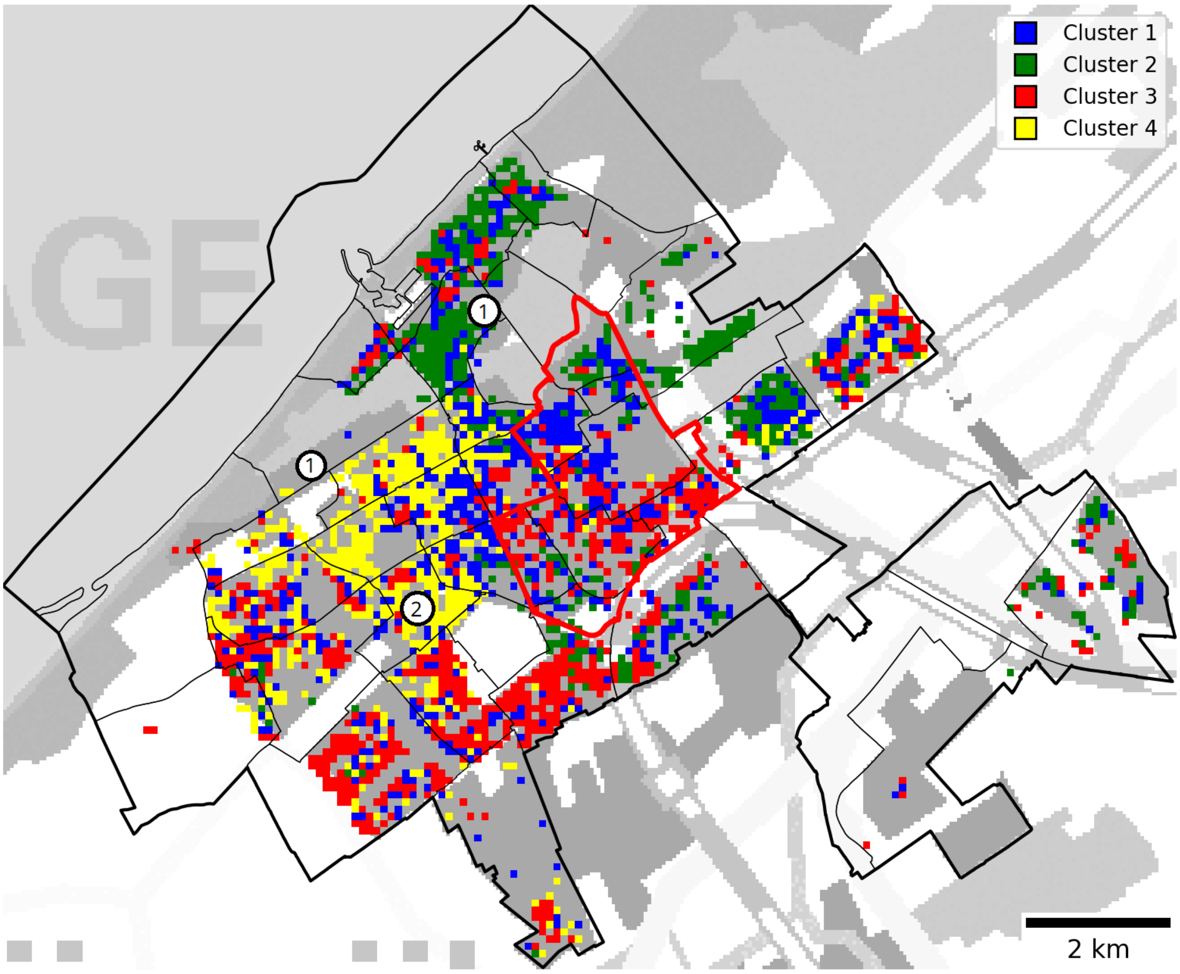
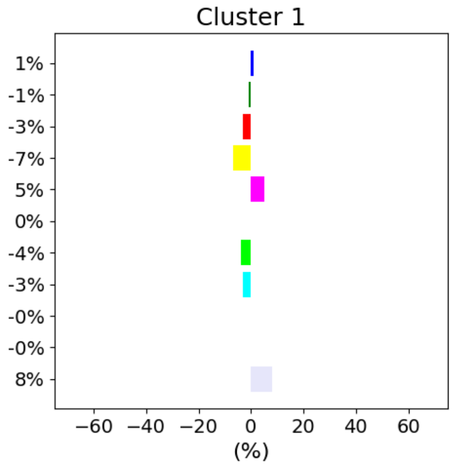


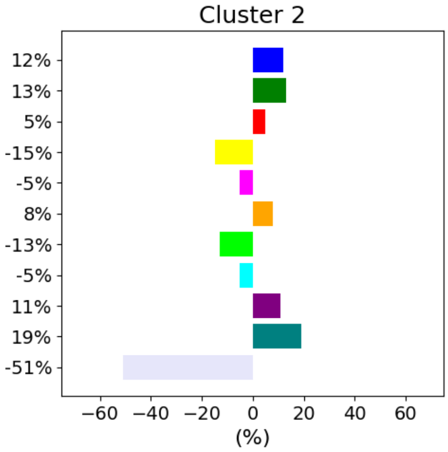
Figure 4.29: Spatial distribution of social clusters in The Hague

4.4.2 cluster characteristics of the combined clusters

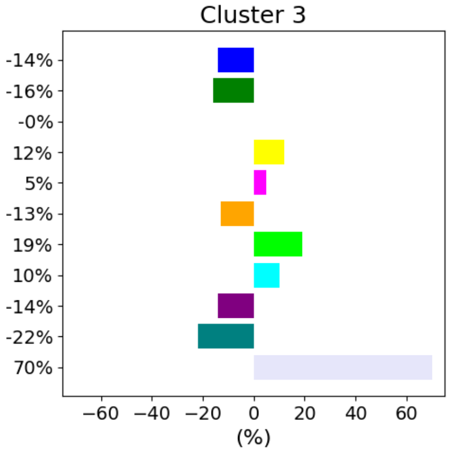
Looking at the individual clusters, it can be seen that cluster 1 does not contain any major deviations from the mean characteristics. On the other hand, cluster 2 does show some differences, with this cluster having a high energy demand, solar capacity, and average income. It also has a low amount of houses and a low amount of rental housing. For the categorical characteristics, it mostly seems to have the same characteristics as cluster 1. Cluster 3 seems to be the opposite of Cluster 2, with a low average energy demand and a high number of houses and rental housing. The categorical characteristics, however, seem quite similar to the other clusters, showing that these clusters were only able to do a limited sorting based on these types of characteristics. Lastly, the fourth cluster mostly seems to be similar to the second cluster with both having similar, energy and social characteristics. The two areas where there are differences are in network availability and the mean yearly gas demand for industry.



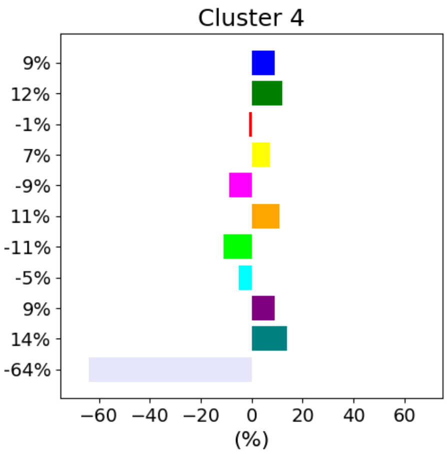
Characteristic	Value
Heat temperature	High temperature
User capacity	Limited
Producer capacity	Available
Land use	Residential
Build year	<1945
Energy label	LABEL_C



Characteristic	Value
Heat temperature	High temperature
User capacity	Limited
Producer capacity	Limited
Land use	Residential
Build year	<1945
Energy label	LABEL_A



Characteristic	Value
Heat temperature	High temperature
User capacity	Limited
Producer capacity	Available
Land use	Residential
Build year	1945-1965
Energy label	LABEL_C



Characteristic	Value
Heat temperature	High temperature
User capacity	Available
Producer capacity	Available
Land use	Residential
Build year	<1945
Energy label	LABEL_C

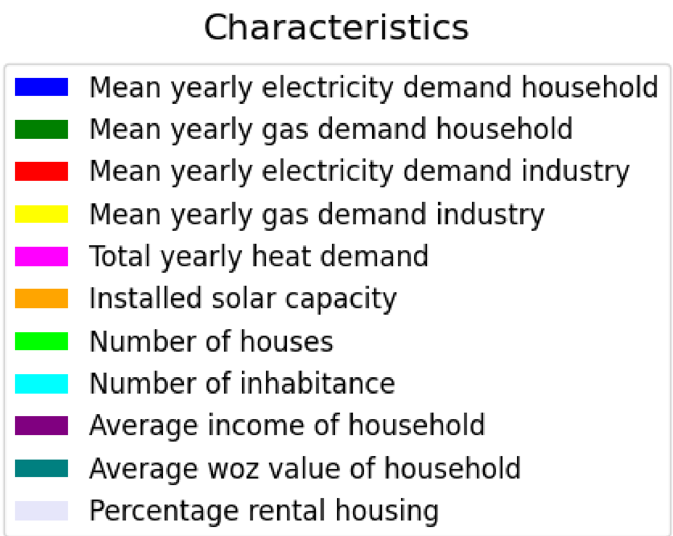


Figure 4.30: Overview of combined clusters

Looking at the clusters, a combination of features of the different cluster types can be seen. For example, similar to the social clusters, major differences between the percentage of rental housing can be seen. Similarly, a clear relationship between mean energy demand and yearly heat demand can still be seen where they are opposite in all clusters. This shows that even though more characteristics were included this does not mean that all properties of the previous analysis were lost.

However, some properties of the previous characteristics are less visible within this analysis. In particular, the categorical characteristics showed fewer deviations between the clusters. Examples of this are energy label and land use which are all almost identical in Figure 4.30. Similarly, differentiation between different network availabilities is also lost within this analysis. Figure 4.31 shows that there are only limited differences in the distributions in the built years of housing between the clusters. Here almost all clusters have <1945 as their prominent built year with a reduction in each subsequent built year. Comparing this to Figure 4.14 which shows this same distribution for the built environment cluster shows that the impact of this characteristic has become less apparent. Combined these findings show that although some features are kept when including more characteristics this is not the case for all. Besides these findings, no new behavior was found within this analysis which was not included within one of the other analyses.

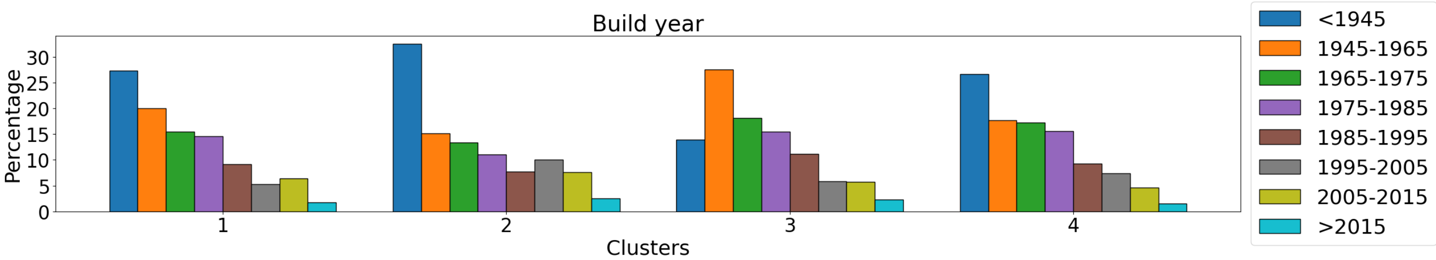


Figure 4.31: Build year of average house per combined cluster

4.4.3 Defining characteristics of combined clusters

Similar to the other analyses, the main characteristic defining the clusters is total yearly heat demand. Besides this, a variety of other characteristics can be seen that define the cluster of an area. Figure 4.32 shows the match between all characteristics and the areas. Here most of the characteristics follow a similar trend to the previous analysis like a relatively high score for the industrial characteristics and a low score for the percentage of rental housing. However, the categorical characteristics score significantly lower in this graph than in previous analyses. This is in line with the findings that showed the differentiation between different clusters to be lacking regarding these characteristics.

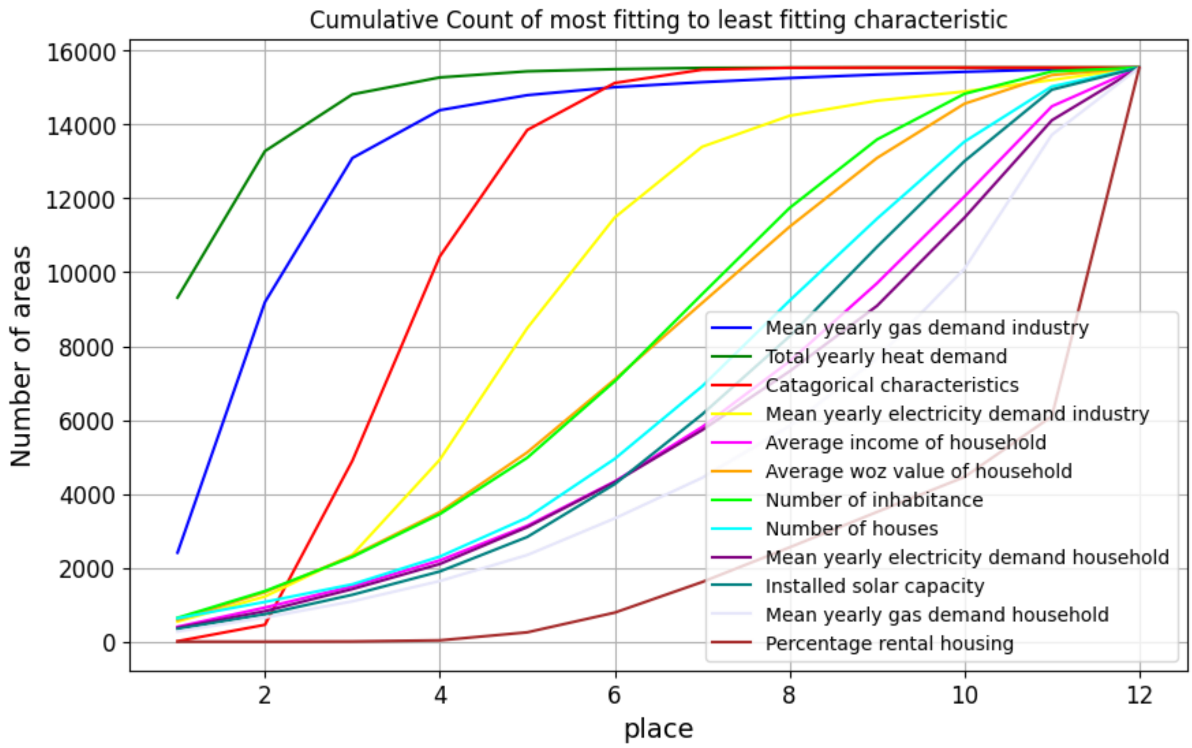


Figure 4.32: A cumulative plot of the match between areas and clusters based on their combined characteristics

Looking at only the primary characteristics again shows that for most areas the relationship with their cluster is the strongest on total yearly energy demand. Similar to the energy characteristics some areas can be in Rotterdam and The Hague that mostly focus on other characteristics like the average income of households and the mean yearly gas demand for industry. Comparing this to the other analyses shows that many areas keep the same primary characteristic even though the clusters themself change. This shows that the clusters are not only determined by the main characteristics but a combination of all characteristics.

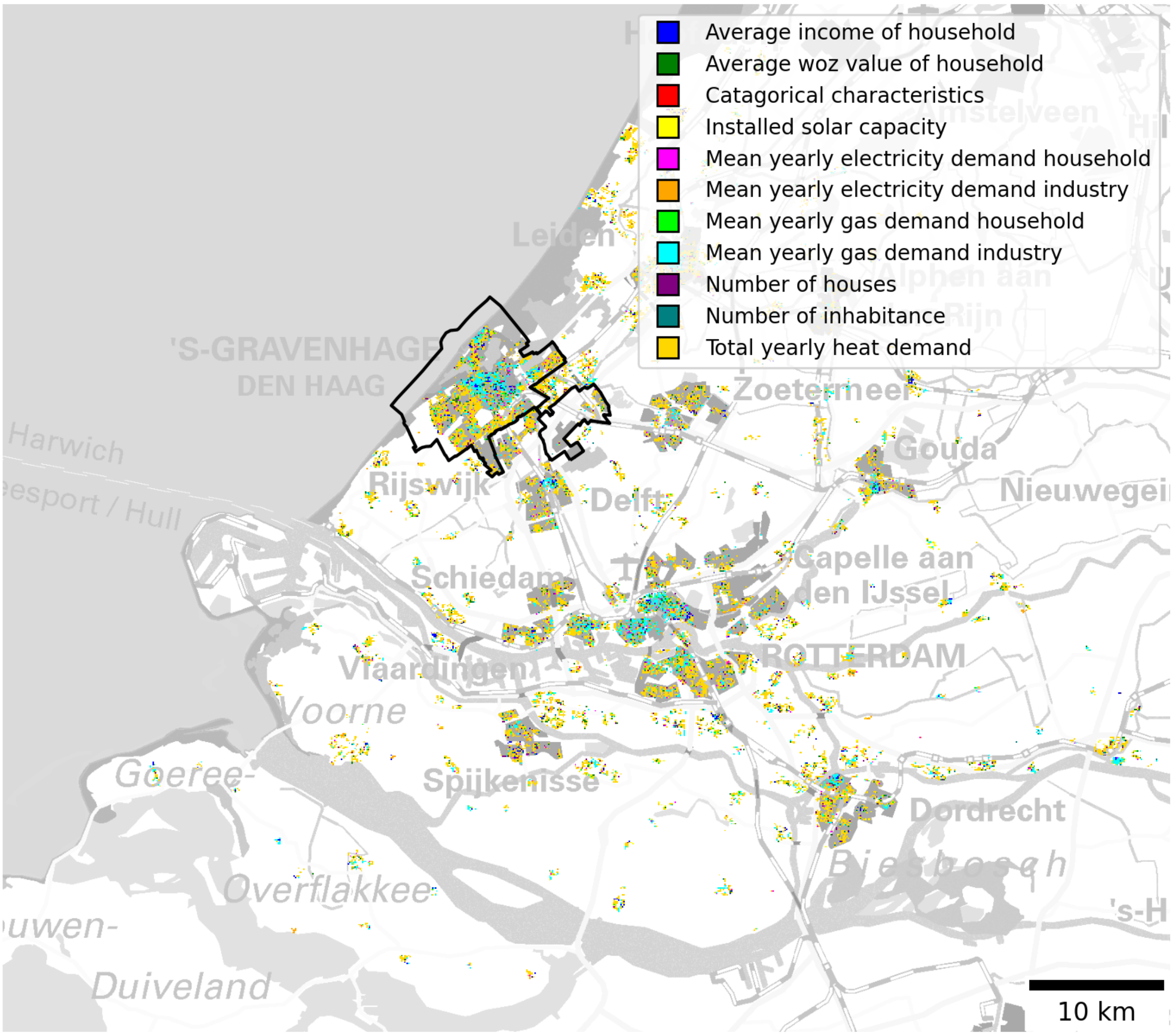


Figure 4.33: Primairy match between areas and combined clusters

5 | Spatial policy insights for clustering

In the previous chapter, the results of the clusters and their insights into the energy and built environment were discussed. This chapter will focus on how the results of this type of cluster analysis can be used in spatial planning. To answer this question, a combination of insights gathered during the modeling process and expert opinions collected during the meetings were used.

Just as in the model iteration section in chapter 3.3 the results were a combination of two meetings. One group discussion with experts and one with the client from the Province of South Holland. During the meetings that were held, the experts were asked two questions. First, they were asked if this type of analysis provided any insights that could help in spatial planning. After that, they were asked how they would use this type of analysis in spatial planning. These expert findings were then combined with the findings from the results and discussion.

Answering the first question both the experts and the client agreed that this type of analysis could be usefull within spatial planning. It was however noted that this would require more input data. For example, an expert in heat infrastructure present at the discussion noted that data for district heating would be required to make it applicable in his field. Similarly, the lack of industrial and rural energy demand was also mentioned as one of the limitations of the current data set. They did however agree that these were mostly data limits and that the methodology itself could be used to provide insights. This already validated the usefulness of this type of analysis. However, it does not yet answer how this method can be used. For this, the second question was used.

Two different uses where identified to answer how to use this type of analysis in spatial planning. The first type of analysis that was mentioned was to use this type of analysis to assess the suitability of certain areas of a specific spatial policy. In this idea, spatial planners would first come up with a policy like switching to district heating. A cluster analysis would then be performed and areas with suitable clusters would then be considered for this policy. Looking at the results of the performed cluster analyses shows that there are some aspects that make the current methodology less practical for this type of implementation. The first limiting factor is the relatively weak clusters that are currently formed by the clustering algorithm. This means that many areas within a cluster will not follow all characteristics from their clusters. As the goal of a suitability analysis would be to find areas with a certain set of characteristics these variations between areas might lead to the wrong conclusion for some areas. Besides this another limiting factor is that cluster analyses do not allow the identification of specific combinations of characteristics. Instead, they search for maximally different sets of characteristics. This limits the usefulness of cluster analysis for finding suitable areas as the clusters might not focus on the characteristics within the policy. Combined these findings point to this type of cluster analysis being relatively unsuitable for identifying suitability.

The second type that was mentioned was to use this form of analysis as a knowledge collection tool for spatial planning. In this case, relations and trends within the energy and built environment will be identified in the cluster analysis. These trends can then be used by decision-makers like the province of South Holland as input for their decision-making process. Looking at the results and discussion does indeed point to cluster analysis being suitable for providing these types of insights. However, there are some types of insights that this method is more suited for than others.

One type of insight that this methodology is particularly suited for is finding differences between areas based on their characteristics. Looking at the results of the cluster analysis show that this method of clustering is able to provide insights into differences between areas based on their energy characteristics. As an example of how this could be used in spatial policy was given by the client. He noted that current spatial planning regarding the energy transition has a tendency to focus on the details of individual neighborhoods without looking at the broader context. He also mentioned that cluster analysis could be a solution for this, saying: "This is also a kind of way to look from a higher level at what broadly are the directions you can take. What are in general the interventions that are defined for a certain area? That is something where this can help with.". This shows that analyzing these differences between areas can be used to aid spatial planning decision-making.

Another type of insight that can be collected using this clustering method is the relationship between different characteristics. An example of this was the relationship between the percentage of rental housing and the electricity demand of households which could be seen in the social cluster analysis. These types of insights can be used to get a better understanding of what relationships might be present within the system. This can then be used by spatial planners to create policies that take into account multiple aspects of the system. This could for example be used to create energy policies that are both focused on the energy transition and certain aspects of the social environment like energy poverty.

Combined this shows that clustering can provide valuable insights that can be used in spatial policy. One thing that is important to mention however is that this method of cluster analysis provides a simplified representation of the real world. Although this is the case for all models this is especially the case for clustering as this method tries to simplify a complex set of characteristics into a better understandable set of clusters. This means that clustering is mostly suitable as an exploratory tool instead of a decision-making tool. The insights of these cluster analyses are therefore not suited to directly be implemented into spatial policies. Instead, they can provide a direction that can guide the spatial planning process or serve as a starting point for a more in-depth analysis of a certain area of characteristic of the system.

Though the exact method of using clusters to provide policy insights for spatial planning is case dependent the methodology used within this analysis of the province of South Holland can be used as a basis for this type of analysis. In Figure 6.1 an overview of this process can be seen. This shows a simplified version of the steps taken within this thesis combined with the remaining steps required to use cluster results within spatial planning.

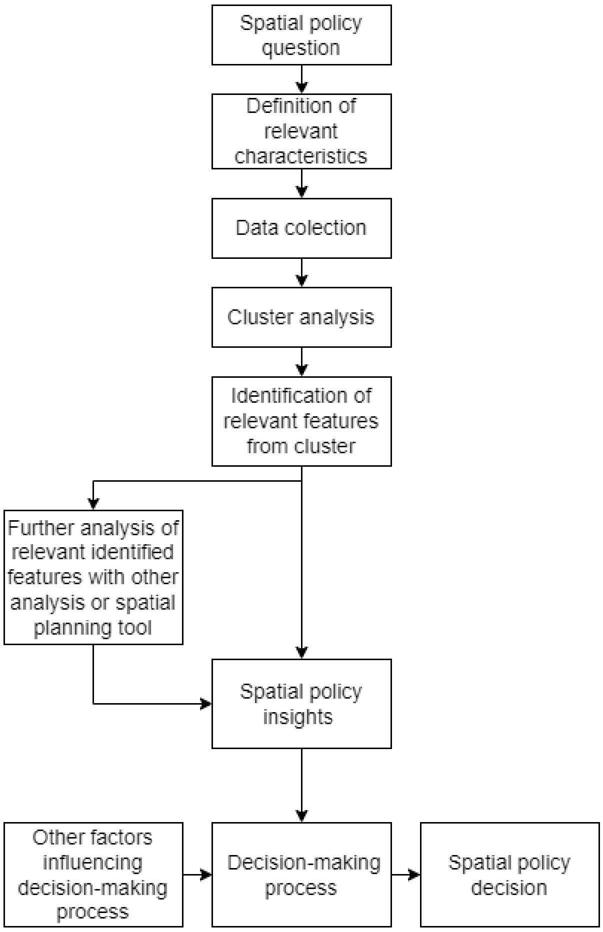


Figure 5.1: Steps for use of cluster analysis within spatial planning

6 | Discussion of methodology and results

The previous four chapters mostly focused on the process and the results which already gave insights into the results for the case of South Holland. This chapter will present a broader analysis of the implications of cluster analysis as a tool for collecting insights into the energy and built environment. This will then be used in the next chapter to analyze the implementation of cluster analysis in spatial planning.

6.1 Result discussion

In chapter 4 some aspects of the results were already discussed. However, this mostly focused on the results themself instead of their implications. One of the main reasons why clusters were used within this thesis was to see whether they are able to provide usefull insight for spatial planning. When comparing the four sets of clusters, differences can be seen depending on the characteristics used within the analysis. Here the energy, built environment, and social cluster sets all showed novel features that were not included within the other datasets. This shows that there is merit in analyzing multiple aspects of the energy system, built environment, and social environment as analyzing only one would not capture all characteristics of the system.

This in itself points to a cluster analysis with a wide set of characteristics combining all aspects being the best approach to get insights for spatial planning. However, multiple findings in the results also point against this approach showing this not being the most suitable approach. The analysis of the combined characteristics showed that although some features of the separate analyses are still visible, many also get lost. Therefore choosing a method that includes all characteristics will remove many of the nuances that might be important for decisionmaking. Besides this, the combined characteristics also provided fewer data points as more areas missed one or more characteristics and therefore were excluded from the dataset. This means that without the collection of extra data points, this type of analysis would also be applicable to fewer areas.

Another reason why including more characteristics in one analysis might not deliver the appropriate results has to do with the weak relationship that was seen between the clusters and their areas. As noted in the experimental setup of chapter 3 this weak relationship means that there is a lot of variation in values between areas in the same clusters. Additionally, there can be overlap between areas of different clusters. This makes it harder to definitively draw conclusions from the results which might hinder the usefulness for spatial planning. This seems to be the most prominent in the cluster analyses with many, possibly uncorrelated, characteristics such as the built environment analysis and the analysis of the combined characteristics. Therefore including more characteristics within one analysis might not provide better insights into the energy and built environment.

The results also show that some of the characteristics are more important than others in the placement of areas within clusters. This points to the possibility of including only characteristics with a higher impact on the overall cluster results. This would reduce the total amount of characteristics while hopefully still including all important aspects of the total set of characteristics. This is possible for some characteristics, for example, energy and gas demand almost always follow the same trend meaning that they can probably be combined. Similarly, housing density and the number of inhabitants will generally lead to similar results as more houses per square acre will also mean more inhabitants. However, there is a limit to this method. This can be seen by the fact that in all analyses the same characteristics had the highest match between clusters and areas but the resulting clusters were significantly different.

Combining these findings shows that within the current clustering model neither including too many nor too few characteristics leads to an outcome that can effectively be used to get a better understanding of the energy system. This mostly seems to be in the trade-off between the accuracy and completeness of the analysis with more characteristics leading to a broader, more complete analysis and fewer characteristics leading to a more narrower but more in-depth analysis. The balance between these trade-offs fits the best is case dependent. For example, exploratory analysis such as this one will often lean more toward a broad analysis while a more focused analysis can better make use of a narrow set of characteristics. In some cases, it might be more useful to run several more specified analyses than one combined. This does not have to mean splitting between energy, built environment, and social cluster as the results show that combining these factors can lead to novel insights. Instead, this can mean focusing on a specific relationship such as household energy demand and energy poverty. Together this shows that for decision-makers like the province of South Holland, it is important to define for what goal they want to use a cluster analysis as this influences which trade-offs can be made and what characteristics fit best with this goal.

Because of these tradeoffs, it is important to note that the list of characteristics used in this paper might not always be suitable for each type of analysis. This is not only the case for analyses of areas outside of South Holland but also for other types of analyses within the province. This is because the relevance of each characteristic is dependent on the analysis itself. Examples of this can be seen in the interviews where multiple characteristics were mentioned that were relevant in the setting of a neighborhood, that could not be implemented in an analysis of the entire province. This is not only limited to geographical scale but also the end goal of the analysis. An analysis of the relation between the social environment and the energy system might look for different results than an analysis of the relation between the built environment and the energy system. Therefore they will also need different characteristics to come to these results. This shows that the characteristics used in this paper can be used as a preliminary set. However, to get the desired results decision-makers like the province of South Holland will have to adjust these characteristics for each individual case to fit the desired goals of their analysis.

Another aspect that should be taken into account with the clustering is that some characteristics are hard to cluster together. An example of this are the industrial characteristics. As the energy demand of industry is significantly higher than household energy demand, it is hard to include them both within the same analysis without losing the nuances within each. This is because the clusters will automatically want to split areas with high and low industrial demand in different clusters independent of any possible household energy demand. Besides this, another problem is that most clustering models require all areas to have all characteristics. This however is hard when including both industrial and residential characteristics. For example, characteristics such as energy labels and percentage of rental housing are only measured for residential housing. This would mean that these types of characteristics can not be included when clustering all industrial areas and residential areas combined. Depending on the characteristics included in the cluster analysis it would therefore probably be more insightful to analyze industry and residential areas separately.

6.2 Limitations

The previous section already mentioned some trade-offs of the clustering methodology. However, there are also some other limitations of the research performed in this thesis. One aspect that influenced the final set of characteristics was the availability of data. Because only open-source datasets were used some characteristics could not be included within the dataset. The results show that the change of one or more characteristics can have an influence on the final clusters. The lack of these datasets therefore may have had an impact on the final clusters shown within the thesis. Although a complete representation of the real world will be impossible, as some aspects will always be simplified, this does mean that the current model has some blind spots with regard to some aspects of the energy and built environment. To reduce this blindspot, more research would be necessary to either acquire these datasets from other parties or to collect this dataset.

As noted in Chapter 3 many of the datasets also lacked data for large areas of the province. Here in particular a lack of data could be found for rural and industrial areas. This meant that although the goal was to define the energy and built environment characteristics of the entire province, the final model mostly shows these environments for the urban areas. Although this probably only has a limited impact on household characteristics, as most households are within these urban areas, it does give a twisted view. This is because some characteristics might only have small differences between urban areas but larger differences between urban and rural areas that are not included in this analysis. The same is also the case for industrial areas, with only a few being included in this analysis. Similar to the previous section it could therefore be interesting to find or collect more data for these regions.

Looking at the clustering method used during the thesis some more observations can be made. One of the main reasons clustering was used is that this method can define similarities between areas mathematically. One advantage is that it limits personal biases that might be present when drawing conclusions by only looking at the data. Although all clustering methods will still have some biases, as the characteristics that are used will influence the outcome of the analysis, the K-prototype has one area where this is especially apparent. This area is the definition of the gamma value that decides the relative importance of the numeric and spatial characteristics. As the methodology defines no clear way to find this value except for manually assessing the balance between these two types of characteristics this leaves a lot of wiggle room. Though this was limited by the testing, this does reduce the overall accuracy of the results of the analysis. Another limitation of the current model is that the results are relatively weak with much overlap between the clusters. This limited the conclusions that could be drawn from the data. Although the previous sections already named some causes and possible solutions the model itself might also be a factor. As this thesis only looked at the K-prototype no comparison could be drawn to other methods. It therefore might be that this model is not optimal for this dataset with lots of outliers and categorical characteristics. Although this does not mean that the K-prototype method is unsuitable for this type of analysis, it might be interesting to look at another algorithm that might not share this disadvantage.

The last thing to note with regard to the clustering method as implemented within this paper is that all characteristics were normalized between a value of zero and one. This means that all characteristics had the same weight. This was done because this would provide an even overview of all characteristics as no previous analysis of this type was performed. However, as could be seen during the testing this did mean that some clusters could not be included because of their dominance on the clustering. It might also not always be the case that all characteristics are of equal importance. For this reason, it can be interesting to include weighted characteristics within this type of analysis.

7 | Conclusion

The goal of this thesis was to find out how the relationships between energy-, and built environment can be mapped in different areas based on a broad set of characteristics. This chapter will conclude whether this goal is reached within this thesis by looking at the main research question and the sub-questions.

The first research question was: *"What characteristics of the energy system and build environment are relevant for defining clusters of the province of South Holland?"*. Based on the findings it can be concluded that different characteristics are relevant depending on the research goal or spatial policy. For the case of the province of South Holland, three different types of characteristics were identified to be important. These were energy, built environment, and social characteristics. From this, it can be concluded that a mix of these types of characteristics is necessary to map the energy and built environment of the province. Within each of these characteristic groups, certain aspects were also defined as relevant. In the case of the energy characteristics these are energy demand, energy supply, network capacity, and energy storage. For each of these, both heating and electricity characteristics should also be included. For the built environment relevant characteristics include land use, characteristics of the buildings, building density, and the supply of renewable energy. Lastly for the social characteristics only characteristics regarding energy poverty and the attitude towards renewable energy were defined as relevant for this case.

The second research question that had to be answered was *"What clustering methodology can be used to identify the relationships between energy and spatial characteristics?"* Based on the findings from chapters 3 and 4 it can be concluded that the methodology used within this paper can be used to identify relationships between energy and built environment. Based on this some steps are required for spatial clustering to be used within this context. A conversion has to take place from the list of characteristics to a dataset. This dataset then has to be combined with a clustering algorithm that can deal with a combination of categorical data and numeric data. Looking at the results it can be concluded that k-prototype can be used for this type of analysis. The results of this analysis can then be used to identify the relationship between the energy system and the built environment in two ways. Firstly it can show differences between areas based on these clusters. Secondly, it can show relationships between characteristics based on their similarities or dissimilarities in different clusters.

Lastly the third research question was *"What kind of policy insights can be identified based on the proposed method of identifying and analyzing energy clusters?"* Here it can be concluded that the type of cluster analysis used with this paper is most suited for exploratory research. This means that the insights that can be gathered are mostly focused on the differences in characteristics between areas and the relations between these characteristics. These insights can then either be used for a more detailed analysis of certain aspects of the energy system or in the definition of spatial policies.

By combining these findings the main question could be answered. The main question of this thesis was: *"How can different sets of characteristics be used in spatial clustering to identify the relationship between the energy system and the built environment for use in spatial planning?"* Based on the subquestions it can be concluded that this paper provides insights into this topic by providing a cluster methodology and by providing an example in the form of the analysis of the province of South Holland.

Appendix 1: Interview protocol

Interview protocol thesis Energy and landscape

Interviewee:
Date:
Time:
Introduction [5 minutes]

1. Introducing the interviewee to the topic.
   (a) Reason for research -> *collaboration with province of South holland*
   (b) Knowlage gab -> *Relationship between the energy system and spatial landscape*
   (c) goal -> *Get insights into the relationship between energy and spatial*
2. Discuss the goal of the interview -> *Get insight into relevant characteristics for defining energy clusters and spatial clusters*
3. Short introduction from the interviewee *(Not for use in the report)*
   (a) What is the role of the department that you are currently working for?
   (b) How long have you been working in your current position?

Reminder for the interview process
Before we begin, a reminder of the university ethics guidelines. All results will be anonymized in such a way that they cannot be directly linked to you. You are allowed to withdraw information from the interview results at any point during or after the interview. If you have any questions about this or any other part of the consent form, please let me know.

General questions [10 minutes]

4. What is the process of developing and updating energy infrastructure plans and/or spatial plans?
5. What types of information do you currently use for decision-making in energy infrastructure planning and/or spatial planning?
   (a) For spatial planners: What types of information regarding the energy system do you take into account during the decision-making process?
   (b) For the energy planners: What type of information regarding spatial planning

do you take into account during the decision-making process?

Current situation [30 min]
Show a list of types of characteristics that are identified for the research.

- Energy characteristics -> *supply and demand*
- Built environment characteristics -> *Land use and housing type*
- Social characteristics -> *demographics, and attitude towards solar energy*

6. Which of these types of characteristics do you currently use within energy infrastructure planning and/or spatial planning?
   (a) How are these characteristics currently being used?
   (b) Why are these types of characteristics currently not being used?
7. Are there any types of characteristics that you think are important that are currently not present within this list?
   (a) How is this type of characteristic currently being used?
   (b) Why is this type of characteristic important for decision-making?

Energy

8. What are the main energy characteristics you currently use in your decision-making process? *Examples: Energy demand, Energy supply*
   (a) How are these characteristics currently being used?
   (b) Why are these characteristics important for the decision-making process?

Show a list of energy characteristics identified in the research.

9. Are there any energy characteristics missing that are important for decision-making in spatial planning and energy infrastructure planning?
   (a) Why are these characteristics important for the decision-making process?

Build environment

10. What are the main geospatial characteristics you currently use in your decision-making process? *Examples: Land use, and types of housing*

(a) How are these characteristics currently being used?

(b) Why are these characteristics important for the decision-making process?

Show a list of built environment characteristics identified in the research.

11. Are there any geospatial characteristics missing that are important for decision-making in spatial planning and energy infrastructure planning?

    (a) Why are these characteristics important for the decision-making process?

Social

12. What are the main social characteristics you currently use in your decision-making process? *Examples: Demographics, Attitude*

    (a) How are these characteristics currently being used?

    (b) Why are these characteristics important for the decision-making process?

Show a list of social characteristics identified in the research.

13. Are there any social characteristics missing that are important for decision-making in spatial planning and energy infrastructure planning?

    (a) Why are these characteristics important for the decision-making process?

14. Are there any characteristics on this list that you think are irrelevant to decision-making in spatial planning and energy infrastructure planning?

Future situation [20 minutes]
Short introduction into why there is a need for future scenarios during the modeling process.

15. What trends do you expect to have a great effect on the future energy system? *Examples: transition to hydrogen, growth in the use of solar energy*

    (a) What effect do you think this will have on the system?

    (b) What energy, built environment, and social characteristics do you think influence this trend?

16. What are the energy characteristics that you expect will become important in the future?

(a) Are there any energy characteristics that are currently not being used that you expect will become important in the future?

17. What are the build environment characteristics that you expect will become important in the future?

    (a) Are there any build environment characteristics that are currently not being used that you expect will become important in the future?

18. What are the social characteristics that you expect will become important in the future?

    (a) Are there any social characteristics that are currently not being used that you expect will become important in the future?

Wrap up [5 minutes]

19. Is there anything you would like to add based on these questions?

20. Are there any answers you would like to withdraw from the interview?

21. Are you willing to participate in the follow-up group meeting where we discuss the results of the model and see what role this method can play in future design spatial policy?

Thank you for your participation. If you have any questions regarding the interview process at any time in the future, please feel free to contact me.

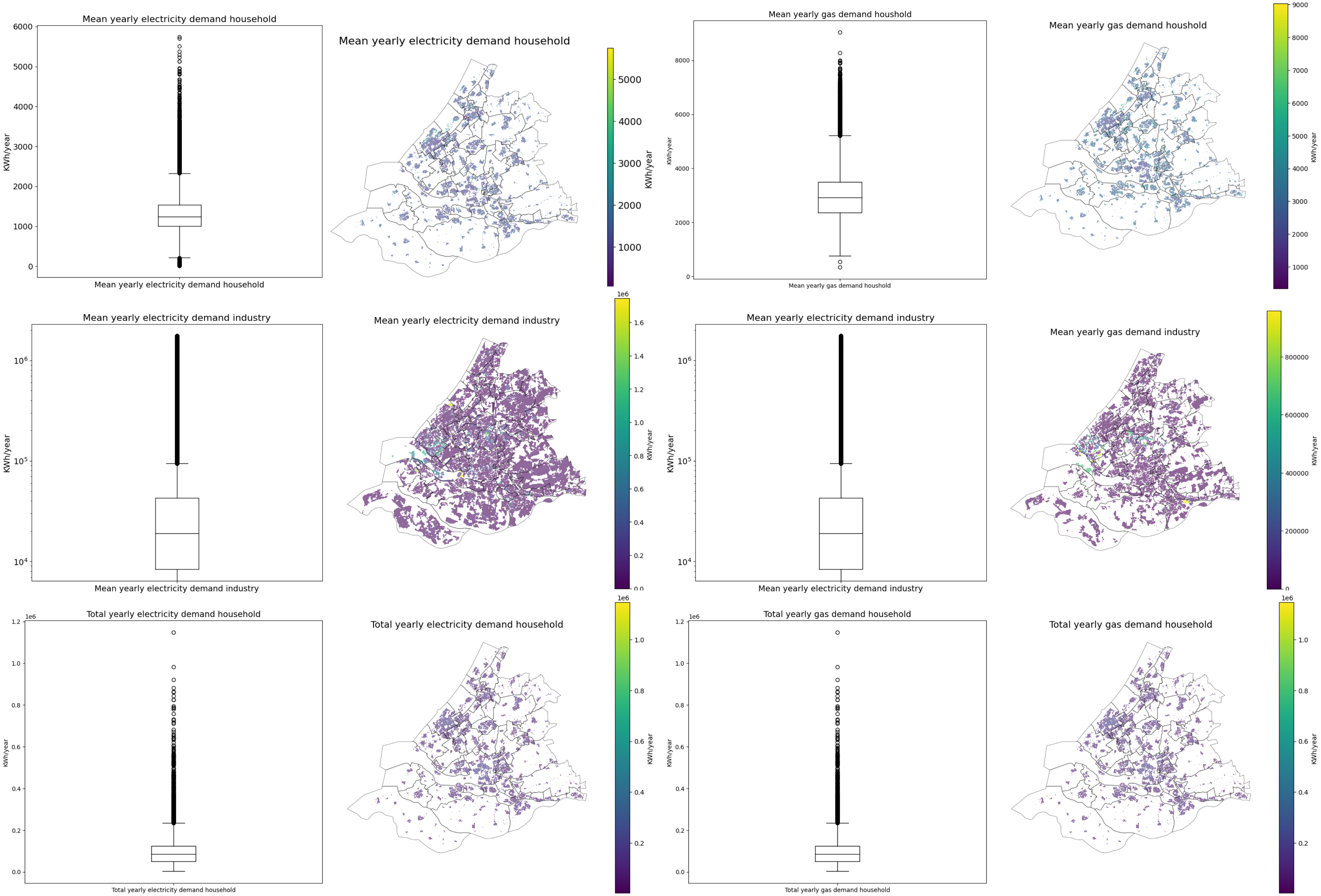
Appendix 2: Literature review characteristics

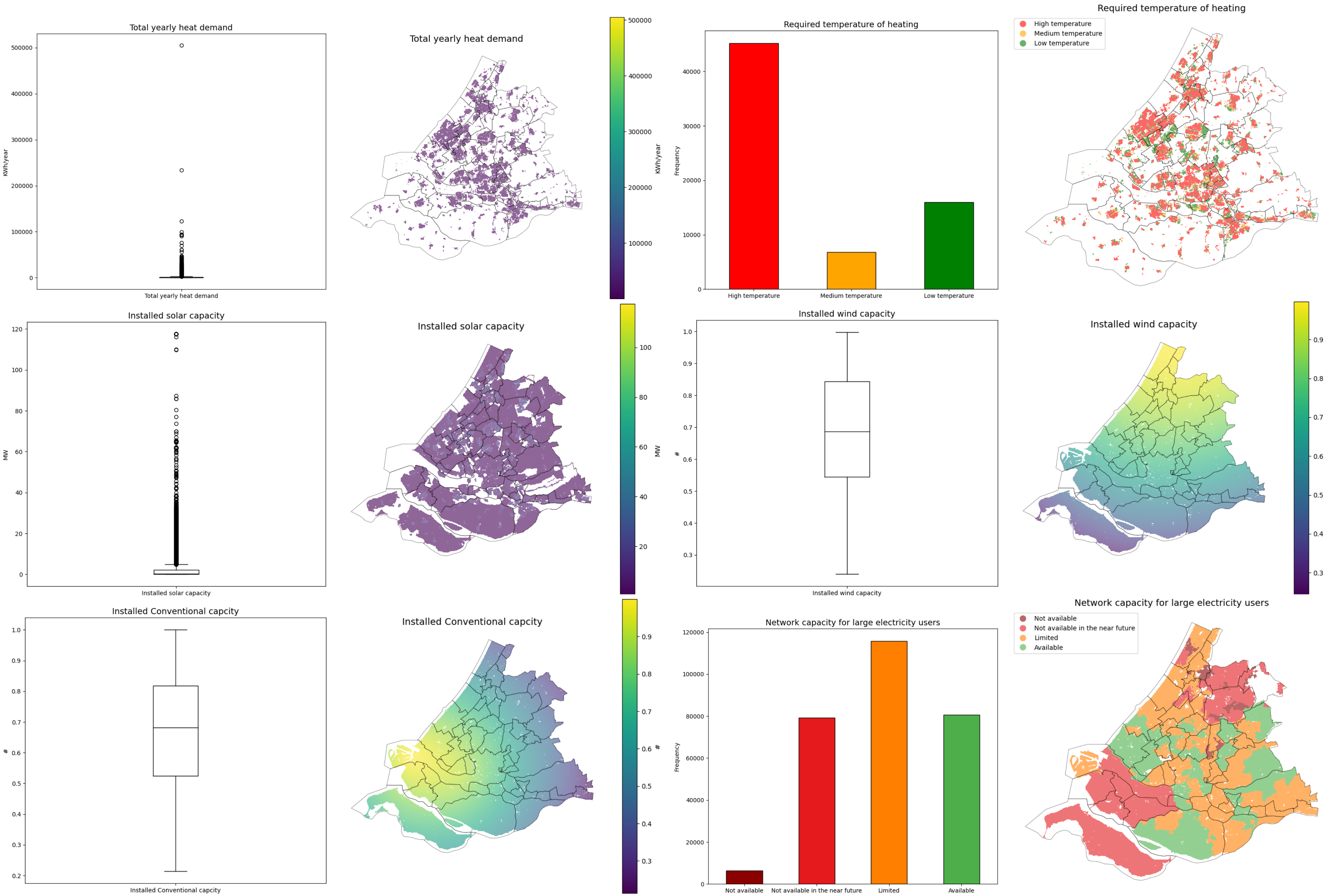
	Energy														Built environment											Social		
	Photovoltaic supply	Wind supply	Geothermal supply	Biomass supply	Supply conventional	Gas supply	Storage	Electricity demand	Peak load Electricity	Heat demand	Peak load heat	Electricity network capacity	Gas network capacity	Heat network capacity	Photovoltaic potential	Wind potential	Geothermal potential	Biomass potential	Land use	Groundtype	Building density	Building age	Building type	Population dencity	Population size	Income	Acceptance of renewable	Employment rate
Liu e.a.Liu et al., 2023	x														x						x		x					
Sahoo e.a.Sahoo et al., 2022	x	x	x	x				x				x		x	x	x	x	x	x				x					
Boa e.a. Bao et al., 2022	x							x							x				x									
De vries and schreyde Vries and Schrey, 2022					x			x	x			x			x	x			x					x	x			
Sahoo Sahoo et al., 2023 e.a.					x	x		x		x					x	x	x	x	x				x					
Zhou Zhou et al., 2021 e.a.					x			x				x	x															
Buckley e.a. Buckley et al., 2021	x				x					x			x									x			x			
Then e.a. Then et al., 2021					x	x		x		x			x										x					
Bao e.a. Bao et al., 2020				x														x	x	x								
Chen e.a. Chen et al., 2020								x		x	x								x		x		x					
Wang e.a.N. Wang, Heijnen, et al., 2020	x	x		x			x	x																				
Wang e.a. N. Wang, Verzijlbergh, et al., 2020	x	x		x	x		x	x		x					x	x		x	x									
Drechsler e.a.Drechsler et al., 2011		x													x					x							x	
Bosch e.a.Bosch et al., 2020															x	x			x									
Eicker e.a. Eicker et al., 2020								x		x											x	x	x					
Fishera e.a. Fichera et al., 2018								x				x							x									
Szarka e.a. Szarka et al., 2018				x				x		x								x	x						x			
Afshari and Friedrich Afshari and Friedrich, 2017								x	x						x	x												
Mutani e.a. Mutani et al., 2016										x									x				x					
Jäger e.a. Jäger et al., 2016	x	x		x	x			x																x	x	x	x	
Nolde e.a. Nolde et al., 2016							x		x							x				x								
Bustos-Turu e.a.Bustos-Turu et al., 2016									x	x				X														
Yamagata and Seya Yamagata and Seya, 2013	x							x											x				x	x	x	x		x
Sum	8	4	1	6	7	2	3	15	4	9	1	4	3	2	8	7	2	5	12	2	3	2	8	3	5	2	2	1

Appendix 3: Dataset overview

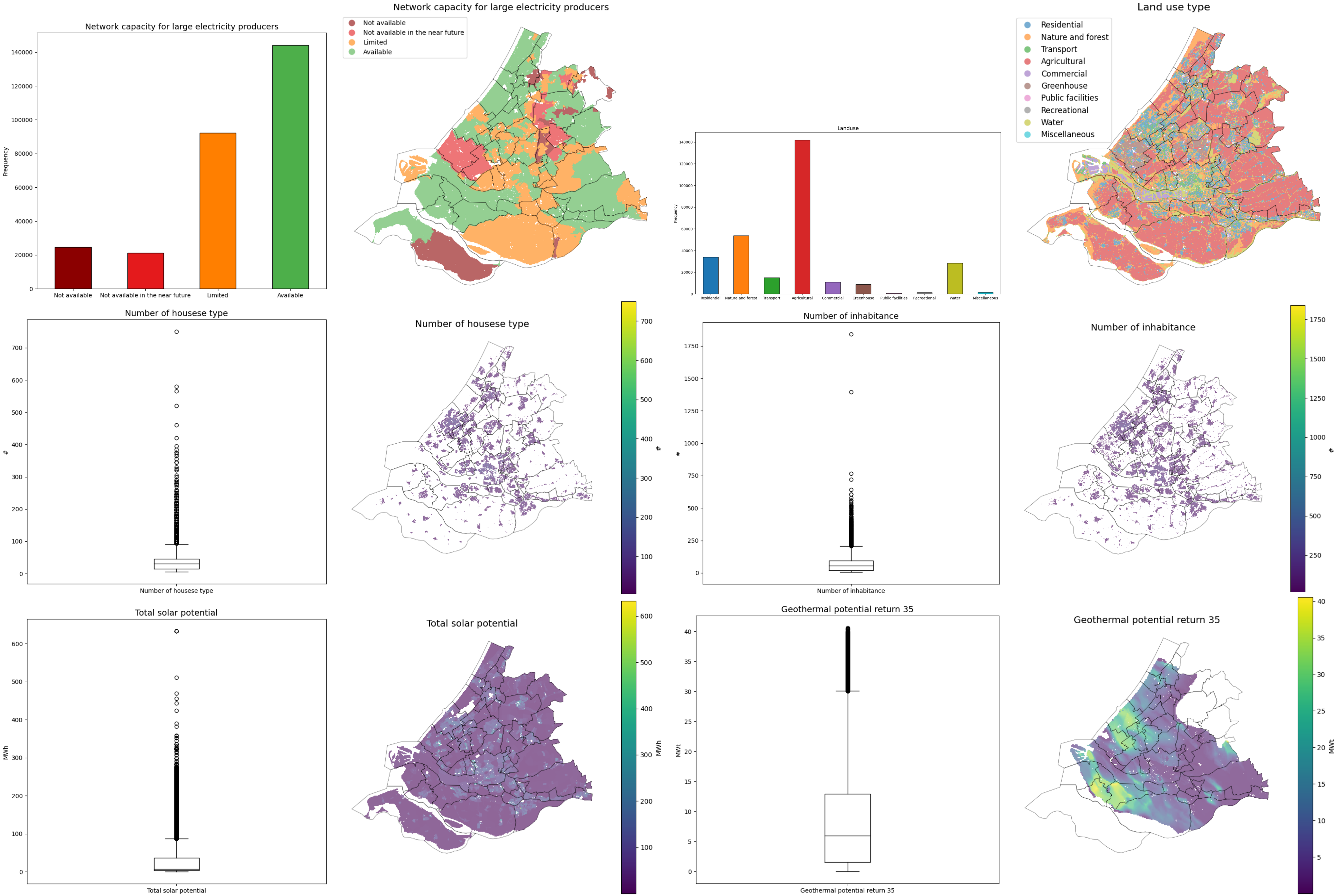
Characteristic	Unit of Measurement	Source	Datatype	Resolution	Transformation
Mean Yearly Electricity Demand (Household)	KWh/year	Centraal Bureau voor de statistiek (2023)	Shapefile	100m	No transformation
Mean Yearly Gas Demand (Household)	KWh/year	Centraal Bureau voor de statistiek (2023)	Shapefile	100m	No transformation
Mean Yearly Electricity Demand (Industry)	KWh/year	Centraal Bureau voor de Statistiek (2019)	Excel	Postcode 6	Shape to raster
Mean Yearly Gas Demand (Industry)	KWh/year	Centraal Bureau voor de Statistiek (2019)	Excel	Postcode 6	Shape to raster
Total Yearly Electricity Demand (Household)	KWh/year	Centraal Bureau voor de statistiek (2023)	Shapefile	100m	No transformation
Total Yearly Gas Demand (Household)	KWh/year	Centraal Bureau voor de statistiek (2023)	Shapefile	100m	No transformation
Total Yearly Heat Demand	KWh/year	Provincie Zuid Holland (2022c)	Shapefile	Building	Shape to raster
Required Temperature for Heating	High/medium/low	Provincie Zuid Holland (2022c)	Shapefile	Building	Shape to raster
Installed Solar Supply	KW	Centraal Bureau voor de Statistiek (2020b) and Centraal Bureau voor de Statistiek (2022)	Shapefile	Neighborhood	Shape to raster
Installed Wind Supply	N/A	Rijksinstituut voor Volksgezondheid en Milieu (2023)	Shapefile	1m	Point to raster
Installed Conventional Supply	N/A	CE Delft et al. (2021)	Excel	1m	Point to raster
Network Capacity for Large Electricity Users	list (Available, limited, not available in the near future, not available)	Netbeheer nederland (2023)	CSV	Postcode 6	Shape to raster
Network Capacity for Large Electricity Producers	list (Available, limited, not available in the near future, not available)	Netbeheer nederland (2023)	CSV	Postcode 6	Shape to raster
Landuse Type	list (Residential, industry, commercial, greenhouse, recreational, public facilities, agricultural, transport, nature and forest, water, miscellaneous)	Provincie Zuid Holland (2022a)	geotiff	0.25m	Raster to raster
Number of Houses	#	Centraal Bureau voor de statistiek (2023)	Shapefile	100m	No transformation
Number of Inhabitants	#	Centraal Bureau voor de statistiek (2023)	Shapefile	100m	No transformation
Mode Build Year of Building List	(<1945, 1945-1965, 1965-1975, 1975-1985, 1985-1995, 1995-2005, 2005-2015, >2015)	Centraal Bureau voor de statistiek (2023)	Shapefile	100m	No transformation
Energy Label of Building	list (A+++,A++,A+,A,B,C,D,E,F)	Centraal Bureau voor de statistiek (2023)	Shapefile	100m	Shape to raster
Total Solar Potential	MWh/year	Zonnewijzer buurt Provincie Zuid Holland, 2017	Shape file	Neighborhood	Shape to raster
Geothermal Potential (Return at 35°C)	MWt	Totaal (COP 15, Tretour 35) Provincie Zuid Holland (2016b)	Geotiff	100m	Raster to raster
Geothermal Potential (Return at 25°C)	MWt	Totaal (COP 10, Tretour 25) Provincie Zuid Holland (2016a)	Geotiff	100m	Raster to raster
Average Income of Household	€	Centraal Bureau voor de Statistiek (2020a) and Centraal Bureau voor de statistiek (2023)	Shapefile	100m	Shape to raster
Average WOZ Value of Household	€	Centraal Bureau voor de statistiek (2023)	Shapefile	100m	No transformation
Percentage of Rental Housing	%	Centraal Bureau voor de statistiek (2023)	Shapefile	100m	No transformation

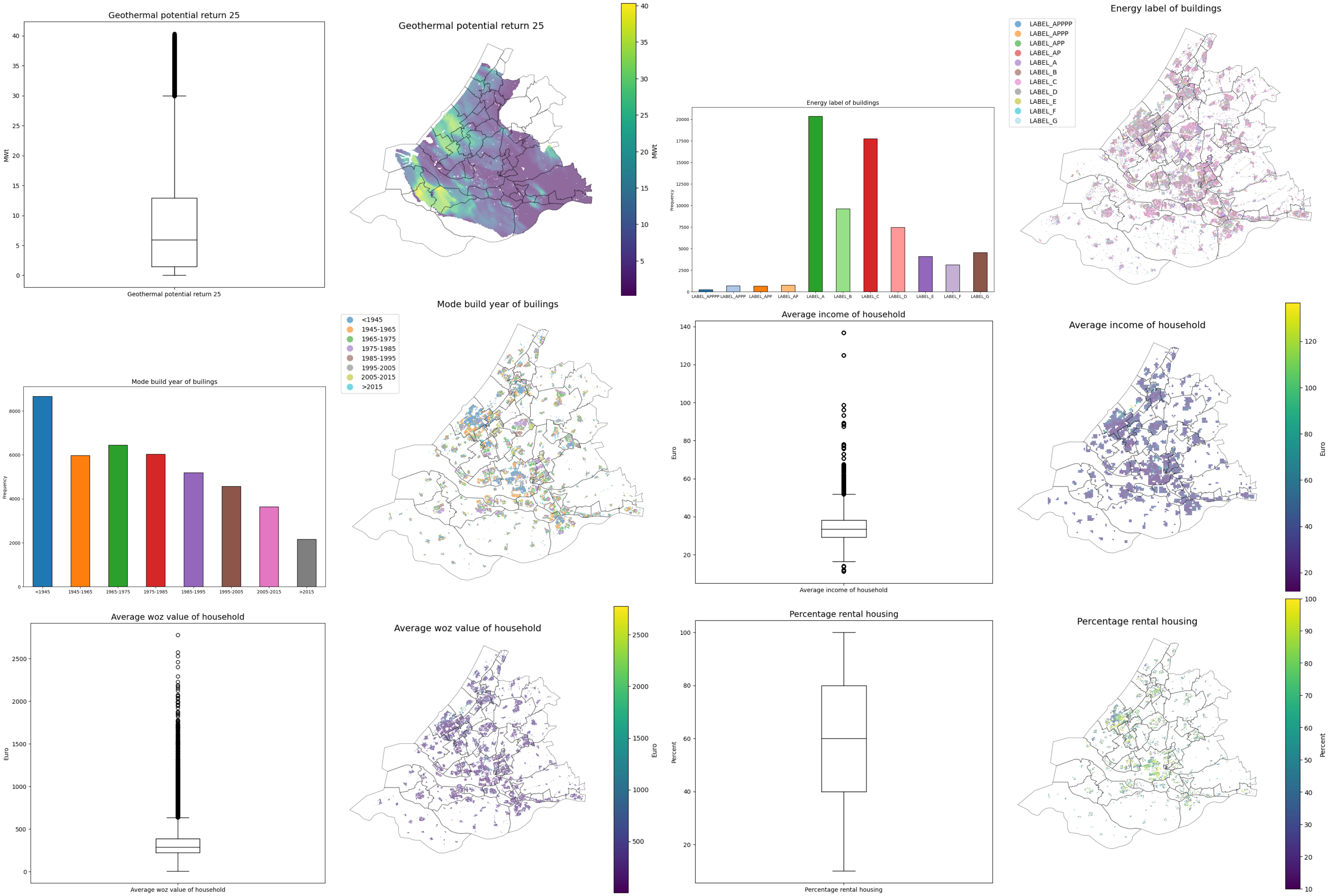
Appendix 4: Plot of dataset





Total yearly heat demand
KWh/year
Total yearly heat demand
Total yearly heat demand
KWh/year
Required temperature of heating
Frequency
High temperature
Medium temperature
Low temperature
Required temperature of heating
High temperature
Medium temperature
Low temperature
Installed solar capacity
MW
Installed solar capacity
Installed solar capacity
MW
Installed wind capacity
#
Installed wind capacity
Installed wind capacity
#
Installed Conventional capcity
#
Installed Conventional capcity
Installed Conventional capcity
#
Network capacity for large electricity users
Frequency
Not available
Not available in the near future
Limited
Available
Network capacity for large electricity users
Not available
Not available in the near future
Limited
Available


Network capacity for large electricity producers
Frequency
Not available
Not available in the near future
Limited
Available
Network capacity for large electricity producers
Not available
Not available in the near future
Limited
Available
Landuse
Frequency
Residential
Nature and forest
Transport
Agricultural
Commercial
Greenhouse
Public facilities
Recreational
Water
Miscellaneous
Land use type
Residential
Nature and forest
Transport
Agricultural
Commercial
Greenhouse
Public facilities
Recreational
Water
Miscellaneous
Number of housese type
#
Number of housese type
Number of housese type
#
Number of inhabitance
#
Number of inhabitance
Number of inhabitance
#
Total solar potential
MWh
Total solar potential
Total solar potential
MWh
Geothermal potential return 35
MWt
Geothermal potential return 35
Geothermal potential return 35
MWt


Geothermal potential return 25
MWt
Geothermal potential return 25
Geothermal potential return 25
MWt
Energy label of buildings
Frequency
LABEL_APPPP
LABEL_APPP
LABEL_APP
LABEL_AP
LABEL_A
LABEL_B
LABEL_C
LABEL_D
LABEL_E
LABEL_F
LABEL_G
Energy label of buildings
Mode build year of builings
Frequency
<1945
1945-1965
1965-1975
1975-1985
1985-1995
1995-2005
2005-2015
>2015
Mode build year of builings
Average income of household
Euro
Average income of household
Average income of household
Euro
Average woz value of household
Euro
Average woz value of household
Average woz value of household
Euro
Percentage rental housing
Percent
Percentage rental housing
Percentage rental housing
Percent

Appendix 5: Experimental setup

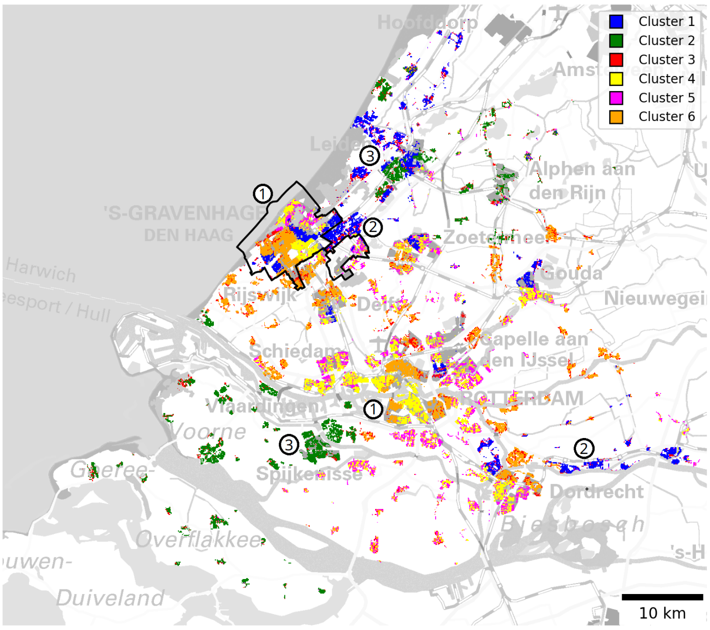
The following parameters were the same for all datasets:

Cluster parameters	
Minimum number of clusters	2
Maximum number of clusters	15
Number of initializations per run	25
Initialization method	Huang
Distance calculation (numeric)	Euclidean
Distance calculation (categoric)	Matching
Outlier filter	3x standard deviation
Seed	30
Maximum number of iterations	1000000

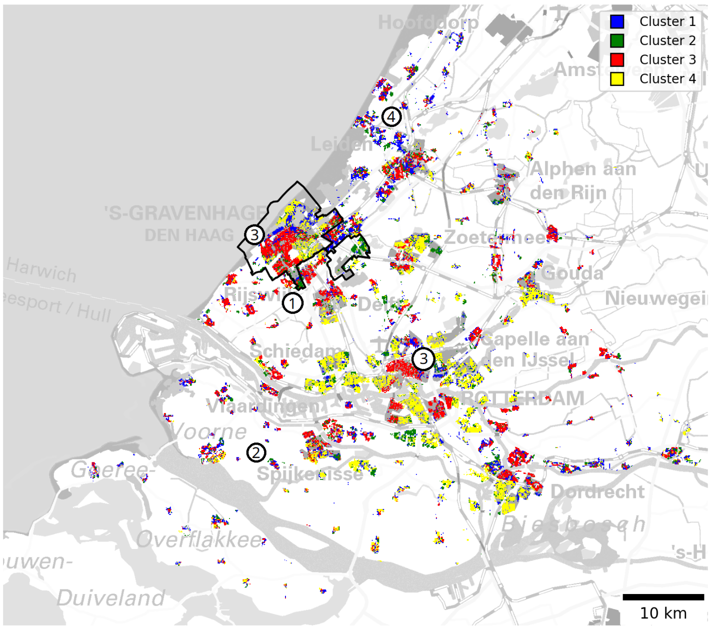
An overview of each of the characteristics and in which of the cluster analyses they were used.

Characteristic	**Energy**	**Energy and built environment**	**Energy and social**	**Total**
Mean Yearly Electricity Demand (Household)	x	x	x	x
Mean Yearly Gas Demand (Household)	x	x	x	x
Mean Yearly Electricity Demand (Industry)	x	x	x	x
Mean Yearly Gas Demand (Industry)	x	x	x	x
Total Yearly Heat Demand	x	x	x	x
Required Temperature for Heating	x	x	x	x
Installed Solar Supply	x	x	x	x
Installed Wind Supply				
Installed Conventional Supply				
Network Capacity for Electricity Users	x	x	x	x
Network Capacity for Electricity Producers	x	x	x	x
Landuse Type		x		x
Number of Houses		x		x
Number of Inhabitants		x		x
Mode Build Year of Building		x		x
Energy Label of Building		x		x
Average Income of Household			x	x
Average WOZ Value of Household			x	x
Percentage Rental Housing			x	x

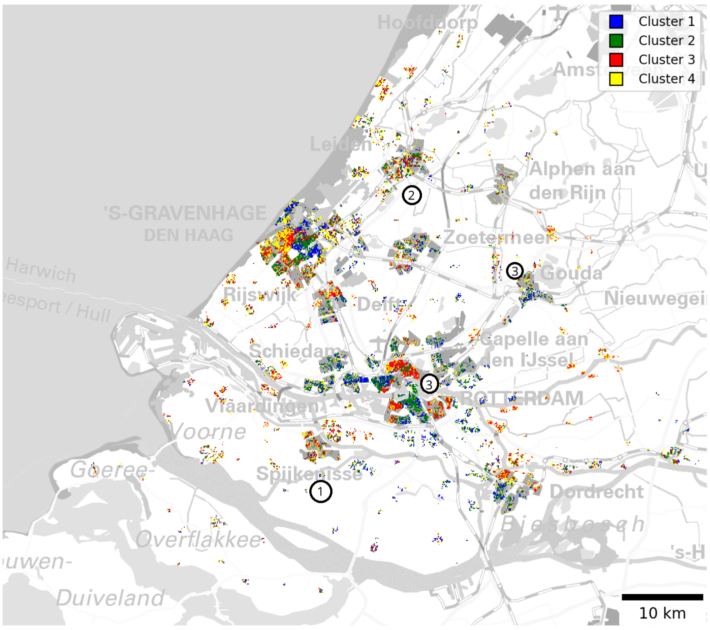
Appendix 6: Side-by-side overview of cluster sets



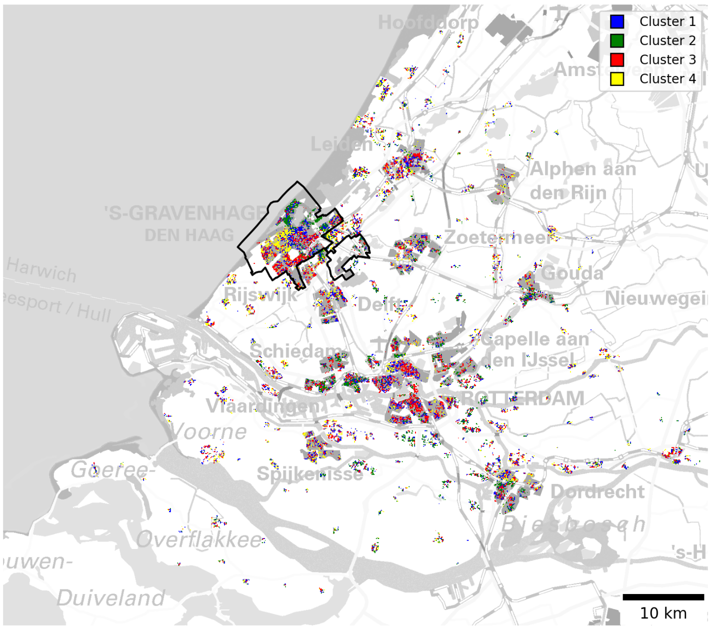
(a) Energy clusters



(b) Energy and built environment clusters



(c) Energy and social clusters



(d) Combined clusters

Bibliography

Afshari, A., & Friedrich, L. A. (2017). Inverse modeling of the urban energy system using hourly electricity demand and weather measurements, part 1: Black-box model. *Energy and Buildings*, *157*, 126–138. https://doi.org/10.1016/j.enbuild.2017.01.053

Asarpota, K., & Nadin, V. (2020). Energy strategies, the urban dimension, and spatial planning. *Energies*, *13*(14). https://doi.org/10.3390/en13143642

Bao, K., Kalisch, L., Santhanavanich, T., Thrän, D., & Schröter, B. (2022). A bottom-up GIS-based method for simulation of ground-mounted PV potentials at regional scale. *Energy Reports*, *8*, 5053–5066. https://doi.org/10.1016/j.egyr.2022.03.187

Bao, K., Padsala, R., Coors, V., Thrän, D., & Schröter, B. (2020). A method for assessing regional bioenergy potentials based on gis data and a dynamic yield simulation model. *Energies*, *13*(24). https://doi.org/10.3390/en13246488

Bosch, S., Schmidt, M., & Kienmoser, D. (2020). On the path to sustainable energy landscapes? the social shaping of energy landscapes in the face of climate protection measures. *Erdkunde*, *74*(4), 263–280. https://doi.org/10.3112/erdkunde.2020.04.03

Brooks, R., & Robinson, S. (2000). Simulation studies: Key stages and processes.

Buckley, N., Mills, G., Letellier-Duchesne, S., & Benis, K. (2021). Designing an energy-resilient neighbourhood using an urban building energy model. *Energies*, *14*(15). https://doi.org/10.3390/en14154445

Bustos-Turu, G., van Dam, K. H., Acha, S., Markides, C. N., & Shah, N. (2016). Simulating residential electricity and heat demand in urban areas using an agent-based modelling approach. https://doi.org/10.1109/ENERGYCON.2016.7514077

Cao, F. (2009). A new initialization method for categorical data clustering. *Expert Systems with Applications*, (36), 10223–10228.

CE Delft, Quintel, & TNO. (2021). Systeemstudie energie-infrastructuur zuid-holland. https://ce.nl/wp-content/uploads/2021/03/CE_Delft_Quintel_TNO_200140_Systeemstudie_Zuid-Holland_DEF.pdf

Centraal Bureau voor de Statistiek. (2019, March 15). Energielevering aan woningen en bedrijven naar postcode [Last Modified: 2019-03-15T16:16:00+01:00]. Retrieved August 22, 2023, from https://www.cbs.nl/nl-nl/maatwerk/2020/33/energielevering-aan-woningen-en-bedrijven-naar-postcode

Centraal Bureau voor de Statistiek. (2023a). *Geografische data* [Centraal Bureau voor de Statistiek] [Last Modified: 15-03-2023T07:11:30]. Retrieved September 15, 2023, from https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data

Centraal Bureau voor de Statistiek. (2018, August 2). *Het industriële landschap van Nederland* [Centraal Bureau voor de Statistiek] [Last Modified: 2018-08-03T00:00:00+02:00]. Retrieved September 6, 2023, from https://www.cbs.nl/nl-nl/achtergrond/2018/31/het-industriele-landschap-van-nederland

Centraal Bureau voor de Statistiek. (2023b). *Inwoners per gemeente* [Centraal Bureau voor de Statistiek] [Last Modified: 25-08-2023T13:34:57]. Retrieved September 6, 2023, from https://www.cbs.nl/nl-nl/visualisaties/dashboard-bevolking/regionaal/inwoners

Centraal Bureau voor de Statistiek. (2020a). Kaart van 500 meter bij 500 meter met statistieken [Last Modified: 02-05-2023T14:58:14]. Retrieved August 22, 2023, from https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/kaart-van-500-meter-bij-500-meter-met-statistieken

Centraal Bureau voor de Statistiek. (2023c, May 1). Statistische gegevens per vierkant en postcode 2022-2021-2020-2019 [Last Modified: 2023-05-02T00:00:00+02:00]. Retrieved August 22, 2023, from https://www.cbs.nl/nl-nl/longread/diversen/2023/statistische-gegevens-per-vierkant-en-postcode-2022-2021-2020-2019

Centraal Bureau voor de Statistiek. (2020b). Wijk- en buurtkaart 2020 [Last Modified: 25-02-2023T00:30:08]. Retrieved August 22, 2023, from https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/wijk-en-buurtkaart-2020

Centraal Bureau voor de Statistiek. (2022, December 15). Zonnestroom; vermogen zonnepanelen woningen, wijken en buurten, 2020 [Last Modified: 2022-12-15T02:00:00+01:00]. Retrieved August 24, 2023, from https://www.cbs.nl/nl-nl/cijfers/detail/85447NED

Centraal Bureau voor de statistiek. (2023, August 22). Kaart van 100 meter bij 100 meter met statistieken [Last Modified: 02-05-2023T15:30:16]. Retrieved August 22, 2023, from https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/kaart-van-100-meter-bij-100-meter-met-statistieken

Chen, H.-C., Han, Q., & De Vries, B. (2020). Modeling the spatial relation between urban morphology, land surface temperature and urban energy demand. *Sustainable Cities and Society*, *60*, 102246. https://doi.org/10.1016/j.scs.2020.102246

Costa, E., Papatsouma, I., & Markos, A. (2022). Benchmarking distance-based partitioning methods for mixed-type data. *Advances in Data Analysis and Classification*. https://doi.org/10.1007/s11634-022-00521-7

de Vos, N. J. (2015). *Kmodes categorical clustering library*. https://github.com/nicodv/kmodes

de Vries, W. T., & Schrey, M. (2022). Geospatial approaches to model renewable energy requirements of the new capital city of indonesia. *Frontiers in Sustainable Cities*, *4*. https://doi.org/10.3389/frsc.2022.848309

Drechsler, M., Ohl, C., Meyerhoff, J., Eichhorn, M., & Monsees, J. (2011). Combining spatial modeling and choice experiments for the optimal spatial allocation of wind turbines. *Energy Policy*, *39*(6), 3845–3854. https://doi.org/10.1016/j.enpol.2011.04.015

Eicker, U., Schumacher, J., Brennenstuhl, M., & Weiler, V. (2020). SWC 2019 renewable energy supply concepts for next generation cities using the integrated urban modeling platform INSEL 4d, 2479–2487. https://doi.org/10.18086/swc.2019.51.02

Fichera, A., Frasca, M., Palermo, V., & Volpe, R. (2018). An optimization tool for the assessment of urban energy scenarios. *Energy*, *156*, 418–429. https://doi.org/10.1016/j.energy.2018.05.114

Fichera, A., Inturri, G., La Greca, P., & Palermo, V. (2016). A model for mapping the energy consumption of buildings, transport and outdoor lighting of neighbourhoods. *Cities*, *55*, 49–60. https://doi.org/10.1016/j.cities.2016.03.011

Fonseca, J. A., & Schlueter, A. (2015). Integrated model for characterization of spatiotemporal building energy consumption patterns in neighborhoods and city districts. *Applied Energy*, *142*, 247–265. https://doi.org/10.1016/j.apenergy.2014.12.068

Foss, A. H., & Markatou, M. (2018). Kamila: Clustering mixed-type data in r and hadoop. *Journal of Statistical Software*, *83*, 1–44. https://doi.org/10.18637/jss.v083.i13

Gupta, S., Kumar, R., Lu, K., Moseley, B., & Vassilvitskii, S. (2017). Local search methods for k-means with outliers. *Proceedings of the VLDB Endowment*, *10*(7), 757–768. https://doi.org/10.14778/3067421.3067425

Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values. proceedings of the first pacific asia knowledge discovery and data mining conference. *Singapore*, *21-34*.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, *2*(3), 283–304. https://doi.org/10.1023/A:1009769707641
Interviewee 1. (2023). Anonymous interview with policy officer heat transition.
Interviewee 2. (2023). Anonymous interview with policy officer provincial energy transition.
Interviewee 3. (2023). Anonymous inteview with policy officer regional energy strategy and spatial policy.
Interviewee 4. (2023). Anonymous interview with policy officer spatial quality and spatial policy.
Interviewee 5. (2023). Anonymous interview with gas network operator.
Interviewee 6. (2023). Anonymous interview with transport system operator.
Interviewee 7. (2023). Anonymous interview with distribution system operator.
Jäger, T., Mckenna, R., & Fichtner, W. (2016). The feasible onshore wind energy potential in baden-württemberg: A bottom-up methodology considering socio-economic constraints. *Renewable Energy*, *96*, 662–675. https://doi.org/10.1016/j.renene.2016.05.013
Jimeno, J., Roy, M., & Tortora, C. (2021). Clustering mixed-type data: A benchmark study on KAMILA and k-prototypes. In T. Chadjipadelis, B. Lausen, A. Markos, T. R. Lee, A. Montanari, & R. Nugent (Eds.), *Data analysis and rationality in a complex world* (pp. 83–91). Springer International Publishing. https://doi.org/10.1007/978-3-030-60104-1_10
Kraaijvanger, C., Verma, T., Doorn, N., & Goncalves, J. (2023, January 4). Evaluating access to solar energy in light of a just energy transition. https://doi.org/10.2139/ssrn.4375481
Leduc, W. R. W. A., & Van Kann, F. M. G. (2013). Spatial planning based on urban energy harvesting toward productive urban regions. *Journal of Cleaner Production*, *39*, 180–190. https://doi.org/10.1016/j.jclepro.2012.09.014
Liu, Z., Liu, X., Zhang, H., & Yan, D. (2023). Integrated physical approach to assessing urban-scale building photovoltaic potential at high spatiotemporal resolution. *Journal of Cleaner Production*, *388*, 135979. https://doi.org/10.1016/j.jclepro.2023.135979
Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2016). *Netcode elektriciteit* [Last Modified: 2022-12-21]. Retrieved February 3, 2023, from https://wetten.overheid.nl/BWBR0037940/2022-12-18#Hoofdstuk7
Ministerie van Economische Zaken en Klimaat. (2022). *Ontwerp Beleidsprogramma Klimaat*. Ministerie van Economische Zaken en Klimaat. Den Haag. https://open.overheid.nl/repository/ronl-53899d440127f31fa5f7382c72d031007894dd2e/1/pdf/Ontwerp_Beleidsprogramma_Klimaat.pdf
Moya, D., Copara, D., Borja, A., Pérez, C., Kaparaju, P., Pérez-Navarro, Á., Giarola, S., & Hawkes, A. (2022). Geospatial and temporal estimation of climatic, end-use demands, and socioeconomic drivers of energy consumption in the residential sector in ecuador. *Energy Conversion and Management*, *261*, 115629. https://doi.org/10.1016/j.enconman.2022.115629
Mutani, G., Delmastro, C., Gargiulo, M., & Corgnati, S. P. (2016). Characterization of building thermal energy consumption at the urban scale. *Energy Procedia*, *101*, 384–391. https://doi.org/10.1016/j.egypro.2016.11.049
Netbeheer nederland. (2023). Capaciteitskaart invoeding elektriciteitsnet. Retrieved August 22, 2023, from https://capaciteitskaart.netbeheernederland.nl

Nolde, M., Schwanebeck, M., Dethlefsen, F., Duttmann, R., & Dahmke, A. (2016). Utilization of a 3d webGIS to support spatial planning regarding underground energy storage in the context of the german energy system transition at the example of the federal state of schleswig–holstein. *Environmental Earth Sciences*, *75*(18). https://doi.org/10.1007/s12665-016-6089-0

Oudes, D., & Stremke, S. (2018). Spatial transition analysis: Spatially explicit and evidence-based targets for sustainable energy transition at the local and regional scale. *Landscape and Urban Planning*, *169*, 1–11. https://doi.org/10.1016/j.landurbplan.2017.07.018

PCW. (2021). *De energietransitie en de financiële impact voor netbeheerders*. https://www.netbeheernederland.nl/_upload/RadFiles/New/Documents/20210407-Finaal%20rapport%20Project%20FIEN.pdf

Provincie Zuid Holland. (2022a, September 22). Klimaatatlas - thema gebiedskenmerken - landgebruik. Retrieved August 22, 2023, from https://opendata.zuid-holland.nl/geonetwork/srv/dut/catalog.search#/metadata/e2025d81-7f32-4464-8ef8-a4657d728fe8

Provincie Zuid Holland. (2023). *Open-dataportaal*. Retrieved September 15, 2023, from https://opendata.zuid-holland.nl/geonetwork/srv/dut/catalog.search#/home

Provincie Zuid Holland. (2016a, October 14). Potentieel geothermie totaal (COP 10, tretour 25). Retrieved August 22, 2023, from https://opendata.zuid-holland.nl/geonetwork/srv/dut/catalog.search#/metadata/CE137282-B2F7-4BC5-A1CF-DA446B5116B8

Provincie Zuid Holland. (2016b, October 14). Potentieel geothermie totaal (COP 15, tretour 35). Retrieved August 22, 2023, from https://opendata.zuid-holland.nl/geonetwork/srv/dut/catalog.search#/metadata/1CE4A9F0-F50F-4023-8700-B70846E495CB

Provincie Zuid Holland. (2022b, December 13). *Startnotitie Provinciaal Miek Zuid Holland* (nieuwsbericht).

Provincie Zuid Holland. (2022c, March 1). Warmteprofielen clusters huidig. Retrieved August 22, 2023, from https://opendata.zuid-holland.nl/geonetwork/srv/dut/catalog.search#/metadata/242a7516-ad64-42ef-9973-ce3ac1213403

Provincie Zuid Holland. (2017). Zonnewijzer - potentie buurt. Retrieved August 22, 2023, from https://opendata.zuid-holland.nl/geonetwork/srv/dut/catalog.search#/metadata/EB253476-C3CB-4994-B526-6C1A0C7BF903

Ramachandra, T., & Shruthi, B. (2007). Spatial mapping of renewable energy potential. *Renewable and Sustainable Energy Reviews*, *11*(7), 1460–1480. https://doi.org/10.1016/j.rser.2005.12.002

Rijksinstituut voor Volksgezondheid en Milieu. (2023). Kaarten | atlas leefomgeving. Retrieved August 22, 2023, from https://www.atlasleefomgeving.nl/kaarten?config=3ef897de-127f-471a-959b-93b7597de188&activateOnStart=info,layermanager&gm-x=155000.00000000006&gm-y=416358.1266850175&gm-z=3.312315746690106&gm-b=1544180834512,true,1;1605001908247,true,1

Rijksoverheid. (2021). *Klimaatmonitor - totaal bekend energieverbruik - 2021 - provincies*. Retrieved September 6, 2023, from https://klimaatmonitor.databank.nl/jive

Sahoo, S., van Stralen, J. N. P., Zuidema, C., Sijm, J., & Faaij, A. (2023). Regionally integrated energy system detailed spatial analysis: Groningen province case study in the northern netherlands. *Energy Conversion and Management*, *277*, 116599. https://doi.org/10.1016/j.enconman.2022.116599

Sahoo, S., van Stralen, J. N. P., Zuidema, C., Sijm, J., Yamu, C., & Faaij, A. (2022). Regionalization of a national integrated energy system model: A case study of the northern netherlands. *Applied Energy*, *306*, 118035. https://doi.org/10.1016/j.apenergy.2021.118035

Szarka, N., Wolfbauer, J., & Bezama, A. (2018). A systems dynamics approach for supporting regional decisions on the energetic use of regional biomass residues. *Waste Management and Research*, *36*(4), 332–341. https://doi.org/10.1177/0734242X18757626

Then, D., Bauer, J., Kneiske, T. M., & Braun, M. (2021). Interdependencies of infrastructure investment decisions in multi-energy systems—a sensitivity analysis for urban residential areas. *Smart Cities*, *4*(1), 112–145. https://doi.org/10.3390/smartcities4010007

TNO. (2021). *ESDL*. Retrieved August 23, 2023, from https://www.esdl.nl/

Unternährer, J., Moret, S., Joost, S., & Maréchal, F. (2017). Spatial clustering for district heating integration in urban energy systems: Application to geothermal energy. *Applied Energy*, *190*, 749–763. https://doi.org/10.1016/j.apenergy.2016.12.136

Wang, N., Heijnen, P. W., & Imhof, P. J. (2020). A multi-actor perspective on multi-objective regional energy system planning. *Energy Policy*, *143*. https://doi.org/10.1016/j.enpol.2020.111578

Wang, N., Verzijlbergh, R. A., Heijnen, P. W., & Herder, P. M. (2020). A spatially explicit planning approach for power systems with a high share of renewable energy sources. *Applied Energy*, *260*. https://doi.org/10.1016/j.apenergy.2019.114233

Wang, Q., M'Ikiugu, M. M., & Kinoshita, I. (2014). A GIS-based approach in support of spatial planning for renewable energy: A case study of fukushima, japan. *Sustainability (Switzerland)*, *6*(4), 2087–2117. https://doi.org/10.3390/su6042087

Wang, S.-H., Huang, S.-L., & Huang, P.-J. (2018). Can spatial planning really mitigate carbon dioxide emissions in urban areas? a case study in taipei, taiwan. *Landscape and Urban Planning*, *169*, 22–36. https://doi.org/10.1016/j.landurbplan.2017.08.001

Yamagata, Y., & Seya, H. (2013). Simulating a future smart city: An integrated land use-energy model. *Applied Energy*, *112*, 1466–1474. https://doi.org/10.1016/j.apenergy.2013.01.061

Yao, H., & Zang, C. (2021). The spatiotemporal characteristics of electrical energy supply-demand and the green economy outlook of guangdong province, china. *Energy*, *214*. https://doi.org/10.1016/j.energy.2020.118891

Zhou, Y., Ma, Y., Wang, J., & Lu, S. (2021). Collaborative planning of spatial layouts of distributed energy stations and networks: A case study. *Energy*, *234*. https://doi.org/10.1016/j.energy.2021.121205