



Wie Watt bespaart, die heeft wat!

*Een onderzoek naar
het energiebesparingspotentieel
van Nederlandse huishoudens.*

Frédérique Lalleman

Masterscriptie Business Analytics

Wie Watt bespaart, die heeft wat!

*Een onderzoek naar
het energiebesparingspotentieel
van Nederlandse huishoudens.*

Frédérique Lalleman

Masterscriptie Business Analytics

Augustus 2013

Vrije Universiteit Amsterdam
Faculteit der Exacte Wetenschappen
Studierichting Bedrijfswiskunde en Informatica
De Boelelaan 1081a
1081 HV Amsterdam

Begeleider: dr. E. Haasdijk
2^e lezer: prof. dr. Mathisca de Gunst

Liander N.V.
Klant & Markt
Utrechtseweg 68
6812 AH Arnhem

Begeleider: Tessa van Doremaele



vrije Universiteit amsterdam



Voorwoord

Ter afronding van de master Business Analytics, voorheen bekend als Business Mathematics and Informatics, is het gebruikelijk een stage te lopen. Binnen deze (bedrijfs)context wordt de afrondende scriptie geschreven waarbij de kennis die tijdens de opleiding in de gebieden bedrijfskunde, wiskunde en informatica is opgedaan, in de praktijk wordt toegepast.

Tijdens mijn stage bij Liander in Arnhem heb ik met veel plezier mogen werken aan een opdracht met betrekking tot energiebesparing; een onderwerp dat van nature al mijn interesse heeft. Dit onderzoek richt zich op het identificeren van (groepen) huishoudens die mogelijk veel energie kunnen besparen en is uitgevoerd binnen de afdeling Klant & Markt. Graag wil ik mijn begeleider van de VU, Evert Haasdijk, bedanken voor alle sparsessies met betrekking tot de aanpak van dit onderzoek dat geen kant-en-klare oplossing had en waarbij de 'standaard' technieken niet 1-2-3 konden worden toegepast. Ook mijn begeleider vanuit Liander, Tessa van Doremaele wil ik bedanken voor de constructieve feedback op de opzet en structuur van dit verslag en voor de verbindingsfactor die zij speelde bij het inwinnen van informatie binnen de organisatie. Daarnaast ben ik mijn team van Markt & Business Intelligence dankbaar voor de hulp bij de dataverzameling en -begrip, het meedenken en de frisse wandelingen in de pauze. Tot slot, wil ik ook de teamleden van project SERI bedanken voor het mogen meekijken en -draaien in het project.

Frédérique Lalleman

Augustus, 2013

Samenvatting

Liander transporteert als netbeheerder de energie van de leverancier naar de gebruiker en heeft in die functie een onafhankelijke rol ten opzichte van de consument omdat de consument de energierekening niet aan haar betaalt. Vanuit deze positie ziet zij het als haar taak om actief consumenten aan te zetten tot energiebesparing. Om dit op een effectieve en efficiënte manier te kunnen doen, wil zij graag weten waar het energiebesparingspotentieel (EBP) van huishoudens zich bevindt. Dit is een complexe vraag omdat nog niet eerder een EBP bepaald is en een norm voor een energiezuinig huishouden lastig te bepalen is. Daarom is onder andere dit onderzoek gestart met de volgende hoofd- en deelvragen:

Hoe kan het EBP van huishoudens worden geschat?

- 1) Wat is de definitie van het EBP?
- 2) Welke factoren zijn van invloed op het elektra- en gasverbruik?

Om een methode te bepalen waarmee het EBP geschat wordt, wordt het EBP als volgt gedefinieerd: *“Een huishouden heeft een energiebesparingspotentieel als zij ten opzichte van een groep vergelijkbare huishoudens een (veel) hoger verbruik heeft.”*

Om vast te stellen of een huishouden, dat bestaat uit het huis en haar bewoners, wel of geen hoog EBP heeft, is het nodig om eerst een vergelijkingsgroep te bepalen en een grens tussen normaal en hoog verbruik binnen die groep. Om deze groepen te vormen is in de literatuur en met behulp van analyses onderzocht wat de factoren zijn die van invloed zijn op het elektra- en gasverbruik. Uit de literatuur komen een aantal kenmerken naar voren, maar deze zijn –in verband met privacy- niet voor elk huishouden beschikbaar. Daarom hebben we acht kenmerken gebruikt die bij benadering bekend zijn en zowel huis- en socio-demografische kenmerken¹ weergeven. Met behulp van een regressie analyse is bepaald welk kenmerk het meest voorspellend is voor het standaardjaarverbruik. Het *elektraverbruik* kan dan voorspeld worden met (1) de oppervlakte, (2) het type woning, (3) de levensfase en (4) het inkomen en het *gasverbruik* met (1) het type woning, (2) de oppervlakte, (3) het bouwjaar en (4) het inkomen.

De methodiek die een antwoord geeft op de hoofdvraag bestaat uit drie stappen. Allereerst worden met het Two-Step cluster algoritme vergelijkbare groepen gemaakt op basis van de acht kenmerken, waarbij het aantal groepen gevarieerd wordt. Met behulp van de Silhouette Coëfficiënt (SC) wordt vervolgens bepaald wat een goed aantal groepen (clusters) zou zijn. Wanneer vergelijkbare groepen worden gemaakt op basis van de bovenstaande invloedsfactoren, dan is door andere kenmerken te verklaren waarom er een verschil bestaat in het verbruik tussen de huishoudens met een normaal verbruik (de massa) en met een hoog verbruik (de potentiëlen). Daarom is de tweede stap in de methodiek ontwikkeld: de homogeniteitstest, die test of de kenmerken van de massa en de potentiëlen binnen een cluster overeenkomen. Is dit het geval, dan is redelijk aan te nemen dat de potentiëlen de huishoudens met een hoog EBP zijn. Er is gekozen de groep potentiëlen te definiëren als 25% van de huishoudens met het hoogste verbruik binnen het cluster. Als derde stap in de methodiek voeren we ter controle stap één uit op subsets van de data en bekijken of dezelfde huishoudens aangemerkt worden als hebbende EBP. Dit is een indicatie dat de procedure consistente clusters oplevert en kan helpen het gevonden resultaat verder te verfijnen.

¹ type woning, bouwjaar, inhoud, oppervlak, inkomen, opleiding, levensfase en eigendom woning (koop/huur)

Er zijn een aantal conclusies te trekken uit het toepassen van de methodiek op verschillende datasets om te onderzoeken hoe deze presteert. Hieruit kan geconcludeerd worden dat er groepen van betere kwaliteit worden gevonden wanneer deze wordt toegepast op een kleinere dataset dan op heel Liander gebied. De clusters zijn dan specifiekere qua kenmerken, waardoor het minder vaak voorkomt dat er een verschil bestaat tussen de massa en potentiëlen. Ook wordt de clustering specifiekere wanneer een deel van de huishoudens als uitschieter aangemerkt mag worden. Daarnaast geldt dat een hoge SC niet betekent dat de gevonden clusters een goede kwaliteit hebben. Zij geeft slechts een indicatie van de kwaliteit die in stap 2 verder getoetst moet worden.

Er komen meerdere aanbevelingen uit dit onderzoek, maar de voornaamste is deze methodiek te gebruiken alvorens een project gericht op energiebesparing te starten om zo te bepalen onder welke huishoudens deze het best kan plaats vinden. Hiermee kan ook in de praktijk worden getoetst of de gevonden huishoudens inderdaad een hoog EBP hebben.

Inhoudsopgave

Inleiding	1
Aanleiding	1
Probleemstelling.....	1
Complexiteit van het onderzoek	2
Over Liander	2
Scope	2
Gebruikte technieken.....	3
Leeswijzer	3
1. De energiemarkt, wetgeving en duurzaamheidsambities	5
1.1 Energiemarkt	5
1.1.1 Partijen in het energieleveringsproces.....	5
1.1.2 Liander en de energiemarkt	6
1.2 Wetgeving.....	7
1.2.1 Wet onafhankelijk beheer	7
1.2.2 Wet bescherming persoonsgegevens	7
1.2.3 Liander en wetgeving	7
1.3 Duurzaamheidsambities op de energiemarkt.....	7
1.3.1 Energietransitie	8
1.3.2 Slimme meter	9
1.3.3 Energielabel.....	10
1.3.4 Liander en duurzaamheidsprojecten	11
1.3.5 In het kort... ..	12
2. Beschrijving van de data.....	13
2.1 Achtergrondkennis data	13
Bisnode	13
Liander	13
2.2 Gebruikte data.....	14
2.2.1 Technische details	14
2.2.2 Data bewerking	14
3. Definitie energiebesparingspotentieel en scope	15
3.1 Definities uit literatuur	15
3.2 Definitie energiebesparingspotentieel.....	15

3.3	Scope en beperkingen definitie.....	16
3.3.1	Beperkingen definitie en onderzoek	16
4.	Invloedsfactoren energieverbruik en onderlinge samenhang	19
4.1	Theoretisch kader.....	19
4.1.1	Invloedsfactoren energieverbruik uit literatuur.....	19
4.2	Methodiek	21
4.2.1	Methodiek trendanalyse elektra- en gasverbruik	21
4.2.2	Methodiek analyse invloedsfactoren energieverbruik	21
4.2.3	Methodiek analyse onderlinge relatie factoren energieverbruik	22
4.3	Resultaten en analyse	24
4.3.1	Resultaten trendanalyse elektra- en gasverbruik	24
4.3.2	Resultaten invloedsfactoren energieverbruik	25
4.3.3	Resultaten en analyse onderlinge relatie factoren	31
4.4	Conclusie	34
5.	Beschrijving veranderde insteek onderzoek	35
5.1	Onderbouwing reële besparing huishoudens alleen op detailniveau mogelijk.....	36
5.2	Standaardverbruik huishoudens niet specifiek genoeg of niet reëel.....	36
6.	Methodiek bepaling energiebesparingspotentieel	37
6.1	Methodiek	37
6.1.1	Stap 1: Bepaal aantal clusters en algehele clusterkwaliteit	38
6.1.2	Stap 2: Test homogeniteit cluster	39
6.1.3	Stap 3: Consistentie uitkomst clusters	41
6.1.4	Toegevoegde waarde clustering	42
6.2	Algoritmes	45
6.2.1	Two-Step clustering.....	45
6.2.2	Silhouette coëfficiënt	47
7.	Resultaten en analyse.....	49
7.1	Resultaten analyses ter bepaling van methodiek	49
7.2	Resultaten en analyse stap 1 en 2 voor verschillende clusteringen	50
7.2.1	Inzichten met betrekking tot de methodiek	55
7.3	Resultaten en analyse stap 3: consistentie clustering.....	56
7.4	Resultaten en analyse toegevoegde waarde clustering.....	57

8. Conclusie en aanbevelingen	61
8.1 Conclusie	61
8.2 Aanbevelingen	62
8.2.1 Aanbevelingen verder onderzoek	62
8.2.2 Aanbevelingen Liander	63
Bijlagen	65
Bijlage A: Begrippenlijst en gebruikte afkortingen	67
Bijlage B. Betekenis categorieën per variabele	69
Bijlage C. Resultaten analyses invloedsfactoren en onderlinge relatie	71
Resultaten factoren elektra	71
Resultaten onderlinge factoren	79
Bijlage D: resultaten clusteranalyse en homogeniteitstest	83
Bibliografie	93

Inleiding

Het energielandschap wereldwijd is sterk veranderd in de afgelopen eeuwen. Door de industriële revolutie zijn veel processen geautomatiseerd en werden fossiele brandstoffen massaal aangeboden om aan de groeiende energievraag te voldoen. De westerse wereld is wederom met een transitie bezig: van fossiele brandstoffen naar een meer duurzame opwekking van energie. Tegelijkertijd groeit ook het bewustzijn om zuinig om te springen met energie, met als gevolg een stijgende aandacht voor energiebesparing.

Aanleiding

Niet iedereen is zich even bewust van zijn energiegebruik of heeft daar evenveel inzicht in. Om dit inzicht kwantitatief te maken heeft Liander, een energie netwerkbeheerbedrijf, onderzoek laten uitvoeren naar de hoeveelheid inzicht dat mensen hebben in hun *energiebesparingspotentieel* (EBP). Het energiebesparingspotentieel was hier gedefinieerd als de hoeveelheid energie die een gebruiker zou kunnen besparen. Uit dit onderzoek bleek dat het *inzicht* in het EBP in de huidige vraagstelling niet is te meten (Liander, 23 februari 2012). Naast dat het inzicht lastig is te meten, is ook een 'potentieel' uitdagend om te berekenen: er is geen harde waarde beschikbaar waaraan je het kan staven. Het is als het voorspellen van een uitkomst, waarbij je niet aan eerdere uitkomsten kan toetsen of de voorspelling redelijk is. Ondanks deze moeilijke factor wil Liander graag een meer kwantitatieve definitie en berekening van dit EBP.

Zij wil dit potentieel kunnen berekenen om inzicht te krijgen in:

- ...bij welke klanten de grootste energiebesparing behaald kan worden;
- ...waar klanten die zelf energie opwekken eerder geneigd zijn om energie terug te leveren aan het net;
- ...waar zij het beste aan de slag kan om mensen te helpen met energiebesparing;
- ...het succes van energiebesparingsacties, waarbij het EBP als maatstaf kan functioneren.

Deze verschillende aspecten laten zien dat het EBP op veel verschillende manieren gebruikt kan worden. Afhankelijk van het doel wordt het EBP verschillend berekend en gelden er andere regels voor de toepassing ervan in verband met de privacy van de gebruiker (zie paragraaf 1.2.2). Het uitgangspunt voor dit onderzoek is het eerste punt: het bepalen van het EBP van huishoudens om te weten waar de grootste energiebesparing behaald kan worden. Waarom dit voor Liander relevant is, wordt in paragraaf 1.1.1 nader toegelicht.

Probleemstelling

Na overweging van de verschillende aspecten van het EBP is gekozen voor de volgende probleemstelling:

Hoe kan het EBP van huishoudens worden geschat?

Om deze vraag te kunnen beantwoorden, onderzoeken we de volgende deelvragen:

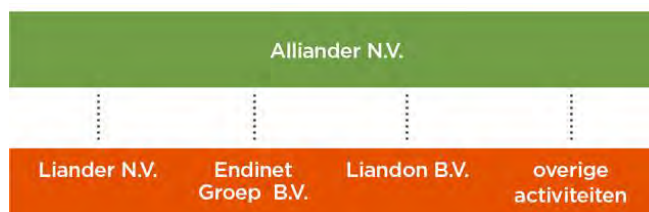
- 1) Wat is de definitie van het EBP?
- 2) Welke factoren zijn van invloed op het elektra- en gasverbruik?

Complexiteit van het onderzoek

Het schatten van het EBP is een uitdagende en complexe taak. Allereerst omdat er niet eerder een EBP is bepaald van een huishouden, waardoor het niet mogelijk is de uitkomsten van dit onderzoek te vergelijken met eerdere onderzoeken. Daarnaast is de potentie tot besparing afhankelijk van de norm voor een energiezuinig huishouden. Kan hierbij een zelfde norm voor alle huishoudens gebruikt worden of niet? En wat bepaalt deze norm? Hoe reëel is het potentieel dat geschat wordt? En met welke huishoudens kan een huishouden vergeleken worden? Deze verschillende vragen waarop geen hapklaar antwoord is te geven, maken dit onderzoek complex.

Over Liander

Dit onderzoek is uitgevoerd binnen het netwerkbedrijf Liander op de afdeling Markt en Business Intelligence. Liander maakt, net als Endinet en Liandon, onderdeel uit van het moederbedrijf Alliander zoals schematisch weergegeven in Figuur 1.



Figuur 1: De organisatiestructuur van Alliander, waar Liander onderdeel van uit maakt.

Deze delen samen zorgen voor onderhoud, vernieuwing, uitbreiding en aanpassing van het energienetwerk. In hoofdstuk 1 komt verder aan bod wat de taken van Liander zijn. Markt & Business Intelligence (MBI) houdt zich bezig met het vergaren van marktkennis en klantinzichten, voorziet collega's van kennis over de markt en operationele prestaties en benut de verkregen informatie proactief om de klantgerichtheid in Liander te vergroten (Velthuis & van Doremalee, 2012).

Scope

De resultaten van dit onderzoek zijn gebaseerd op de klanten van Liander en de gebieden waarin zij actief is. Daarnaast beperkt het EBP zich tot huishoudens en laat zij zakelijke klanten en grootverbruikers buiten beschouwing.

Het energieverbruik wordt beïnvloed door een vaste en een variabele component. Onder het vaste deel worden de huishoudkenmerken als type woning, isolatie, etc. bedoeld. Het variabele deel is energieverbruik door gedrag, bijvoorbeeld hoe hoog de temperatuur in huis is. In dit onderzoek wordt bij het verbruik van een *huishouden* deze vaste en variabele component tezamen bedoeld: het huis mét de inwoners. Deze wordt ook gezamenlijk onderzocht.

Het EBP van een huishouden heeft alleen betrekking op de woonruimte en de activiteiten die zich daarin afspelen. Woningkenmerken en socio-demografische gegevens als opleidingsniveau, inkomen, gezinssamenstelling, plaats etc. worden meegenomen en energie die gebruikt wordt bij woon-werk verkeer en reizen worden buiten beschouwing gelaten.

In hoofdstuk 3 over de definitie van het EBP wordt nader ingegaan op de details van de randvoorwaarden en beperkingen van dit onderzoek.

Gebruikte technieken

De belangrijkste bijdrage van dit onderzoek is de ontwikkelde methodiek om huishoudens met een hoog EBP te onderscheiden. In de methodiek is gebruik gemaakt van statistische data analyse en het Two-Step cluster algoritme dat zowel het BIRCH algoritme implementeert en hiërarchisch clusteren toepast. Deze algoritmes worden in paragraaf 6.2 ‘Algoritmes’ verder toegelicht. De ontwikkelde methodiek is gevalideerd en de resultaten van deze validatie staan in hoofdstuk 7.

Leeswijzer

Om een globaal beeld te krijgen van hoe het energielandschap in Nederland in elkaar steekt, wordt hier in hoofdstuk 1 bij stil gestaan. Ook worden hier recente ontwikkelingen besproken die van belang zijn voor dit onderzoek en de hieraan gerelateerde projecten binnen Liander. De lezer die hier al kennis over heeft, kan dit hoofdstuk overslaan of alleen het kopje ‘In het kort’ lezen. Te allen tijde kan ook bijlage A met hierin een begrippenlijst geraadpleegd worden. Hoofdstuk 2 bespreekt de oorsprong en kwaliteit van de gebruikte data. In hoofdstuk 3 wordt een antwoord gegeven op de eerste deelvraag ‘Wat is de definitie van het EBP?’ Daarna wordt in hoofdstuk 4 de tweede deelvraag beantwoord door een literatuuronderzoek en analyses naar de invloedsfactoren op het energieverbruik en hun onderlinge relatie. Hoofdstuk 5 beschrijft de inzichten en resultaten die voor een veranderde insteek van dit onderzoek hebben gezorgd, waarna in hoofdstuk 6 een antwoord op de hoofdvraag wordt gegeven door de ontwikkelde methodiek en de gebruikte algoritmes te beschrijven. In hoofdstuk 7 staan de resultaten van de analyses die gedaan zijn om meer inzicht in het toepassen van deze methodiek te krijgen. Hoofdstuk 8 sluit af met de conclusie van dit onderzoek, aanbevelingen voor verder onderzoek en aanbevelingen voor de implementatie van de methodiek.

1. De energiemarkt, wetgeving en duurzaamheidsambities

In dit hoofdstuk wordt in paragraaf 1.1 een beeld gegeven van hoe de huidige energiemarkt in elkaar steekt en welke wetgeving relevant is voor dit onderzoek (paragraaf 1.2). Daarnaast wordt er in paragraaf 1.3 aandacht besteed aan de duurzaamheidsambities op de energiemarkt. Aan het eind van elke sectie wordt kort ingegaan op de relatie van Liander met het onderwerp. Tot slot wordt in 'In het kort...' een overzicht gegeven van de onderwerpen die genoemd worden waarmee het EBP samenhangt.

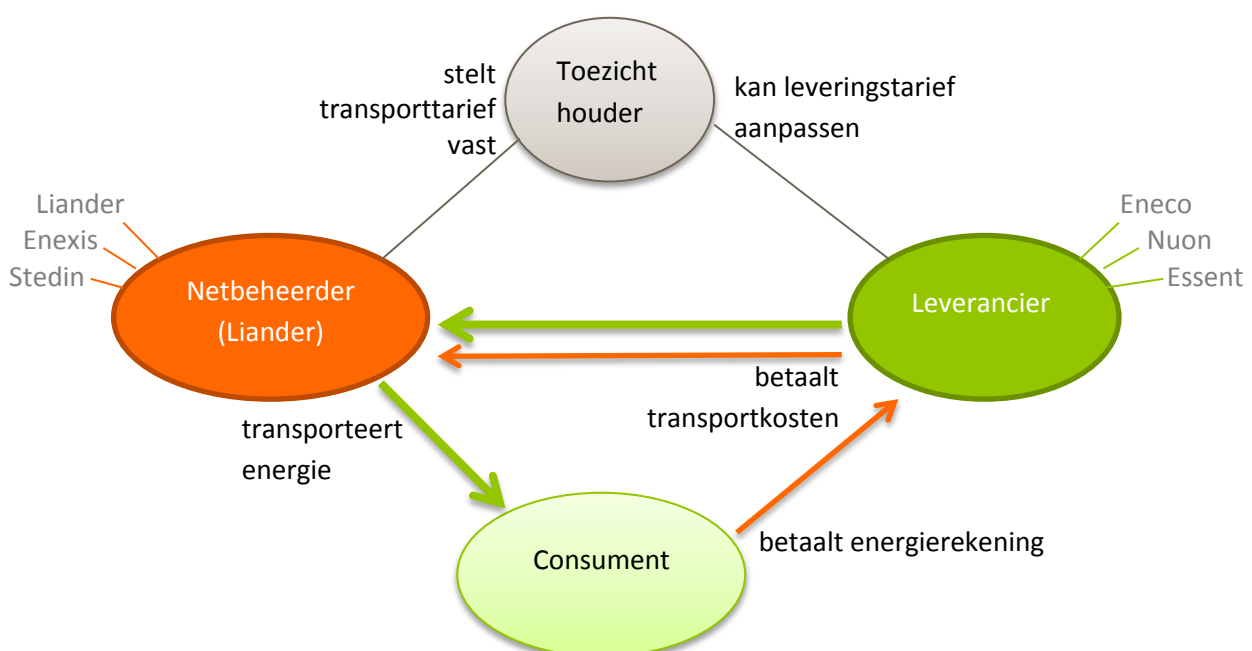
1.1 Energiemarkt

Onbewust speelt energie een belangrijke rol in ons leven: we gebruiken het met koken, wassen, verwarmen van ons huis en nog veel meer, maar wat gebeurt er voor het thuis uit het stopcontact komt?

Allereerst zijn er grofweg twee verschillende manieren om energie op te wekken: groen en grijs. Groen staat voor duurzame energie wat betekent dat het energie is "waarover de mensheid voor onbeperkte tijd kan beschikken en waarbij, door het gebruik ervan, het leefmilieu en de mogelijkheden voor toekomstige generaties niet worden benadeeld." Grijs energie daarentegen is opgewekt door bijvoorbeeld kolencentrales of gas (Wikipedia).

1.1.1 Partijen in het energieleveringsproces

Nadat de energie is opgewekt, wordt het vanaf de centrale via transformaties van hoog naar laagspanning via kabels getransporteerd naar de gebruiker (groene pijlen in onderstaande figuur). Consumenten die zelf energie opwekken laten we buiten beschouwing. Bij dit transportproces zijn verschillende partijen betrokken waarvan de hoofdspelers hieronder in Figuur 2 schematisch zijn weergegeven: de leverancier, de netbeheerder, de consument en de toezichthouder.



Figuur 2: De partijen in het energieleveringsproces en hun taken, waarbij de groene pijlen de energiestromen weergeven en de oranje pijlen de betaalstroom. Dat een consument ook energie kan opwekken is niet weergegeven.

De *leverancier* levert de groene of grijze energie en werkt deze eventueel ook op. Via het netwerk transporteert zij vervolgens de energie naar de *consument* die haar hiervoor de rekening betaalt. In deze rekening zitten ook de transportkosten die de leverancier aan de netbeheerder moet betalen voor het transporteren van de stroom.

De *netbeheerder* op haar beurt, is verantwoordelijk voor het onderhoud, de uitbreiding en de vervanging van het net. Ook het onderhoud en de vervanging van de meter bij de consument valt onder haar beheer. Storingen in het net worden ook door haar verholpen. Voor deze verschillende taken ontvangt zij het transporttarief dat door de *toezichthouder* wordt vastgesteld. Naast het vaststellen van het transporttarief, kan de toezichthouder ook het leveringstarief van de leverancier aanpassen. Dit toezicht wordt in Nederland door de Nederlandse Mededingingsautoriteit² (NMa) gehouden “om ervoor te zorgen dat de kwaliteit van het transport goed is en dat de consument daar een redelijke prijs voor betaalt (NMa, Regulerings regionale netbeheerders) (Netbeheer-Nederland, ECN, & Energie-Nederland, 2012).

Liander heeft als netbeheerder een onafhankelijke rol ten opzichte van de consument omdat de consument de energierekening niet aan haar betaalt. Vanuit deze positie ziet zij het als haar taak om actief consumenten aan te zetten tot energiebesparing. Om dit op een effectieve en efficiënte manier te kunnen doen, wil zij graag weten waar het EBP van huishoudens zich bevindt.

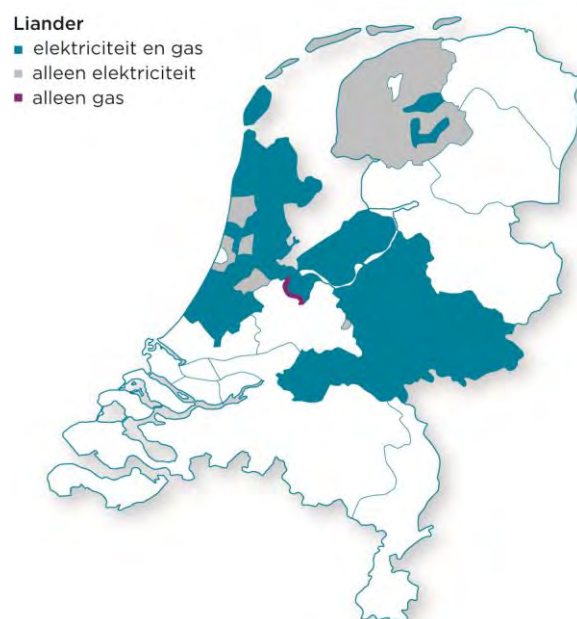
Bovenstaand model geeft de huidige situatie weer en is aan verandering onderhevig door wetgeving van de overheid en door de ontwikkelingen op het gebied van duurzame energie opwekking. Door deze ontwikkelingen gaan ook steeds meer consumenten energie terugleveren en loopt de energiestroom niet meer één richting op.

Regionaal gebonden netbeheer

Het elektriciteitsnet in Nederland is onderverdeeld in regionale distributienetten die door de minister van Economische Zaken, Landbouw en Innovatie elk zijn toegewezen een netbeheerder (NMa, Regulerings regionale netbeheerders). Om deze reden zijn netbeheerders, waaronder Stedin, Enexis en Liander, onderling geen concurrenten. Netbeheerders kunnen zowel voor gas, elektra of beide verantwoordelijk zijn in hun regio.

1.1.2 Liander en de energiemarkt

De netbeheerder heeft een onafhankelijke rol: zij staat tussen de leverancier en de consument in. Vanuit deze positie is zij de aangewezen persoon om consumenten aan te zetten tot energiebesparend gedrag. Liander wil, naast haar dagelijkse taken, zich actief met deze maatschappelijke rol bezig houden. Zoals in Figuur 3 te zien is, is Liander vooral in West & Oost Nederland actief en deels in Friesland. De



Figuur 3: Transportgebied gas en elektra door Liander.

² Tijdens dit onderzoek is het NMa met andere partijen samengevoegd tot één toezichthouder: 'Autoriteit Consument en Markt' (ACM).

data voor dit onderzoek komt dan ook zowel uit dicht stedelijk gebied als landelijk gebied.

1.2 Wetgeving

Er bestaan een twee wetten die van invloed zijn op dit onderzoek. Allereerst de Wet onafhankelijk netbeheer (WON) die de splitsing van levering en netbeheer binnen een bedrijf verplicht en daarnaast de Wet bescherming persoonsgegevens (Wbp).

1.2.1 Wet onafhankelijk beheer

De “Wet onafhankelijk netbeheer bepaalt dat netbeheerders geen andere (commerciële) activiteiten mogen uitvoeren dan het beheren van de elektriciteits- en gasnetten” (NMa, Energiewetten). Deze wet bepaalt sinds 2006 dat de leveranciers- en de beheerderstaken wettelijk van elkaar gescheiden moeten zijn (Overheid) (Eerste Kamer).

1.2.2 Wet bescherming persoonsgegevens

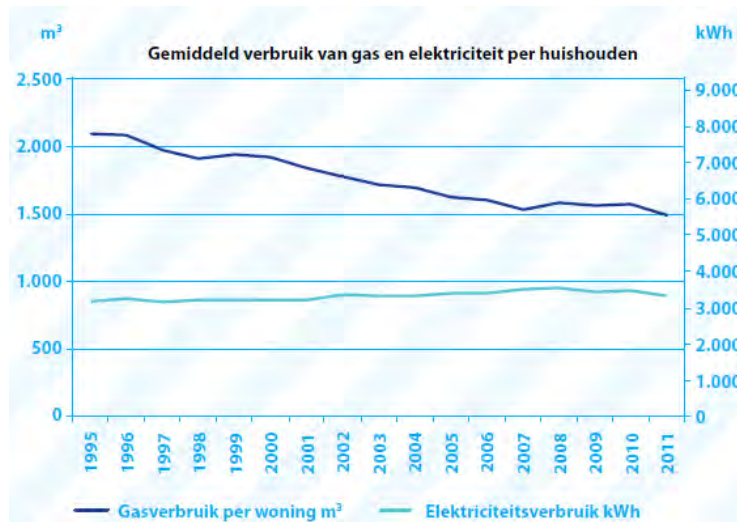
De Wet bescherming persoonsgegevens (Wbp) regelt, ter bescherming van de privacy, wat er allemaal wel en niet mag met de persoonsgegevens van een persoon. In de Wbp staat wat de rechten zijn van de consument als zijn gegevens gebruikt worden (Rijksoverheid, Wat regelt de Wbp?).

1.2.3 Liander en wetgeving

Vanwege de WON is het vroegere N.V. Nuon gesplitst in het nieuw opgerichte netwerkbedrijf Alliander en het productie- en leveringsbedrijf Nuon Energy (Alliander, Geschiedenis Alliander). Binnen Liander wordt er eerst een Privacy & Security proces doorlopen voordat de data gebruikt wordt voor onderzoek, om te zorgen dat aan de Wpb voldaan wordt.

1.3 Duurzaamheidsambities op de energiemarkt

Het huishoudelijk energieverbruik in Nederland daalt de afgelopen jaren gestaag. Het gasverbruik daalde de afgelopen 15 jaar en ook daarvoor daalde zij al. Dit wordt voornamelijk veroorzaakt door beter geïsoleerde woningen en efficiëntere verwarmingsketels. Waar het elektriciteitsverbruik voorheen toenam, stabiliseerde dit de afgelopen jaren. De eerdere stijging van het elektriciteitsverbruik is te verklaren door de toename van elektrische apparaten in het huishouden. In 2011 neemt het elektriciteitsverbruik zelfs af zoals in onderstaande Figuur 4 te zien is; het gemiddeld gasverbruik is dan 1484 m³ en het elektriciteitsverbruik is 3312 kWh (Netbeheer-Nederland, ECN, & Energie-Nederland, 2012).



Figuur 4: Gemiddeld verbruik van gas en elektriciteit per huishouden; bron: Energietrends 2012, Netbeheer-Nederland.

Wat betekent dit voor de toekomst? Zet deze daling zich voort? Enerzijds zou het energieverbruik de komende jaren weer kunnen gaan stijgen door de mogelijke doorbraak van de elektrische auto. Anderzijds, het groeiende bewustzijn om zuiniger om te gaan met de beperkte natuurlijke bronnen kan, naast de verduurzaming van het energieverbruik, ook zorgen voor een verdere daling van het verbruik. Hieronder worden drie onderwerpen besproken die hieraan gerelateerd zijn en relevant zijn voor dit onderzoek, te weten: het begrip Energietransitie, de slimme meter en het energielabel.

1.3.1 Energietransitie

De overheid heeft de doelstelling om in 2020 20% van de energie duurzaam op te wekken, 20% minder CO₂ uit te stoten en 20% minder energie te verspillen. Dit worden de 20-20-20 doelstellingen genoemd (Rijksoverheid, Europa 2020 | Europese Unie). Om deze reden wil zij overgaan van een energievoorziening gebaseerd op fossiele brandstoffen naar een volledig duurzame energievoorziening. Dit overgangsproces wordt de Energietransitie genoemd (intranet Alliander) en kan grofweg worden opgesplitst in twee sporen met elk hun eigen implicaties voor Liander die hieronder worden toegelicht:

- (A) Energiebesparing;
- (B) Duurzame energieopwekking.

A. Energiebesparing

De doelstelling van 20% energiebesparing heeft als gevolg dat de Nederlandse overheid 2% energie per jaar wil besparen. Aangezien de overheid in de persoon van provincies en gemeenten aandeelhouder is van Alliander heeft deze doelstelling directe invloed op Liander. (Alliander, jaarverslag 2012, 2012) Vanuit deze doelstelling en vanuit haar maatschappelijke positie wil Liander klanten aanmoedigen tot energiebesparing. Dit is ook een reden waarom dit onderzoek plaats vindt.

B. Duurzame energieopwekking

Naast energieleveranciers en bedrijven worden ook consumenten gestimuleerd om zelf energie op te wekken om de 20% doelstelling te behalen. Dit heeft impact op het energienetwerk.

Impact op energienetwerk

Voor consumenten wordt het financieel interessanter om energie op te wekken om hun eigen energierekening te verlagen. De consument wordt *prosumer*: een consument die naast energie afneemt, ook energie opwekt. Dit betekent dat er veel meer (kleine) energieleveranciers bij zullen komen waar de netbeheerder rekening mee zal moeten houden. Verder is een kenmerk van duurzame energie als wind en zon het fluctuerende karakter. De wind waait niet altijd en de zon laat zich niet altijd zien. (vrij naar intranet Alliander). Om deze twee redenen zal het netwerk twee grote veranderingen gaan ondervinden: scherpere fluctuaties in de hoeveelheden energie die getransporteerd worden en een verschuiving van eenrichtings- naar tweerichtingsverkeer van energie door het netwerk.

Er zijn verschillende manieren om de grote fluctuaties in het netwerk op te vangen. Een eerste manier is om de netten te verzwaren door meer bekabeling in de grond te leggen, maar dit is een dure oplossing. Een andere manier is de zogeheten smart grids: “slimme netwerken welke vraag en aanbod kunnen sturen en zodoende pieken kunnen afvlakken.” Smart grids vallen buiten dit onderzoek omdat ze meer van toepassing zijn op het gebruik van het net dan op energiebesparing.

1.3.2 Slimme meter

Naast de bovengenoemde 20% doelstellingen heeft de overheid zich ten doel gesteld dat in 2020 in minimaal 80% van de huishoudens een *slimme meter* geplaatst is. Dit is conform de norm van de Europese Unie (Kamp, 2013).

Een slimme meter houdt net als een traditionele meter het elektra- en gasverbruik bij. Dit wordt echter niet meer analoog gedaan, maar digitaal. Dit maakt het mogelijk dat de meter zelf de meterstanden doorgeeft aan de netbeheerder, tenzij de gebruiker hier bezwaar tegen maakt. Naast deze mogelijkheid, kunnen er ook verschillende toepassingen op de meter worden aangesloten die ‘slim’ omgaan met energie mogelijk maken. Om deze reden wordt het een ‘slimme’ meter genoemd. (Rijksoverheid, De slimme meter, 2011)

De slimme meter geeft de netbeheerder meer inzicht in de hoogte en de fluctuaties van het energieverbruik van consumenten, waardoor netbeheerders als Liander hun netten beter in kunnen richten om de hoeveelheid energie goed te transporteren. De netbeheerder mag de meterstanden zes keer per jaar op afstand uitlezen om een verbruiksoverzicht naar de consument te sturen. Deze data mag zij ook gebruiken voor analyses. Mocht de gebruiker niet willen dat de meterstanden automatisch op afstand worden doorgegeven, dan kan zij de meter laten uitzetten.

Door de frequentere meterstandoverzichten en de toepassingsmogelijkheden van de slimme meter kan de consument meer inzicht in haar verbruik krijgen, wat haar kan helpen besparen (Rijksoverheid, De slimme meter, 2011).

Maatschappelijke business case slimme meter

In 2012 en 2013 is begonnen met de kleinschalige aanbidding (KSA) van de slimme meter. “Tijdens de kleinschalige uitrol wordt de slimme meter aangeboden in de situaties die de Europese richtlijn energie-efficiëntie³ voorschrijft, zoals bij nieuwbouw, renovatie, op eigen verzoek van de consument (zogenaamde prioriteitsplaatsingen) en bij reguliere vervanging. Het voornemen is om vanaf begin

³ Richtlijn 2006/32/EG van het Europees Parlement en de Raad van 5 april 2006 betreffende energie-efficiëntie bij het eindgebruik en energiediensten en houdende intrekking van Richtlijn 93/76/EEG van de Raad

2014 over te gaan tot het grootschalig aanbieden (GSA) van slimme meters door de netbeheerders aan alle huishoudens in Nederland” (Kamp, 2013).

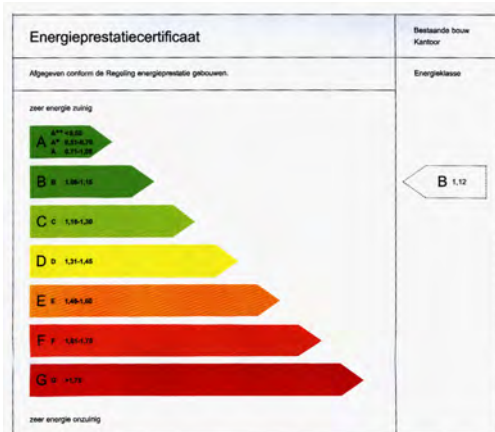
Tijdens de KSA wordt ook gemonitord in hoeverre de slimme meter effect heeft op de besparing van energie bij huishoudens door de feedback die zij, of haar applicaties, geeft over het verbruik. In het algemeen wordt een onderscheid gemaakt tussen verschillende soorten feedback: direct en indirect. “Directe feedback (via een display in de woning waarop ook het momentane verbruik te zien is) is effectiever dan indirecte feedback (via websites, verbruiksoverzichten en dergelijke)” (Kema N.V., 2010). In het onderzoek ‘Intelligente meters in Nederland’ worden op pagina 57 overwegingen beschreven over de invloed van directe en indirecte feedback op het energieverbruik bij verschillende soorten gebruikers. Op basis van deze overwegingen, studies en experimenten zijn landelijke besparingspercentages bepaald die door de indirecte feedback behaald zouden moeten worden. Dit is 3,7% voor gas en 3,2% elektra (Kema N.V., 2010).

Deze norm-percentages worden ook door Liander gehanteerd om te meten wat het effect is van haar campagnes om mensen met een slimme meter tot energiebesparing aan te zetten. Onder ‘KPI Besparingspotentieel’ (zie paragraaf 1.3.4) wordt hier verder op ingegaan.

1.3.3 Energielabel

Om huurders en kopers van woningen te helpen inzicht te krijgen in het verwachte energieverbruik van een woning heeft de overheid naar Europese richtlijnen een *energielabel* voor woningen

ingesteld dat sinds 1 januari 2008 verplicht is (AgentschapNL, 2012). Een voorbeeld van dit label ziet u hiernaast in Figuur 5, waarbij geldt dat een woning met een hoger label, bijvoorbeeld ‘A’, een lager verwacht energieverbruik heeft dan een woning met een ‘D’ label. “Het label drukt het energiegebruik uit in de energie-index. Deze energie index, die wordt berekend op basis van de gebouweigenschappen, de gebouwgebonden installaties en een gestandaardiseerd bewonersgedrag, geeft energiegebruik aan per m² gebruiksoppervlak in MJ” (TU Delft & Stimuleringsfonds Volkshuisvesting, Perspectieven voor energiebesparing in de particuliere woningvoorraad, 2009).



Figuur 5: Voorbeeld van een energielabel.

Uit onderzoek door de TU Delft blijkt dat het werkelijke verbruik hier veel van kan afwijken. “Het theoretische elektriciteitsverbruik per label is veel kleiner dan het werkelijk gebruik (gemiddeld 2.5 keer kleiner, bijna onafhankelijk van het label), wat komt doordat het niet-gebouwgebonden elektriciteitsverbruik niet meegenomen wordt in de theoretische berekening.” Ook is “het theoretische gasverbruik (licht) te laag in de betere labels en sterk te hoog in de slechtere labels. (...) Het label duidt op de energetische kwaliteit van de woning zelf maar kan nooit aangeven hoe de woning in werkelijkheid gebruikt zal worden.”

Naast dit energielabel bestaat er nog een andere indicatie van het verbruik: de energetische kwaliteitsscore. De TU Delft heeft in opdracht van het Stimuleringsfonds Volkshuisvesting onderzocht hoe deze labels in verhouding tot elkaar staan. De verschillen wijzen erop “dat het energielabel een beter beeld geeft van de feitelijke energetische kwaliteit dan de energetische kwaliteitsscore” (TU

Delft & Stimuleringsfonds Volkshuisvesting, Perspectieven voor energiebesparing in de particuliere woningvoorraad, 2009). De energetische kwaliteitsscore wordt in de praktijk weinig gebruikt. Om deze redenen en omdat de data van energielabels meer voorhanden is, laten we de energetische kwaliteitsscore in dit onderzoek verder buiten beschouwing.

1.3.4 Liander en duurzaamheidsprojecten

Zoals eerder beschreven is dit onderzoek uitgevoerd op de afdeling Markt & Business Intelligence dat onderdeel uitmaakt van de afdeling Klant & Markt. Markt & Business Intelligence (MBI) houdt zich bezig met het vergaren van marktkennis en klantinzichten, voorziet collega's van kennis over de markt en operationele prestaties en benut de verkregen informatie proactief om de klantgerichtheid in Liander te vergroten (Velthuis & van Doremaele, 2012). Deze focus komt onder andere naar voren in een tweetal projecten die nauw verwant zijn aan dit onderzoek: de Key Performance Indicator (KPI) Besparingspotentieel en het project SERI.

KPI Besparingspotentieel

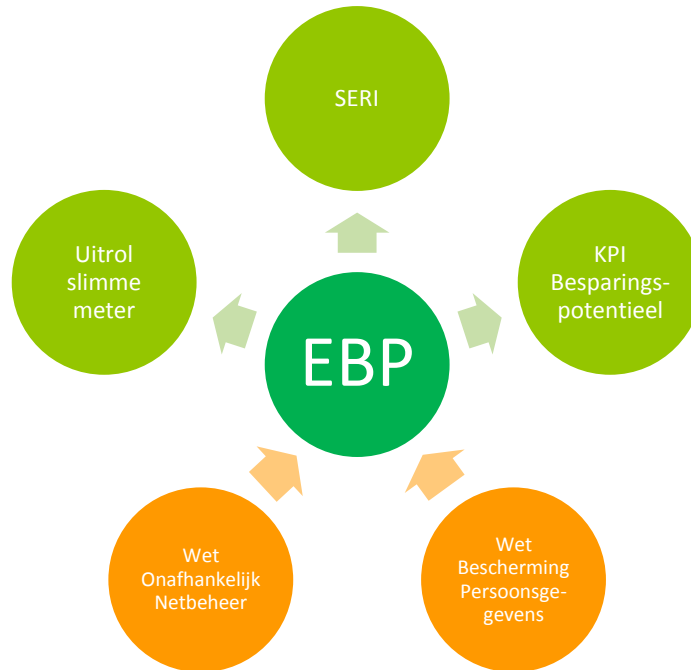
“De definitie van de KPI is: ‘besparingspotentieel aantonen bij consumenten met een slimme meter die door Liander worden gestimuleerd om energie te besparen.’ Deze KPI maakt inzichtelijk wat het effect is van de inspanningen die Liander verricht om besparingsgedrag van consumenten te beïnvloeden” (Alliander, RvB vergaderstuk, 2013). Het verschil tussen deze KPI en het EBP is als volgt: de KPI kijkt naar het procentuele verschil in het energieverbruik van de consument voor en na een actie gericht op energiebesparing, waar het EBP aangeeft waar mogelijk besparingspotentieel zit op basis van een aantal factoren, los van acties.

Project SERI

SERI staat voor Smarter Energy Research Institute, een onderdeel van het IBM Research lab in New York waar Alliander sinds oktober 2012 in een tweejarig project samen met andere partijen mee samen werkt. Voor Alliander wordt dit project gebruikt om te leren over de analyse van grote hoeveelheden en complexe data en om modellen te ontwikkelen die kunnen helpen bij het voorspellen van zaken als onder andere het EBP, de acceptatie van de slimme meter en de bereidheid van mensen om energie te besparen. Bij het EBP van SERI wordt gekeken naar een voorspelling op huishoudniveau, waarbij de uitkomsten en inzichten van dit onderzoek kunnen gebruikt worden als input voor de verdere ontwikkeling van eerdergenoemde modellen. Dit onderzoek is daarom ook aan het IBM SERI team gepresenteerd.

1.3.5 In het kort...

Uit de verscheidenheid van de hierboven beschreven aspecten als Wet Onafhankelijk Netbeheer, privacy, slimme meter, KPI Besparingspotentieel en SERI blijkt dat het EBP met veel andere zaken raakvlakken en samenhang heeft. Deze worden hieronder in Figuur 6 schematisch weergegeven en kort toegelicht.



Figuur 6: Factoren van invloed op het EBP (oranje) en de relatie van het EBP met onderwerpen (groen).

- + Het EBP zou kunnen helpen om het effect van de slimme meter te meten.
- + De opgedane kennis kan in het leertraject van SERI toegepast worden.
- + Zij kan helpen om de KPI Besparingspotentieel mogelijk verder te verscherpen.
- De Wet bescherming persoonsgegevens beperkt de mate waarin het EBP gebruikt kan worden.
- De Wet Onafhankelijk Netbeheer beperkt de mogelijkheden van het gebruik van het EBP binnen andere afdelingen binnen Liander.

Waarom is Liander geïnteresseerd in het EBP van huishoudens?

Liander heeft als netbeheerder een onafhankelijke rol ten opzichte van de consument omdat de consument de energierekening niet aan haar betaalt. Vanuit deze positie ziet zij het als haar taak om actief consumenten aan te zetten tot energiebesparing. Om dit op een effectieve en efficiënte manier te kunnen doen, wil zij graag weten waar het EBP van huishoudens zich bevindt.

2. Beschrijving van de data

Dit hoofdstuk geeft achtergrondkennis over de data (paragraaf 2.1) en beschrijft welke data gebruikt is en hoe deze bewerkt is (paragraaf 2.2).

2.1 Achtergrondkennis data

De dataset die gebruikt is voor dit onderzoek is samengesteld uit Liander en Bisnode data, aangevuld met algemeen bekende gegevens als in welke wijk/buurt een huis staat. Vanuit Liander zijn het verbruik van gas en elektra bekend, evenals hoeveel energie een huishouden opwekt en terug levert, het type meter etc. Bisnode data bevat huiseigenschappen als het type woning, huur of koop, bouwjaar etc. en persoonsgegevens als levensfase, inkomen en opleiding.

Bisnode

Bisnode gebruikt veel verschillende bronnen waar zij hun gegevens op baseren waaronder het CBS en Dataland. Ook heeft zij gegevens die verkregen zijn door een enquête⁴ onder 700.000 huishoudens. Bisnode data bestaat zodoende uit of werkelijke waardes van één persoon binnen het huishouden of geschatte waardes op basis van voorspellingsmodellen en beslisregels. De betrouwbaarheid van de schatting verschilt per kenmerk (Timmermans) .

Liander

De data van het gas- en elektraverbruik bestaat uit geschatte en daadwerkelijke gemeten data. De daadwerkelijke data kan zowel doorgegeven zijn door de consument zelf of worden opgemeten. Dit laatste wordt eens per 3 jaar door een meteropnemer gedaan of bij het verhuizen gecontroleerd. Het geschatte jaarverbruik wordt geschat op basis van eerdere verbruiken, of in het geval van nieuwbouw, op basis van de aansluitwaarde. Er zijn twee verschillende manieren om het verbruik uit te drukken:

- 1) Werkelijk verbruik
- 2) Standaardjaarverbruik

Het *werkelijke* verbruik per jaar voor gas wordt erg beïnvloed door de weersomstandigheden. Immers, in een strenge winter zal de kachel vaker aanstaan en er meer verbruikt worden. Om de verschillende jaren met elkaar te kunnen vergelijken wordt het *standaardjaarverbruik* (SJV) gebruikt waarbij het werkelijke verbruik is gecorrigeerd voor de temperatuur. Voor elektra wordt ook een SJV bepaald omdat de standen niet alleen aan het eind van het jaar worden geregistreerd en dit ook fluctueert.

In de praktijk worden veel verschillende type meters gebruikt met elk een verschillende aansluitwaarde. Een aansluitwaarde geeft aan hoeveel Ampère elektriciteit of m³ gas er door de aansluiting kan. Binnen de verschillende type aansluitwaardes voor elektra en gas zit en logische ordening. Naarmate de aansluitwaarde groter is, kan een huishouden in theorie meer verbruiken. De aansluitwaarde voor huishoudens kan variëren van 1 zekering van 16A (kortweg 1x16A) tot 3 zekeringen van 25A (3x25A).

⁴ De Grote Consumenten Enquête, Bisnode Nederland

2.2 Gebruikte data

2.2.1 Technische details

Gebruikte data voor de analyses is gedateerd op 20-2-2012 en bestaat uit 44 parameters per huishouden. Hierbij zitten onder andere de verbruiken van 2007 -2012 voor gas, elektra en teruglevering, de type meters, geografische gegevens als postcode, huisnummer etc. Deze zijn afkomstig van Liander. Ook zijn er acht huishoudkenmerken (afkomstig van Bisnode) gebruikt welke steeds met een hoofdletter weergegeven zijn als er over de variabele gesproken wordt. Deze zijn: Type woning, Eigendom (huur/koop), Bouwjaar, Oppervlakte en Inhoud woning, Levensfase, Inkomen en Opleiding. Elk van deze variabelen bestaat uit een aantal categorieën waartoe een huishouden kan behoren. Voor een compleet overzicht van de variabelen en hun categorieën, zie bijlage B. De dataset is opgebouwd uit alle kleinverbruik particuliere klanten.

2.2.2 Data bewerking

Een huishouden heeft gebruikelijk een G4 of G6 gas meter en een 1x25A of 3x25A elektra meter. Er is slechts een enkeling is die een grotere meter aanvraagt in verband met bijvoorbeeld een jacuzzi. De hoge uitschieters zitten voornamelijk in huishoudens met een metaaraanluiting van hoger dan 3x25A voor elektra en een G10, G16 of G25 meter voor gas. Deze type aansluitwaardes worden in de regel niet voor 'normale' huishoudens gebruikt en zijn daarom uit deze data set verwijderd.

Uit de ruwe data analyse blijkt dat het SJV van gas en SJV van elektra beide naast hele hoge uitschieters, ook lage uitschieters hebben. Na overleg is besloten, de verbruiken van gas met minder dan 20m³ en van elektra met minder dan 50kWh uit de analyses weg te laten⁵ omdat deze waarden hoogstwaarschijnlijk bij leegstaande panden horen en bij garages. Het gaat hier om 11.676 en 34.118 huishoudens (cases) voor respectievelijk elektra en gas.

Van de 2,5 miljoen mensen zijn er zo'n half miljoen die geen gasaansluiting hebben. Vooral bij de categorie flat/appartement heeft een groot deel dit niet, omdat zij vaker aangesloten zijn op stadsverwarming. Deze zijn logischerwijs bij de analyses voor gas weggelaten; de groep voor de bepaling van het EBP-gas is dan ook kleiner.

De data die Liander heeft over het energieverbruik is gekoppeld aan een klantnummer en aan een aansluitobject, zoals een huis. Soms heeft een consument twee elektra meters omdat hij bijvoorbeeld een vakantiehuis heeft of een extra meter voor de garage. *Er is alleen met consumenten gewerkt die één elektra meter hebben.* Ook kan het voorkomen dat persoon X het gas en persoon Y de elektra voor een aansluitobject betaalt. Zij hebben dan verschillende klantnummers, die beide in de analyse zijn meegenomen. *Dit maakt voor dit onderzoek niet uit omdat gas en elektra gescheiden geanalyseerd worden.* Het uitgangspunt voor dit onderzoek is het aansluitobject en niet het klantnummer, omdat vooral het verbruik in grote mate wordt beïnvloed door het gebouw waarin geleefd wordt. Er zijn zo'n 405.000 klanten waarbij de Bisnode data niet gekoppeld kan worden. Deze zijn ook niet meegenomen. Uiteindelijk is met een dataset gewerkt van 1.526.130 cases (59% van het totaal).

Er is geen rekening gehouden met verschillende bewoners op hetzelfde adres door verhuizingen.

⁵ Pairwise deletion; hierbij wordt niet de hele case verwijderd bij de analyse, maar wordt de case alleen weggelaten wanneer de variabele gebruikt wordt waarvan die waarde mist.

3. Definitie energiebesparingspotentieel en scope

In dit hoofdstuk wordt een antwoord gegeven op de eerste deelvraag: *‘Wat is de definitie van het EBP?’* Hiertoe wordt eerst een overzicht gegeven van definities die in de literatuur worden gehanteerd (paragraaf 3.1). Vervolgens wordt de definitie van het EBP beschreven zoals deze in dit onderzoek is gebruikt (paragraaf 3.2). Tot slot worden in paragraaf 3.3 de beperkingen van deze definitie aangegeven en de scope van dit onderzoek verder toegelicht.

3.1 Definities uit literatuur

Vanuit de theorie en praktijk zijn verschillende definities in omloop van het EBP. In de praktijk is het *‘de hoeveelheid energie die bespaard kan worden’*. Deze definitie is heel breed en dat blijkt ook uit de verscheidenheid van onderzoeken waarin zij naar voren komt: van het EBP van de slaapstand van kantoorartikelen (Kawamoto, Shimoda, & Mizuno, 2004) tot een Europees onderzoek over het EBP binnen landen. Dit laatste is het eindrapport *‘Study on the Energy Savings Potentials in EU Member States, Candidate Countries and EEA Countries’*⁶. In dit onderzoek wordt de definitie van het EBP zeer verdiept en wordt bijvoorbeeld onderscheid gemaakt tussen de technische potentie *‘Wat is er mogelijk?’* en economische potentie *‘Is het economisch haalbaar dit te realiseren?’* (p 24 e.v.) Dit komt naar voren in verschillende definities als:

- *“How much energy would be saved, if all end-use technology, buildings etc, existing today (in the base year) would at once be replaced by the best available technology (BAT)?”*
- *“How much MORE energy would be saved by the year 20xx (compared to frozen efficiency), if during renovation cycles or in new installations each year between now and 20xx a certain part of investments in end-use technology, buildings etc. would be in BAT.”*

Deze definities laten zien dat het EBP ook voorwaardelijk gedefinieerd kan zijn.

3.2 Definitie energiebesparingspotentieel

In dit onderzoek is er geen voorwaardelijke factor waarvoor het EBP wordt onderzocht, zoals *‘Hoeveel energie wordt bespaard als een huis van label F naar C gaat?’*, maar wordt gezocht welke huishoudens in het Liander gebied mogelijk een hoog EBP hebben. De definitie is dan ook niet gericht op het directe effect is van een maatregel, maar wil een onderscheid kunnen maken tussen huishoudens/groepen met en zonder hoog EBP. Hierbij wordt de bereidheid van mensen om te besparen buiten beschouwing gelaten, maar alleen gekeken naar de geschatte technische potentie. De definitie van het EBP was in eerste instantie een schatting van de hoeveel een huishouden zou kunnen besparen, maar is om verschillende redenen aangepast om te onderscheiden of een huishouden wel of geen hoog EBP heeft. De redenen hiervoor zijn in hoofdstuk 5 toegelicht.

⁶ Fraunhofer-Institute for Systems and Innovation Research; ENERDATA; Institute of Studies for the Integration of Systems ISIS; Technical University (Vienna, Austria); Wuppertal Institute for Climate, Environment and Energy WI; 2009

Het EBP voor een huishouden is dan als volgt gedefinieerd:

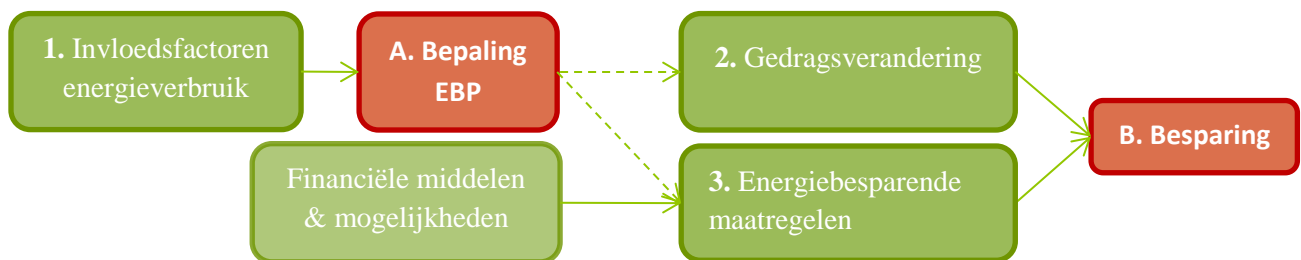
Een huishouden heeft een energiebesparingspotentieel als zij ten opzichte van een groep vergelijkbare huishoudens een (veel) hoger verbruik heeft.

Uit deze definitie volgt dat er voor het bepalen van het EBP een tweetal vragen beantwoord moet worden: ‘Wat bepaalt het energieverbruik?’ en ‘Hoe worden huishoudens onderling vergeleken?’

De antwoorden op deze vragen komen in de volgende hoofdstukken aan bod, waarbij in hoofdstuk 4 de invloedsfactoren op het energieverbruik zijn beschreven en hun onderlinge samenhang. Hoofdstuk 5 beschrijft het proces en de resultaten om tot vergelijkbare huishoudens te komen.

3.3 Scope en beperkingen definitie

Zodra het woord energiebesparingspotentieel genoemd wordt, komen vaak direct bezwaren boven als ‘Fijn dat je weet wat het potentieel is, maar hoe weet je of het ook te realiseren is? Besparing hangt toch ook af van of mensen willen besparen en of ze er iets aan kunnen doen?’ Om een duidelijk kader aan dit onderzoek en dit soort vragen te geven zijn deze verschillende zaken in de context van het EBP in onderstaande Figuur 7 weergegeven, met de bijbehorende toelichting daaronder.



Figuur 7: Plaats van het EBP binnen verschillende factoren met betrekking tot energiebesparing.

Er zijn verschillende rapporten en onderzoeken die inzicht geven in de factoren die van invloed zijn op het verbruik (1), zie paragraaf 4.1. Met behulp van deze invloedsfactoren, kan het EBP bepaald worden (A). Daarnaast is er onder andere door Abrahamse (2005, 2009) gekeken naar manieren die mensen aanzetten tot energiebesparing: de psychologische kant van besparen (2). Ook de invloed van energiebesparende maatregelen op het verbruik (3), zoals isolatie, zijn onderzocht door onder andere Bakos (2000). De huishoudens die een hoog EBP hebben, kunnen besparing behalen door gedragsverandering (2) en het toepassen van energiebesparende maatregelen zoals isolatie (3). Deze maatregelen kunnen alleen gedaan worden indien hier de financiële middelen voor zijn en als de consument deze veranderingen ook kan/mag doorvoeren, bijvoorbeeld in het geval van een huurwoning. De uiteindelijke behaalde besparing (B) wordt door (2) en (3) bepaald⁷.

3.3.1 Beperkingen definitie en onderzoek

Dit onderzoek richt zich op de invloedsfactoren energieverbruik (1 in Figuur 7 hierboven) en het EBP (A in Figuur 7). Daarnaast wordt in de definitie geen rekening gehouden met of een huishouden deze

⁷ Ook door het zelf opwekken van energie kan een huishouden besparen op haar energierekening.

besparing financieel zou kunnen realiseren of wat de bereidheid is van gebruikers om energiebesparende aanpassingen te doen, zowel op gedrags- als op investeringsniveau.

Daarnaast is dit onderzoek op huishoudniveau uitgevoerd, wat een aantal beperkingen met zich meebrengt die hieronder zijn verwoord.

- Er wordt geen onderscheid gemaakt tussen besparing die behaald kan worden door maatregelen of door gedrag. De besparing door gedrag kan wel indirect naar voren komen als tussen woningen met gelijke kenmerken een verschil zit qua energieverbruik, omdat over het algemeen gelijke kenmerken zorgen voor een gelijk basis gasverbruik. Ook zijn gedragsinvloeden indirect verwerkt in het opleidingsniveau en de gezinssamenstelling.
- Door het gebruiken van een groep vergelijkbare huishoudens komt wellicht niet alle besparing aan bod. Stel in een groep met flatwoningen heeft niemand extra isolatie. Dan is het verbruik van de hele groep hoger en zijn er geen huishoudens die aanduiden dat een lager verbruik voor deze groep mogelijk is. Het aantal huishoudens met EBP kan dan hoger zijn dan geschat.
- Een laatste beperking is dat de data slechts bij benadering beschikbaar is, waardoor een specifieke bepaling van het EBP niet mogelijk is. In hoofdstuk 5 wordt dit verder toelicht.

4. Invloedsfactoren energieverbruik en onderlinge samenhang

In het vorig hoofdstuk hebben we de eerste deelvraag beantwoord door het vaststellen van de definitie van het EBP: *‘Een groep huishoudens heeft een energiebesparingspotentieel als zij ten opzichte van een groep vergelijkbare huishoudens een (veel) hoger verbruik heeft.’*

In dit hoofdstuk geven we een antwoord op de tweede deelvraag ‘Welke factoren zijn van invloed op het elektra- en gasverbruik?’ We geven eerst een overzicht van de factoren die in de literatuur als invloedsfactoren worden aangemerkt (paragraaf 4.1). In de volgende paragrafen beschrijven we de methodiek (paragraaf 4.2), de resultaten (paragraaf 4.3) en de conclusie (paragraaf 4.4.) van dit onderzoek naar deze factoren in het Liander gebied.

We onderzoeken de invloedsfactoren om met dit inzicht vervolgens beter groepen van vergelijkbare huishoudens kunnen maken om het EBP binnen de groep te onderscheiden.

4.1 Theoretisch kader

In binnen- en buitenland zijn veel verschillende onderzoeken gedaan in relatie tot energiebesparing van huishoudens. Sommige richten zich op de besparing die behaald wordt door het doorvoeren van maatregelen, andere op de impact van gedragsverandering op het energieverbruik. Daarnaast zijn er ook onderzoeken die inzicht geven in de factoren die van invloed zijn op het verbruik. Zoals in de scope al is aangegeven (paragraaf 3.3) richten we ons hier op dit laatste. De eerste twee bepalen in hoeverre de geschatte besparing ook behaald wordt.

4.1.1 Invloedsfactoren energieverbruik uit literatuur

Om te bepalen wat het besparingspotentieel van een huishouden is, is het van belang om te weten welke factoren van invloed zijn op zowel het gas- als het elektriciteitsverbruik.

Er zijn verschillende onderzoeken gedaan naar de relatie tussen huishoudkenmerken en socio-demografische factoren, en het energieverbruik, zoals dat van Abrahamse & Steg (2009) en Gram-Hanssen, Kofod, & Nærvig Petersen (2004). De factoren die uit de Nederlandse studies voortkomen, komen redelijk overeen, terwijl buitenlandse studies ook andere factoren aandragen. Deze verschillen zijn te verklaren door een verschil in gebruik van energie. In de Verenigde Staten bijvoorbeeld, worden huizen vaak elektrisch verwarmd of gekoeld, terwijl dit in Nederland vooral met gas gebeurt. De grootte van het huis is in de VS dan ook verklarend voor het elektraverbruik terwijl zij in Nederland verklarend is voor het gasverbruik.

Er is gekozen om van de Nederlandse onderzoeken uit te gaan omdat dit onderzoek ook op Nederlandse huishoudens betrekking heeft.

Netbeheer Nederland geeft jaarlijks een overzicht van de energietrends en in haar rapport over 2012 schrijft het: “De spreiding van energieverbruik over huishoudens is aanzienlijk, zowel bij gas als bij elektriciteit. Belangrijke factoren zijn gezinsgrootte, woninggrootte, apparaatbezit, isolatiegraad, woning- en apparaatgebruik en gedrag” (Netbeheer-Nederland, ECN, & Energie-Nederland, 2012). Voor het gas- en elektraverbruik zijn verschillende factoren afzonderlijk van invloed die hieronder worden beschreven.

Factoren elektra

Het Nibud heeft in 2009 onderzoek gedaan naar de verschillen in energielasten tussen Nederlandse huishoudens. In datzelfde jaar publiceerde het Stimuleringsfonds Volkshuisvesting (SVn) in samenwerking met TU Delft 'De perspectieven voor energiebesparing in de particuliere woningvoorraad'. Beide onderzoeken laten zien dat de factoren die het meest van invloed zijn op het elektraverbruik *de vloeroppervlakte, de gezinsgrootte en het inkomen* zijn. "Met deze variabelen kan ongeveer 35% tot 56% van de spreiding in elektriciteitsgebruik worden verklaard" (TU Delft & Stimuleringsfonds Volkshuisvesting, Perspectieven voor energiebesparing in de particuliere woningvoorraad, 2009). Ook Abrahamse concludeert in haar dissertatie dat inkomen en huishoudgrootte van invloed zijn op het energieverbruik (Kema N.V., 2010).

In met name het onderzoek van het Nibud wordt nog specifiek ingegaan op deze (en andere) factoren en op de onderlinge samenhang. Netbeheer Nederland geeft aan dat ook het apparaatbezit en -gebruik een belangrijke factor is voor het elektraverbruik. Deze informatie is per huishouden echter niet publiekelijk beschikbaar. Het vloeroppervlak is daarom een factor die dit aspect indirect meeneemt in de verklaring van het elektraverbruik. Ook heeft het vloeroppervlak samenhang met het inkomen. Over het algemeen wonen mensen met een hoger inkomen in een groter huis en hebben zij meer apparaten. De invloed van het aantal apparaten is ook terug te vinden in de samenstelling van het gezin. Een gezin met kinderen ouder dan 12 jaar verbruikt meer energie dan een gezin met kinderen van 6 jaar of jonger (Nibud, 2009).

Factoren gas

In het onderzoek van SVn en TU Delft is gebruik gemaakt van veel bouwtechnische gegevens die uit het WOOn onderzoek, dat door het VROM in 2005/2006 is uitgevoerd, zijn verkregen. Het gaat hier om zo'n 5000 gebouwen binnen de Energiemodule van dit onderzoek (TU Delft & Stimuleringsfonds Volkshuisvesting, Perspectieven voor energiebesparing in de particuliere woningvoorraad, 2009). Om deze reden zijn hier meer specifieke gebouwkenmerken gebruikt dan voor algemeen gebruik beschikbaar zijn zoals de isolatiegraad, wijze van verwarming en ventilatie, en de energetische kwaliteit van de woning. Belangrijke verklaringen uit dit onderzoek voor het gasverbruik van een woning zijn de *isolatiegraad* en de *grootte* van de woning. Het Nibud onderzoek, dat veel specifiek en gedetailleerder is, geeft als top 5 van factoren die het meeste invloed hebben op het gasverbruik: *type woning, oppervlakte woonkamer, aantal vertrekken, bouwjaar en leeftijd van het hoofd van het huishouden* (ofwel jonger of ouder dan 65 jaar).

Net als bij het elektriciteitsverbruik geldt ook hier dat deze factoren onderling samenhangen. De grootte van de woning wordt bepaald door het type, het oppervlak van de woonkamer en het aantal vertrekken.

Samenvattend, op het oppervlak van de woning na zijn de factoren die van invloed zijn op elektra- en gasverbruik onderling verschillend. Het elektraverbruik wordt veel sterker beïnvloed door gedrag dan het gasverbruik en er zal voor besparing op beide gebieden voor elk een andere aanpak gekozen moeten worden. Om het gasverbruik te verlagen zijn meer investeringsmaatregelen als isolatie nodig, terwijl bij elektra de winst in blijvende gedragsverandering zit. Vanwege het verschil in invloedsfactoren op het elektra- en gasverbruik, was de eerste insteek van dit onderzoek om aparte vergelijkbare groepen te maken voor elektra en gas. Later is gekozen die niet te doen vanwege resultaten uit onderzoek die in paragraaf 7.1 onder 'Inzicht 2' worden toegelicht.

In de eerder genoemde onderzoeken zijn meer specifieke kenmerken gebruikt dan Liander tot haar beschikking heeft. In onderstaande analyses onderzoeken we daarom welke van de kenmerken die beschikbaar zijn, gebruikt kunnen worden om vergelijkbare groepen te maken om het EBP te bepalen. De kenmerken per huishouden zijn bij benadering.

4.2 Methodiek

Binnen deze analyses kan een onderscheid gemaakt worden tussen drie hoofdanalyses die nodig zijn om te bepalen welke factoren van invloed zijn op het gas- en elektraverbruik:

1. trendanalyse gas- en elektraverbruik;
2. analyse invloedsfactoren;
3. analyse onderlinge samenhang factoren.

De trendanalyse is gedaan om te bepalen of in verdere analyses één jaar als representatief voor het verbruik van huishoudens kan worden beschouwd. De gebruikte methodiek wordt hieronder per analyse toegelicht.

4.2.1 Methodiek trendanalyse elektra- en gasverbruik

Om het verbruik (het SJV) tussen de verschillende jaren (2007-2012) met elkaar te vergelijken zijn met behulp van het programma R⁸ dichtheidsfuncties en boxplots gemaakt. Beide laten zien wat de spreiding van het verbruik is en waar zwaartepunten liggen, maar geven dit op een verschillende manier weer. In de dichtheidsfunctie is ook goed te zien welke verbruiken veel voorkomen.

4.2.2 Methodiek analyse invloedsfactoren energieverbruik

Zoals in hoofdstuk 3 staat beschreven, is er data van Liander en Bisnode beschikbaar voor dit onderzoek. De data van Liander bestaat onder andere uit de jaarverbruiken voor gas en elektra, welke type meter gebruikt wordt en of er teruggeleverd wordt. Behalve het type meter zijn dit continue variabelen. De data van Bisnode echter, is categorische data, zowel ordinaal als nominaal. (Deze begrippen worden in bijlage A uitgelegd.) Vanwege de aard van deze variabelen, die niet continu zijn, is bij deze analyses met dichtheidsfuncties en boxplots gewerkt⁹. De reden is om te ontdekken wat de verschillen in verbruik zijn tussen de categorieën bij nominale data en of er een relatie tussen het verbruik en de variabele zit bij de ordinale data.

Daarnaast is met SPSS Statistics¹⁰ een voorwaartse lineaire regressieanalyse gedaan om te ontdekken in welke mate de huishoudenkenmerken (met al hun afzonderlijke categorieën) voorspellend zijn voor het verbruik en welke meer voorspellend zijn ten opzichte van de rest. Bij een voorwaartse lineaire regressie begint het model zonder variabelen en voegen we steeds de variabele toegevoegd die de grootste verbetering in R^2 het model geeft om het verbruik te verklaren. Dit geeft inzicht in welke variabele het meest verklarend is. We gebruiken de R^2 ook om de kwaliteit van het gehele model te bepalen. De R^2 geeft een mate van hoe goed het verbruik wordt geschat door de variabelen en kan tussen de 0 en 1 liggen, waarbij 1 de perfecte schatting is.

⁸ R is een gratis software omgeving waarin statistische berekeningen en grafieken gemaakt kunnen worden. Het is beschikbaar via <http://www.r-project.org/>

⁹ Een uitleg van deze type grafieken staat in bijlage A.

¹⁰ SPSS Statistics is een statistisch software programma van IBM.

Herocodering data: De variabelen uit de vorige analyse die geen lineaire relatie met het verbruik zijn gecodeerd als dummy variabelen. In onderstaande Tabel 1 zijn deze weergegeven, waarbij in de rechterkolom de referentiegroep staat en de reden waarom deze gekozen is. De variabelen Eigendom woning is binair en als 0/1 gecodeerd. De analyse is met ‘pairwise exclude’ uitgevoerd, wat betekent dat het huishouden alleen wordt weggelaten uit de analyse wanneer bij de variabele de categorie onbekend is (deze is als missend gecodeerd). Ook zijn de dummy variabelen steeds als een blok toegevoegd aan het model om de verbetering te bekijken.

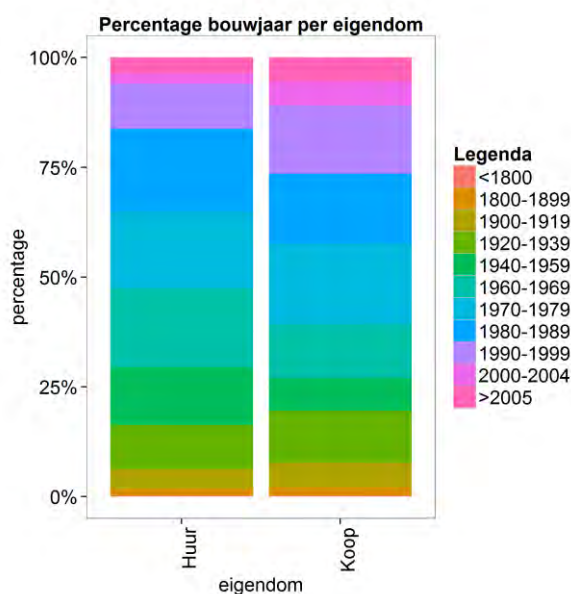
Tabel 1: Dummy variabelen in de regressie analyse met de bijbehorende referentie categorie en de reden voor de keuze.

Dummy variabele	Referentie categorie (reden)
Type woning	Rijteswoning (meest voorkomend)
Levensfase	Jonge alleenstaanden (willekeurig)
Bouwjaar	Na 2005 (waarschijnlijk meer energiezuinig)

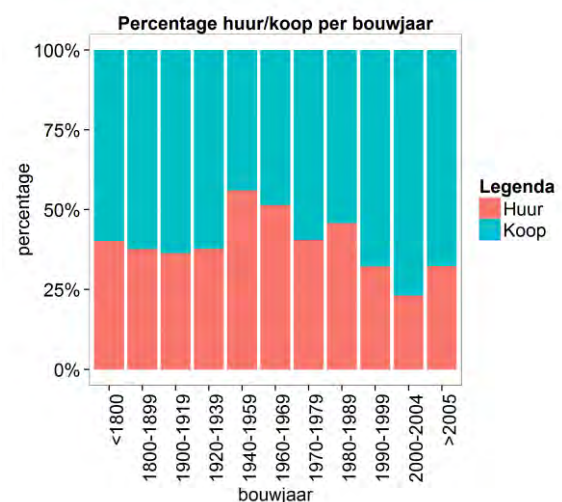
4.2.3 Methodiek analyse onderlinge relatie factoren energieverbruik

Ook bij de analyse van de relatie tussen de variabelen geldt dat we met niet-continue variabelen te maken hebben. Om deze reden hebben we hierbij met procentuele gestapelde staafdiagrammen gewerkt. We kunnen dan per categorie van een variabele A zien welk percentage per categorie van variabele B daarin valt. Als deze percentages van B per categorie nagenoeg gelijk zijn, dan geeft het weten van A niet meer inzicht over wat B kan zijn. Of als een categorie van B gelijk toeneemt bij de toenemen van categorie A (bij ordinale data) dan laat dit een lineaire relatie tussen de variabelen zien. Een tweede reden om de relatie met plots te analyseren is dat een statistische test als de chi-kwadraat test laat zien of een relatie bestaat, maar geen inzicht geeft over wat de relatie tussen specifieke categorieën is.

Ter voorbeeld: in Figuur 8 is voor huur (type 1) en voor koop (type 2) een onderverdeling gemaakt naar de verschillende bouwjaren. Type 0 betekent ‘onbekend’. In de Figuur 9 is dit andersom: per categorie bouwjaar is de verdeling huur/koop te zien. Op deze manier zijn de relaties tussen de variabelen onderzocht.

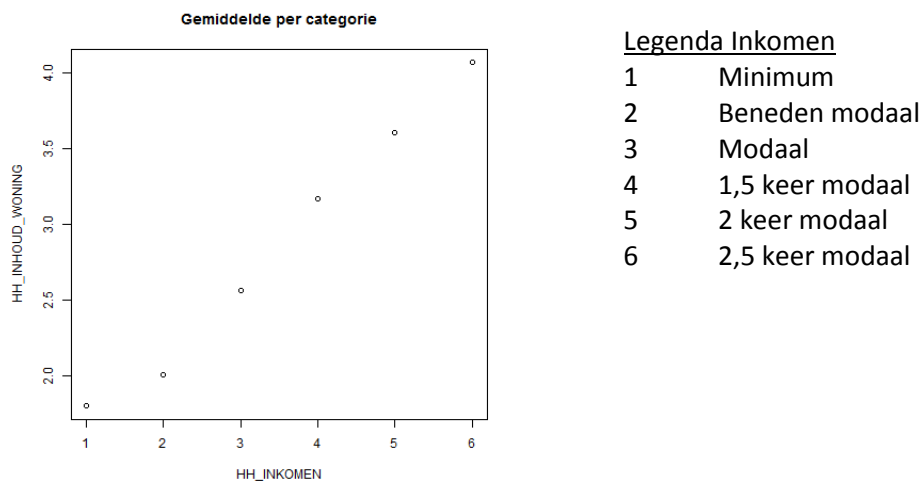


Figuur 8: De verhouding tussen de verschillende bouwjaren per type eigendom woning (1: huur en 2: koop).



Figuur 9: De verhouding tussen huur- en koopwoningen per bouwjaar.

Om de lineaire relatie tussen de ordinale variabelen beter te onderzoeken is een scatterplot geplott door het gemiddelde per categorie te nemen en deze punten te plotten. Dit is voor elk van de variabelen gedaan. Hierdoor is te zien of er verbanden zijn tussen de variabelen. Een voorbeeld van een lineaire relatie is in Figuur 10 te zien.



Figuur 10: Voorbeeld van hoe de relatie tussen twee categorische variabelen kan worden onderzocht. Dit voorbeeld laat een lineaire relatie zien.

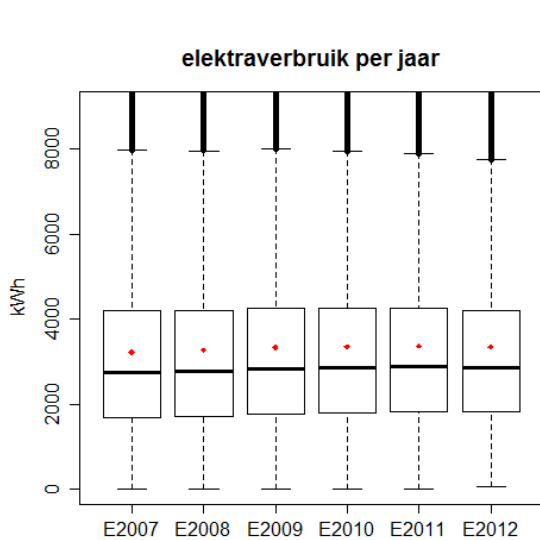
4.3 Resultaten en analyse

Hieronder worden de belangrijkste resultaten en inzichten uit de hierboven beschreven analyses weergegeven. Voor de invloedsfactoren van elektra (paragraaf 5.3.2) en de onderlinge relatie (paragraaf 5.3.3) geven we in bijlage C meer resultaten weer dan hieronder staan.

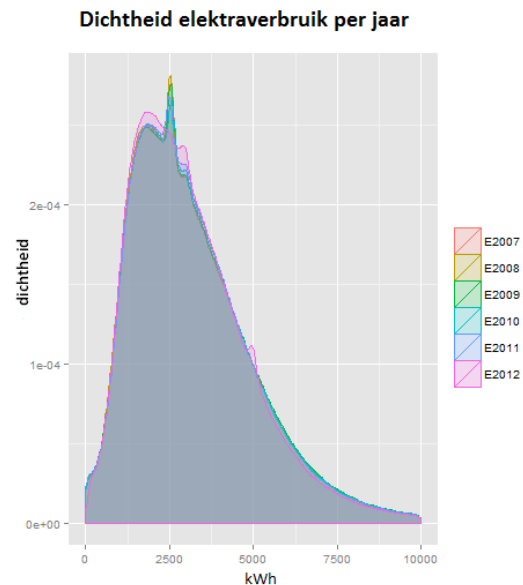
4.3.1 Resultaten trendanalyse elektra- en gasverbruik

In deze trendanalyses is steeds gebruikt gemaakt van het standaardjaarverbruik voor gas en elektra.

Elektraverbruik over de tijd



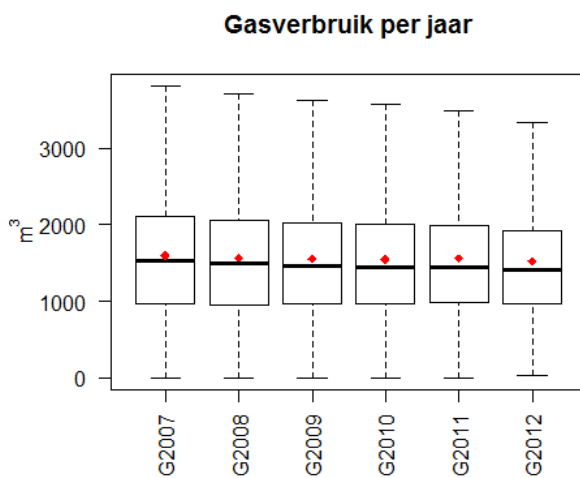
Figuur 11: Spreiding van het elektraverbruik per jaar dat een gelijke spreiding over de jaren heeft en een gemiddelde (rode punten) dat tot 2011 stijgt en daarna daalt.



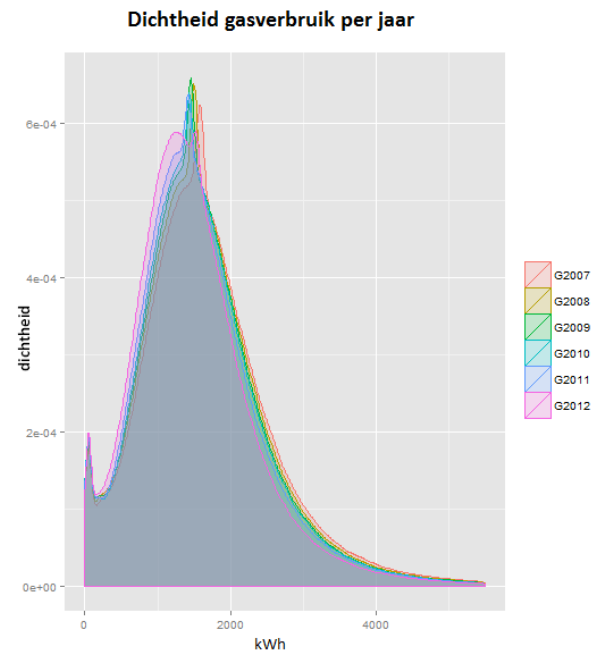
Figuur 12: De verdeling van het elektraverbruik van de huishoudens per jaar.

Uit de rechter grafiek (Figuur 12) maken we op dat het verbruik een zelfde verloop laat zien over de verschillende jaren, behalve een piek rond 2500 kWh in de jaren 2007-2011 en de kleine piek bij 5000 kWh in 2012 waar geen verklaring voor is behalve dat deze door het verbruik van huurwoningen wordt veroorzaakt. In deze grafiek is de lange staart (waarden hoger dan 10.000 kWh) weggelaten ten bate van de leesbaarheid. Door deze hoge waarden (36.428 huishoudens) is het gemiddelde (rode punt in Figuur 11) per jaar hoger dan de mediaan (de dikke zwarte lijn; de middelste waarneming). Ook is te zien dat het verbruik de afgelopen jaren steeds toenam, maar in 2012 licht daalde. De toename is te verklaren door een stijgend aantal apparaten in een huishouden. De daling kan te verklaren zijn door de economische crisis en dat apparaten energiezuiniger zijn geworden.

Gasverbruik over de tijd



Figuur 13: Spreiding van het gasverbruik per jaar. Per jaar neemt de spreiding van het verbruik en het gemiddelde (rode punten) af.



Figuur 14: De verdeling van het gasverbruik van de huishoudens per jaar, waarbij er een grote groep huishoudens een verbruik rond de 1800 m³ heeft (de extra piek).

Het gasverbruik laat ook eenzelfde verloop zien over de verschillende jaren, waarbij in de boxplot (

Figuur 13) te zien is dat de spreiding afneemt over de jaren. Ook het gemiddelde verbruik daalt licht. Er is geen verklaring gevonden waarom het verbruik op de hoge piek nog een extra piek heeft in Figuur 14. De lage piek in het verbruik is door verschillende dingen te verklaren als huizen die leeg staan, vakantiehuis, nieuwbouw of huizen die met elektriciteit verwarmen en alleen met gas koken.

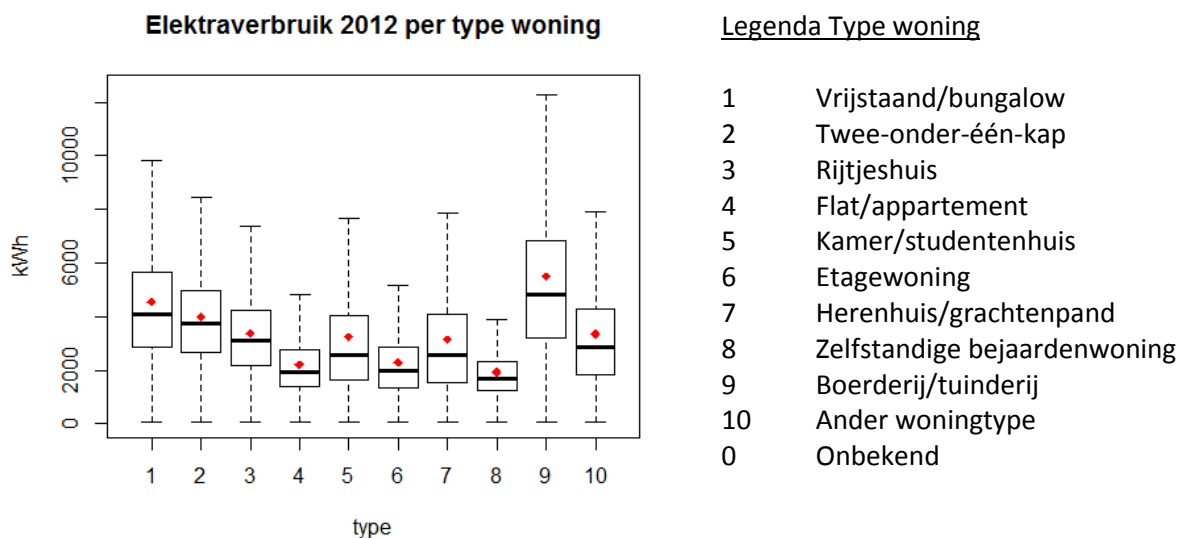
4.3.2 Resultaten invloedsfactoren energieverbruik

Factoren elektra

De resultaten van deze analyses laten gelijkenis zien met de resultaten die uit de literatuurstudie naar voren kwamen. We geven ze hieronder kort en bondig weer, waarbij de rest van de grafieken in bijlage B staan.

TYPE WONING

Het type woning maakt veel verschil uit voor de spreiding en het gemiddelde van het elektraverbruik (zie Figuur 15). Huizen met een kleiner oppervlakte (type 4, 6, 8) zoals flat, etagewoning en zelfstandige bejaardenwoning hebben een kleinere spreiding van hun elektra verbruik en een lagere mediaan.



Figuur 15: Het elektraverbruik per type woning, waarbij de rode punten het gemiddelde per type is. Huizen met een groter oppervlak (type 1, 2, 9) hebben een hoger verbruik.

Het gemiddelde ligt in alle gevallen hoger, door de lange rechterstaart zoals in paragraaf 4.3.1 al is opgemerkt. Bij het type kamer/studentenhuis, herenhuis/grachtenpand en boerderij (type 5, 7 en 9) ligt het gemiddelde echter veel hoger dan de mediaan. Dit betekent dat er binnen deze groepen huizen zijn met een veel hoger verbruik.

EIGENDOM WONING

De spreiding in het verbruik van huurwoningen is groter dan bij koopwoningen. Ook is het gemiddelde hoger.

BOUWJAAR

Het verbruik laat over de verschillende jaren redelijk dezelfde mediaan en gemiddelde zien, wat toont dat het weten van het bouwjaar weinig zegt over het elektraverbruik. De spreiding neemt tot 1969 af, waarna deze weer toeneemt. Een mogelijke verklaring hiervoor is dat de spreiding van het oppervlak van de woningen over deze categorieën ook eerst afneemt en vervolgens weer toeneemt, waarbij het oppervlak wel een verklarende factor is voor het verbruik.

INHOUD EN OPPERVLAKTE

Naarmate de woning groter wordt, neemt de spreiding en het verbruik toe.

LEVENSFASE

Gezinnen verbruiken meer dan paren die weer meer dan alleenstaanden verbruiken. Binnen de groepen hebben alleenstaanden/paren van middelbare leeftijd hoger verbruik dan de jonge en oudere alleenstaanden/paren. Voor gezinnen geldt dat een gezin met gemengd jonge en oude kinderen een hoger verbruik heeft dan gezinnen met alleen jonge of oudere kinderen.

INKOMEN EN OPLEIDING

Het energieverbruik neemt toe (zowel mediaan als gemiddelde) naarmate het inkomen/opleiding toeneemt en ook de spreiding wordt groter.

AANSLUITWAARDE

Een grotere aansluitwaarde betekent dat er meer verbruikt kan worden en zoals te verwachten is, neemt het elektriciteitsverbruik toe naarmate de aansluitwaarde van de meter toeneemt.

Kleinverbruikers met een aansluitwaarde kleiner of gelijk aan 3x25A hebben een spreiding tot maximaal 10.000 kWh.

Regressie analyse elektra

We hebben de variabele elektra voorspeld door steeds die variabele toe te voegen aan het lineaire model die de grootste verandering in R^2 gaf. We geven tussen haakjes de verhoging van de R^2 van het model weer. De volgende variabelen hebben in afnemende mate invloed op het elektraverbruik:

Oppervlakte (0,195), Type (0,020), Levensfase (0,013), Inkomen (0,008) en Bouwjaar (0,002) en zijn allemaal significant. Het model heeft een R^2 van 0,239, wat dit model een matige voorspeller maakt voor het gebruik omdat slechts 24% van de variatie in het verbruik met deze factoren wordt verklaard. Voegen we de overige twee variabelen toe, dan wordt de R^2 0,241. Voor dit onderzoek is dit geen belemmering omdat we alleen inzicht willen krijgen in welke factoren als indicatie van het verbruik gebruikt kunnen worden en niet het verbruik exact proberen te voorspellen.

De coëfficiënten van dit model staan in Tabel 2, waarbij de niet-gestandaardiseerde B bij de dummy variabelen aangeeft wat het verschil in verbruik is met de referentiegroep¹¹. Bijvoorbeeld, jonge paren zonder kinderen verbruiken ongeveer 80 kWh minder dan de referentiegroep van jonge alleenstaanden en een gezin met jonge en oude kinderen verbruikt ongeveer 536 kWh meer.

Tabel 2: De factoren en hun bèta's van het lineaire regressie model dat het elektraverbruik voorspeld. De verandering in R² geeft de verbetering van het model aan door het toevoegen van die variabele.

Elektra	Niet-gestandaardiseerde coëfficiënten		Gestandaardiseerde coëfficiënten	Verandering in R ²	
	B	Std. Error	Bèta		
(Constante)	1399,860	7,549		-	
Oppervlakte woning	356,826	1,275	,277	0,195	
Vrijstaand_Bungalow	511,783	5,143	,081	0,020	
Twee_onder_een_kap	218,491	4,794	,035		
Flat_Appartement	-373,988	4,533	-,071		
Kamer_studentenhuis	184,299	36,359	,004		
Etagewoning	-315,200	6,253	-,040		
Herenhuis_grachtenpand	-107,155	16,254	-,005		
Zelfst_bejaardenwoning	-401,070	10,849	-,028		
Boerderij_Tuinderij	1370,716	15,726	,064		
Ander_woningtype	15,749	1,331	,009		
Middelbare_alleenstaanden	-100,853	6,751	-,017		0,013
Oude_alleenstaanden	-324,855	7,086	-,051		
Gezin_jonge_kinderen	259,886	6,560	,052		
Gezin_jonge_en_oude_kinderen	535,538	10,737	,042		
Gezin_oude_kinderen	289,232	6,800	,054		
Jonge_paren_zonder_kinderen	-79,727	9,835	-,007		
Middelb_paren_zonder_kinderen	26,201	6,838	,005		
Oudere_paren_zonder_kinderen	-234,366	7,623	-,032		
Inkomen	154,178	1,220	,109	0,008	
Voor1800	343,431	30,046	,008	0,002	
Tussen1800en1899	300,635	13,057	,019		
Tussen1900en1919	219,595	9,206	,024		
Tussen1920en1939	252,665	7,947	,040		
Tussen1940en1959	215,527	8,126	,032		
Tussen1960en1969	213,613	7,626	,037		
Tussen1970en1979	207,634	7,385	,039		
Tussen1980en1989	347,403	7,458	,065		
Tussen1990en1999	365,129	7,581	,062		
Tussen2000en2004	372,893	9,489	,037		

Factoren gas

TYPE EN EIGENDOM WONING

Ook bij het gasverbruik verschilt het gemiddelde en de spreiding veel per type woning. Huizen met een kleiner oppervlakte (type 4, 6, 8) zoals flat, etagewoning en zelfstandige bejaardenwoning hebben een kleinere spreiding van hun gasverbruik en een lagere mediaan. Het gemiddelde ligt ook

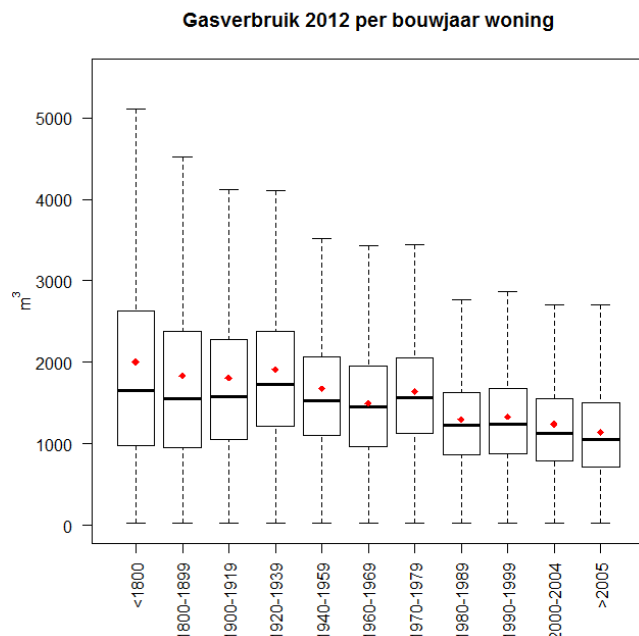
¹¹ Type woning: Rijtjeswoning; Levensfase: Jonge alleenstaanden; Bouwjaar: Na 2005 (paragraaf 4.2.1)

hier in alle gevallen hoger, waarbij voor type kamer/studentenhuis, herenhuis/grachtenpand en boerderij wederom geldt dat de huizen met een hoog verbruik, een veel hoger verbruik hebben dan bij de andere type woningen.

De huurwoningen hebben een hoger verbruik dan de koopwoningen en een iets grotere spreiding, zo'n 200m³ meer.

BOUWJAAR WONING

Het gemiddelde verbruik, de rode stip in (en mediaan) laten een daling zien naarmate het huis minder oud is. Een uitzondering hierop vormen de huizen die tussen 1970 en 1979 gebouwd zijn. Zij hebben een iets hoger verbruik.

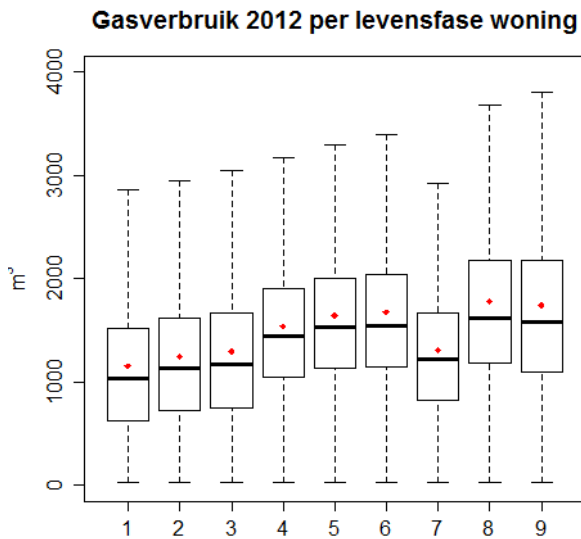


Figuur 16: De spreiding van het verbruik en het gemiddelde (rode stip) neemt af naarmate een recenter bouwjaar heeft.

INKOMEN/OPLEIDING EN INHOUD/OPPERVLAK WONING

Naarmate het inkomen en opleidingsniveau stijgt en de inhoud/oppervlakte van de woning groter is, neemt ook het gasverbruik toe. Ditzelfde is ook te zien bij het elektra verbruik.

LEVENSFASE



Legenda levensfases

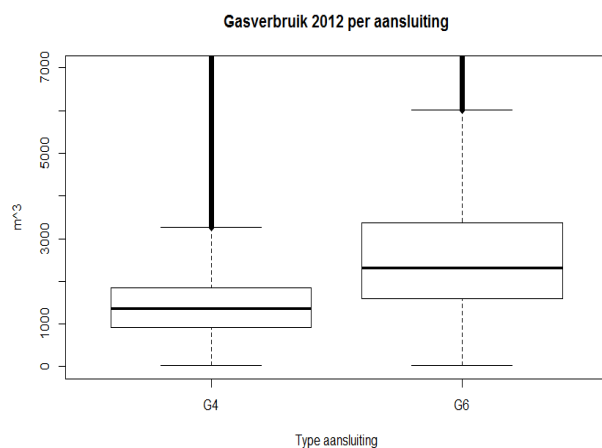
- 1 Jonge alleenstaanden
- 2 Middelbare alleenstaanden
- 3 Oudere alleenstaanden
- 4 Gezin met alleen jonge kinderen
- 5 Gezin met jonge en oude kinderen
- 6 Gezin met alleen oude kinderen
- 7 Jonge paren zonder kinderen
- 8 Middelbare paren zonder kinderen
- 9 Oudere paren zonder kinderen

Figuur 17: Het gasverbruik neemt toe naarmate er meer bewoners zijn. Ook verbruiken oudere mensen meer dan jongere.

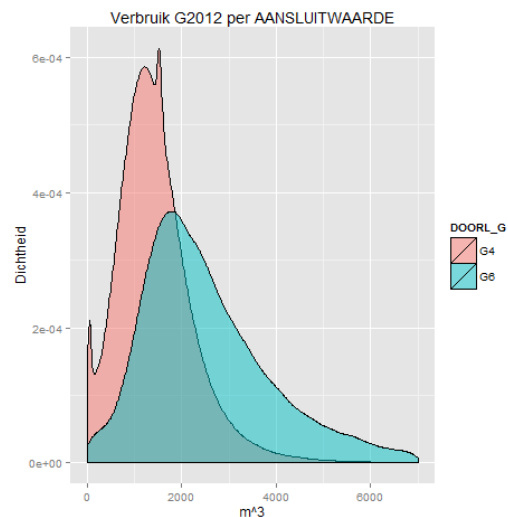
In Figuur 17 zien we het gasverbruik per levensfase. Bij gezinnen met kinderen (categorie 4-6) neemt het gasverbruik toe naarmate de kinderen ouder zijn. Jonge paren zonder kinderen verbruiken (categorie 7) veel minder dan middelbare of oudere paren (type 8 en 9). Bij alleenstaanden (categorie 1-3) geldt ook dat jongere minder verbruiken dan oudere alleenstaanden.

AANSLUITWAARDE

Zoals te verwachten is, neemt het gemiddelde verbruik toe naarmate de aansluitwaarde van de gasaansluiting groter is. De spreiding is bij G6 aansluitingen groter dan bij G4 aansluitingen zoals Figuur 18 en Figuur 19 laten zien. In Figuur 19 zijn dezelfde pieken te zien als eerder opgemerkt.



Figuur 18: Het gasverbruik per type aansluiting, waarbij huishoudens met een grotere meter (G6) ook een hoger verbruik hebben en een grotere spreiding.



Figuur 19: De huishoudens met een G6 meter hebben een meer gespreid verbruik; het verbruik van G4-huishoudens ligt voornamelijk rond de 1400 m³.

Regressie analyse gas

Met behulp van de regressie analyse is gekeken naar welke factoren het gasverbruik kunnen voorspellen. Hieruit blijkt dat Type woning (0,256), Oppervlakte (0,047), Bouwjaar (0,057), Inkomen (0,006) en Levensfase (0,001) in afnemende mate voorspellende factoren zijn voor het gasverbruik. De verhoging in R^2 van het model staat tussen haakjes en ook hier zijn alle variabelen significant. De coëfficiënten voor de vier factoren¹² staan in onderstaande Tabel 3 en het gehele model heeft een R^2 van 0,365. Dit betekent dat deze vier variabelen ongeveer 37% van de variatie in het gasverbruik verklaren. Worden de overige drie variabelen toegevoegd, dan wordt de R^2 0,367.

De niet-gestandaardiseerde B laat zien dat bijvoorbeeld het verbruik van een flat/appartement ongeveer 249 m³ lager ligt dan de referentiegroep 'rijtjeswoning' en een vrijstaande woning/bungalow 577 m³ meer. Bij de bouwjaaren is de referentiegroep de huizen die na 2005 zijn gebouwd.

Tabel 3: De factoren en hun bèta's van het lineaire regressie model dat het elektraverbruik voorspeld. De verandering in R^2 geeft de verbetering van het model aan door het toevoegen van die variabele.

Gas	Niet-gestandaardiseerde Coëfficiënten		Gestandaardiseerde coëfficiënten	Verandering in R^2
	B	Std. Error	Bèta	
(Constante)	106,533	3,833		-
Vrijstaand_Bungalow	576,825	2,164	,200	0,256
Twee_onder_een_kap	209,362	2,006	,074	
Flat_Appartement	-249,257	1,867	-,104	
Kamer_studentenhuis	171,757	15,063	,007	
Etagewoning	-110,699	2,567	-,031	
Herenhuis_grachtenpand	65,069	6,819	,006	
Zelfst_bejaardenwoning	64,577	4,431	,010	
Boerderij_Tuinderij	650,173	6,648	,066	
Ander_woningtype	19,001	,553	,023	
Oppervlakte woning	185,987	,543	,317	
Voor1800	780,941	12,472	,042	0,057
Tussen1800en1899	710,839	5,405	,100	
Tussen1900en1919	706,332	3,807	,172	
Tussen1920en1939	789,905	3,285	,272	
Tussen1940en1959	750,159	3,360	,244	
Tussen1960en1969	578,797	3,150	,220	
Tussen1970en1979	525,653	3,043	,219	
Tussen1980en1989	353,653	3,071	,145	
Tussen1990en1999	215,268	3,129	,080	
Tussen2000en2004	18,607	3,936	,004	
Inkomen	58,386	,511	,090	0,006

4.3.3 Resultaten en analyse onderlinge relatie factoren

Er zijn veel verschillende dingen op te merken uit de resultaten van de onderlinge factoren. Hieronder staan de algemene of meest opvallende conclusies, waarbij gekozen is diegene weer te geven waarbij uit de bovenstaande analyses bleek dat deze een relatie hadden met het energieverbruik. Ook wordt soms de relatie tot huur en koopwoningen benoemd, omdat dit –bij het

¹² Het toevoegen van de variabele 'Levensfase' aan het model, zorgt voor een verhoging in R^2 van slechts 0,001.

gebruik van het EBP- invloed heeft op de mogelijkheden die mensen hebben om een besparing te realiseren. Zo zijn bij huurwoningen vaak minder technische aanpassingen mogelijk op individueel niveau, tenzij de woningcorporatie/eigenaar deze kosten op zich neemt. De overige inzichten staan in bijlage C onder 'Resultaten onderlinge factoren'.

Inkomen

- De hoogte van het inkomen en het soort levensfase, of het een koop/huur woning is en wat het oppervlak is, is redelijk onafhankelijk van de verschillende bouwjaren. Hogeropgeleiden wonen voornamelijk in oude huizen van voor 1920 of in nieuwbouw. Lager opgeleiden wonen voornamelijk in huizen uit de tussenperiode. Deze huizen zijn vooral woningen uit 1960-1989 en worden in 60% van de gevallen bewoond door huurders met minimum inkomen.
- Mensen met een hoger inkomen hebben vaak een hogere opleiding genoten, vaker een koopwoning en een groter oppervlak van hun huis.
- De verschillende inkomensniveaus zijn redelijk gelijk verspreid over de verschillende bouwjaren, al wonen mensen met een laag inkomen vooral in woningen van middelbare leeftijd.
- Eerder bleek al dat de woningen met een klein oppervlak eenzelfde spreiding in verbruik hadden. Van de bewoners heeft 80-90% een laag inkomen en wonen zij in 70-90% van de gevallen in een huurhuis.

Bouwjaar

- Mensen met een hoger inkomen wonen vaker in een oudere woning, of een woning uit 1990-1999.
- In woningen uit 1980-1989 leven voornamelijk mensen met een minimum inkomen; ongeveer 60% van deze klasse woont in huizen uit 1960-1989.
- Oppervlakte van woning is groter bij huizen van voor 1920.

Oppervlak woning

- Naarmate het oppervlak van de woning toeneemt, neemt ook het percentage paren toe (van 5% bij een klein oppervlak naar 40% bij een groter oppervlak). Ook het percentage gezinnen neemt toe naarmate het oppervlak toeneemt (tot 150m³).

Type woning

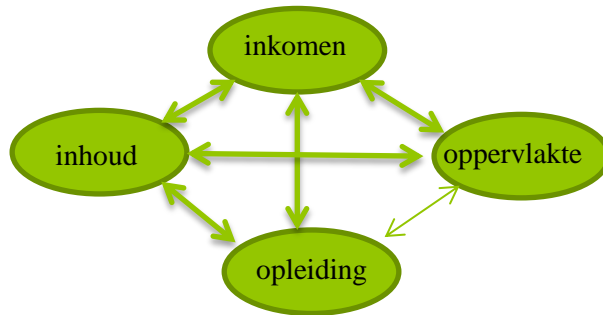
- Behalve de relatie met het oppervlak en sommige categorieën van het inkomen, heeft het type woning weinig andere duidelijke relaties met factoren.

Levensfase

- 50% van de mensen met een minimum inkomen is alleenstaand, dit neemt af naar 15% bij een stijging van het inkomen.
- Van middelbare paren en gezinnen met jonge kinderen woont plusminus 20% in vrijstaand/twee-onder-een-kap/rijtjeshuis. In een zelfstandige bejaarden woning wonen logischerwijs veel oudere alleenstaanden; in kamer/etage/herenwoning is plusminus 50% alleenstaand.

Resultaten lineariteit van de relatie tussen variabelen

Uit de scatterplots op basis van het gemiddelde per categorie blijkt dat de factoren die in Figuur 20 staan lineair stijgend gecorreleerd zijn, waarbij een dikke pijl een sterke relatie en een dunne pijl een minder sterke relatie aangeeft. De scatterplots staan in bijlage C. Vanwege het bestaan van deze relaties, moet bij verdere (cluster)analyses opgelet worden dat deze factoren niet samen worden gebruikt of moet er bewust mee worden om gegaan. Bij de overige factoren is geen duidelijke relatie te zien.



Figuur 20: De factoren die informatie geven over de hoogte van het verbruik die onderling een sterke (dikke lijn) of minder sterke (dunne lijn) lineaire relatie laten zien.

4.4 Conclusie

Uit de resultaten en analyse komt naar voren dat zowel het elektra als het gasverbruik een zelfde spreiding over de afgelopen 5 jaar laat zien. Het gasverbruik daalt ieder jaar en het elektraverbruik nam eerst toe door een hoger gebruik van apparaten, maar daalt nu licht.

Het *elektraverbruik* wordt het best voorspeld door (1) de oppervlakte, (2) het type woning, (3) de levensfase en (4) het inkomen. Dit uit zich als volgt:

1. Huishoudens met een grotere aansluiting, hebben een hoger elektraverbruik.
2. Bij het type woning speelt vooral het oppervlak een rol: grotere huizen verbruiken meer elektra dan huizen met een kleiner oppervlak.
3. Naarmate een gezin groter is, is het verbruik hoger. Ook de leeftijd speelt een rol: alleenstaanden, paren van middelbare leeftijd verbruiken meer binnen dan jongere of oudere alleenstaanden of paren. En ook gezinnen met zowel jonge als oude kinderen verbruiken meer dan gezinnen met of alleen jonge of alleen oude kinderen.
4. Het verbruik van mensen met een hoger inkomen ligt over het algemeen ook hoger.

Het *gasverbruik* wordt het best voorspeld door (1) het type woning, (2) de oppervlakte, (3) het bouwjaar en (4) het inkomen. Binnen elk van deze factoren is het volgende op te merken:

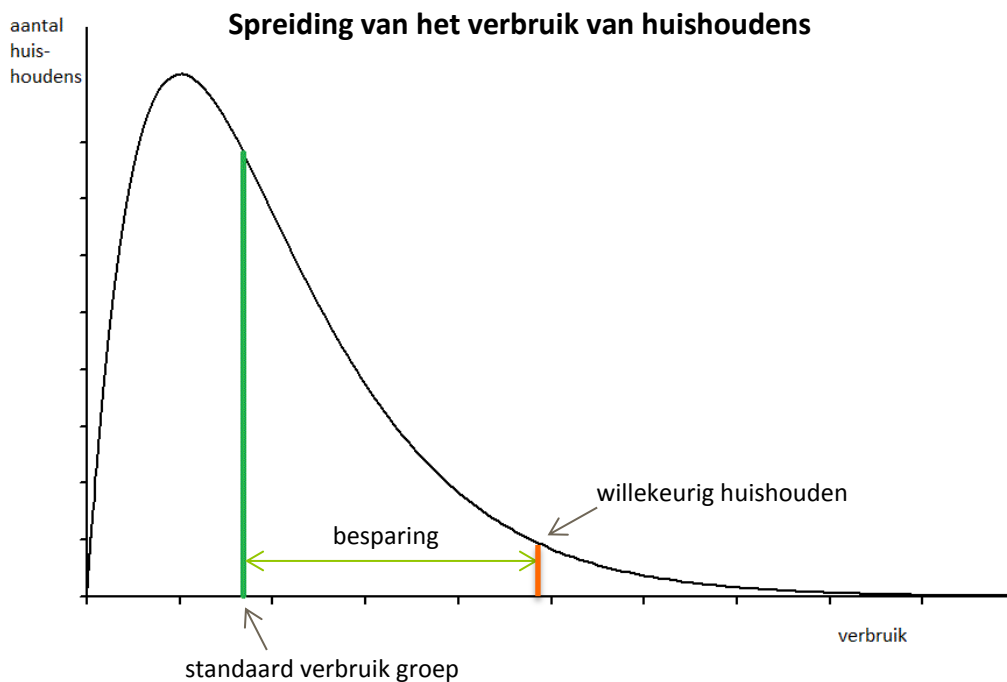
1. Huizen met een kleiner oppervlak (flat, etagewoning, zelfstandige bejaardenwoning) hebben een onderling een gelijke spreiding. Ten opzichte van de andere type huizen hebben zij een kleinere spreiding en een lagere mediaan.
2. Oudere huizen hebben een hoger verbruik dan nieuwere huizen.
3. Het gasverbruik van woningen met een G4 aansluiting ligt lager dan met een G6 aansluiting.
4. Hoe groter de oppervlakte van een huis, hoe meer het huishouden verbruikt.

Tussen deze verschillende factoren komen een aantal relaties naar voren, waarbij tussen de vier - factoren inkomen, inhoud, oppervlakte en opleiding sterke lineair stijgende relaties bestaan met uitzondering van oppervlakte en opleiding, waartussen deze relatie zwak is. Verder valt op dat woningen met een klein oppervlak in 70-90% van de gevallen huurwoningen zijn en dat 80-90% van de inwoners een laag inkomen heeft. Ook neemt bij een toenemend oppervlakte van een woning het percentage paren en gezinnen toe. Tot slot hebben mensen met een hoger inkomen vaak een hogere opleiding genoten en hebben zij vaker een koopwoning en een groter oppervlak van hun huis.

5. Beschrijving veranderde insteek onderzoek

Dit hoofdstuk beschrijft de inzichten die geleid hebben tot een andere insteek van dit onderzoek.

Zoals in paragraaf 3.2 al beschreven is was in eerste instantie het doel van dit onderzoek om per huishouden een (percentage) EBP aan te kunnen geven. Hiervoor is op zoek gegaan naar (1) een onderbouwing voor een reële besparing per huishouden en (2) een maat voor een standaard energieverbruik per groep huishoudens waarmee het huidige verbruik vergeleken kan worden. De insteek bij (1) was om per huishouden op basis van haar kenmerken te schatten hoeveel ze zou kunnen besparen. Bij (2) was de insteek om per groep vergelijkbare huishoudens een benchmark vast te stellen (een standaard verbruik) op basis waarvan het EBP van elk van de huishoudens binnen die groep gemakkelijk bepaald kon worden. Deze insteek is in onderstaande Figuur 21 weergegeven. De besparing die het willekeurige huishouden kan behalen is het verschil van zijn verbruik ten opzichte van het standaard verbruik.



Figuur 21: Illustratie eerste insteek onderzoek: op basis van een standaard verbruik per groep kan per huishouden het EBP (de besparing) bepaald worden.

Om inzicht te krijgen in een reële maat voor besparing of een standaardverbruik, hebben we verschillende gesprekken met Alliander medewerkers gevoerd en adviesbureaus¹³ op het gebied van energiebesparing benaderd. Hieruit bleek dat bovenstaande manieren niet haalbaar zijn. Door de inzichten uit dit veldonderzoek is besloten om de probleemstelling te wijzigen naar het bepalen van *groepen* die mogelijk EBP hebben in plaats van het bepalen van een EBP op huishoudniveau. We geven hieronder kort de inzichten weer.

¹³ Vabi en ISSO

5.1 Onderbouwing reële besparing huishoudens alleen op detailniveau mogelijk

In paragraaf 3.3 is al genoemd dat de te behalen besparing afhankelijk van de maatregelen die worden doorgevoerd en van gedragsverandering. Om een onderbouwing per huishouden te geven voor een reële besparing is het nodig dit voor beide aspecten te doen. Op basis van huishoudkenmerken als bouwjaar en type woning is een inschatting te maken van de besparing die behaald kan worden door maatregelen. Besparing door gedragsverandering is op deze manier lastiger in te schatten. Eenzelfde verandering in gedrag kan verschillende invloeden hebben: de verwarming een graad lager zetten in een kleinere woning bijvoorbeeld heeft in absolute zin minder effect dan in een grotere woning. Om het EBP te schatten dat behaald kan worden door gedrag, is gedetailleerde informatie over het gedrag van een huishouden nodig, wat voor dit onderzoek niet beschikbaar is. Daarnaast is de data die vanuit Bisnode beschikbaar is slechts een benadering van de kenmerken van een huishouden, waardoor een model op huishoudniveau extra onzekerheid bevat over de basiskenmerken.

5.2 Standaardverbruik huishoudens niet specifiek genoeg of niet reëel

Als bij een woning geen historische verbruik bekend is, wordt binnen Liander op basis van type meter het verbruik ingeschat. Een maat voor het standaardverbruik wordt gegeven door het energielabel van een woning.

Standaardverbruik afhankelijk van type meter

Bij het inschatten van het verbruik op basis van de meter, worden de overige kenmerken van een huishouden niet meegenomen. Het standaardverbruik voor de elektrameters gebruikt in dit onderzoek (tot en met 3x 25A) is 3000 kWh en voor de G4 en G6 gasmeters is dit 2000m³ (afd. Uitvoering). Dit is niet specifiek genoeg voor dit onderzoek.

Standaardverbruik afhankelijk van energielabel

Op het energielabel van een woning wordt ook het standaardverbruik voor gas en elektra vermeld, zoals beschreven in paragraaf 1.3.3. Majcan, Itard en Visscher (2013) hebben aangetoond dat deze verbruiken niet reëel zijn voor het huishouden. Elektra wordt standaard te laag geschat omdat dit factoren als het aantal bewoners en het gebruik van de woning (gedrag) buiten beschouwing laat. Het gasverbruik wordt in de hoge labels (A⁺⁺, A⁺, A en B) te laag ingeschat en in de lagere labels te hoog. Het label heeft wel een voorspellende waarde, maar zegt niets over het werkelijke standaardverbruik. Naar aanleiding hiervan hebben we onderzocht wat de berekeningsmethodiek van het energielabel is: deze is te vinden in ISSO-publicatie 82.3. Voor dit onderzoek was het niet zinvol om een diepte studie te doen naar de bepaling van een standaardverbruik van een huishouden, omdat we hier voornamelijk met benaderde data werken en geen informatie op detailniveau van een woning hebben. Wanneer detailinformatie wel beschikbaar is, is het mogelijk wel interessant om te onderzoeken wat binnen deze methodiek de weging is van de verschillende factoren die invloed hebben op het verbruik bij het bepalen van het standaardverbruik van de woning.

Kortom, voor aanpak (1) is een onderbouwing van een reële besparing van huishoudens alleen op detailniveau mogelijk, en voor aanpak (2) is een standaardverbruik beschikbaar dat of niet specifiek genoeg is, of niet reëel. Om deze reden is in overleg met de begeleiders van dit onderzoek besloten het onderzoek aan te passen naar het identificeren of een huishouden wel of geen hoog EBP heeft in plaats van een EBP op huishoudniveau te schatten.

6. Methodiek bepaling energiebesparingspotentieel

In dit hoofdstuk wordt een antwoord gegeven op de probleemstelling: ‘Hoe kan het EBP van huishoudens worden geschat?’ Vanuit de definitie van het EBP zijn we op zoek naar huishoudens met een hoog verbruik ten opzichte van een groep vergelijkbare huishoudens. Drie vragen staan hierbij centraal:

- (1) In hoeveel groepen moeten we de data verdelen?
- (2) Bevatten deze groepen huishoudens met gelijke kenmerken?
- (3) Zijn deze groepen reproduceerbaar?

De methodiek om deze huishoudens te onderscheiden bestaat uit drie stappen die op deze vragen een antwoord geven. In stap één verdelen we met behulp van het Two-Step algoritme de dataset in groepen op basis van een aantal factoren. We variëren hierbij het aantal groepen (paragraaf 6.1.1) die we onderling vergelijken met de Silhouette Coëfficiënt (SC) om te bepalen wat het beste aantal groepen is om de data in te verdelen. De SC en het Two-Step algoritme om de groepen (clusters) te maken wordt in paragraaf 6.2 toegelicht. In stap twee valideren we vervolgens met een homogeniteitstest of de huishoudens met een hoog verbruik binnen elk cluster gelijke kenmerken hebben ten opzichte van de andere huishoudens binnen het cluster. Deze test wordt in paragraaf 6.1.2 toegelicht. We onderbouwen hier eveneens de keuze voor de grens tussen een normaal en hoog verbruik.

Blijkt uit stap twee dat de clusters homogeen zijn, dan valideren we in stap drie of deze clusters reproduceerbaar zijn. De centrale vraag hierbij is of dezelfde huishoudens die eerder gevonden zijn, aangemerkt worden als de groep met EBP (paragraaf 6.1.3). Onafhankelijk van welk algoritme gekozen wordt om groepen te maken in stap één, is het van belang de controles in stap twee en drie uit te voeren. Tot slot onderzoeken we wat de toegevoegde waarde is om met behulp van een clustering de groep huishoudens met EBP binnen Liander gebied te onderscheiden (paragraaf 6.4).

6.1 Methodiek

Voordat we stap één van de methodiek beschrijven, onderbouwen we eerst de keuze voor het gebruik van het Two-Step algoritme om de dataset te clusteren.

In de literatuur maakt men onderscheid tussen twee manieren waarop een model ontwikkeld kan worden: supervised en unsupervised. Supervised betekent dat er een uitkomstwaarde bekend is en dat het model op basis van een aantal parameters zo goed mogelijk probeert deze waarde te benaderen. Bij een unsupervised manier is dit niet het geval en probeert het model een onderliggende structuur in de data te ontdekken en natuurlijke groepen (clusters) te vormen (Duda, Hart, & and Stork, 2001).

In het verleden is nog niet eerder bepaald welke groepen EBP hebben, waardoor er geen historische data is om uitkomsten van een model op te toetsen. Om deze reden is het soort modellen waaruit we kunnen kiezen beperkt tot unsupervised modellen.

Voorbeeld: Een supervised model is de eerder gebruikte lineaire regressie om het verbruik per huishouden te schatten op basis van een aantal factoren, waarbij het verbruik per huishouden bekend is. Een unsupervised model verdeelt bijvoorbeeld een groep huishoudens in segmenten op basis van hun houding ten opzichte van energiebesparing, waarbij niet van tevoren bekend is tot welke groep een huishouden behoort.

Naast de keuze voor een supervised of unsupervised model spelen ook andere factoren een rol, zoals het *type data*. In onze dataset zijn de acht eigenschappen van huishoudens (inkomen, opleiding, oppervlakte/inhoud woning, etc.) nominale en ordinale variabelen. Dit beperkt het aantal mogelijke algoritmes sterk, omdat de meeste algoritmes uitgaan van continue variabelen.

Een ander punt ter overweging is de *grootte van de dataset*, zowel qua aantallen huishoudens als qua waarden die de categorische variabelen kunnen aannemen. Als we bijvoorbeeld groeperen op basis van de verschillende combinaties tussen de variabelen levert de combinatie Type woning, Bouwjaar en Levensfase bijvoorbeeld al maximaal 990 groepen op¹⁴. Alle combinaties tussen de categorieën maken is dan ook niet werkbaar. Met behulp van een beslissingsboom, die met behulp van hiërarchische clustering gevormd wordt, wordt dit aantal groepen kleiner. Met een dergelijke boom kan gemakkelijk per huishouden bepaald worden tot welke groep hij behoort. Echter, elk van de acht variabelen heeft meerdere categorieën en de dataset is groot, waardoor het algoritme er lang over doet om een boom te genereren¹⁵. Dit maakt het in de praktijk minder geschikt wanneer regelmatig bepaald moet worden wat een groep is met EBP omdat het met de huidige technologie nog steeds veel tijd in beslag neemt.

We hebben daarom voor de Two-Step clustering methode van IBM SPSS Statistics gekozen omdat zij met categorische data kan werken en goed met een grote dataset overweg kan. Two-Step clustering vormt in de eerste stap een heleboel kleine clusters van huishoudens die op elkaar lijken, waarna ze in de tweede stap hiërarchisch clusteren toepast om dit aantal te reduceren. Dit gaat sneller dan normaal hiërarchisch clusteren omdat de grootte van de dataset is gereduceerd tot een beperkt aantal subclusters. Een uitgebreidere beschrijving van het Two-Step algoritme wordt gegeven in paragraaf 6.2.

Met dit algoritme hebben we de data voor verschillende sets variabelen en verschillende parameters geclusterd, waarvan de resultaten in hoofdstuk 7 te vinden zijn.

6.1.1 Stap 1: Bepaal aantal clusters en algehele clusterkwaliteit

In de eerste stap van onze methodiek verdelen we de data in een aantal cluster. De centrale vraag hierbij is: 'Wat is het beste aantal?'

Er zijn grofweg twee manieren om te evalueren hoe goed een clustering is: een interne en externe evaluatie. Een historisch EBP is niet bekend, waardoor deze vergelijking niet mogelijk is. Daarom evalueren we de clustering met een interne methode: op basis van de gebruikte data.

¹⁴ Type woning, bouwjaar en levensfase hebben respectievelijk 10, 11 en 9 categorieën, wat $10 \times 11 \times 9$ combinaties oplevert. Het zijn maximaal 990 groepen, omdat enkele combinaties niet voorkomen.

¹⁵ Ze heeft namelijk een hoge complexiteit: $\mathcal{O}(n^3)$ voor een *agglomerative* aanpak waarbij elk huishouden een apart cluster voorstelt en herhaaldelijk steeds twee clusters die veel op elkaar lijken worden samengevoegd; en $\mathcal{O}(2^n)$ voor een *divisive* aanpak waarbij gestart wordt met de gehele dataset en deze herhaaldelijk in kleinere clusters wordt gesplitst.

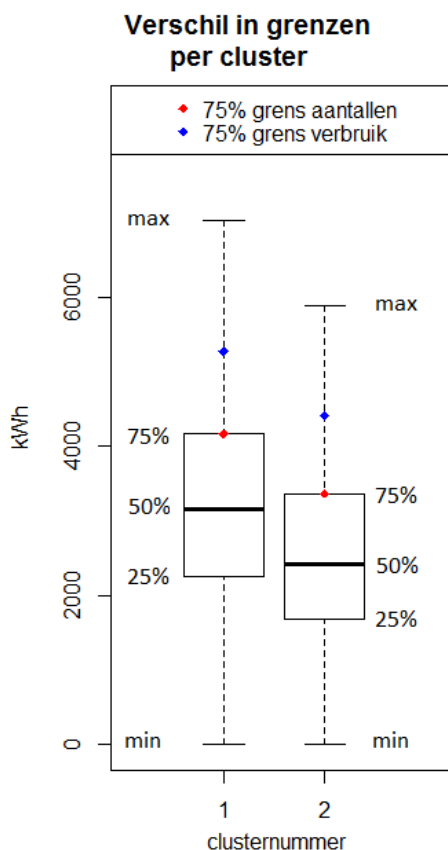
We kijken hierbij naar twee aspecten: 1) cluster cohesion en 2) cluster separation. Cluster cohesion geeft weer hoe goed de samenhang van de data binnen een cluster is en cluster separation laat juist zien hoe onderscheidend het cluster is van andere clusters. Beide zijn een maat voor hoe goed de groepen binnen de data worden gevormd en kunnen elk apart gebruikt worden om te controleren hoe goed een clustering is. Een evaluatiemetriek die beide aspecten meeneemt is de SC. Om deze reden gebruiken we de SC om de clustering te evalueren en te bepalen wat het beste aantal clusters is. Een andere reden is dat de SC beschikbaar is binnen SPSS Statistics.

De SC neemt een waarde aan tussen de -1 en 1, waarbij -1 een zeer slecht model weergeeft en 1 een perfect model. Volgens Kaufman & Rousseeuw (1990) duidt een SC groter dan 0.5 op redelijke verdeling van de data in groepen; is deze kleiner dan 0.2 dan laat de data geen cluster structuur zien. In paragraaf 5.2 is de SC in detail toegelicht.

Met behulp van de SC krijgen we een idee van wat een goed aantal clusters is. Uit de praktijk blijkt echter dat binnen een clustering met een hoge SC met een laag aantal clusters ook clusters kunnen voorkomen die geen gelijke kenmerken bevatten. Om deze reden passen we stap twee van de methodiek toe, waarbij we de interne kwaliteit van de gevonden clusters testen.

6.1.2 Stap 2: Test homogeniteit cluster

Om te bepalen of de huishoudens met een hoog verbruik binnen een cluster gelijke kenmerken hebben ten opzichte van de rest van de cluster, moet eerst een grens tussen normaal en hoog verbruik worden vastgesteld. Vervolgens kunnen we testen of deze groepen overeenkomen. We nemen aan dat elk huishouden EBP heeft, maar binnen dit onderzoek zijn we op zoek naar de huishoudens met een hoog EBP.



Grens tussen normaal en hoog verbruik

Binnen elk cluster stellen we de grens tussen normaal en hoog verbruik op dezelfde manier vast, te weten met een gelijk percentage. Dit percentage kan op twee manieren bepaald worden: (1) op basis van het aantal huishoudens en (2) op basis van het verbruik. Beide zijn in Figuur 22 respectievelijk met rood en blauw aangegeven. We kiezen voor een grens op basis van aantallen om de volgende reden.

Stel we sorteren de huishoudens van laag naar hoog op basis van hun verbruik. Vervolgens delen we hen op in vier gelijke groepen van elk 25%. We zien dan dat de hoogste 25% van de huishoudens (vanaf het rode punt naar boven) bijna een zelfde spreiding heeft als de overige 75%. Kijken we vervolgens naar een grens op basis van het verbruik (de blauwe punten) dan zien we dat deze grens sterk beïnvloed wordt door de 25% van de huishoudens met het hoogste verbruik. Als we een grens op basis van de spreiding van het verbruik kiezen dan beïnvloedt juist de groep die we willen onderscheiden deze grens. Hoe hoger het verbruik van deze huishoudens is, hoe kleiner de groep met een hoog EBP wordt. Omdat dit niet gewenst is,

Figuur 22: Een grens tussen massa en potentiëlen op basis van verbruik binnen het cluster, zorgt voor een kleinere groep dan wanneer een grens op basis van aantallen bepaald wordt.

kiezen we voor een grens op basis van aantallen.

De groep die onder de grens ligt noemen we de *massa*; de huishoudens die daarboven liggen noemen we de *potentiëlen*.

Vanwege de grote spreiding van het verbruik van de hoogste 25% stellen we de grens vast op 25% van de huishoudens met het hoogste verbruik binnen het cluster. Deze grens leggen we niet lager op bijvoorbeeld de mediaan (50% van de huishoudens) of het gemiddelde verbruik, omdat een deel van de potentiëlen dan slechts een geringe besparing heeft. Uit de praktijk blijkt dat mensen bij een geringe besparing minder bereid om maatregelen te treffen of hun gedrag aan te passen om deze te behalen. We kiezen daarom voor een grens waarbij aannemelijker is dat de huishoudens daarboven hoog EBP hebben. In vervolgonderzoek kan gekeken worden naar een eventuele cluster-specifieke grens, waarbij kenmerken van het cluster de grens bepalen.

Uit praktijkonderzoek moet blijken of de potentiëlen inderdaad energie kunnen besparen. Echter, als er een goede reden is waarom de potentiëlen een hoog verbruik hebben, bijvoorbeeld door een groter woonoppervlak, dan kan op voorhand al gezegd worden dat het cluster niet uit vergelijkbare huishoudens bestaat. Dit bleek uit analyses die in 7.1 'Inzicht 2' worden toegelicht en daarom onderzoeken we met behulp van onderstaande homogeniteitstest of het verschil in verbruik tussen de massa en de potentiëlen te verklaren is.

Homogeniteit clustering

De meest voor de hand liggende test om een verschil in verhoudingen van een categorische variabele tussen twee sets te bepalen is de Chi-Kwadraat test. Echter, deze test neemt aan dat elke waarde binnen categorie verwacht wordt minstens 5 keer voor te komen (Agresti & Franklin, 2007). Wanneer de massa en de potentiëlen zodanig van elkaar verschillen dat van een variabele in een van beide groepen een categorie ontbreekt, dan kan deze test niet uitgevoerd worden. Om deze reden gebruiken we een andere manier: de Wilcoxon rank test¹⁶.

We testen hiermee of het gemiddelde rangnummer van twee steekproeven overeen komt. Is dit het geval, dan impliceert dit gelijke verhoudingen tussen de twee steekproeven (Agresti & Franklin, 2007). De nul- en alternatieve hypothese zijn als volgt:

H_0 : Identieke steekproef verdeling tussen de massa en potentiëlen

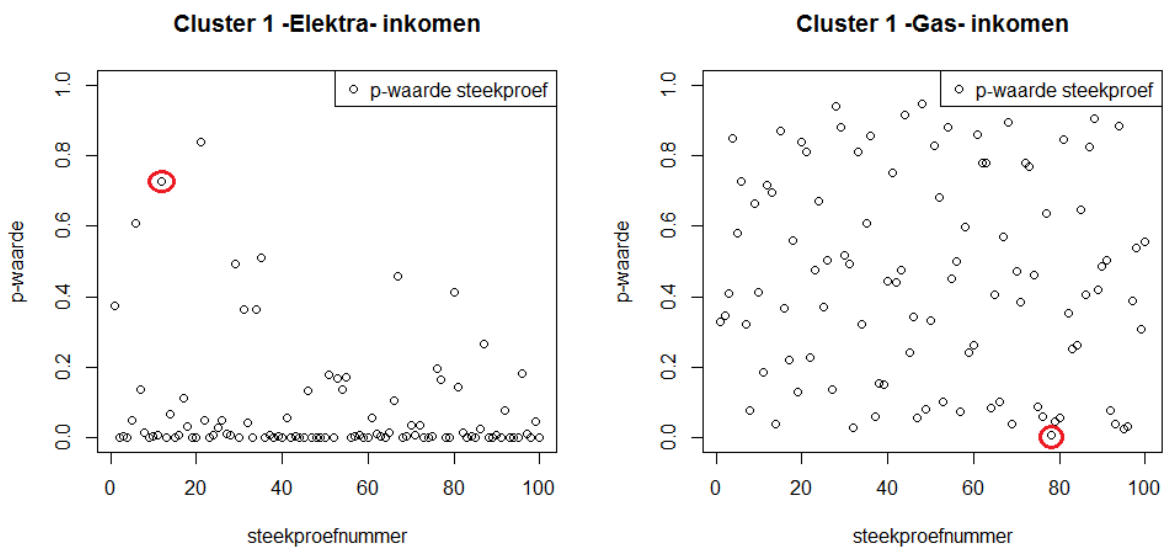
H_1 : Verschillende steekproef verdeling tussen de massa en potentiëlen

We hebben deze test als volgt uitgevoerd: we nemen 100 keer een steekproef van 1000 waarnemingen uit zowel de massa als de potentiëlen zonder terugleggen en voeren de Wilcoxon test uit. Dit levert 100 p-waardes op. We verwerpen de nulhypothese als meer dan 50 keer $p \leq 0.05$; oftewel er zit een verschil tussen de massa en de potentiëlen. We voeren deze test zowel voor gas en elektra per categorie uit.

We gebruiken 100 steekproeven omdat de massa en potentiëlen niet uit gelijke groepen bestaan en de toevalsfactor bij de steekproef bepalend kan zijn of een test verworpen wordt. Bijvoorbeeld, in de linker figuur kan de nulhypothese onterecht verworpen worden door een hoge p-waarde en in de

¹⁶ Ook wel Mann-Whitney test genoemd

rechter figuur kan deze onterecht niet verworpen worden. Om deze gevallen (zoals rood omcirkeld in Figuur 23) uit te sluiten gebruiken we 100 steekproeven.



Figuur 23: Ter voorbeeld: de spreiding van de p-waarden van 100 keer Wilcoxon rank test voor de variabele ‘inkomen’ voor elektra (links) en gas (rechts). Als er slechts één steekproef genomen zou worden, dan kan in het roodomcirkelde geval een andere conclusie getrokken worden dan het algemene beeld van de steekproef weergeeft .

We beschouwen een cluster als ‘goed’ wanneer minder dan 5 categorieën worden verworpen of aan de volgende voorwaarden wordt voldaan:

- elektra: oppervlakte en levensfase worden niet beide verworpen
- gas: oppervlakte of inhoud, en levensfase worden niet beide verworpen

We hebben gekozen voor deze extra voorwaarden, omdat dit twee factoren zijn die veel invloed hebben op het verbruik: meer personen in een huishouden en een groter oppervlak/inhoud.¹⁷ Deze methode om de kwaliteit van het clusters te evalueren, noemen we de *homogeniteitstest*.

Deze twee stappen, het bepalen van een aantal clusters met de SC waarna de homogeniteit getest wordt, vormen de methodiek om mogelijk EBP onder huishoudens op te sporen.

6.1.3 Stap 3: Consistentie uitkomst clusters

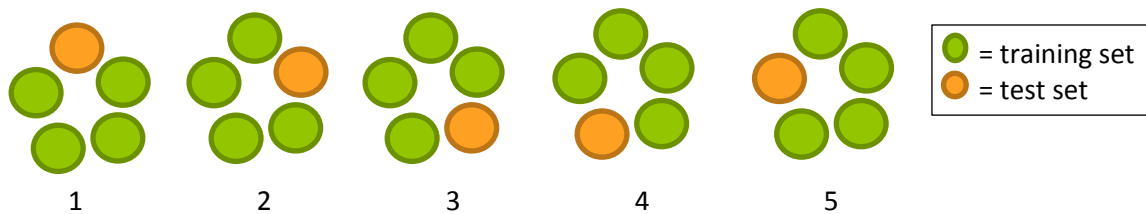
Als uit de vorige test blijkt dat de clustering voldoende¹⁸ homogene clusters oplevert, moet ook gecontroleerd worden of dit inderdaad de natuurlijke groepen binnen de data vormen en dat zij reproduceerbaar zijn, of dat ze op toeval berusten.

Een methode die veel gebruikt wordt om een uitkomst te valideren is *k-fold-cross-validation*. Bij deze methode wordt de dataset willekeurig in *k* delen gedeeld en worden *k* paren van test- en trainingsets gevormd waarbij één van de *k* delen als testset wordt genomen en de rest als training wordt

¹⁷ Het bouwjaar geeft een indicatie over de mate van isolatie (en dus het verbruik), maar op basis hiervan kan geen goed onderscheid gemaakt worden tussen groepen, vandaar dat we deze niet als extra voorwaarde hebben toegevoegd.

¹⁸ Dit is afhankelijk van de keuze van de gebruiker.

gebruikt. De testresultaten worden dan gemiddeld over de k testsets. In Figuur 24 is dit visueel weergegeven voor $k=5$. Op basis van de groene training sets wordt steeds een model gemaakt, dat op de oranje test set getest wordt.



Figuur 24: Visualisatie 5-fold-cross-validation, waarbij steeds met 4/5 deel van de data een model gemaakt wordt dat met het overige 1/5 deel getest wordt.

Dit concept kan uitgebreid worden om de kwaliteit van clusters te evalueren. Mirkin (2005) deelt de verschillende type uitbreidingen op in twee groepen: een machine learning aanpak of een data mining aanpak. De machine learning aanpak kijkt naar de consistentie van het clusteralgoritme en de data mining aanpak evalueert de cluster resultaten. In ons geval is een data mining aanpak van toepassing. Hierbij wordt eerst een clustering gedaan op de volledige dataset voordat deze in k delen gesplitst wordt. De trainingsets worden dan op de gebruikelijke manier gemaakt en gebruikt om de resultaten van de clustering op de gehele dataset te verifiëren. Hiertoe wordt het clusteralgoritme toegepast op elk van de k trainingsets en de gevonden resultaten worden vergeleken met de resultaten uit de clustering op de gehele dataset. Wij gebruiken deze aanpak en delen de dataset op in vijf willekeurige gelijke groepen.

Om te zien hoeveel de clustering op de trainingsets verschillen van de clustering op de gehele set, kijken we naar de huishoudens die per test set als ‘potentieel’ aangemerkt worden. Een indicatie dat de clusters redelijk overeen komen, is wanneer ‘potentiëlen’ uit de originele set in de verschillende test sets ook als potentieel worden aangemerkt en idem dito voor de massa. Een aanbeveling is om in vervolgonderzoek de clusterkwaliteit ook te evalueren met behulp van een van de methodes zoals beschreven door Mirkin. Deze hebben we niet toegepast vanwege het ontbreken hiervan in SPSS.

Ongeacht het clusteralgoritme dat in stap 1 van gebruikt wordt, is het van belang de homogeniteitstest in stap 2 en bovenstaande test op de consistentie van de clustering uit te voeren.

6.1.4 Toegevoegde waarde clustering

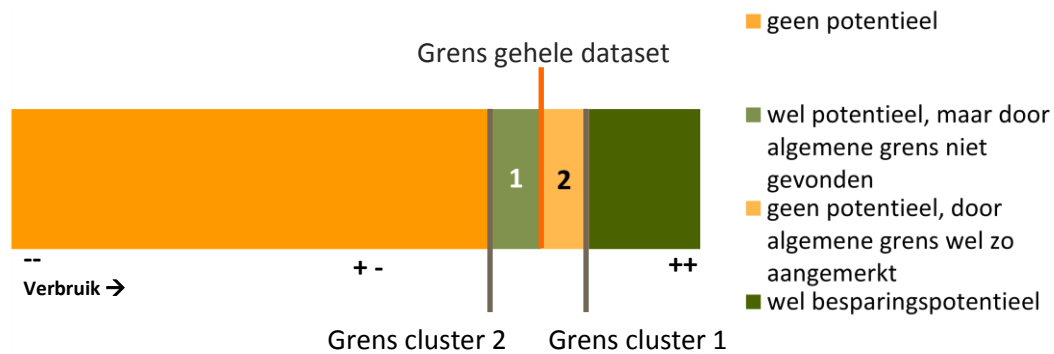
Dit onderzoek wordt pas relevant als een clustering van de dataset ook daadwerkelijk een beter inzicht in het EBP oplevert. Daarom willen we weten in hoeverre potentiëlen beter worden onderscheiden door clustervorming.

Hiertoe vergelijken we twee type huishoudens:

- a) de huishoudens die potentieel hebben ten opzichte van hun cluster en;
- b) de huishoudens die potentieel hebben ten opzichte van de hele data set.

Door deze vergelijking kunnen we vier afzonderlijke groepen identificeren die in Figuur 25 zijn gevisualiseerd. In twee groepen zijn we vooral geïnteresseerd:

- 1) de huishoudens die in vergelijking tot hun vergelijkingsgroep een hoog verbruik hebben, maar niet in vergelijking tot alle huishoudens en;
- 2) de huishoudens die een hoog verbruik hebben ten opzichte van het geheel, maar ten opzichte van vergelijkbare huishoudens een normaal verbruik.

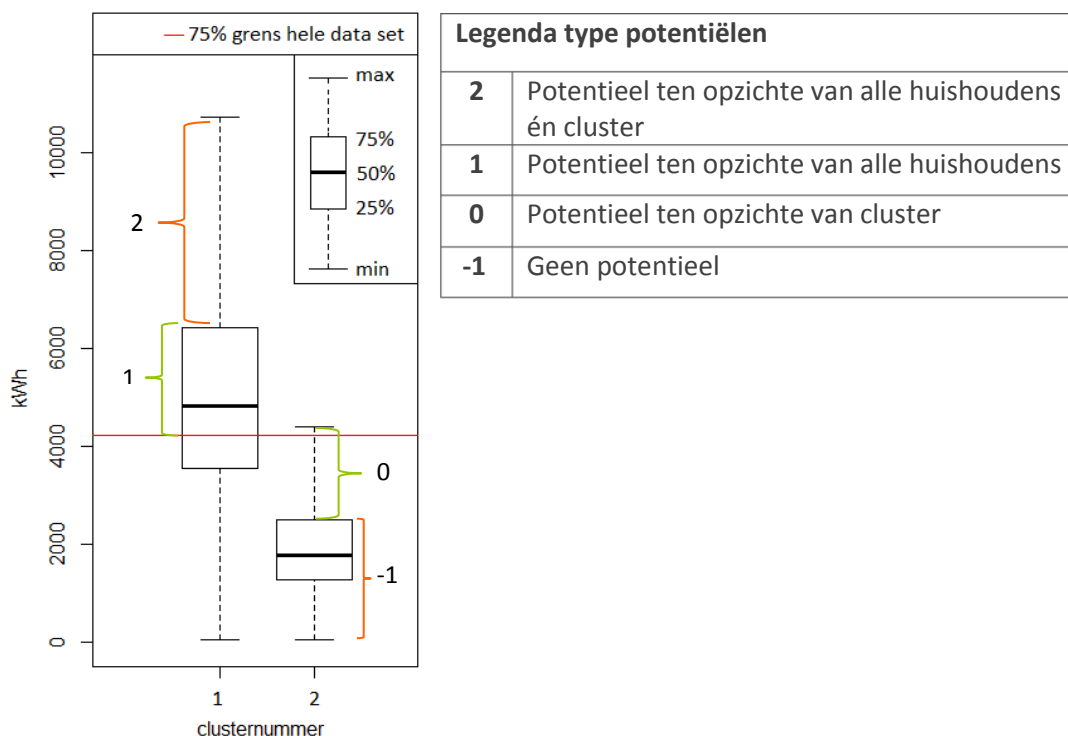


Figuur 25: Op basis van het verbruik van huishoudens kunnen verschillende grenzen gekozen worden om de groep met EBP te onderscheiden. Een grens op basis van alle data laat twee groepen buiten beschouwing die met clustervorming wel worden opgemerkt.

Deze groepen hebben we op de volgende manier onderscheiden. We hebben eerst bepaald of een huishouden tot de massa of tot de potentiële behoort, de massa-potentieel (MP) indicatie. Vervolgens hebben we deze MP indicatie zowel per cluster als voor alle huishoudens samen bepaald en op basis hiervan een nieuwe variabele gemaakt om deze te vergelijken. Hierbij zijn de cases die in een restcluster¹⁹ zijn ingedeeld buiten beschouwing gelaten. Deze variabele geeft aan in welke gevallen een huishouden als potentieel is aangemerkt; dit is in Figuur 26 grafisch weergegeven.

¹⁹ Een restcluster ontstaat wanneer ruis wordt toegelaten bij de clustering. Cases die qua kenmerken te veel afwijken van de clusters vormen het restcluster.

Indicatie van type potentiëlen op basis van clustering of gehele dataset



Figuur 26: Vergelijking tussen de huishoudens die ten opzichte van een clustering aangemerkt worden als potentieel of ten opzichte van de gehele dataset.

De percentages in de legenda geven aan tot waar de spreiding in het verbruik van x-% van de huishoudens loopt; de rode lijn geeft de MP grens op basis van de gehele dataset weer. De huishoudens in categorie **2** hebben potentieel in beide gevallen: met of zonder clustering. De categorie **-1** is de massa in beide gevallen. De type potentiëlen die beter onderscheiden worden door middel van clustering zijn de huishoudens die ten opzichte van hun groep:

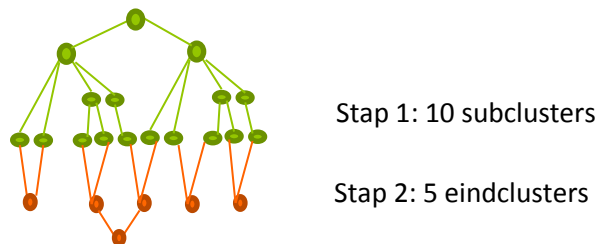
- onterecht *wel* als potentieel worden aangeduid (categorie 1) of;
- onterecht *niet* als potentieel worden aangeduid (categorie 0).

De cases in categorie 1 worden onterecht als potentieel aangeduid terwijl zij op basis van haar vergelijkingsgroep standaard een hoger verbruik heeft. De cases in categorie 0 worden onterecht niet meegenomen, terwijl zij -binnen de groep van kleine huishoudens- juist een relatief hoog verbruik heeft. De toegevoegde waarde van de clustering wordt behaald doordat de huishoudens in categorie 0 en 1 ook als EBP aangemerkt worden. In de resultaten geven we weer wat de percentages van deze categorieën zijn.

6.2 Algoritmes

6.2.1 Two-Step clustering

Het clusteren van de data met het Two-Step algoritme gebeurt in twee fases zoals in Figuur 27 zijn weergegeven: pre-clustering en clustering. In de eerste stap verdeelt een algoritme de huishoudens over vele vergelijkbare groepen, die in stap twee met een ander algoritme op basis van hun gezamenlijke kenmerken weer samengevoegd worden tot een kleiner aantal clusters. Dit wordt hieronder in meer detail beschreven.



Figuur 27: In stap 1 maakt het Two-Step cluster algoritme een groot aantal subclusters die in stap 2 samengevoegd worden tot het gewenste aantal eindclusters.

Stap 1: Voor het maken van de pre-clusters wordt een boom structuur gebruikt (groen in de figuur hierboven) die met behulp van het BIRCH algoritme (Zhang, Ramakrishnon, & Livny, 1996) wordt geconstrueerd. Het BIRCH algoritme scant de cases één voor één en voegt cases samen in een subcluster die binnen een afstandsmaat van elkaar liggen. Deze afstandsmaat wordt later beschreven. De knopen van de boom bevatten elk een 'clustering feature'²⁰ (CF) wat de belangrijkste informatie over het subcluster bevat. Het opslaan van het CF als samenvatting van het subcluster, neemt minder geheugen in dan het opslaan van alle kenmerken en daardoor kan de afstandsmaat van een nieuwe case tot het subcluster sneller bepaald worden.

Wanneer een nieuwe case aan de boom wordt toegevoegd, loopt deze door de takken van de boom om bij het dichtstbijzijnde subcluster uit te komen. Als blijkt dat zij toch verder dan de drempelwaarde van dit cluster af ligt of de maximale grootte van een subcluster bereikt is (parameter: maxbranch), wordt het cluster gesplitst. De CF van de bovenliggende knopen wordt bij elke nieuw toegevoegde case geüpdatet. Wordt door de splitsing de boom groter dan toegelaten (parameter: maxlevel) dan wordt de boom herbouwd tot een kleinere boom waardoor weer ruimte is voor nieuwe cases.

De parameter ruis bepaalt hoe er met uitschieters omgegaan wordt. Een blad wordt bij het herbouwen van de boom als ruis beschouwd als de ratio van het aantal cases in het cluster ten opzichte van het aantal cases in het grootste blad kleiner is dan de mate van ruis. De bladeren die als ruis worden beschouwd, vormen samen het restcluster en worden niet meegenomen in de tweede stap van het Two-Step algoritme.

Omdat dit algoritme de boom dynamische bouwt, is het gevoelig voor de volgorde waarin de data wordt gescand. Om deze reden is het belangrijk de data niet te sorteren, maar willekeurig te gebruiken.

²⁰ Bij alleen categorische variabelen bestaat het CF uit het aantal cases van de knoop en de frequentie van elke categorie van een variabele.

Stap 2: In de tweede stap worden de vele subclusters samengevoegd (zie oranje in de figuur hierboven) door middel van *agglomerative* hiërarchische clustering. Hierbij worden steeds de twee clusters die het dichtst bij elkaar liggen (op basis van de afstandsmaat) samengevoegd net zolang tot het gewenst aantal eindclusters is bereikt.

Afstandsmaat: De afstandsmaat die gekozen is, is de log-likelihood, omdat deze ook met categorische variabelen kan werken om de afstand te bepalen. De log-likelihood neemt aan dat categorische variabelen multinomiaal verdeeld en onafhankelijk zijn. Ook de verschillende invoer cases worden onderling onafhankelijk verondersteld. Uit empirisch onderzoek (intern bij IBM) blijkt dat de methode redelijk robuust is tegen het schenden van deze aannames. De log-likelihood is een afstandmaat die op kansen gebaseerd is. De afstand $d(i,j)$ tussen twee clusters i en j is dan gedefinieerd als de log-likelihood van elk van de clusters opgeteld minus de log-likelihood van deze clusters wanneer ze samengevoegd zijn:

$$d(i,j) = -N_i - N_j + N_{\langle i,j \rangle}$$

Waarbij in het geval van alleen categorische variabelen (IBM SPSS Statistics):

$$d_v = -N_v \sum_{k=1}^{K^B} \hat{E}_{vk}$$

$$\hat{E}_{vk} = - \sum_{l=1}^{L_k} \frac{N_{vkl}}{N_v} \log \frac{N_{vkl}}{N_v}$$

- K^B Totaal aantal categorische variabelen
- L_K Aantal categorieën voor de k^{de} categorische variabele
- N Aantal cases in totaal
- N_k Aantal cases in cluster k
- N_{jkl} Aantal cases in cluster j waarvan de k^{de} categorische variabele de l^{de} categorie is
- N_{kl} Aantal cases van de k^{de} categorische variabele die de l^{de} categorie zijn
- $\langle j,s \rangle$ Index die aangeeft dat het cluster gevormd is door cluster j en s samen te voegen.

6.2.2 Silhouette coëfficiënt

De Silhouette coëfficiënt (SC) wordt gebruikt om te valideren hoe goed het Two-Step algoritme de clusters onderscheid. De SC van een case i is gedefinieerd als

$$\frac{\min\{D_{ij}, j \in C_{-i}\} - D_{ic_i}}{\max(\min\{D_{ij}, j \in C_{-i}\}, D_{ic_i})}$$

Waarbij:

C_{-i} de clusterlabels waarvan i geen deel van uit maakt

C_i het clusterlabel waartoe i behoort

D_{ij} de afstand tussen case i en de centroïde van cluster j

Gebaseerd op deze individuele data is de gemiddelde SC:

$$SC = \frac{1}{N} \sum_{i=1}^N \frac{\min\{D_{ij}, j \in C_{-i}\} - D_{ic_i}}{\max(\min\{D_{ij}, j \in C_{-i}\}, D_{ic_i})}$$

Als $\max(\min\{D_{ij}, j \in C_{-i}\}, D_{ic_i})$ gelijk is aan 0, wordt case i niet meegenomen in de berekening van het gemiddelde.

Het berekenen van deze exacte coëfficiënt voor een dataset met meer dan een miljoen regels kost veel rekentijd en om deze te verminderen, wordt dit versimpeld door D_{ic_i} en $\min\{D_{ij}, j \in C_{-i}\}$ als volgt te definiëren:

D_{ic_i} de afstand van de invoer tot de centroïde van zijn cluster;
 $\min\{D_{ij}, j \in C_{-i}\}$ de minimale afstand van de invoer tot de centroïdes van elk ander clusters.

De afstand wordt hierbij Euclidisch berekend, waarbij de categorische variabelen opnieuw gecodeerd worden met één-uit- k codering om dit te kunnen berekenen. Als een variabele k categorieën heeft, dan wordt deze als k vectoren opgeslagen, met als eerste categorie als $(1,0,\dots,0)$, de tweede $(0,1,0,\dots,0)$ en als laatste $(0,0,\dots,0,1)$.

7. Resultaten en analyse

Het resultaat van dit onderzoek is de methodiek zoals deze in hoofdstuk 6 beschreven is. In dit hoofdstuk worden allereerst enkele resultaten van analyses gegeven die bepalend zijn geweest bij de ontwikkeling van de methodiek en verdere analyse (paragraaf 7.1). Vervolgens volgen de resultaten van analyses voor verschillende soorten clusterings van de data en de inzichten die hieruit volgen met betrekking tot de methodiek (paragraaf 7.2). Tot slot volgen de resultaten van het bepalen van de consistentie van een clustering (paragraaf 7.3) en de resultaten van het onderzoek naar de toegevoegde waarde van clusteren bij het bepalen van het EBP van huishoudens.

7.1 Resultaten analyses ter bepaling van methodiek

Het gas- en elektraverbruik wordt elk door een aantal verschillende hoofdkenmerken beïnvloed (hoofdstuk 4). De eerste insteek van dit onderzoek was om voor het EBP van gas en elektra afzonderlijk te analyseren, waartoe twee datasets zijn gemaakt. In de eerste zitten alle cases waarvan het elektraverbruik en type meter bekend is; in de tweede die waarbij dit voor gas het geval is. Huishoudens waarvan zowel het elektra- als gasverbruik bekend is, komen in beide sets voor.

Hieronder volgt een overzicht van de inzichten bij een clustering van de gas-set op basis van de variabelen Type en Bouwjaar. Deze twee variabelen zijn gekozen omdat zij in de praktijk veel gebruikt worden om het gasverbruik van huishoudens te vergelijken. De resultaten van de analyses zijn te vinden in bijlage D.

Inzicht 1: De aantallen per categorie hebben veel invloed op de clustervorming

Wanneer ruis wordt toegelaten, worden slechts een aantal type woningen meegenomen in de clustering. Dit is te verklaren doordat deze types vaak voorkomen. Zo zijn er 750.000 rijtjeswoningen (type 3) die over een aantal clusters verdeeld worden met elk verschillende bouwjaren. In latere analyses hebben we deze type woningen apart geclusterd, net als de resterende dataset, om te onderzoeken of dit een verbetering opleverde.

Inzicht 2: Het onderscheid tussen de massa en potentiëlen is te verklaren

Een groep die zeer waarschijnlijk besparingspotentieel heeft, zijn de huishoudens met een verbruik hoger dan de grens tussen normaal en hoog verbruik. Voor een willekeurig cluster, dat alleen type 1 en 3 woningen bevat, is het verschil tussen de massa en deze groep potentiëlen als volgt:

Massa: huurwoning, **15% heeft grootste inhoud**, 5% heeft grootste oppervlak, hoogste inkomen klasse is 30%, opleidingsniveau gelijkmatig, **woning type 3: 70%**;

Potentiëlen: koopwoning, **80% heeft grootste inhoud**, 50% heeft grootste oppervlak, alleen hoogste klasse inkomen, hoogste opleidingsniveau, **woning type 1: 82%**.

Het hoge gasverbruik van de potentiëlen is te verklaren doordat zij vooral huizen bevat met een grotere inhoud: er moet meer verwarmd worden dan bij de huishoudens in de massa. Kijken we naar het elektraverbruik in deze groep, dan zien we eenzelfde verdeling qua eigenschappen tussen de massa en de potentiëlen en is ook hier een verschil te verklaren. Alleen clusteren op basis van Type woning en Bouwjaar zorgt niet voor voldoende onderscheid tussen de groepen.

Om deze reden zijn we op zoek gegaan naar een clustering van de data waarbij het idealiter niet mogelijk is een verklarende factor te vinden die het hoge verbruik van de potentiëlen verklaard. Dit heeft drie gevolgen:

- i) In plaats van op een select aantal variabelen te clusteren, clusteren we in principe op alle acht kenmerken, tenzij er een selectie van huizen op basis van een kenmerk wordt gemaakt.
- ii) Omdat nu alle kenmerken gebruikt worden om vergelijkbare groepen te maken, is het overbodig om twee verschillende clusteringen voor gas en elektra te maken. Het werken met twee aparte sets is dan ook overbodig, zoals in paragraaf 4.1.1 als is vermeld.
- iii) Omdat het verschil in verbruik tussen de massa en potentiëlen te verklaren was, is stap 2 van de methodiek, 'de homogeniteitstest' ontwikkeld, zoals in paragraaf 6.1.2 is toegelicht. Hiermee kan de kwaliteit van een cluster beoordeeld worden.

Passen we deze test toe op deze clustering, dan blijkt dat de test in 90% van de gevallen verworpen wordt. Er is slechts één cluster dat voor gas redelijk scoort. De p-waardes van de test per variabele van dit cluster zijn in Tabel 4 weergegeven; de overige resultaten staan in bijlage D.

Uitleg tabel: De rood gearceerde waardes zijn diegene waarvoor de nulhypothese verworpen wordt. Dit betekent dat de kenmerken van die variabele bij de massa binnen dat cluster niet gelijk zijn aan de kenmerken van de potentiëlen. Een 'NA' (Not Applicable) betekent dat er slechts één categorie binnen dat cluster valt en dat de test niet uitgevoerd kan worden.

Tabel 4: De p-waardes per variabele van de homogeniteitstest voor één cluster; bij een waarde groter dan 0.05 verschillende de kenmerken tussen de massa en potentiëlen significant.

	Type	Eigendom	Bouwjaar	Inhoud	Oppervlak	Levensfase	Inkomen	Opleiding
elektra	NA	0.36	NA	0.00	0.00	0.00	0.01	0.37
gas	NA	0.03	NA	0.24	0.53	0.12	0.04	0.05

7.2 Resultaten en analyse stap 1 en 2 voor verschillende clusteringen

In het proces om een goede clustering van de data te vinden, zijn verschillende clusteringen onderzocht waarvan de resultaten hieronder zijn weergegeven. We hebben hierbij steeds eerst op basis van de SC het beste aantal clusters gekozen, waarna de clusterkwaliteit werd gecontroleerd door middel van de homogeniteitstest. Een kwalitatief goed cluster heeft weinig verschil in kenmerken tussen de massa en potentiëlen en kan gebruikt worden om de huishoudens met EBP te onderscheiden. Van potentiëlen uit een cluster met slechte kwaliteit kan niet gezegd worden dat zij EBP hebben, omdat zij geen gelijke kenmerken als de massa heeft. Bevat een clustering veel goede clusters, dan is er binnen de groep een goede onderverdeling van vergelijkbare huishoudens.

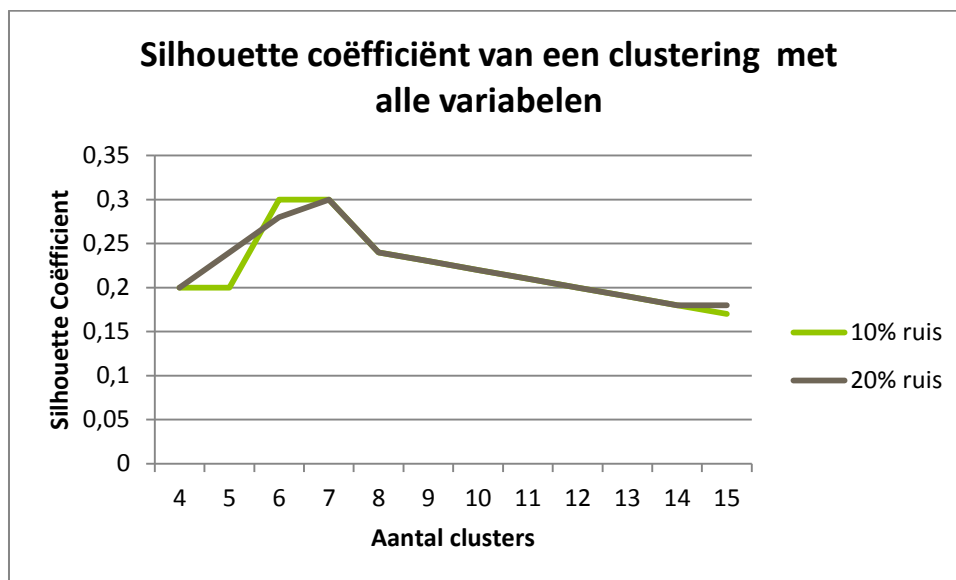
Clustering alle variabelen

Op basis van alle acht variabelen hebben we twee type clusteringen²¹ gemaakt: een deel met 10% ruis en een deel met 20% ruis. Bij een kleinere percentage ruis wordt meer data meegenomen bij de clustering dan bij een groter percentage: 10% neemt 2/3 deel van de cases mee en 20% neemt 1/3

²¹ Maxbranch: 20, maxlevel: 4

deel mee. De reden hiervoor is dat meer huishoudens als uitschieter aangemerkt mogen worden. Een vergelijking tussen deze twee clusterings, laat zien wat het verschil is van een variatie in ruis.

In Figuur 28 is het verloop van de SC voor de twee verschillende sets te zien. Hieruit blijkt de SC niet hoger dan 0,3 te worden en op basis hiervan kiezen we voor de een clustering met 6 variabelen voor de 10% ruis clustering en voor 7 clusters voor de 20% clustering. In het vervolg zullen we over de 6-clustering en 7 clustering spreken, waarbij de 7 clustering meer specifieke clusters heeft omdat zij maar 1/3 deel van de data meeneemt.



Figuur 28: De Silhouette Coëfficiënt per clustering met een verschillend aantal clusters en ruis; bij 10% ruis heeft een clustering met 6 variabelen de hoogste kwaliteit, bij 20% ruis is dit bij 7 clusters.

We gebruiken de homogeniteitstest om de kwaliteit van deze clusterings te controleren. Deze test hebben we voor elke variabele in elk cluster uitgevoerd en voor gas en elektra apart. De homogeniteitstest onderzoekt of de massa en potentiëlen gelijke kenmerken hebben of niet. In Tabel 5: De p-waardes van de homogeniteitstest bij de 7-clustering, waarbij de rood gearceerde variabelen met een p-waarde lager dan 0.05 duiden op geen gelijke kenmerken tussen de massa en de potentiëlen. Het hogere verbruik zou hierdoor verklaard kunnen worden. Tabel 5 staan de p-waardes die uit deze test kwamen voor beide clusterings, waarbij de cellen zijn gearceerd waarvoor de test wordt verworpen. Wanneer de test wordt verworpen, verschillen de massa en potentiëlen in kenmerken. De 'Aantal geen overeenkomst' kolom geeft aan bij hoeveel categorieën in het cluster de nulhypothese wordt verworpen en de 'Slecht cluster' kolom of een cluster slecht is op basis van de voorwaarden gegeven in paragraaf 6.1.2:

- elektra: oppervlakte en levensfase worden beide verworpen
- gas: oppervlakte of inhoud, en levensfase worden beide verworpen

De variabelen waarom een cluster als slecht wordt beschouwd zijn grijs gearceerd.

Tabel 5: De p-waardes van de homogeniteitstest bij de 7-clustering, waarbij de rood gearceerde variabelen met een p-waarde lager dan 0.05 duiden op geen gelijke kenmerken tussen de massa en de potentiëlen. Het hogere verbruik zou hierdoor verklaard kunnen worden.

elektra	Type	Eigendom	Bouwjaar	Inhoud	Oppervlak	Levensfase	Inkomen	Opleiding	Slecht cluster	Aantal geen overeenkomst
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	xxx	8
2	0.02	0.07	0.08	0.42	0.27	0.45	0.05	0.47		2
3	0.19	0.17	0.02	0.53	0.13	0.51	0.03	0.58		2
4	0.42	0.00	0.01	0.19	0.30	0.07	0.00	0.51		3
5	0.00	0.25	0.00	0.32	0.00	0.00	0.40	0.52	xxx	4
6	0.08	0.10	0.01	0.28	0.00	0.50	0.01	0.47		3
7	0.01	0.04	0.31	0.43	0.24	0.00	0.00	0.47		4

gas	Type	Eigendom	Bouwjaar	Inhoud	Oppervlak	Levensfase	Inkomen	Opleiding	Slecht cluster	Aantal geen overeenkomst
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	xxx	8
2	0.00	0.59	0.00	0.32	0.33	0.00	0.37	0.00		4
3	0.01	0.34	0.00	0.00	0.44	0.25	0.14	0.27		3
4	0.06	0.01	0.00	0.13	0.16	0.53	0.08	0.18		2
5	0.43	0.47	0.00	NA	0.00	0.01	0.05	0.35	xxx	4
6	0.00	0.01	0.00	0.52	0.50	0.29	0.06	0.29		3
7	0.00	0.08	0.00	0.49	0.14	0.27	0.16	0.12		2

Voor cluster 1 wordt de nulhypothese voor alle categorieën verworpen: de eigenschappen van de massa en potentiëlen komen niet overeen. Kijken we naar de eigenschappen, dan zien we dat er zowel huishoudens met een grote en kleine inhoud/oppervlak in dit cluster zitten; ook bij de andere variabelen in dit cluster liggen de categorieën verspreid. Cluster 5 wordt ook als 'slecht' bestempeld omdat zij de test op Oppervlak en Levensfase verwerpen. In totaal wordt 46%²² van de testen voor zowel gas als elektra verworpen, maar levert deze clustering vijf redelijk goede clusters om het EBP van te bepalen. Deze clustering levert zodoende niet inzicht in het EBP onder alle huishoudens, maar slechts van een deel.

²² Dit is per gas en elektra per clustering als volgt berekend: $\frac{\sum_{i=1}^{\text{aantal clusters}} \text{aantal geen overeenkomst}}{\text{aantal clusters} \times \text{aantal cluster variabelen}}$

Tabel 6: De p-waardes van de homogeniteitstest bij de 6-clustering, waarbij de rood gearceerde variabelen met een p-waarde lager dan 0.05 duiden op geen gelijke kenmerken tussen de massa en de potentiëlen. Het hogere verbruik zou hierdoor verklaard kunnen worden.

elektra	Type	Eigendom	Bouwjaar	Inhoud	Opper- vlak	Levens- fase	Inkomen	Opleiding	Slecht cluster	Aantal geen overeenkomst
1	0.17	NA	0.00	0.00	0.00	0.41	0.01	0.14		4
2	0.10	0.00	0.00	0.16	0.53	0.25	0.02	0.42		3
3	0.00	0.00	0.03	0.44	0.50	0.17	0.00	0.35		4
4	0.00	0.45	0.31	0.11	0.00	0.26	0.00	0.49		3
5	0.00	0.03	0.00	0.01	0.00	0.00	0.02	0.29	xxx	7
6	0.00	0.00	0.52	0.26	0.45	0.01	0.00	0.19		4

gas	Type	Eigendom	Bouwjaar	Inhoud	Opper- vlak	Levens- fase	Inkomen	Opleiding	Slecht cluster	Aantal geen overeenkomst
1	0.47	NA	0.00	0.00	0.01	0.05	0.46	0.00	xxx	5
2	0.00	0.01	0.00	0.13	0.23	0.52	0.16	0.49		3
3	0.00	0.40	0.00	0.00	0.33	0.00	0.00	0.00	xxx	6
4	0.00	0.32	0.00	0.05	0.00	0.00	0.00	0.20	xxx	6
5	0.00	0.35	0.00	0.01	0.00	0.00	0.50	0.42	xxx	5
6	0.00	0.05	0.00	0.37	0.39	0.14	0.50	0.19		3

Bij deze 6-clustering zijn meer cases meegenomen omdat er door een lager percentage ruis (10%) minder huishoudens als uitschieter aangemerkt worden. Hierdoor zit er ook meer variatie in de clusters, wat in Tabel 6 is terug te zien in het percentage variabelen dat geen overeenkomst heeft tussen de massa en uitschieters: respectievelijk 60% en 67% voor elektra en gas. Deze clustering is niet geschikt om het EBP gas mee te bepalen per cluster, omdat 2/3 deel van de clusters hiervoor ongeschikt is. Voor elektra zou dit voor de cases uit alle clusters behalve 5 wel kunnen. Vergelijken we deze clustering met de 7-clustering, dan heeft het de voorkeur om een clustering met meer ruis te maken. Omdat we echter graag meer huishoudens meenemen in onze analyse, kiezen we toch voor 10% ruis.

Het verloop van de SC bij de verschillende clusteringen is in bijlage D te vinden, evenals de tabellen met p-waardes waarover gesproken wordt.

Clustering per type woning apart: rijtjeshuis, flat/appartement en boerderij

We hebben eerder opgemerkt dat de clustering beïnvloed wordt door de frequenties van categorieën, waaronder de driekwart miljoen rijtjeshuizen (type 3). Om deze reden clusteren²³ we deze groep apart om te zien of binnen één type woning een beter onderscheid gemaakt kan worden tussen de woningen op basis van de overige zeven kenmerken. Dit blijkt niet het geval als we naar de SC kijken: deze blijft rond de 0,2 voor alle clusteringen van 3 tot 10 clusters. Bekijken we de flats/appartementen (type 4) ook apart, dan zien we hetzelfde: ook hierbij blijft de SC rond de 0,2. Omdat een eerdere clustering met 7 clusters een redelijke clustering gaf, bekijken we ook de homogeniteit van 7 clusters voor beide types met 10% ruis. Bij de clustering die gevonden wordt

²³ Maxbranch: 40, maxlevel:4, ruis:10

voor rijtjeshuizen, wordt voor elektra slechts één cluster verworpen en voor gas twee. In totaal wordt 41% van de tests verworpen, wat lager is dan de 7-clustering met 20% ruis. Voor de flats wordt slechts 37% (33%) voor elektra (gas) verworpen, maar zijn respectievelijk 3 (2) clusters slecht omdat de massa en potentiëlen zowel op oppervlak als levensfase niet overeenkomen.

Het gebruiken van een subset van de data op basis van type woning maakt dat minder tests verworpen worden, maar levert niet noodzakelijkerwijs betere clusters op; dit hangt van het type af. Daarnaast moet naast de SC ook het gezond verstand gebruikt blijven worden om een goed aantal clusters te kiezen wat tot een betere homogeniteit kan leiden.

Kijken we naar het type boerderij/tuinderij, dan vinden we over het algemeen een SC van 0,2 behalve bij 5 clusters (0,3). De homogeniteitstest op deze clusters laat zien dat geen enkele test voor elektra wordt verworpen en slechts 2 (6%) voor gas. Het aantal cases voor de boerderij (flat/appartement) bedroeg ongeveer 15.000 (290.000). Een veel kleinere subset lijkt dan ook voor een betere clustering te zorgen.

Clustering zonder rijtjeshuis en flat/appartement

Laten we huishoudens met een hoge frequentie weg (rijtjeshuis en flat/appartement), dan is ook hier een beste SC voor een clustering²⁴ met 3 clusters; de homogeniteitstest hiervan verwerpt echter 71% (79%) van de gevallen voor elektra (gas) en alle clusters worden hierdoor als ‘slecht’ aangeduid. Wellicht zou een clustering met 7 clusters hier ook een beter resultaat opleveren.

Clustering hoofdkenmerken gas en elektra: inhoud & bouwjaar, levensfase & oppervlakte

In het zoeken naar een clustering met zoveel mogelijk homogeniteit binnen de clusters, zijn we weer naar het begin teruggekeerd om te onderzoeken of een clustering op basis van meerdere hoofdkenmerken voor gas en elektra toch ook gebruikt zou kunnen worden. Hierbij wordt een redelijk goede clustering²⁵ met een SC van 0,4 gevonden bij 5 clusters. Uit de homogeniteitstest, blijkt dat voor elektra goed op bouwjaar, inhoud en levensfase wordt geclusterd; voor gas is dit alleen inhoud. Kijkend naar alle variabelen levert deze clustering slechte clusters.

Clustering binnen gelijke sociale kenmerken

In het onderzoek naar de relatie tussen variabelen (hoofdstuk 4) kwamen onder andere een tweetal groepen naar voren die ook te onderscheiden zijn in de eerdere clustering op basis van alle variabelen. In deze 6- en 7-clustering liet de homogeniteitstest zien dat er bij veel variabelen een verschil zit in verdeling van de kenmerken tussen de massa en de potentiëlen. Om deze redenen bekijken we of het maken van subgroepen op basis van ‘sociale kenmerken’ voor een verbetering van de clusterkwaliteit zorgt. We maken de groepen Hoog en Laag met als kenmerken²⁶:

Hoog: hoogopgeleid, koopwoning, inkomen boven modaal

Laag: laagopgeleid, huurwoning, inkomen modaal of lager

In beide groepen wordt de beste clustering op basis van de SC gevonden bij 3 clusters met een SC van 0,4 voor Laag en 0,3 voor Hoog. Ook is de clusterkwaliteit van Laag steeds net iets hoger dan Hoog bij een gelijk aantal clusters. Kijken we echter naar de homogeniteitstest, dan wordt 67-90%

²⁴ Maxbranch: 40, maxlevel:4, ruis:10

²⁵ Maxbranch: 40, maxlevel:4, ruis:10

²⁶ De huishoudens waarbij een van deze variabelen onbekend was (categorie 0) zijn buiten beschouwing gelaten. Ook huishoudens met een gemiddeld opleidingsniveau zijn buiten beschouwing gelaten.

van de testen verworpen en 2/3 van de clusters. Clusteren we net als eerder met 7 clusters, dan wordt 62-67% verworpen, ofwel 3 van de 7 clusters voor elektra van Hoog en 1 van de 7 voor gas en geen van de clusters bij de Laag clustering. Huishoudens met sociale eigenschappen 'Laag' worden dan ook beter geclusterd.

Net als bij de type woningen levert ook hier een clustering met een lagere SC op basis van de homogeniteitstest een betere clustering op. Het kan dan ook verstandig zijn, wanneer de SC bij lage cluster aantallen hoog is, toch voor een clustering te kiezen met meerdere clusters als de homogeniteitstest hier een beter resultaat geeft.

Deze twee groepen vragen bij het verlagen van de energierekening een andere aanpak vanwege de onder andere financiële mogelijkheden die zij hebben. Als gekozen wordt om binnen een dergelijke groep het EBP te bepalen, dan blijkt uit deze analyse dat dit hierbinnen een goede clustering mogelijk is.

7.2.1 Inzichten met betrekking tot de methodiek

Over het algemeen kunnen we een aantal zaken opmerken met betrekking tot de gevolgde methodiek.

- Een clustering op basis van slechts een aantal (primaire) kenmerken die invloed hebben op het verbruik, levert geen betere clusters dan wanneer alle variabelen worden meegenomen.
- Door het toestaan van ruis wordt de kwaliteit van de clusters over het algemeen beter, maar worden minder huishoudens met EBP gevonden omdat meer huishoudens als uitschieter aangemerkt mogen worden. Dit is geen probleem, maar een afweging die de gebruiker zelf kan maken.
- De homogeniteit van clusters wordt ook hoger wanneer er op een subset van de data geclusterd wordt, zoals het type woning. De clusters worden hiermee meer specifiek.
- Bij het kiezen van een optimaal aantal clusters kan niet alleen naar de SC gekeken worden: het kan voorkomen dat een clustering met meer clusters en een lagere SC meer homogene clusters oplevert.
- Net zoals 'clustering is in the eye of the beholder', geldt dit ook voor de homogeniteit. Afhankelijk van de reden waarom binnen een groep huishoudens de groep met mogelijk EBP geïdentificeerd wordt, kunnen andere (extra) eisen aan de homogeniteit worden gesteld.

Aanpassingen die niet tot een beter resultaat leiden

- De variabele Levensfase kan grofweg ingedeeld worden in drie hoofdcategorieën: alleenstaand, paren of gezin met als subcategorie jong, middelbaar of oud. Om te onderzoeken of een verlaging van het aantal categorieën tot een betere clustering leidde, hebben we de subcategorieën samengevoegd binnen haar hoofdcategorie en daarmee de 6-clustering opnieuw gedaan. Dit leverde geen verbetering op.
- Een clustering²⁷ op Inhoud en Levensfase met 8 clusters levert slechte clusters.
- Het verhogen van maxlevel naar bijvoorbeeld 8 maakt dat het algoritme er langer over doet terwijl de kwaliteit bij gelijkblijvende andere parameters niet beter wordt.
- Het transformeren van ordinale variabelen naar continue variabelen levert een verbetering op in de SC, maar de clusters worden slechter interpreteerbaar (zie analyse bijlage D).

²⁷ Maxbranch: 40, maxlevel:4, ruis:10

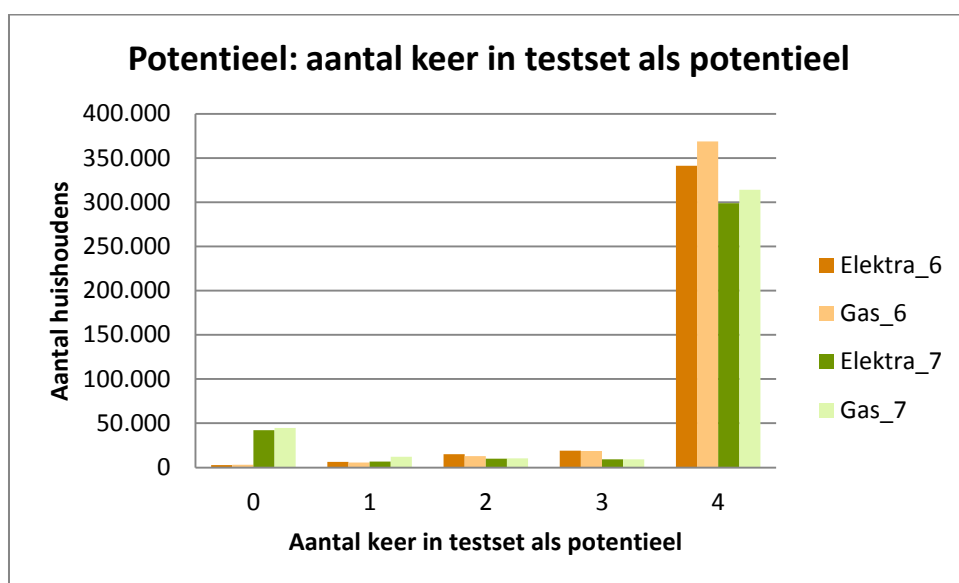
- Voegen we de variabele Aansluitwaarde aan de acht clustervariabelen toe en clusteren we de 7-clustering opnieuw, dan geeft dit eenzelfde SC voor de 7 clusters en treedt er geen verbetering op in maken van vergelijkbare huishoudens.

Verbeteringen onderzoek

Bij een latere analyse van de 7-clustering werd duidelijk dat wanneer de parameter maxbranch verhoogd wordt naar 50, de SC al bij een clustering met 5 clusters 0,3 wordt. Waarschijnlijk zal de homogeniteit van deze clusters niet heel hoog zijn, maar dat zal uit verder onderzoek moeten blijken. Andere, niet parameter specifieke, aanbevelingen voor verder onderzoek staan in het volgende hoofdstuk onder 'Aanbevelingen'.

7.3 Resultaten en analyse stap 3: consistentie clustering

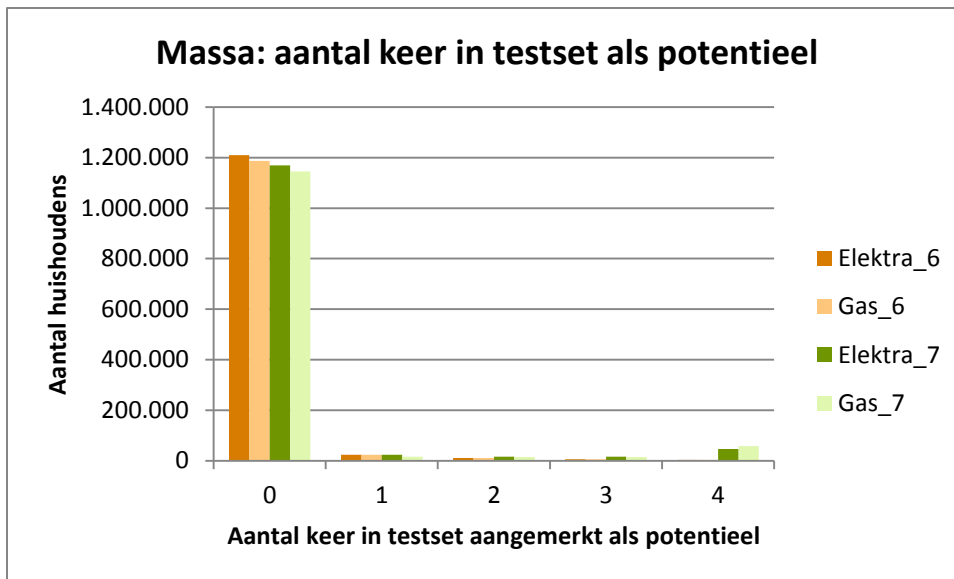
We hebben voor elk van de vijf test sets steeds de potentiëlen per cluster aangemerkt om de consistentie van de clustering te controleren. Vervolgens hebben we gekeken hoe vaak een huishouden als potentieel binnen elk van de tests werd aangemerkt en dit vergeleken met de potentiëlen op basis van de gehele clustering. Dit is maximaal 4 keer. Ook is gekeken of huishoudens uit de massa consistent als zodanig worden aangemerkt. De resultaten voor zowel de clustering met 6 clusters als de clustering met 7 clusters zijn in Figuur 29 in aantallen weergegeven.



Figuur 29: Aantal keren dat een huishouden met EBP uit de algehele clustering in de test sets als zodanig wordt aangemerkt.

Van de potentiëlen in de clustering op de gehele set wordt 80-90% viermaal in elk van de test sets aangemerkt als hebbende potentieel (Figuur 29). Dit betekent dat consistent dezelfde huishoudens aangemerkt worden als potentiëlen. Slechts een kleine groep is bij de 7-clustering op de gehele dataset waarschijnlijk onterecht tot de potentiëlen gerekend, omdat zij in de subsets geen enkele keer als potentieel aangemerkt worden (43.000 cases (11 %) van de potentiëlen²⁸).

²⁸ Voor gas en elektra gemiddeld



Figuur 30: Aantal keren dat een huishouden zónder EBP uit de algehele clustering in de test sets als hebbende EBP wordt aangemerkt.

Van de huishoudens die in de algehele clustering als massa worden aangemerkt, is 92-96% dit ook in alle test sets (zie Figuur 30). Ook hier is een kleine groep van de massa van de 7-clustering (4 %) die steeds als potentieel wordt aangemerkt.

Deze analyse toont aan de clustering op de gehele dataset een goed onderscheid maakt tussen de massa en de potentiëlen. In de verschillende test sets worden de huishoudens in deze groepen namelijk herhaaldelijk als zodanig aangemerkt. Slechts voor een kleine groep bij een meer specifieke clustering is dit niet het geval: de 7-clustering. De groepen massa en potentieel kunnen mogelijk nog aangescherpt worden door deze ‘foutief’ aangemerkte huishoudens alsnog in als massa/potentieel aan te merken. We bevelen aan dit vooral bij een meer-specifieke clustering te doen.

7.4 Resultaten en analyse toegevoegde waarde clustering

Om te bepalen in hoeverre potentiëlen beter onderscheiden worden door clustervorming, hebben we de potentiëlen verkregen uit de clustering vergeleken met potentiëlen op basis van de gehele dataset.

- In Tabel 7 staan de resultaten van deze analyse, waarbij in de gearceerde cellen het percentage huishoudens staat dat door de clustering beter wordt onderscheiden. Dit zijn namelijk de huishoudens die:
 - een hoog verbruik hebben ten opzichte van het geheel, maar ten opzichte van vergelijkbare huishoudens een normaal verbruik (categorie 1) of;
 - in vergelijking tot hun vergelijkingsgroep een hoog verbruik hebben, maar niet in vergelijking tot alle huishoudens (categorie 0).

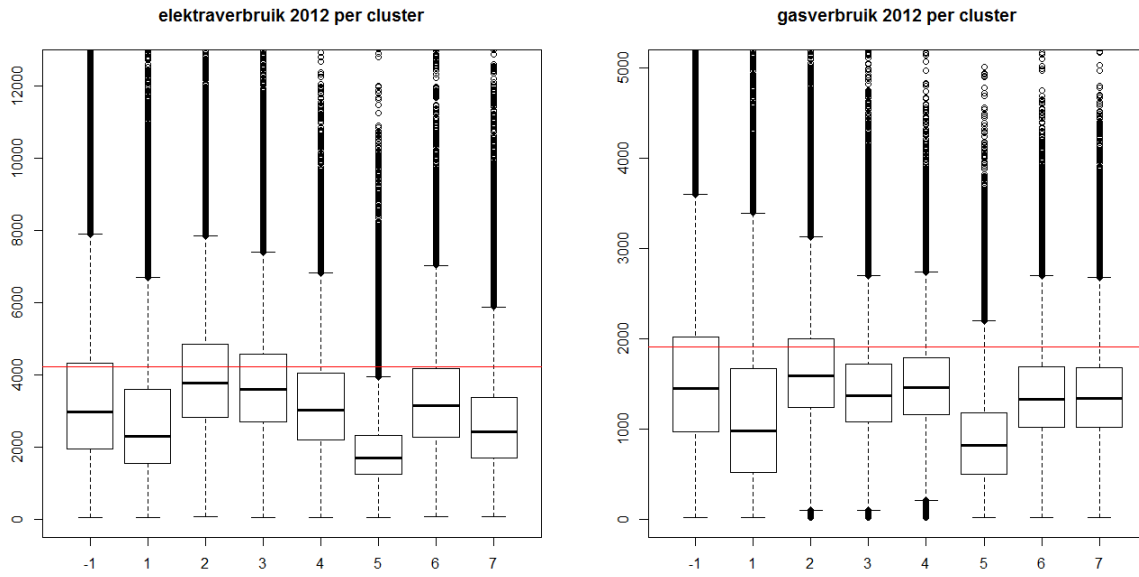
Tabel 7: Vergelijking percentages huishoudens die door clustering beter worden onderscheiden als potentieel dan zonder clustering (gearceerde cellen).

7-CLUSTERING; 75% grens					6-CLUSTERING; 75% grens				
elektra	-1	0	1	2	elektra	-1	0	1	2
aantal	1.221.696	32.056	32.044	352.453	aantal	1.182.536	71.216	71.230	313.267
%	75%	2%	2%	22%	%	72%	4%	4%	19%
gas	-1	0	1	2	gas	-1	0	1	2
aantal	1.183.839	45.649	45.405	363.356	aantal	1.142.097	87.391	87.191	321.570
%	72%	3%	3%	22%	%	70%	5%	5%	20%

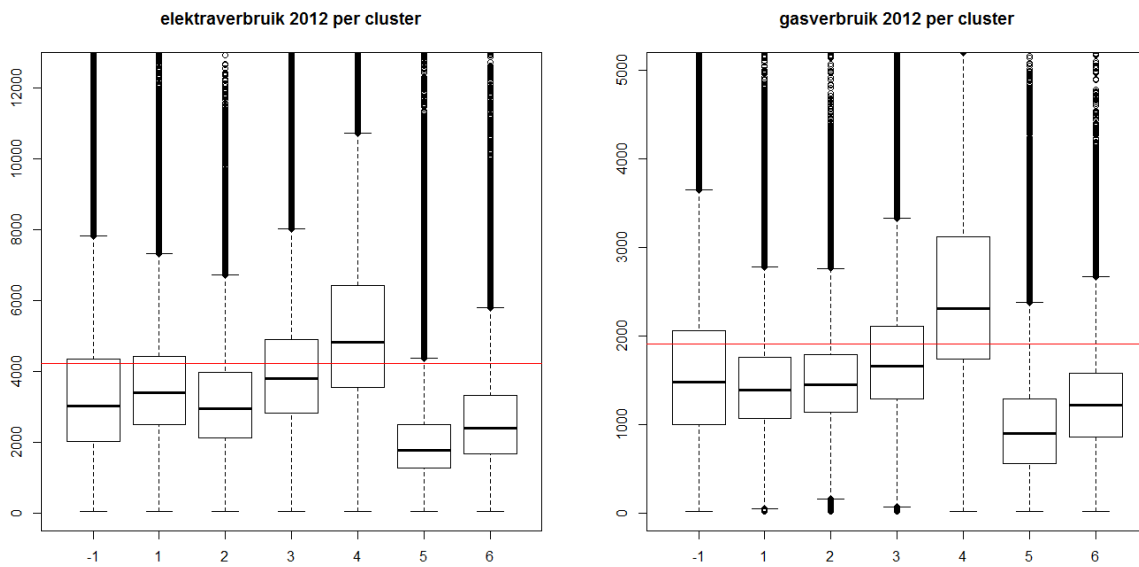
Kijken we naar de 7-clustering dan wordt hierin 2% (3%) van de huishoudens voor elektra (gas) extra aangemerkt als hebbende EBP (categorie 0). Eenzelfde percentage (categorie 1) wordt door clustering juist niet meer onderscheiden. De clustering in 6 clusters levert een groter verschil op: 8% (elektra) en 10% (gas); dit is te verklaren door het verschil in type clustering. In de 6-clustering zijn namelijk meer huishoudens meegenomen bij de clustervorming, waardoor het verschil in de spreiding tussen de clusters groter is. In 22% van de gevallen wordt een huishouden sowieso als potentieel aangemerkt, of er nu clustering wordt toegepast of niet.

In de praktijk betekent dit bijvoorbeeld het volgende. Een groep huishoudens wordt aangeschreven om mee te doen aan een energiebesparingspilot voor elektra. Kiezen we alleen diegenen met een hoog verbruik, dan zijn daar 32.000 huishoudens (7-clustering) die ten opzichte van een vergelijkingsgroep geen extreem hoog verbruik hebben. Daarnaast zijn er ook 32.000 huishoudens die meer baat hebben door deelname aan deze pilot, maar gemist worden. Door clustering wordt de groep dus gericht benaderd.

Vergelijken we de 75% grenzen per cluster met de 75% grens op alle gebruikte data (rode lijn in Figuur 31 en Figuur 32) dan zien we dit verschil ook terug: de 75% grens ligt in de 7-clustering 2 tot 3 clusters lager dan de clustergrens, waarbij dit bij de 6-clustering 3 tot 4 clusters zijn. Het toepassen van een clustering lijkt hiermee zin te hebben.



Figuur 31: Een vergelijking van de individuele clustergrens van 75% van de 7-clustering ten opzichte van de 75%-grens (rode lijn) van alle gebruikte data voor zowel elektra (links) als gas (rechts).



Figuur 32: Een vergelijking van de individuele clustergrens van 75% van de 6-clustering ten opzichte van de 75%-grens (rode lijn) van alle gebruikte data voor zowel elektra (links) als gas (rechts).

8. Conclusie en aanbevelingen

In deze scriptie stonden de volgende probleemstelling en deelvragen centraal:

Hoe kan het energiebesparingspotentieel (EBP) van huishoudens worden geschat?

- 1) Wat is de definitie van het EBP?
- 2) Welke factoren zijn van invloed op het elektra- en gasverbruik?

In dit hoofdstuk geven we de antwoorden op elke deelvraag, waarna we de probleemstelling beantwoorden. Ook de aanbevelingen voor verder onderzoek en het gebruik van de methodiek komen aan bod.

8.1 Conclusie

Tijdens dit onderzoek is de volgende definitie van het EBP gebruikt:

Een groep huishoudens heeft een EBP als zij ten opzichte van een groep vergelijkbare huishoudens een (veel) hoger verbruik heeft.

Om deze vergelijkbare groepen te vormen is onderzocht wat de factoren zijn die van invloed zijn op het elektra- en gasverbruik. Het *elektraverbruik* van de data wordt het best voorspeld door (1) de oppervlakte, (2) het type woning, (3) de levensfase en (4) het inkomen en het *gasverbruik* door (1) het type woning, (2) de oppervlakte, (3) het bouwjaar en (4) het inkomen.

Het resultaat van dit onderzoek, en daarmee het antwoord op de hoofdvraag, is de methodiek die ontwikkeld is om te schatten of een huishouden een hoog EBP heeft of niet. De drie stappen van de methodiek zijn als volgt. In stap 1 clusteren we de data met behulp van het Two-Step clusteralgoritme voor een aantal clusters en meten de kwaliteit hiervan met de Silhouette Coëfficiënt (SC). Op basis hiervan kiezen we een clustering met de hoogste clusterkwaliteit, welke verder onderzocht wordt op haar homogeniteit in stap 2. In de tweede stap bepalen we met behulp van de homogeniteitstest of de eigenschappen van de massa en de potentiële overeenkomsten en het cluster inderdaad vergelijkbare huishoudens bevat. In stap 3 testen we de consistentie van de clustering die in stap 1 gevonden is door het toepassen van 5-fold-cross-validation. Met de uitkomsten kan de groep potentiële eventueel nog aangepast worden. Het resultaat is een selectie van 25% van de huishoudens die een hoog EBP hebben.

Deze methodiek geeft een richting om goede clusters te vinden binnen de data. Echter, er kan niet blindelings één weg gevolgd worden die de beste clustering oplevert. De volgende bevindingen die gevonden zijn door op verschillende datasets te clusteren, geven handvatten bij het bepalen van de weg.

- Een clustering met een hoge SC levert niet noodzakelijkerwijs meer homogene clusters; het kan voorkomen dat uit stap 2 blijkt dat een clustering met meer clusters en een lagere SC meer homogene clusters oplevert. Over het algemeen leveren 7 clusters al goede resultaten.

- Een clustering van alle huishoudens op basis van de acht beschikbare variabelen²⁹ levert een betere clustering dan wanneer er met minder variabelen geclusterd wordt.
- Wanneer een bepaalde mate van huishoudens wordt weggelaten bij de clustering worden de clusters specifieker. Hierdoor worden op basis van de homogeniteitstest minder clusters als 'slecht' aangeduid omdat er minder verschillende kenmerken binnen een cluster zitten. Het gevolg hiervan is dat de groep potentiëlen kleiner is omdat minder data gebruikt wordt en de clusters beter van kwaliteit zijn. Het is beter een kleine mate van ruis toe te passen omdat dit de kwaliteit ten goede komt.
- De clusters worden ook beter van kwaliteit wanneer op een subset van de data, als type woning of sociale klasse, geclusterd wordt dan wanneer clusters van alle huishoudens in het Liander gebied gemaakt worden.

Om deze laatste reden kan een onderzoek naar de huishoudens met EBP het beste gedaan worden met een duidelijk doel of een duidelijke groep voor ogen. Op basis hiervan kan geclusterd worden op een subset van de data. Dit komt de kwaliteit van de clusters ten goede en de kans is groter dat de potentiëlen daadwerkelijk EBP bezitten.

8.2 Aanbevelingen

8.2.1 Aanbevelingen verder onderzoek

In de voorgaande analyses zijn meerdere punten naar voren gekomen die een aanbeveling voor verder onderzoek verdienen. Het gaat om zowel methodische aanbevelingen als aanbevelingen met betrekking tot de algehele insteek van het onderzoek. Ook zijn er een aantal aanbevelingen die voortkomen uit bestudeerde literatuur of uit ideeën voor de eerste aanpak van dit onderzoek.

- Mogelijk kan een hogere kwaliteit clusters gevormd worden wanneer aan de clustervariabelen een verschillend gewicht wordt meegegeven die de mate van invloed op het verbruik weergeeft. Hierbij kan de methodiek die leidt tot de bepaling van het energielabel verder onderzocht worden om inzicht te krijgen in een mogelijke weging van verschillende variabelen (aanbeveling vanuit paragraaf 5.2).
- Het bepalen van de grens tussen de massa en de potentiëlen verdient verder onderzoek. In dit onderzoek is een algemene grens genomen, gelijk voor alle clusters, maar wellicht levert een cluster-specifieke grens nauwkeurigere of meer betrouwbare resultaten op, zoals in paragraaf 6.1.2 al genoemd is.
- In paragraaf 6.1.3 merkten we al op dat Mirkin (2005) ook andere methodes aandraagt om de consistentie van een clustering te controleren. Om inzicht te krijgen in de beste methode kunnen deze verder worden onderzocht.
- Tijdens dit onderzoek bleek er ook een andere classificatie van type huishoudens te bestaan zoals tussenwoning, hoekwoning, flat etc. Onderzocht kan worden of deze typering voor een beter onderscheid tussen de groepen zorgt.
- Later in dit onderzoek stuiten we op een andere methode om de relatie tussen de verschillende variabelen te onderzoeken: Multi Correspondence Analysis (MCA) (Abdi & Valentin, 2007). Hiermee kunnen patronen van relaties tussen verschillende categorische

²⁹ type woning, inhoud/oppervlakte woning, eigendom woning, bouwjaar, opleiding, levensfase en inkomen

(nominale) variabelen geanalyseerd worden. Het kan gezien worden als een generalisatie van Principal Component analysis (PCA). Hiermee kunnen bijvoorbeeld verschillende type woningen als één algemeen type meegenomen kunnen worden.

- In het algemeen wordt aangenomen dat postcode-6 gebieden huishoudens bevatten die redelijk overeenkomen qua kenmerken. Een aanpassing op het huidige onderzoek zou kunnen zijn om in plaats van het Two-Step algoritme preclusters te laten vormen, de postcode gebieden als preclusers te gebruiken, hiervan de CF te bepalen en deze samen te voegen tot grotere clusters.

8.2.2 Aanbevelingen Liander

Binnen Klant en Markt en de afdeling Energiebesparing worden veel verschillende pilots gedraaid om te onderzoeken hoe goed een tool werkt of wat de impact van het doorvoeren van maatregelen is. Op dit moment is er nog geen duidelijke strategie over welke huishoudens voor deze projecten geselecteerd worden, maar worden mensen bijvoorbeeld via Twitter geworven of wordt er bij partijen gepeild waar animo is als het om een project met een andere partij gaat. Hierdoor doen vooral mensen mee die affiniteit hebben met energiebesparing.

Dit onderzoek richt zich op het ontwikkelen van een methodiek om groepen huishouden te onderscheiden waar mogelijk (veel) energiebesparingspotentieel zit ten opzichte van een vergelijkbare groep. Met het toepassen van deze methodiek bij het selecteren van pilotgebieden kan er gericht gewerkt worden. Vaak is er al een doelgroep waarbij de pilot zou passen, bijvoorbeeld huishoudens in Arnhem, of woningen van een bepaalde woningcoöperatie. Uit deze subset kunnen dan de huishoudens geselecteerd worden die een hoog verbruik hebben. De motivatie om mee te doen aan een energiebesparingsproject kan hierdoor ook hoger zijn.

Een eerste aanbeveling is om de clustering bijvoorbeeld te beperken tot subgroepen die een verschillende aanpak behoeven om besparing te bewerkstelligen, zoals de eerder genoemde Hoog-Laag clustering op sociale kenmerken, omdat de aanpak om tot besparing aan te zetten verschillend is. Op deze manier sluiten de inspanningen die gedaan worden beter aan bij de doelgroep.

Het mes snijdt zo aan twee kanten: enerzijds worden huishoudens die fors kunnen besparen geholpen om dit te realiseren en anderzijds kan het nut van een tool hiermee beter aangetoond worden omdat deze gericht bij een doelgroep ingezet kan worden. Een kanttekening is dat de doelgroep waarop getest wordt, bevooroordeeld is omdat er sowieso energiebesparing te behalen valt.

Er bestaan tools die inzicht geven in het verbruik van huishoudens, zoals bijvoorbeeld tools voor gemeenten en woningcoöperaties. Een tweede aanbeveling is om te onderzoeken of het onderscheid tussen huishoudens met of zonder EBP toegevoegd kan worden aan dergelijke tools en of dit onderscheid bij meerdere tools van toegevoegde waarde is.

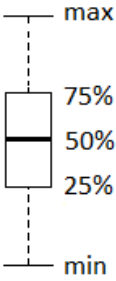
Een derde aanbeveling is dat deze methodiek in de praktijk moet worden toegepast om zichzelf te bewijzen. Zo kan gericht gekozen worden voor een pilotgroep en kunnen krachten gebundeld worden om tot meer inzicht te komen in het EBP dat bij huishoudens behaald kan worden. Ook kan verder onderzoek naar de reden van het EBP bij een groep leiden tot meer inzicht in hoe bij deze huishoudens bespaard kan worden.

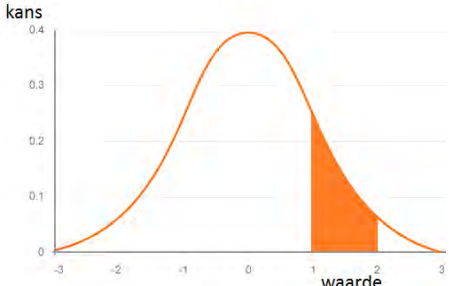
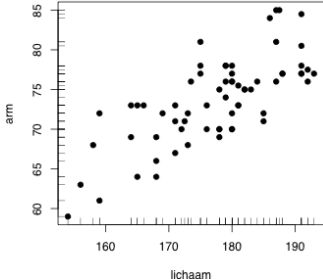
Bijlagen

Bijlage A: Begrippenlijst en gebruikte afkortingen

<i>Aansluitwaarde</i>	De aansluitwaarde van een meter geeft aan hoeveel stroom/gas een meter maximaal kan doorlaten.
<i>EBP</i>	Energiebesparingspotentieel bij een groep huishoudens die ten opzichte van een vergelijkingsgroep energie zouden kunnen besparen.
<i>Energie</i>	Verzamelbegrip voor zowel gas als elektra.
<i>Energietransitie</i>	De overgang van een energievoorziening gebaseerd op fossiele brandstoffen naar een volledig duurzame energievoorziening.
<i>KPI Inzicht</i>	Een Key Performance indicator (KPI) die gebruikt wordt om besparingspotentieel aan te tonen bij consumenten met een slimme meter die door Liander worden gestimuleerd om energie te besparen.
<i>KSA</i>	Kleinschalige aanbidding. Deze periode loopt van 1-1-2012 tot en met 31-12-2013. Tijdens deze periode zijn de netbeheerders verplicht slimme meters te plaatsen in de wettelijk verplichte categorieën (prioriteitplaatsingen, nieuwbouw, grootschalige renovaties, wijziging energielabel (meer met minder) en reguliere vervangingen).
<i>Netbeheerder</i>	De partij die verantwoordelijk is voor het onderhoud, de uitbreiding en de vervanging van het net en de stroom/gas van de leverancier naar de consument transporteert.
<i>Prosumant</i>	Een consument die energie afneemt en produceert.
<i>SJV</i>	Standaardjaarverbruik; het voor het weer gecorrigeerde gasverbruik van een huishouden.
<i>Slimme meter</i>	Een meter die het verbruik bijhoudt, net als een traditionele meter, maar dan niet meer analoog, maar digitaal.
<i>Stadsverwarming</i>	Woningen hebben geen eigen CV ketel, maar worden centraal verwarmd waarbij warm water of warmte vanuit een centraal punt naar de woningen in de wijk worden gebracht (Woonwebsite).
<i>Wob</i>	Wet onafhankelijk beheer; deze wet stelt dat Liander als netbeheerder geen andere (commerciële) activiteiten mag uitvoeren dan het beheren van de elektriciteits- en gasnetten.
<i>Wpb</i>	Wet bescherming persoonsgegevens; deze wet bepaalt wat er allemaal wel en niet mag met de gegevens van een persoon, om de privacy te waarborgen.

Uitleg wiskundige begrippen

<i>Boxplot</i>	 <p>Een boxplot geeft vereenvoudigd de spreiding van de data weer, zoals hiernaast is weergegeven. De dikke streep is de <i>mediaan</i>, de middelste waarneming. De doos daaromheen geven de 25% van de waarnemingen die daaronder en 25% die daarboven liggen weer en dit wordt de <i>interkwartielafstand</i> genoemd (IKA). Het minimum en het maximum zijn de waarnemingen die op maximaal 1,5xIKA van de randen van de doos verwijderd liggen. Waarnemingen die daarbuiten vallen heten <i>uitschieters</i>.</p>
<i>Centroïde</i>	De gemiddelde positie van alle punten voor alle variabelen.

<p><i>Dichtheidsfunctie</i></p>		<p>Grofweg laat een dichtheidsfunctie de kans zien dat een variabele een bepaalde waarde aanneemt. De totale kans onder de curve is opgeteld 1, dus het oppervlak tussen twee waardes geeft de kans weer dat de variabele die waarde aanneemt.</p>
<p><i>Mediaan</i></p>	<p>De middelste waarneming: 50% van de huishoudens liggen hieronder en 50% hierboven.</p>	
<p><i>Nominaal</i></p>	<p>De variabele heeft geen natuurlijke ordening, zoals bijvoorbeeld bij het type woning. Nominale variabelen kunnen niet op een schaal weergegeven worden, ordinale variabelen wel.</p>	
<p><i>Ordinaal</i></p>	<p>De variabele bezit een natuurlijke ordening in de categorieën. Een voorbeeld is de onderverdeling van bouwjaren in groepen; de ene groep is dan lager of hoger dan de andere.</p>	
<p><i>Scatterplot</i></p>		<p>In een scatterplot kun je de relatie of samenhang tussen twee variabelen onderzoeken. Voor elke waarneming van variabele 1 wordt de bijbehorende waarde van variabele 2 gegeven, waardoor inzichtelijk wordt of er een relatie tussen de twee bestaat. Deze kan bijvoorbeeld lineair zijn, zoals hiernaast.</p>

Bijlage B. Betekenis categorieën per variabele

Type woning

Omschrijving

- 1 Vrijstaand/bungalow
- 2 Twee-onder-één-kap
- 3 Rijtjeshuis
- 4 Flat/appartement
- 5 Kamer/studentenhuis
- 6 Etagewoning
- 7 Herenhuis/grachtenpand
- 8 Zelfstandige bejaardenwoning
- 9 Boerderij/tuinderij
- 10 Ander woningtype
- 0 Onbekend

Koop/huur

Omschrijving

- 1 Huur
- 2 Koop
- 0 Onbekend

Bouwjaar

Omschrijving

- 1 Voor 1800
- 2 Tussen 1800 - 1899
- 3 Tussen 1900 - 1919
- 4 Tussen 1920 - 1939
- 5 Tussen 1940 - 1959
- 6 Tussen 1960 - 1969
- 7 Tussen 1970 - 1979
- 8 Tussen 1980 - 1989
- 9 Tussen 1990 - 1999
- 10 Tussen 2000 - 2004
- 11 Later dan 2005
- 0 Onbekend

Inhoud woning

Omschrijving

- 1 Minder dan 250 m³
- 2 Tussen 250 - 325 m³
- 3 Tussen 325 - 375 m³
- 4 Tussen 375 - 475 m³
- 5 Meer dan 475 m³
- 0 Onbekend

Oppervlakte woningOmschrijving

- 1 Minder dan 60 m²
- 2 Tussen 60 - 80 m²
- 3 Tussen 80 - 100 m²
- 4 Tussen 100 - 120 m²
- 5 Tussen 120 - 150 m²
- 6 Tussen 150 - 200 m²
- 7 200 m² of meer
- 0 Onbekend

LevensfaseOmschrijving

- 1 Jonge alleenstaanden
- 2 Middelbare alleenstaanden
- 3 Oudere alleenstaanden
- 4 Gez. met alleen jonge kinderen
- 5 Gez. met jonge en oude kinderen
- 6 Gez. met alleen oude kinderen
- 7 Jonge paren zonder kinderen
- 8 Middelbare paren zonder kinderen
- 9 Oudere paren zonder kinderen
- 0 Onbekend

OpleidingOmschrijving

- 1 Laag
- 2 Middelbaar
- 3 Hoog
- 0 Onbekend

InkomenOmschrijving

- 1 Minimum
- 2 Beneden modaal
- 3 Modaal
- 4 1,5 keer modaal
- 5 2 keer modaal
- 6 2,5 keer modaal
- 0 Onbekend

Bijlage C. Resultaten analyses invloedsfactoren en onderlinge relatie

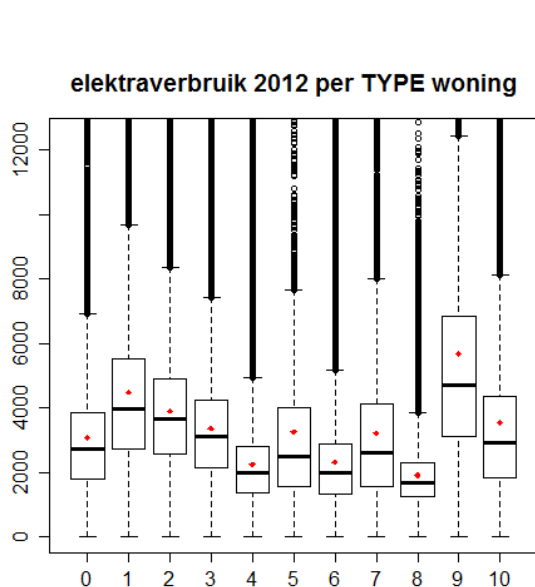
Deze bijlage geeft een overzicht van de verkregen resultaten uit de analyse van de factoren die invloed hebben op het energieverbruik zoals beschreven in sectie 4.3.2. Eerst wordt de relatie van het elektraverbruik met de verschillende socio-demografische eigenschappen en type aansluiting gegeven. Vervolgens zijn de socio-demografische factoren onderling vergeleken.

Voor alle analyses geldt dat categorie 0 'onbekend' betekent en de y-as als schaal kWh heeft voor elektra. Het bekijken van de boxplots moet met enige voorzichtigheid gedaan worden omdat de aantallen binnen de verschillende categorieën niet gelijk zijn. De rode punten in de boxplots geven het gemiddelde van elk cluster aan.

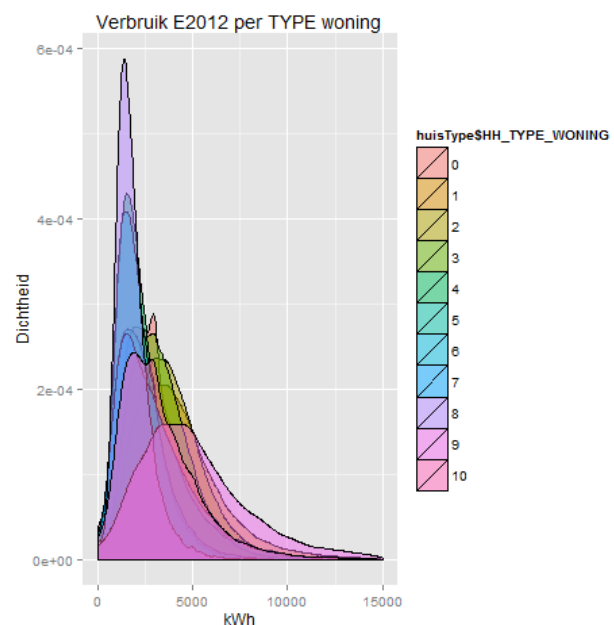
Resultaten factoren elektra

Verbruik elektra per type woning

Huizen met een kleiner oppervlakte (type 4,6,8) zoals flat, etagewoning en zelfstandige bejaardenwoning hebben een kleinere spreiding van hun elektra verbruik en een lagere mediaan. Een kamer (type 5) vormt hier qua spreiding een uitzondering op, terwijl zij ook een lage mediaan heeft. Het gemiddelde ligt in alle gevallen hoger (zie Figuur 33). Dit wordt veroorzaakt door de lange rechterstaart.



Figuur 33: Spreiding van het verbruik per type woning, waarbij de rode punt het gemiddelde aangeeft.

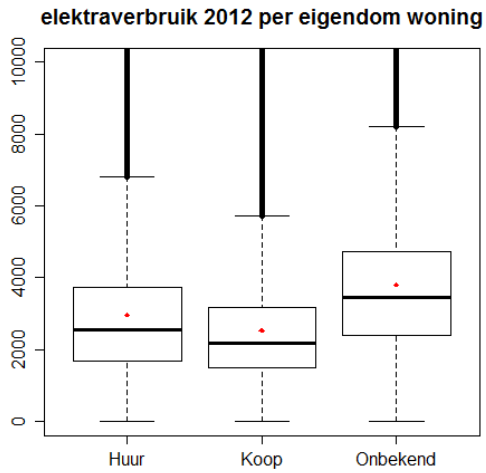


Figuur 34: De dichtheid per type woning, waarbij de zelfstandige bejaardenwoning (type 8) de hoogste piek heeft.

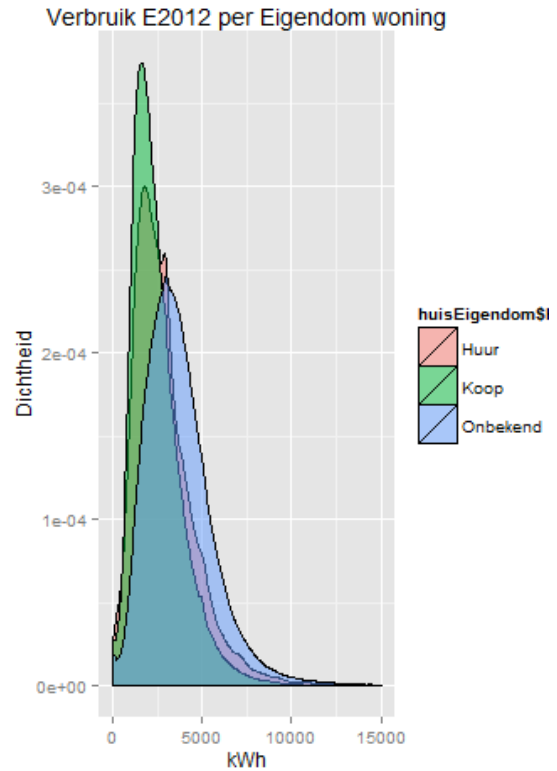
Figuur 34 met de benaderde verdelingsfunctie, laat zien dat de verdeling binnen de verschillende groepen niet vergelijkbaar is. Er zijn een drietal type woningen (4,6,8) die een hoge piek bij een laag verbruik hebben; de rest heeft een meer gespreid elektra verbruik.

Verbruik elektra per eigendom woning

De spreiding van het verbruik bij huurwoningen is een stuk groter dan die van koopwoningen, zoals in Figuur 35 te zien is. Ook ligt de mediaan en het gemiddelde hoger. In Figuur 36 is te zien dat huur en koop de grootste piek rond hetzelfde punt hebben.



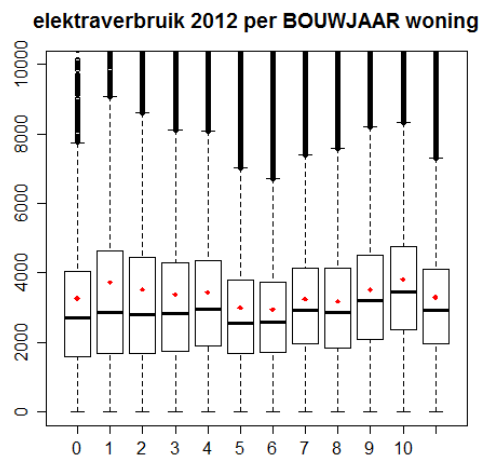
Figuur 35: Spreiding verbruik per eigendom, waarbij huurwoningen een grotere spreiding hebben dan koopwoningen.



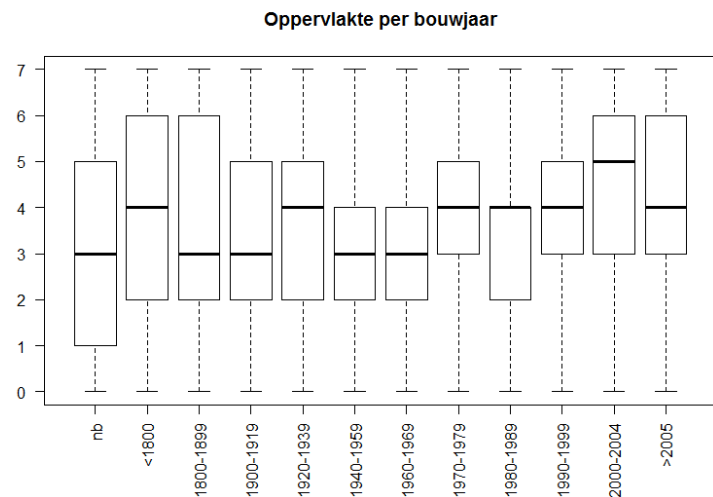
Figuur 36: De dichtheid van koop en huurwoningen, waarbij bij huurwoningen dezelfde extra piek in verbruik is te zien als bij het algemene elektraverbruik die door huurwoningen veroorzaakt.

Verbruik elektra per bouwjaar woning

De verschillende bouwjaren laten eenzelfde verdeling zien. Figuur 37 laat zien dat de spreiding van het elektraverbruik bij de woning gebouwd tussen 1800 en 1969 afneemt over de tijd (t/m categorie 6), waarna deze spreiding weer toeneemt. Ook neemt het gemiddelde verbruik weer toe. Een mogelijke verklaring hiervoor is dat de spreiding van het oppervlak van de woningen over deze categorieën ook eerst afneemt en vervolgens weer toeneemt (zie Figuur 38), waarbij het oppervlak wel een verklarende factor is voor het verbruik. Huizen die na 2005 gebouwd zijn vormen hier een uitzondering op. Zij hebben een vergelijkbare spreiding als huizen uit de periode 1970-1979 (type 7).



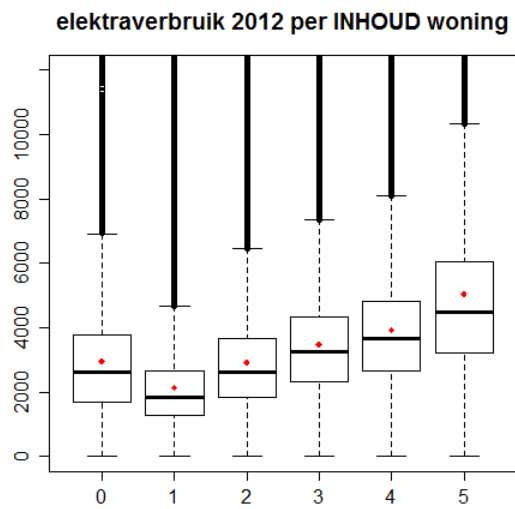
Figuur 37: Spreiding elektraverbruik per bouwjaar.



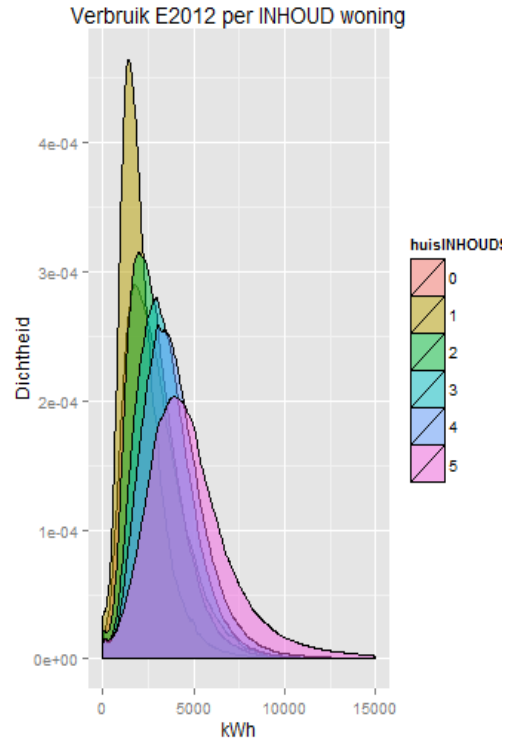
Figuur 38: Spreiding van de huishoudens over de verschillende categorieën van oppervlakte per bouwjaar.

Verbruik elektra per inhoud/oppervlakte woning

Op de onbekende woninginhoud (type 0) na, neemt het elektra verbruik toe en wordt de spreiding groter naarmate de woning groter wordt (zie Figuur 39). Ook laten ze redelijk dezelfde verdeling zien (zie Figuur 40). De oppervlakte laat hetzelfde verloop zien als de inhoud.

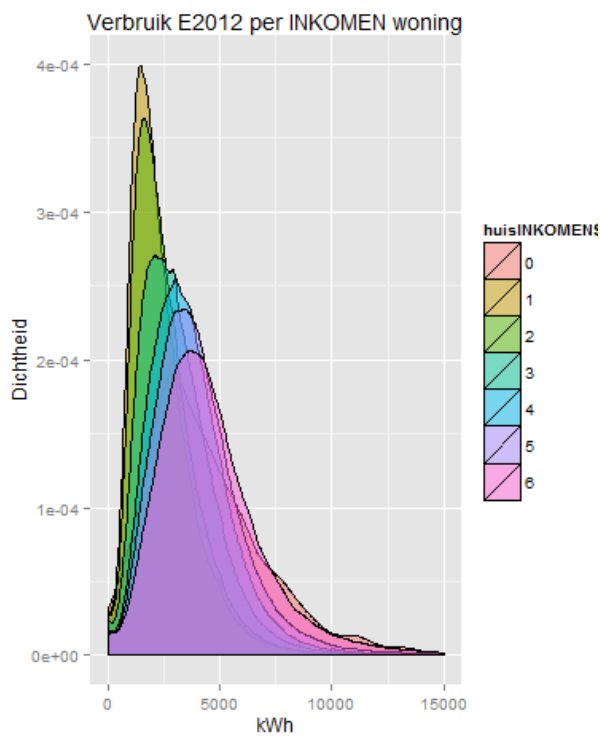


Figuur 39: De spreiding van het verbruik per inhoud woning. Het verbruik neemt toe, naarmate de inhoud groter is.



Figuur 40: Spreiding per inhoud. Hier is te zien dat de kans op een laag verbruik afneemt, naarmate de inhoud toeneemt.

Verbruik elektra per inkomen

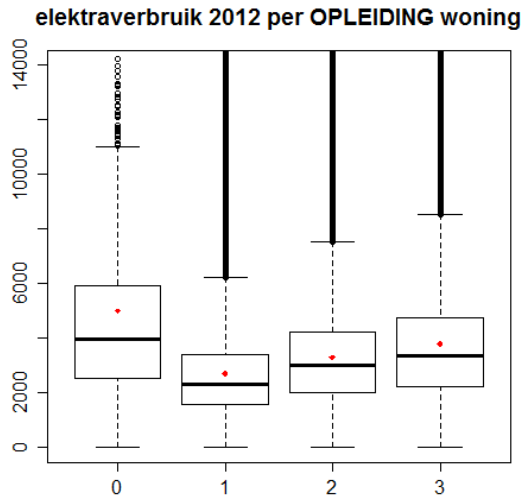


Het energieverbruik neemt toe (zowel mediaan als gemiddelde) naarmate het inkomen toeneemt en ook de spreiding wordt groter, zie Figuur 41.

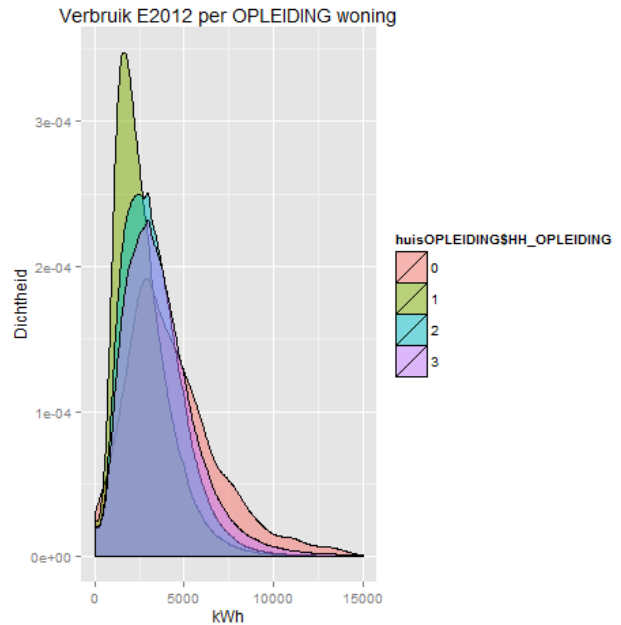
Figuur 41: Het verbruik per inkomenscategorie. Huishoudens met een hoger inkomen, hebben over het algemeen een hoger verbruik.

Verbruik elektra per opleiding

Bij het opleidingsniveau is hetzelfde te zien als bij het inkomen. Naarmate het opleidingsniveau toeneemt, neemt het gemiddelde verbruik toe en wordt spreiding groter (zie Figuur 42). Dit is te verklaren omdat mensen met een hogere opleiding vaak in grotere woningen wonen. Ook in Figuur 43 is dit te zien.



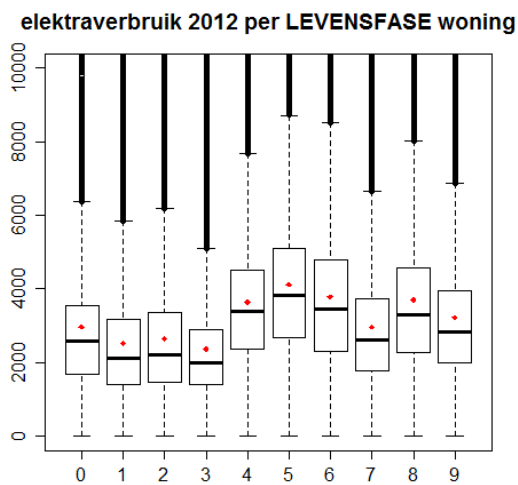
Figuur 42: Het verbruik per opleidingscategorie. Naarmate het opleidingsniveau toeneemt, wordt de spreiding groter en het gemiddelde hoger.



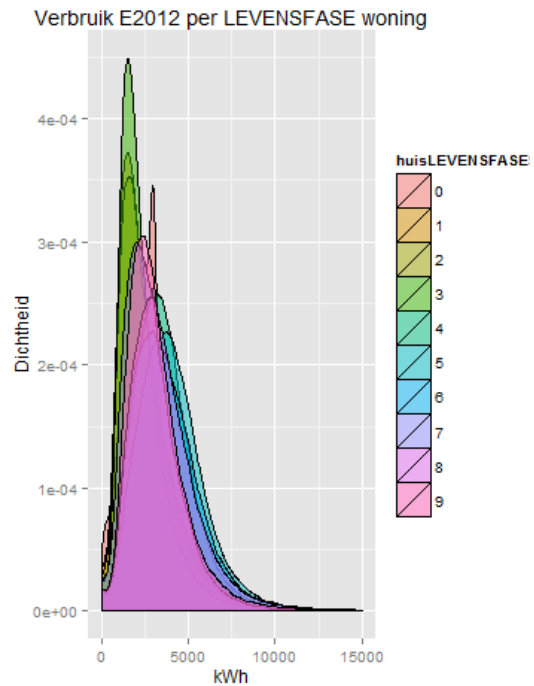
Figuur 43: De verdelingsfunctie van het verbruik per opleidingsniveau. Huishoudens met een hoge opleiding verbruiken meer elektra.

Verbruik elektra per levensfase

Gezinnen (type 4,5,6) hebben een hoger gemiddeld/mediaan verbruik dan paren (type 7,8,9) die weer meer verbruiken dan alleenstaanden (type 1,2,3) zoals in Figuur 44 te zien is. Binnen de groepen hebben alleenstaanden/paren van middelbare leeftijd een hoger verbruik (respectievelijk type 2 en 8) dan de jonge en oudere alleenstaanden/paren (type 1&3 en 7&8). Voor gezinnen geldt dat een gezin met gemengd jonge en oude kinderen (type 5) een hoger verbruik heeft dan gezinnen met alleen jonge of oudere kinderen (type 4 en 6). Dit laat de dichtheidsfunctie in Figuur 45 ook zien.



Figuur 44: Verbruik per levensfase van het huishouden. Gezinnen (type 4-6) hebben een hoger verbruik dan paren (type 7-8) die weer een hoger verbruik dan alleenstaanden (type 1-3) hebben.

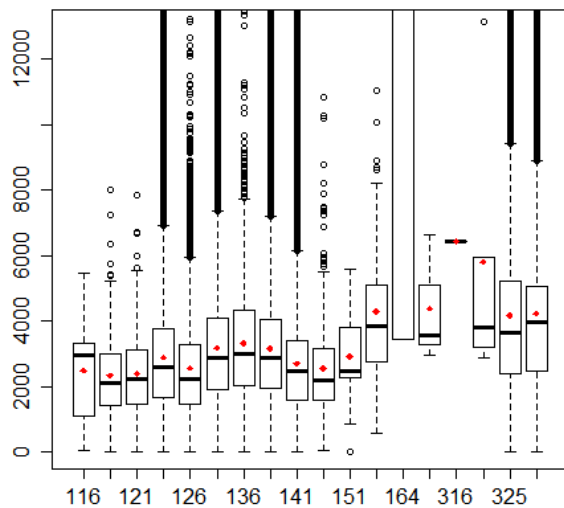


Figuur 45: Spreiding van het verbruik per levensfase, waarbij te zien is dat alleenstaanden voornamelijk een laag verbruik hebben.

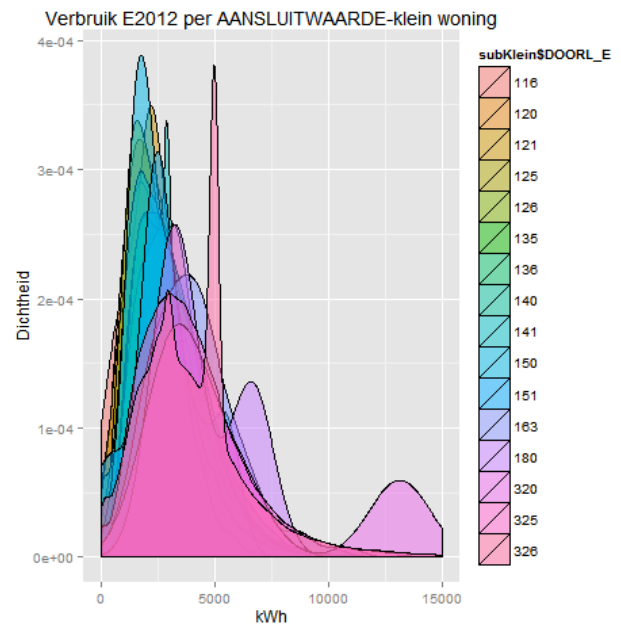
Verbruik elektra per aansluitwaarde

Bij de verbruiken tot 3x25A (Figuur 46; as-label 325) schommelt de mediaan (en vaak ook het gemiddelde) tussen de 2000-4000 kWh. De spreiding schommelt hier op dezelfde manier. De meters met een aansluitwaarde vanaf 1x64A tot 3x25A vertonen een aparte spreiding omdat dit over 2-5 aansluitingen gaat. Opvallend bij de dichtheidsplots (Figuur 47) zijn de twee pieken die in het verbruik bij de hoogste aansluitingen vanaf 3x20A voorkomen. Het is onbekend wat dit veroorzaakt.

elektraverbruik 2012 per AANSLUITWAARDE wor



Figuur 46: Elektraverbruik per aansluitwaarde elektrameter. Met een grotere meter kan meer verbruikt worden. Het verbruik van huishoudens met een aansluitwaarde kleiner dan 3x25A ligt ook lager.



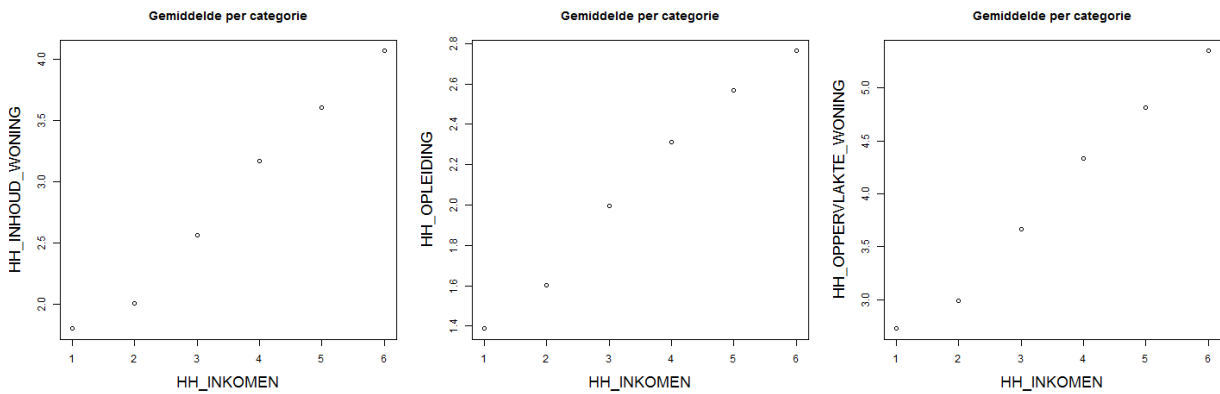
Figuur 47: De dichtheidsfunctie voor de verschillende type aansluitwaarde laat zien dat de lage aansluitwaardes een verwachte piek hebben bij een lager verbruik en de hogere aansluitingen (groter dan 180) bij een hoger verbruik.

Resultaten onderlinge factoren

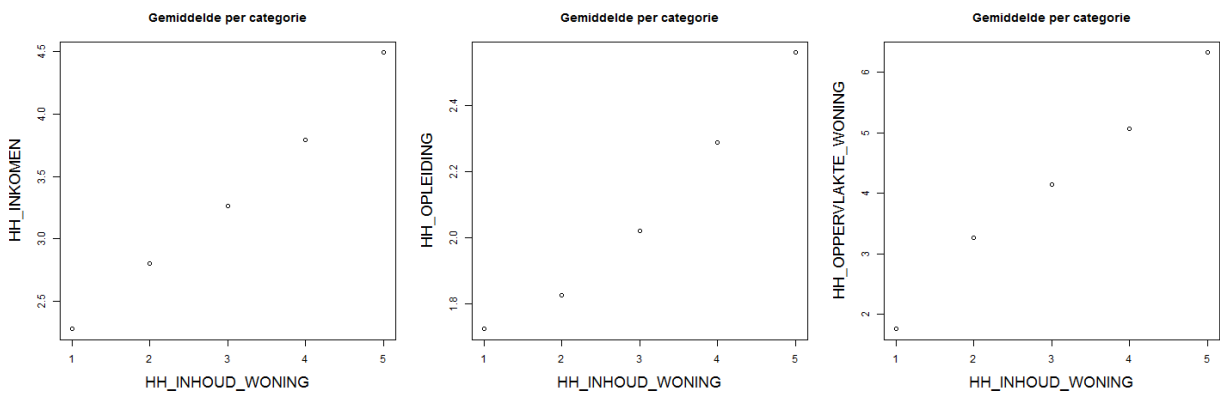
Lineariteit onderlinge relatie

Hieronder staan de scatterplots van de variabelen waarbij een lineaire relatie tussen de variabelen is te zien, op basis van het gemiddelde per categorie. De betekenis van de aslabels staat in bijlage B.

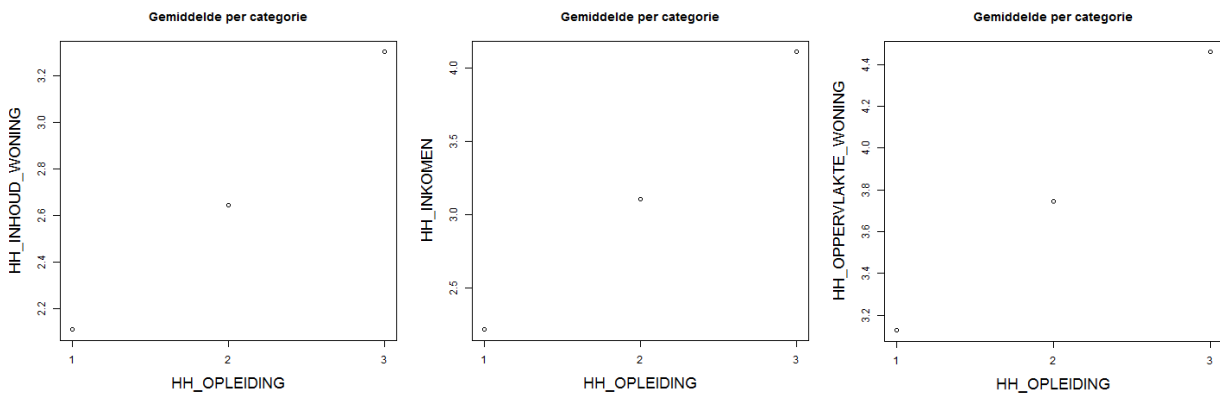
Inkomen



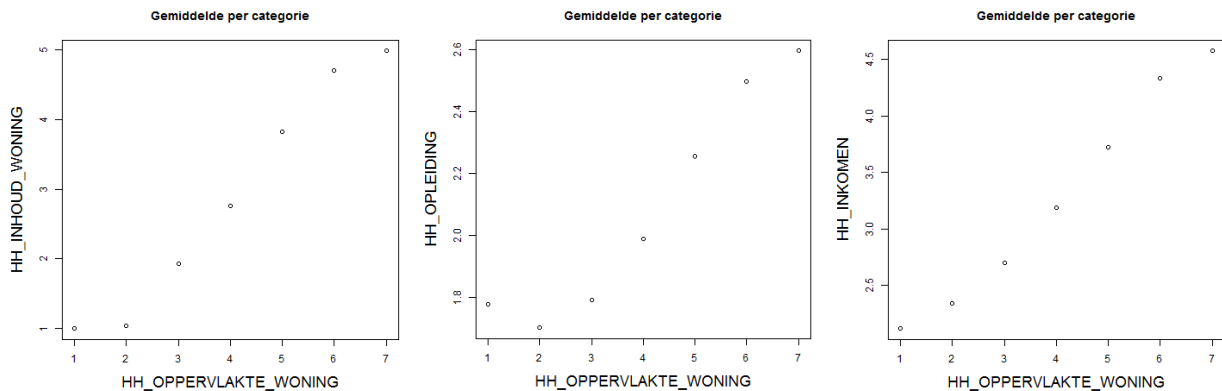
Inhoud



Opleiding



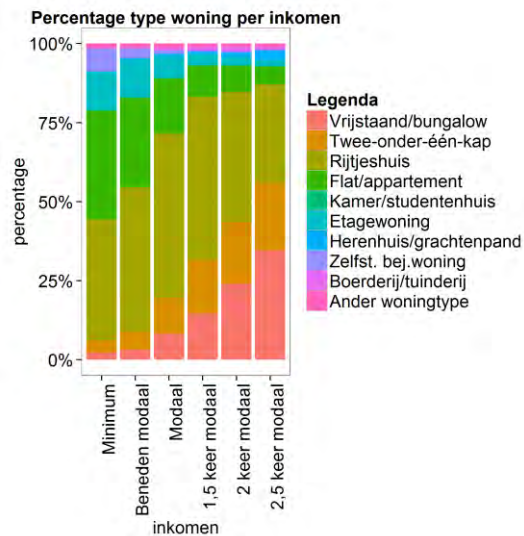
Oppervlakte



Inzichten uit de analyse naar de onderlinge relatie tussen categorieën van variabelen

Hieronder wordt steeds de verdeling van één factor over de categorieën van andere factoren vergeleken met behulp van gestapelde staafdiagrammen, zoals is uitgelegd in de methodiek (paragraaf 4.2.3) en waarvan een voorbeeld in Figuur 48 te zien is. Deze factoren zijn steeds respectievelijk:

1. Bouwjaar
2. Eigendom
3. Inhoud
4. Inkomen
5. Levensfase
6. Opleiding
7. Oppervlakte



Figuur 48: Een voorbeeld van hoe de relatie tussen de categorieën van variabelen is onderzocht.

1. Bouwjaar

1.1. ---

1.2. *Huurhuizen voornamelijk uit periode '60-'89 (klasse 6-8)*

1.3. *1960-1969 heeft vooral een oppervlakte van 325-475m³*

1.4. *- inkomen en bouwjaar → mensen met een hoger inkomen wonen vaker in een oudere woning, of een woning uit 1990-1999*

- in woningen uit 1980-89 leven voornamelijk mensen met een minimum inkomen (klasse 1); ongeveer 60% v deze klasse woont in huizen uit '60-'89

1.5. *Per levensfase niet veel verschil te zien*

- 1.6. Hoge opleiding woont 30% in huizen uit voor 1920, bij laag is dit 10%
- 1.7. Oppervlakte van woning is groter bij huizen van voor 1920
- 1.8. Boerderij/tuinderij vaak veel ouder

2. Eigendom woning

- 2.1. Bouwjaar maakt geen verschil verdeling koop/huur
- 2.2. ---
- 2.3. 90% van de woningen met een grote inhoud zijn koopwoningen | | Naarmate de woning groter wordt, is een groter percentage koopwoning
- 2.4. Inkomen omhoog, groter percentage koop
- 2.5. Levensfase → verdeling koop/huur voor alleenstaand 40-60, bij de rest is dit ongeveer 60/70%-40/30%
- 2.6. Hogere opleiding = groter percentage koopwoning
- 2.7. Oppervlakte groter= groter percentage koopwoningen
- 2.8. Vrijstaand & boerderij/tuinderij is voornamelijk koop; bij (4,5,6,8) flat, kamer, etagewoning, bejaarden is 70-90% huur

3. Inhoud woning

- 3.1. Nieuwere huizen hebben een kleiner percentage woningen met een inhoud <250m³
- 3.2. 40% (70%) van de huurhuizen heeft inhoud <250 (325) m³, bij koop is dit 10% (30%)
- 3.3. –
- 3.4. Inkomen omhoog, percentage inhoud <250 (325) m³ daalt sterk (vergeleken met: min inkomen 80% <325, modaal 55%, max 18%)
- 3.5. Van de alleenstaanden woont een 45-60% in een <250 woning dan anderen. Gezinnen en paren 10-18%, behalve jonge paren, daarvan woont 30% klein
- 3.6. Hogere opleiding = groter percentage grotere woningen (lage opleiding heeft 5% een grote woning vergeleken met 25% van de hoogopgeleiden) 35% van mensen met lage opleiding woont <250m³
- 3.7. Inhoud-oppervlak sterk gecorreleerd
- 3.8. 45-60% van type 4-8 klein oppervlak (appartement etc), vrijstaand & boerderij (45 & 60%) heel groot

4. Inkomen

- 4.1. Redelijk verspreid over de bouwjaren, wel laag inkomen minder in oude huizen en minder in super nieuwbouw, vooral in bouwjaar (5-8)= 1940-1989
- 4.2. 90% huurwoningen wordt door mensen met inkomen ≤modaal bewoond t.o.v. 50% van de koopwoningen
- 4.3. Grotere inhoud = kleiner percentage laag inkomen (vergeleken: bij <250 heeft 60% <modaal inkomen, bij 325-375m³ is dit 10% en verder naar 5%)
- 4.4. –
- 4.5. 60% vd oudere alleenstaanden heeft beneden modaal inkomen, alleenstaanden 45-60% beneden modaal, voor gezin en paren is dit rond de 20%, boven modaal is hier rond de 40%
- 4.6. Hogere opleiding = groter percentage boven modaal (vergeleken: laag=10% boven modaal, hoog=65%)
- 4.7. Idem 4.3

4.8. type 4,5,6,8 hebben een laag inkomen (80-90% v deze woningen) (en zijn huur)

5. Levensfase

- 5.1. Redelijk verspreid over de bouwjaren, wel zitten voornamelijk alleenstaanden in oude huizen
- 5.2. Meer alleenstaanden wonen in een huurwoning (50%) dan in koop (25%)
- 5.3. Vooral alleenstaanden in <math><250\text{m}^3</math> woning (60%) vanaf 250-325 daalt dit naar 30%
- 5.4. 50% van de mensen met minimum inkomen is alleenstaand, dit neemt af naar 15% bij een stijging van het inkomen; percentage paren neemt toe naarmate het inkomen stijgt; n.b. er zijn weinig jonge paren zonder kinderen
- 5.5. –
- 5.6. 50% van de alleenstaanden heeft lage opleiding, 20% van de hoog opgeleiden heeft gezin met jonge kinderen (klasse 4) en 20% middelbaar paar zonder kinderen (klasse 8)
- 5.7. Naarmate het oppervlak van de woning toeneemt, neemt ook het percentage paren toe (van 5% naar 40%) Het percentage gezinnen neemt toe naarmate het oppervlak toeneemt (tot 150m^3)
- 5.8. Middelbare paren en gezinnen met jonge kinderen zitten van elk plusminus 20% in vrijstaand/twee-onder-één-kap/rijtjeshuis. In zelfstandige bejaarden woning zitten veel oudere alleenstaanden; kamer/etage/herenwoning plusminus 50% alleenstaand

6. Opleiding

- 6.1. 'normaalcurve' van laag opgeleiden over de bouwjaren met 40% als piek bij 1940-1960, omgekeerd voor hoogopgeleiden, van de oude huizen voor 1800 is 60% hoogopgeleiden, bij nieuwste woningen is 40% dat.
- 6.2. Hoger percentage hoogopgeleid bij koopwoningen (40%) t.o.v. 10% laagopgeleid, bij huur is 50% laagopgeleid
- 6.3. Hogeropgeleiden zitten vaak in een grotere woningen, lager vice versa
- 6.4. Opleiding vs inkomen lineair; meer lager inkomen bij lagere opleiding en vice versa
- 6.5. Van de oudere alleenstaanden is 70% laagopgeleid. Bij de overige levensfase is rond de 25% hoogopgeleid. Per levensfase is 15 tot 30% laagopgeleid.
- 6.6. ---
- 6.7. Groter oppervlak = hoger percentage hoogopgeleid
- 6.8. In herenhuis/grachtenpand 90% hoogopgeleid, in zelfstandige bejaardenwoning vooral laagopgeleid

7. Oppervlakte woning

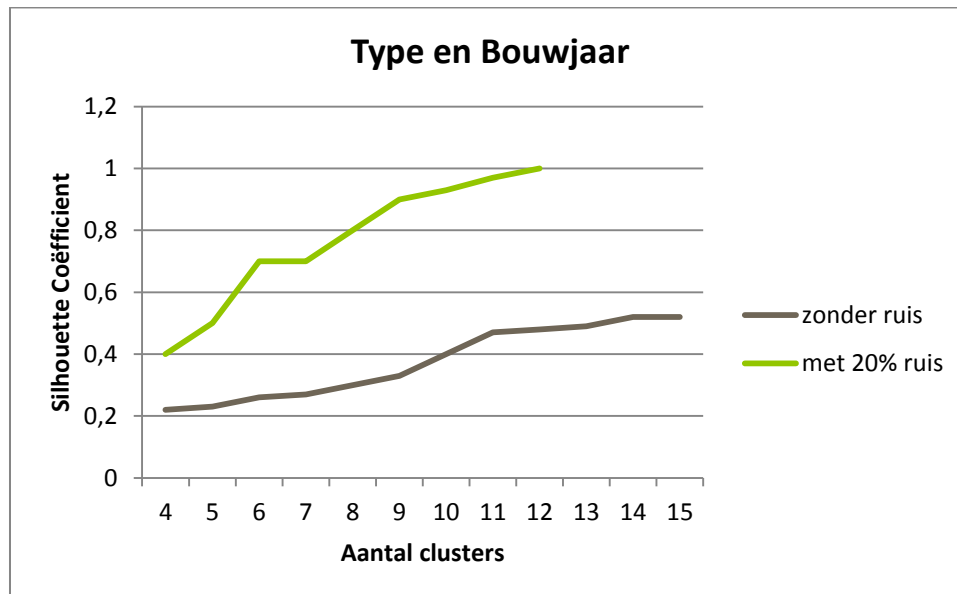
- 7.1. De verschillende oppervlaktes zijn redelijk verspreid over de bouwjaren
- 7.2. 70% van de huur woningen is <math><100\text{m}^2</math>, bij koop is dit 30%
- 7.3. gecorreleerd
- 7.4. hoger inkomen = groter oppervlak
- 7.5. naarmate alleenstaanden ouder zijn neem het percentage dat op <math><80\text{m}^2</math> leeft af (van 60-45%), voor de rest ongeveer zie inhoud
- 7.6. van de hogeropgeleiden zit 30% op <math><100\text{m}^2</math>, van de laag is dit 70%
- 7.7. --
- 7.8. idem als inhoud

Bijlage D: resultaten clusteranalyse en homogeniteitstest

In deze bijlage worden de extra resultaten weergegeven van de clusteranalyse. Hierbij worden dezelfde kopjes gebruikt als in de hoofdtekst vermeld staan.

Clustering Type en Bouwjaar

We hebben voor een verschillend aantal clusters (van 4 tot 15) de clusterkwaliteit bekeken met gelijke parameters maxbranch^{30} 20 en maxlevel^{31} 4 en met en zonder dat een ruis van 20% werd toegelaten.



Figuur 49: Een clustering van de gehele dataset met en zonder toepassing van ruis, laat zien dat de clusters beter onderscheiden worden wanneer ruis wordt toegelaten. Dit is te zien aan de hogere Silhouette Coëfficiënt.

De kwaliteit van de clustering neemt af naar mate er minder clusters toegelaten worden, maar een clustering van 12 clusters met ruis gaf bijna eenzelfde kwaliteit als de clustering met 9 clusters: een SC van 1 ten opzichte van 0,9. Hierbij hadden in de 12-clustering de woning types 1 en 2 een eigen cluster, waarbij deze bij de 9-clustering in zijn geheel zijn toegevoegd aan een ander cluster met eenzelfde bouwjaar. Omdat we de voorkeur geven aan minder clusters en de kwaliteit niet veel achteruit gaat, kiezen we voor 9 clusters.

Met 20% ruis, waardoor uitschieters worden weggelaten, worden veel betere clusters gevonden (SC van 0,9 bij 9 clusters ten opzichte van een SC van 0,3), maar worden slechts een aantal type woningen en bouwjaren meegenomen: alleen types 1-4 (Vrijstaand, Twee-onder-één-kap, Rijtjeshuis en Flat/Appartement) en bouwjaren categorie 4-9 (1920-1999).

Uit de homogeniteitstest, waarvan de resultaten in Tabel 8 staan, blijkt ook dat de massa en potentiëlen binnen veel clusters een verschil in eigenschappen hebben.

³⁰ Maximum aantal subclusters in een tak

³¹ Maximale diepte van de boom in de eerste stap

Tabel 8: p-waardes van de clustering op basis van Type en Bouwjaar met 20% ruis voor elektra en gas, waarbij de massa en potentiëlen in de meeste clusters geen gelijke eigenschappen hebben (een p-waarde kleiner dan 0.05).

Elektra	Type	Eigendom	Bouwjaar	Inhoud	Opp	Levensfase	Inkomen	Opl	Slecht cluster	Aantal verworpen
1	NA	0.00	NA	0.00	0.00	0.48	0.00	0.00		5
2	NA	0.00	NA	0.00	0.00	0.49	0.00	0.00		5
3	0.00	0.00	NA	0.00	0.00	0.02	0.00	0.00	xxx	7
4	NA	0.00	NA	0.00	0.00	0.17	0.00	0.00		5
5	NA	0.00	NA	0.00	0.00	0.01	0.00	0.00	xxx	6
6	0.00	0.00	NA	0.00	0.00	0.00	0.00	0.00	xxx	7
7	NA	0.36	NA	0.00	0.00	0.00	0.01	0.37	xxx	4
8	NA	0.00	0.18	0.00	0.00	0.00	0.00	0.00	xxx	6
9	NA	0.00	NA	0.00	0.00	0.00	0.00	0.00	xxx	6

Gas	Type	Eigendom	Bouwjaar	Inhoud	Opp	Levensfase	Inkomen	Opl	Slecht cluster	Aantal verworpen
1	NA	0.00	NA	0.00	0.00	0.44	0.00	0.00		5
2	NA	0.00	NA	0.00	0.00	0.05	0.00	0.00	xxx	6
3	0.00	0.00	NA	0.00	0.00	0.00	0.00	0.00	xxx	7
4	NA	0.00	NA	0.00	0.00	0.14	0.00	0.00		5
5	NA	0.00	NA	0.00	0.00	0.01	0.00	0.00	xxx	6
6	0.00	0.00	NA	0.00	0.00	0.00	0.00	0.00	xxx	7
7	NA	0.03	NA	0.24	0.53	0.12	0.04	0.05		3
8	NA	0.00	0.41	0.00	0.00	0.07	0.00	0.00		5
9	NA	0.05	NA	0.00	0.00	0.00	0.00	0.21	xxx	5

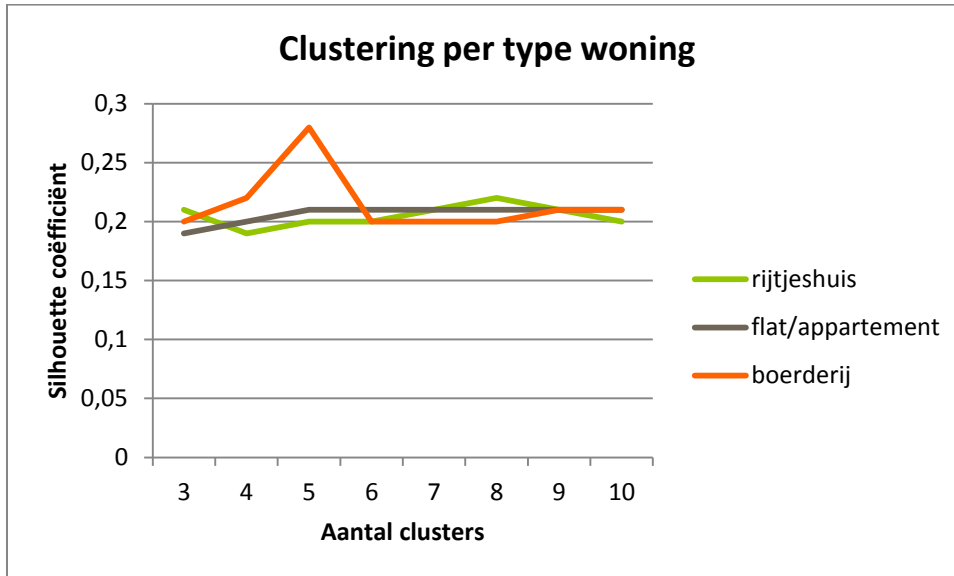
Clustereigenschappen 7-clustering

In Tabel 9 staan van links naar rechts de clusters in afnemende grootte. In de staafdiagrammen is elke staaf het percentage van die eigenschap die in het cluster valt ten opzichte van het aantal huishoudens met die eigenschap in de gehele set. Een hogere staaf van een categorie binnen het cluster, betekent dan ook niet dat die categorie van de variabele het meest voorkomt in het cluster.

Tabel 9: Clustereigenschappen voor de 7-clustering, waarbij 20% ruis is toegelaten. Elke staaf stelt het percentage huishoudens van de gehele dataset voor die in die categorie valt.



Clustering rijtjeshuis, flat/appartement en boerderij



Figuur 50: De SC van de clustering voor rijtjeshuizen en flat/appartement blijft redelijk gelijk voor het verschillend aantal clusters. Bij de boerderij is een duidelijk piek te zien bij 5 clusters. Waarschijnlijk komt dit omdat deze dataset kleiner is.

Tabel 10: p-waardes van de clustering van type 3 woningen: rijtjeshuis (753972 cases). In de meeste clusters bevatten de massa en de potentiële gelijke eigenschappen. Dit is te zien aan het kleine aantal eigenschappen met een p-waarde kleiner dan 0.05.

Elektra	Eigendom	Bouwjaar	Inhoud	Oppervlak	Levensfase	Inkomen	Opleiding	Slecht cluster	Aantal verworpen
1	NA	0.04	NA	NA	0.49	0.04	0.52		2
2	NA	0.00	NA	0.00	0.12	0.02	0.07		3
3	0.00	0.07	0.40	0.01	0.54	0.00	0.55		3
4	NA	0.00	0.00	0.00	0.06	0.00	0.07		4
5	0.00	0.00	0.40	0.49	0.37	0.01	0.54		3
6	0.00	0.01	NA	NA	0.18	0.00	0.14		3
7	0.43	0.51	0.33	0.00	0.00	0.07	0.55	xxx	2

Gas	Eigendom	Bouwjaar	Inhoud	Oppervlak	Levensfase	Inkomen	Opleiding	Slecht cluster	Aantal verworpen
1	NA	0.00	NA	NA	0.25	0.41	0.36		1
2	NA	0.00	NA	0.51	0.25	0.31	0.01		2
3	0.43	0.00	0.00	0.00	0.16	0.45	0.49		3
4	NA	0.00	0.00	0.00	0.00	0.00	0.04	xxx	6
5	0.02	0.00	0.02	0.55	0.50	0.20	0.51		3
6	0.00	0.00	NA	NA	0.49	0.49	0.42		2
7	0.59	0.00	0.56	0.00	0.00	0.46	0.39	xxx	3

Tabel 11: p-waardes van de clustering van type 4 woningen: flat/appartement (289.492 cases). In de meeste clusters bevatten de massa en de potentiële gelijke eigenschappen (p-waarde kleiner dan 0.05). Bij clusters 3-5 verschillen het oppervlak en de levensfase echter waardoor deze als 'slecht cluster' aangemerkt worden voor elektra en clusters 4&5 voor gas.

Elektra	Eigendom	Bouwjaar	Inhoud	Oppervlak	Levensfase	Inkomen	Opleiding	Slecht cluster	Aantal verworpen
1	NA	0.00	0.01	0.28	0.00	0.22	0.33		3
2	0.00	0.11	0.00	0.00	0.12	0.00	0.39		4
3	0.53	0.06	NA	0.00	0.00	0.57	0.49	xxx	2
4	0.38	0.56	NA	0.00	0.00	0.18	0.10	xxx	2
5	NA	0.01	NA	0.00	0.00	NA	NA	xxx	3
6	0.58	0.01	NA	0.46	0.01	0.04	0.22		3
7	0.56	0.02	NA	NA	0.15	0.52	0.52		1

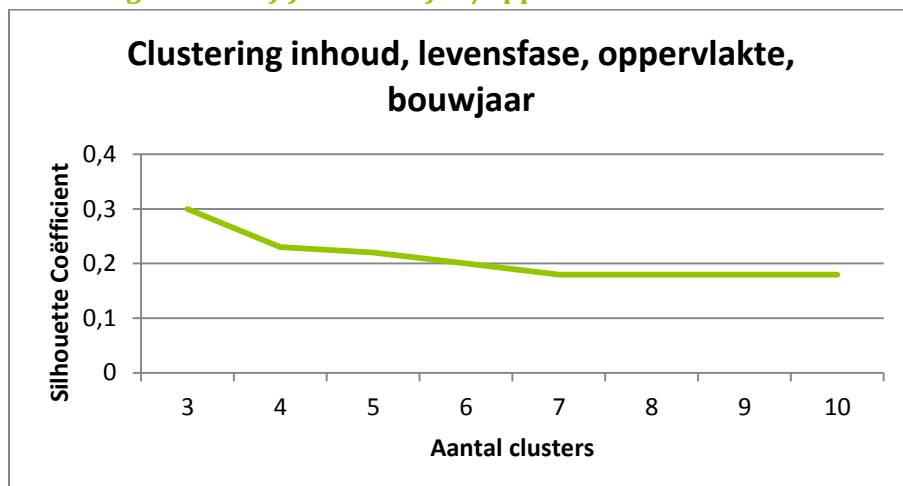
Gas	Eigendom	Bouwjaar	Inhoud	Oppervlak	Levensfase	Inkomen	Opleiding	Slecht cluster	Aantal verworpen
1	0.58	0.44	0.27	0.06	0.20	0.48	0.11		0
2	0.06	0.00	0.00	0.00	0.40	0.00	0.23		4
3	0.14	0.00	NA	0.00	0.31	0.39	0.18		2
4	0.49	0.00	NA	0.03	0.05	0.42	0.54	xxx	3
5	NA	0.00	NA	0.00	0.01	NA	NA	xxx	3
6	0.45	0.00	NA	0.51	0.05	0.35	0.36		2
7	0.39	0.00	NA	NA	0.04	0.06	0.46		2

Tabel 12: p-waardes van de clustering van type 9 woningen: boerderij/tuinderij (14.745 cases). Op cluster 5 voor gas na bevatten de massa en de potentiële gelijke eigenschappen.

Elektra	Eigendom	Bouwjaar	Inhoud	Oppervlak	Levensfase	Inkomen	Opleiding	Slecht cluster	Aantal verworpen
1	NA	0.48	NA	0.46	0.48	0.45	0.43		0
2	0.47	0.58	NA	NA	0.46	0.22	0.40		0
3	NA	0.55	NA	NA	0.44	0.47	NA		0
4	NA	0.54	NA	0.47	0.42	0.58	0.49		0
5	0.17	0.12	NA	0.12	0.56	0.23	NA		0

Gas	Eigendom	Bouwjaar	Inhoud	Oppervlak	Levensfase	Inkomen	Opleiding	Slecht cluster	Aantal verworpen
1	NA	0.35	NA	0.18	0.53	0.20	0.55		0
2	0.47	0.34	NA	NA	0.44	0.16	0.45		0
3	NA	0.30	NA	NA	0.49	0.33	NA		0
4	NA	0.49	0.42	0.51	0.47	0.44	0.52		0
5	0.04	0.01	NA	0.33	0.36	0.54	0.57		2

Clustering zonder rijtjeshuis en flat/appartement



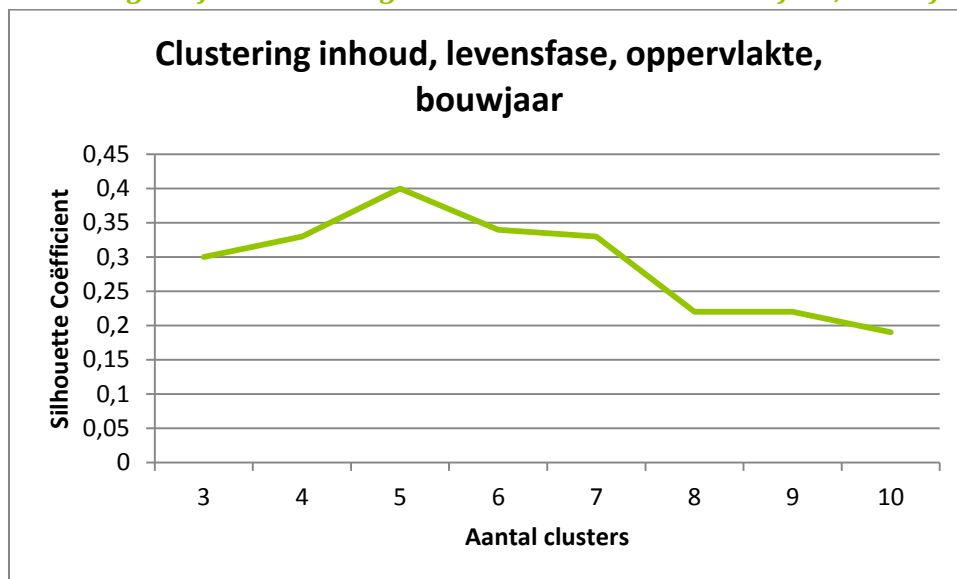
Figuur 51: De Silhouette Coëfficiënt voor een verschillend aantal clusters op basis van inhoud, levensfase, oppervlakte en bouwjaar op een dataset waarbij rijtjeshuizen en flats/appartementen niet zijn meegenomen.

Tabel 13: p-waarden van de clustering zonder rijtjeshuis en flat/appartement. Een clustering met 3 clusters levert een slechte kwaliteit clusters op.

Elektra	Type	Eigendom	Bouwjaar	Inhoud	Opp	Levensfase	Inkomen	Opl	Slecht cluster	Aantal verworpen
1	0.00	NA	0.31	0.01	0.00	0.27	0.00	0.02		5
2	0.47	0.48	0.03	0.00	0.00	0.00	0.00	0.40	xxx	5
3	0.00	0.36	0.01	0.00	0.00	0.01	0.00	0.05	xxx	7

Gas	Type	Eigendom	Bouwjaar	Inhoud	Opp	Levensfase	Inkomen	Opl	Slecht cluster	Aantal verworpen
1	0.00	NA	0.00	0.00	0.00	0.00	0.0	0.00	xxx	7
2	0.00	0.41	0.00	0.00	0.01	0.00	0.0	0.00	xxx	7
3	0.00	0.43	0.00	0.00	0.00	0.00	0.5	0.38	xxx	5

Clustering hoofdkenmerken gas en elektra: inhoud & bouwjaar, levensfase & oppervlakte



Figuur 52: De Silhouette Coëfficiënt van de clustering van de gehele dataset op basis van inhoud, levensfase, oppervlakte en bouwjaar. Een vijftal clusters lijkt het beste aantal te zijn.

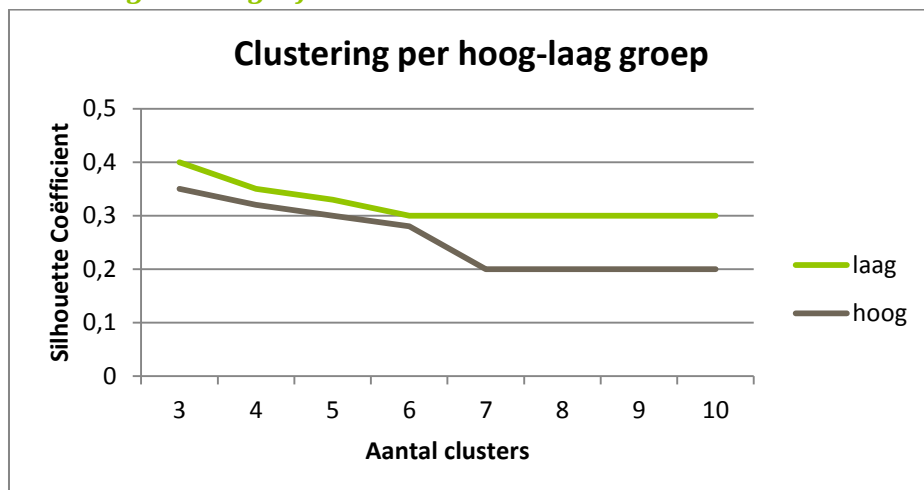
Hierbij worden vijf clusters gekozen. De homogeniteitstest op deze vijf clusters geeft de volgende p-waardes, waarbij de vetgedrukte kolomnamen de variabelen zijn waarop geclusterd is.

Tabel 14: De p-waardes van de clustering van de gehele dataset op basis van de vetgedrukte kenmerken. De meeste clusters zijn van slechte kwaliteit.

Elektra	Type	Eigendom	Bouwjaar	Inhoud	Opp	Levensfase	Inkomen	Opl	Slecht cluster	Aantal verworpen
1	0.00	0.00	0.01	0.03	0.00	0.43	0.00	0.00		7
2	0.00	0.00	0.46	NA	0.01	0.52	0.00	0.07		4
3	0.00	0.09	0.21	0.02	0.00	0.43	0.00	0.01		5
4	0.00	0.00	0.00	NA	0.00	0.00	0.00	0.00	xxx	7
5	0.00	0.00	0.09	NA	NA	0.00	0.00	0.00	xxx	5

Gas	Type	Eigendom	Bouwjaar	Inhoud	Opp	Levensfase	Inkomen	Opl	Slecht cluster	Aantal verworpen
1	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	xxx	7
2	0.00	0.01	0.00	NA	0.00	0.00	0.00	0.00	xxx	7
3	0.00	0.40	0.00	0.00	0.00	0.00	0.00	0.00	xxx	7
4	0.00	0.00	0.00	NA	0.00	0.00	0.01	0.13	xxx	6
5	0.00	0.00	0.00	NA	NA	0.00	0.01	0.01		6

Clustering binnen gelijke sociale kenmerken



Figuur 53: De Silhouette coëfficiënt van de clustering op Hoog en Laag sociale kenmerken.

Tabel 15: De p-waardes van de clustering op sociale eigenschappen 'Hoog' met 3 clusters.

Elektra	Type	Bouwjaar	Inhoud	Opp	Levensfase	Inkomen	Slecht cluster	Aantal verworpen
1	0.00	0.45	0.00	0.00	0.36	0.00		4
2	0.00	0.50	0.00	0.00	0.39	0.00		4
3	0.00	0.03	NA	NA	0.01	0.00		4

Gas	Type	Bouwjaar	Inhoud	Opp	Levensfase	Inkomen	Slecht cluster	Aantal verworpen
1	0.00	0.00	0.00	0.00	0.00	0.00	xxx	6
2	0.00	0.00	0.00	0.00	0.00	0.03	xxx	6
3	0.00	0.00	NA	NA	0.00	0.00		4

Tabel 16: De p-waardes van de clustering op sociale eigenschappen 'Laag' met 3 clusters.

Elektra	Type	Bouwjaar	Inhoud	Opp	Levensfase	Inkomen	Slecht cluster	Aantal verworpen
1	0.00	0.15	0.00	0.00	0.00	0.00	xxx	5
2	NA	0.00	0.02	0.02	0.56	0.02		4
3	0.00	0.00	NA	0.00	0.00	0.02	xxx	5

Gas	Type	Bouwjaar	Inhoud	Opp	Levensfase	Inkomen	Slecht cluster	Aantal verworpen
1	0.00	0.00	0.00	0.00	0.00	0.44	xxx	5
2	NA	0.00	0.00	0.00	0.48	0.13		3
3	0.00	0.00	NA	0.00	0.00	0.47	xxx	4

Tabel 17: De p-waardes van de clustering op sociale eigenschappen 'Hoog' met 7 clusters.

Elektra	Type	Bouwjaar	Inhoud	Opp	Levensfase	Inkomen	Slecht cluster	Aantal verworpen
1	NA	0.48	NA	0.00	0.00	0.00	xxx	3
2	0.00	0.48	0.00	0.00	0.03	0.00	xxx	5
3	0.10	0.11	0.00	NA	0.00	0.00		3
4	0.04	0.01	0.49	0.01	0.07	0.00		4
5	0.00	0.00	0.00	0.00	0.03	0.00	xxx	6
6	0.17	0.37	NA	NA	0.01	0.00		2
7	0.50	0.00	NA	NA	0.01	0.00		3

Gas	Type	Bouwjaar	Inhoud	Opp	Levensfase	Inkomen	Slecht cluster	Aantal verworpen
1	NA	0.00	NA	0.00	0.25	0.00		3
2	0.00	0.00	0.00	0.00	0.00	0.00	xxx	6
3	0.00	0.00	0.00	NA	0.43	0.00		4
4	0.00	0.00	0.21	0.37	0.00	0.00		4
5	0.00	0.00	0.00	0.00	0.23	0.53		4
6	0.00	0.00	NA	NA	0.20	0.00		3
7	0.00	0.00	NA	NA	0.00	0.00		4

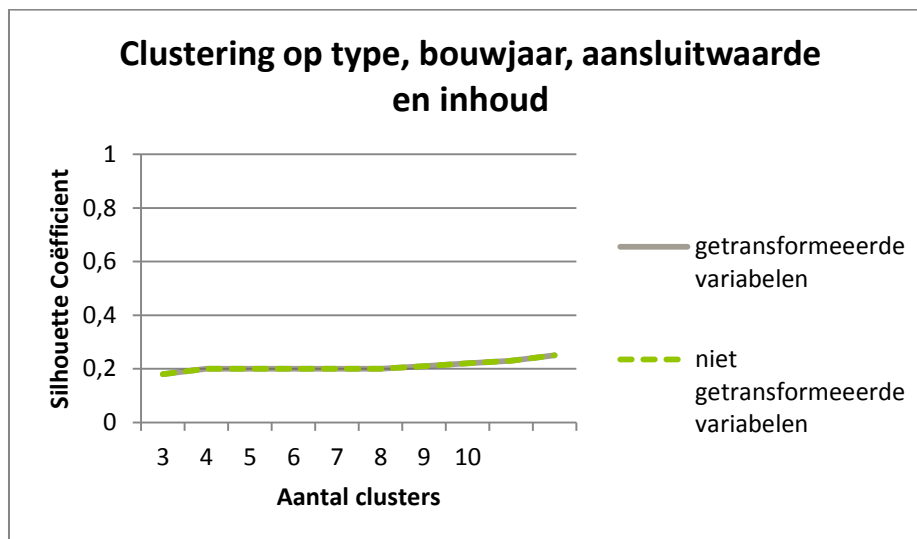
Tabel 18: : De p-waardes van de clustering op sociale eigenschappen 'Laag' met 7 clusters.

Elektra	Type	Bouwjaar	Inhoud	Opp	Levensfase	Inkomen	Slecht cluster	Aantal verworpen
1	0.01	0.02	NA	0.01	0.21	0.00		4
2	0.00	0.00	0.01	NA	0.00	0.17		4
3	NA	0.00	0.02	0.02	0.38	0.02		4
4	0.49	0.30	NA	NA	0.00	0.31		1
5	0.00	0.42	NA	0.00	NA	0.01		3
6	0.27	0.01	NA	NA	0.04	0.45		2
7	NA	0.63	NA	NA	0.00	0.00		2

Gas	Type	Bouwjaar	Inhoud	Opp	Levensfase	Inkomen	Slecht cluster	Aantal verworpen
1	0.00	0.00	NA	0.08	0.49	0.29		2
2	0.00	0.18	0.51	NA	0.38	0.00		2
3	NA	0.00	0.00	0.00	0.43	0.24		3
4	0.21	0.00	NA	NA	0.05	0.01		3
5	0.00	0.00	NA	0.01	NA	0.22		3
6	0.17	0.00	NA	NA	0.03	0.39		2
7	NA	0.00	NA	NA	0.01	0.27		2

Transformeren van ordinale variabelen naar continue variabelen

Uit de analyses in hoofdstuk 4 blijkt dat de factoren die invloed hebben op het gasverbruik, het Type woning, het Oppervlak, het Bouwjaar en de Aansluitwaarde zijn. Op basis van deze vier factoren zijn clusters van vergelijkbare huishoudens gevormd. Op dit moment in het proces werd alleen nog de SC gebruikt om de kwaliteit van de clustering te evalueren en werd de homogeniteit van een cluster nog niet bekeken. De standaard parameters voor het model zijn gebruikt: maxbranch 20, ruis= 0, maxlevel=4. Dit model levert een matige clustering van de data op met een SC van 0,2, zoals in Figuur 54 te zien is. Dit betekent dat de clusters in de dataset niet goed zijn te onderscheiden.³²



Figuur 54: De Silhouette Coëfficiënt van de clustering van de gehele dataset op type, bouwjaar, aansluitwaarde en inhoud. Het transformeren van ordinale variabelen naar continue variabelen levert geen verbetering in de kwaliteit op.

Om dit model te verbeteren is onderzocht of de kwaliteit van het model beter werd als de ordinale variabelen Bouwjaar en Oppervlak getransformeerd werden naar continue variabelen op de schaal 0 tot 1. Zoals in Figuur 54 te zien is, leverde dit geen verbetering van de clusterkwaliteit op. Om deze reden is met niet-getransformeerde data gewerkt.

³² Een model met een SC kleiner dan 0,2 laat geen duidelijke structuur zien.

Bibliografie

- Abdi, H., & Valentin, D. (2007). *Encyclopedia of Measurement and Statistics*. Neil Salkind.
- Abrahamse, W., & Steg, L. (2009). How do socio-demographic and psychological factors relate to households' direct and indirect energy use and savings. *Journal of Economic Psychology*, 711-720.
- Abrahamse, W., Steg, L., Vlek, C., & Rothengatter, T. (2005). A review of intervention studies aimed at household energy conservation. *Journal of Environmental Psychology* 25, 273-291.
- Acquah, H. d.-G. (2010). Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship. *Journal of Development and Agricultural Economics*, Vol. 2(1) pp. 001-006.
- afd. Uitvoering Liander. Henk Brasz.
- AgentschapNL. (2012, 03 11). *EPBD/ Energielabel/ Agentschap NL*. Opgehaald van AgentschapNL Ministerie van Economische Zaken: <http://www.agentschapnl.nl/programmas-regelingen/epbd-energielabel>
- Agresti, & Franklin. (2007). *Statistics, the art and science of learning from data*. Ney Jersey: Pearson Education, Inc.
- Alliander. (2012). *jaarverslag 2012*.
- Alliander. (2013, 01 10). RvB vergaderstuk. Tessa van Doremaele.
- Alliander. *Geschiedenis Alliander*. Opgeroepen op 26-2- 2013, van <http://www.alliander.com/nl/alliander/over-alliander/onze-organisatie/geschiedenis.htm>
- Bacher, J., Wenzig, K., & Vogler, M. (2004). *SPSS Two-Step Cluster - A first evaluation*.
- Bakos, G. (2000). Insulation protection studies for energy saving in residential and tertiary sector. *Energy and Buildings (Elsevier) Volume 31, Issue 3,, 251-259*.
- Bartiaux, F., & Gram-Hanssen, K. (2005). *Socio-political factors influencing household electricity consumption- a comparison between Denmark and Belgium*. ECEEE.
- CBS. (2013, 02 22). *CBS*. Opgehaald van <http://www.cbs.nl>
- CBS. (2013). *Toelichting Wijk-en buurtkaart 2010*. Den Haag: CBS.
- CDA, D66, PvdA, SGP, ChristenUnie, GroenLinks, et al. (2010, 3 17). *Een partijoverstijgend voorstel voor een Deltaplan Nieuwe Energie*. Opgeroepen op 2 27, 2013, van <http://www.duurzaamheidsoverleg.nl/> of <http://www.bngvermogensbeheer.nl/eCache/SUB/68/681.html>
- Duda, R. O., Hart, P. E., & and Stork, D. G. (2001). In *Unsupervised Learning and Clustering, Pattern classification (2nd edition)* (p. 17). New York, NY: Wiley, ISBN 0-471-05669-3.

- Eerste Kamer. *Splitsing van energiebedrijven*. Opgeroepen op 2 26, 2013, van http://www.eerstekamer.nl/wetsvoorstel/30212_splitsing_van
- Fraunhofer-Institute for Systems and Innovation Research; ENERDATA; Institute of Studies for the Integration of Systems ISIS; Technical University (Vienna, Austria); Wuppertal Institute for Climate, Environment and Energy WI;. (2009). *Study on the Energy Savings Potentials in EU Member States, Candidate Countries and EEA Countries for the European Commission Directorate-General Energy and Transport*. Karlsruhe/Grenoble/Rome/Vienna/Wuppertal.
- Gram-Hanssen, K., Kofod, C., & Nærvig Petersen, K. (2004). "Different Everyday Lives – Different Patterns of Electricity.
- IBM SPSS Statistics 21.
- intranet Alliander. *Energietransitie*. Opgeroepen op 2 27, 2013, van <http://kenniscafe.alliander.local/wiki/EnergieTransitie>
- Kamp, H. v. (2013, 2 18). Stand van zaken uitrol slimme meter-kamerbrief. Den Haag: briefkenmerk: DGETM-EM / 13021663.
- Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: John Wiley and Sons.
- Kawamoto, K., Shimoda, Y., & Mizuno, M. (2004, 02 19). Energy saving potential of office equipment power management. *Energy and Buildings*, pp. 915-923.
- Kema N.V. (2010). *Intelligente meters in Nederland*. Arnhem.
- Kuha, J. (2004). AIC and BIC : Comparisons of Assumptions and Performance. *Sociological Methods & Research*, 188-229.
- Liander. (23 februari 2012). *Energiebesparingspotentieel-Van inzicht naar actie*.
- Majcan, D., Itard, L., & Visscher, H. (2013, januari). Energielabels en werkelijk energieverbruik. *TVVL Magazine*, pp. 4-9.
- Mirkin, B. (2005). *Clustering: A data recovery approach*. Boca Raton: Chapman & Hall.
- Netbeheer-Nederland, ECN, & Energie-Nederland. (2012). *Energietrends 2012*.
- Nibud. (2009). *Energielastenschouwing: Verschillen in energielasten tussen huishoudens nader onderzocht*. Utrecht.
- NMa. *Energiewetten*. Opgeroepen op 2 2013, 26, van http://www.nma.nl/wet__en_regelgeving/energiewetten/default.aspx
- NMa. *Regulering regionale netbeheerders*. Opgeroepen op 2 22, 2012, van [www.nma.nl: http://www.nma.nl/regulering/energie/elektriciteit/regulering_regionale_netbeheerders/default.aspx](http://www.nma.nl/regulering/energie/elektriciteit/regulering_regionale_netbeheerders/default.aspx)

- Overheid. *Overheid* Opgeroepen op 2 26, 2013, van Wet van 23 november 2006:
http://wetten.overheid.nl/BWBR0020608/geldigheidsdatum_26-02-2013
- Rijksoverheid. (2011, 2). De slimme meter. Den Haag, Publicatienummer: 13PD2011G005.
- Rijksoverheid. *Europa 2020 | Europese Unie*. Opgeroepen op 2 27, 2013, van
<http://www.rijksoverheid.nl/onderwerpen/europese-unie/europa-2020>
- Rijksoverheid. *Wat regelt de Wbp?* Opgeroepen op 2 26, 2013, van www.rijksoverheid.nl:
<http://www.rijksoverheid.nl/onderwerpen/persoonsgegevens/vraag-en-antwoord/wat-regelt-de-wet-bescherming-persoonsgegevens-wbp.html>
- stuurgroep 'Vorbereiding GSA'. (2012). *Programmaplan voorbereiding GSA*.
- Timmermans. Bisnode, persoonlijk contact.
- TU Delft, & Stimuleringsfonds Volkshuisvesting. (2009). *Perspectieven voor energiebesparing in de particuliere woningvoorraad*.
- Velthuis, C., & van Doremaele, T. (2012). *Trendanalyse Alliander*.
- Wikipedia*. Opgeroepen op 2 26, 2013, van Wikipedia:
http://nl.wikipedia.org/wiki/Duurzame_energie
- Woonwebsite*. Opgehaald van <http://www.woonwebsite.nl/topics/encyclopedie-verwarming/21>
- Zhang, T., Ramakrishnon, R., & and Livny, M. (1996). BIRCH: An efficient data clustering method for very large databases. *Proceedings of the ACM SIGMOD Conference on Management of Data*. Montreal, Canada: ACM.