

Energieverbruik in de gemeente Amsterdam

Demi van Dijk
Marlon de Jong
Mathias Kirkeng
Nina Verheijen

28 juni 2018

1 Inleiding

Energiebedrijven maken tegenwoordig steeds meer gebruik van data die verschillende eigenschappen in kaart brengen over het energieverbruik van een bepaald gebied of huishouden. Dit gebruiken ze om hun verdienmodel te maximaliseren. Uit deze data kan men bijvoorbeeld aflezen van welk soort energie er gebruik wordt gemaakt, hoeveel gebruik er van gas of stroom wordt gemaakt en hoeveel aansluitingen er zijn op een stroom- of gasnet. Dit is belangrijke informatie, want hierdoor kan de levering van stroom of gas beter op de wensen van de klant worden afgestemd.

Dit onderzoek richt zich op het stroom- en gasverbruik in de gemeente Amsterdam. Hoofdzakelijk richt het onderzoek zich op de vraag of er interessante informatie uit de data kan worden gehaald van de netbeheerder Liander. Verwacht wordt dat er voornamelijk voor relatief grote gebieden interessante informatie is te vinden, dit omdat de datasets zijn opgedeeld in kleine groepjes van postcodes. Wanneer blijkt dat deze data bruikbaar is wordt gekeken naar het energieverbruik in bepaalde periodes. Gepoogd wordt om deze informatie te koppelen aan een bepaalde gebeurtenis, bijvoorbeeld *bitcoin mining*. Dit zal een uitdaging vormen, aangezien het lastig is om een causaal verband aan te tonen. Daarnaast geven de datasets alleen inzicht in het energieverbruik per jaar en niet per maand of dag. Hierdoor zijn er wellicht te weinig datapunten om te koppelen aan een gebeurtenis als bitcoin mining. Ook wordt er gekeken naar ongewone patronen die uit de datasets zijn te halen. Hierbij is er ook grote kans dat er specifieke gebeurtenissen aan te koppelen zijn. Zo niet, dan kan er naar een verklaring worden gezocht hoe het komt dat er bijvoorbeeld uitschieters zijn in sommige gebieden of periodes. Ten slotte wordt het energieverbruik per gebied in Amsterdam nog geanalyseerd. Ook kunnen uit deze gegevens een aantal conclusies worden getrokken. De verwachting is dat welvarende wijken over het algemeen meer energie verbruiken dan arme wijken.

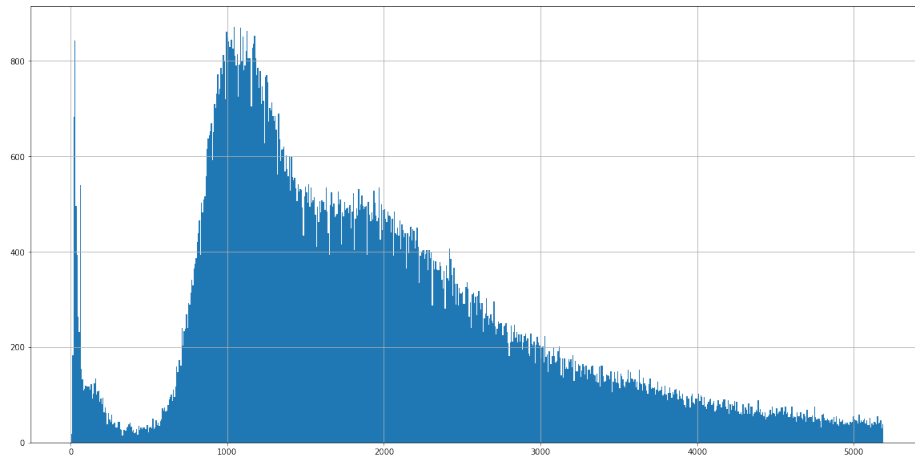
Wanneer er een antwoord kan worden gevonden op deze vragen, zal de technologie waarvan gebruik wordt gemaakt bij het leveren van stroom zich ook verder kunnen ontwikkelen. Men kan hierdoor ook zorgen voor steeds meer duurzaam energieverbruik, wat het milieu weer ten goede komt.

2 Methode

De datasets van Liander van 2011 tot en met 2017 worden gebruikt voor dit project. De dataset van 2018 is niet compleet en zal dus niet gebruikt worden. De datasets van 2011 tot en met 2017 staan in excel bestanden. Om bruikbare datasets te hebben worden de excel bestanden omgezet naar CSV-bestanden. Ten tweede worden alle irrelevante kolommen verwijderd. Deze kolommen zijn: "NETGEBIED", "LANDCODE", "Definitieve aansl (NRM)". De kolom "LANDCODE" is verwijderd omdat deze voor alle nodige data hetzelfde is. Omdat Amsterdam hetzelfde "NETGEBIED" heeft was deze kolom niet relevant. "Definitieve aansl (NRM)" had overal dezelfde waarde en voegde hierdoor niets toe aan de dataset. Verder moeten alle rijen die niet de waarde AMSTERDAM als woonplaats hebben verwijderd worden. Hierna kan ook de kolom "WOONPLAATS" worden verwijderd. Om een completere dataset te maken wordt de dataset opgedeeld in stadsdelen. Amsterdam heeft acht stadsdelen. Ook worden de inwonersaantallen per stadsdeel toegevoegd aan de dataset. Hiermee kan de data worden gerelativeerd.

Om de dataset verder "schoon" te maken moeten de *outliers* worden verwijderd. Outliers worden gevonden middels box- en whisker plot. De set wordt eerst geordend. Vervolgens wordt de mediaan, het middelpunt, van de set gevonden ($Q2$). De mediaan van het deel van de set onder $Q2$ is $Q1$. De mediaan van het deel van de set boven $Q2$ is $Q3$. De minimale waarde is de laagste waarde van de set. De maximale waarde is de hoogste waarde van de set. De 50 procent in het midden, de box ofwel de IQR , is $Q3 - Q1$. Om te weten hoe groot je fences zijn bereken je $1.5 * IQR$. Alle datapunten die kleiner zijn dan $Q1 + (1.5 * IQR)$ (je fence) zijn outliers. Alle datapunten groter dan $Q3 + (1.5 * IQR)$ (je fence) zijn outliers.

Bij het plotten van de histogram viel het op dat er een kleine piek is tot 450. Alle datapunten onder 450 zijn niet reële datapunten, het is immers zeer onwaarschijnlijk dat een verzameling van huizen minder dan 450 kilowattuur gemiddeld per jaar gebruikt. Bij het plotten van de histogrammen van de jaren 2011 tot en met 2017 is overal deze kleine piek te zien, dus de ondergrens wordt op 450 gezet.



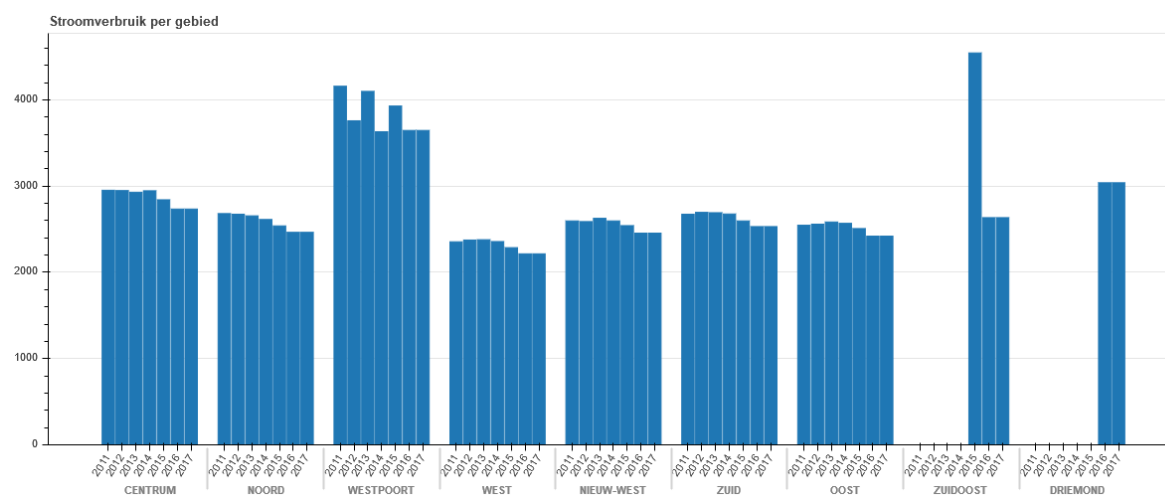
Figuur 1: piek onder 450

Vervolgens is de dataset klaar voor *exploratory data analysis*. Er zijn vier categorieën waarvoor minimaal twee soorten plots worden gemaakt als het mogelijk is. De categorieën zijn: *univariate graphical*, *multivariate graphical*, *univariate non-graphical*, en *multivariate non-graphical*. Aan de hand van de gemaakte plots kan een deel van de vragen worden beantwoord. Niet elke categorie is geschikt voor het beantwoorden van elke vraag. Bijvoorbeeld als men de stroomrichting in kaart wil brengen is *univariate graphical* in eerste opzicht geschikt, maar als men de data bekijkt ziet men dat minder dan 3 procent van de data onder de 100 procent ligt. Deze data is dus niet geschikt voor een histogram, maar wel voor technieken die passen bij *univariate non graphical*.

Om de data analyse af te sluiten hoort er *in-depth analysis* plaats te vinden. Met andere woorden wordt toekomstige data voorspeld aan de hand van de dataset. Dit kan gedaan worden middels algoritmes, bijvoorbeeld k-means. Echter zijn de beschikbare algoritmes beperkt en is de dataset niet geschikt voor de meeste aanpakken.

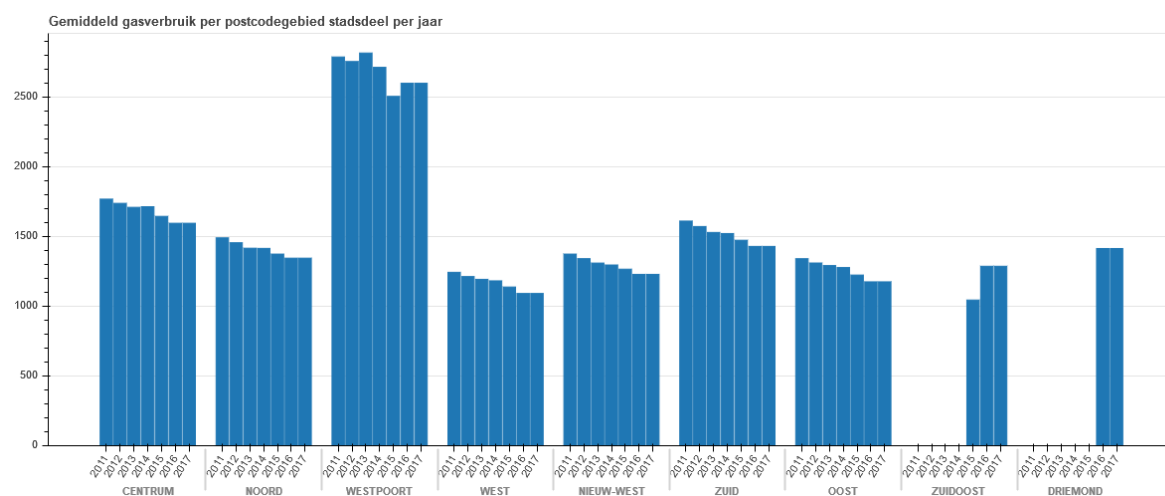
3 Resultaten

Het gemiddelde stroomverbruik per postcodegebied, per stadsdeel, per jaar, over de jaren 2011 tot en met 2017.



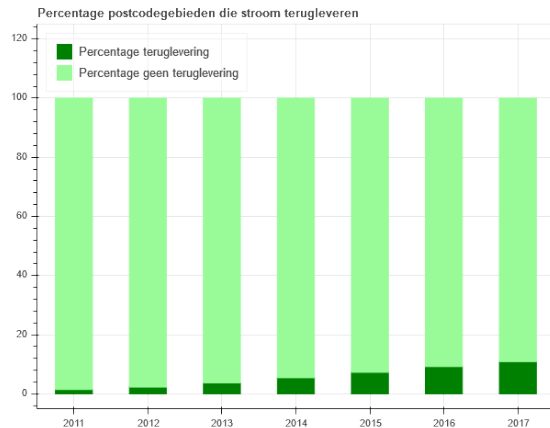
Figuur 2: gemiddeld stroomverbruik per stadsdeel

Het gemiddelde gasverbruik per postcodegebied, per stadsdeel, per jaar, over de jaren 2011 tot en met 2017.



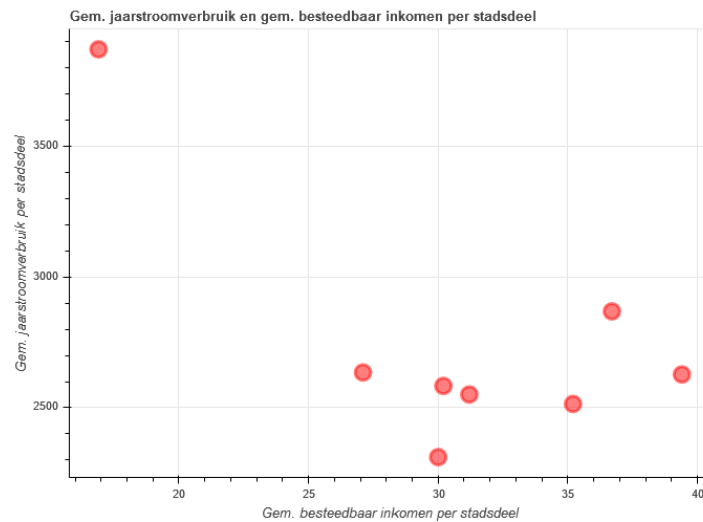
Figuur 3: gemiddeld gasverbruik per stadsdeel

De leveringsrichting geeft aan of er ook groene stroom van gebruikers wordt geleverd. Dit is een indicatie van duurzaamheid.



Figuur 4: leveringsrichting percentage

Op de x-as staat het gemiddelde inkomen. Op de y-as staat het verbruik per postcodegebied.



Figuur 5: inkomen afgezet tegen stroomverbruik

4 Conclusie

Doordat de dataset relatief beperkt bleek te zijn was het merendeel van de gestelde vragen niet of niet volledig te beantwoorden. Zo was de vraag of mensen die meer stroom gebruiken ook meer gas gebruiken niet te beantwoorden met de beperkte dataset van Liander. De datapunten bestaan uit meerdere huishoudens die zijn samengevoegd, dit is gedaan uit privacy overwegingen. Omdat niet ieder huishouden zowel gas als elektra gebruikt zijn de clusters van postcodes ook verschillend. Hierdoor kan men niet vergelijken of mensen die relatief veel stroom gebruiken ook relatief veel gas gebruiken. Ook de vraag of evenementen zoals Bitcoin mining of koude/warme winters kunnen worden gevonden binnen de data bleek onmogelijk om te beantwoorden. De data geeft namelijk alleen het gemiddelde energieverbruik over een jaar. De winter data is verspreid over twee jaren. Hierdoor is het niet mogelijk om een uitspraak te doen over warme of koude winters. Het is immers zo dat als een koude winter een milde winter opvolgt het gemiddeld een normale winter is qua verbruik. Verder bleek mining van cryptocurrency te recent om data over te vinden. Als hier wel een causaal verband was gevonden was het echter nog steeds niet mogelijk om met zekerheid te zeggen dat de gebeurtenis door bitcoin mining was veroorzaakt.

De andere vragen konden wel beantwoord worden, al was er soms nog steeds last van ruis door de incomplete dataset. De vraag of er verschil was tussen energieverbruik in wijken binnen Amsterdam kan bevestigend worden beantwoord. Figuur 2 en 3 tonen aan dat de datapunten van westpoort overal bovenuit steken. Dit is opvallend, omdat Westpoort de minste inwoners heeft. Men kan dus verwachten dat het stroom- en gasverbruik lager is. Echter gaat de data van Liander over kleingebruikers, onder kleinverbruikers vallen ook kleine bedrijven. Als Westpoort relatief veel kleine bedrijven heeft verklaart dat waarom Westpoort vaak afwijkt. Verder is er gekeken of wijken waar het inkomen gemiddeld hoger lag ook meer consumeerde in vergelijking met wijken waar het inkomen gemiddeld lager ligt. Uit figuur 5 kan men concluderen dat er geen verband is tussen stroomverbruik en inkomen. De consumptie van het stadsdeel met het hoogste inkomen had niet een significant hoger consumptie niveau. Bij deze conclusie is Westpoort buiten beschouwing gehouden vanwege eerder genoemde reden. Tenslotte is er gekeken of de toenemende interesse in duurzaamheid terug te zien is in de data van liander. Uit figuur 4 kan men aflezen dat er een stijgende trend is met betrekking tot terugleveringsrichting. Dit kan veroorzaakt zijn door zonne- en windenergie, maar kan ook andere oorzaken hebben. Of dit het gevolg van de interesse in duurzaamheid is kan niet causaal worden aangetoond.

5 Discussie

Op de meeste vragen is geen duidelijk of bevestigend antwoord gevonden. Dit komt voornamelijk doordat de gebruikte datasets incompleet en beperkt zijn. De informatie die nodig is zoals het stroomverbruik per maand of per dag ontbreekt. Daarnaast lopen de kolommen die informatie bevatten over het gas- en stroomverbruik door elkaar. De gegeven dataset was voor het beantwoorden van de meeste vragen nietszeggend, dit omdat bij veel kolommen de waarden hetzelfde zijn. De datasets waren ook niet specifiek genoeg om de locaties te kunnen bepalen als het om gebieden ging in steden. Dit komt door de postcodes die in groepjes zijn samengevoegd in de datasets. Dit zorgt ook voor problemen, omdat postcodegebieden bij sommige jaren zijn aangepast. Hierdoor kunnen bepaalde postcodes van gebied verwisselen en dat zorgt ervoor dat het bepalen van een vast gebied moeilijker is. Ten slotte is er te weinig data over het aantal slimme meters en het leveringsvermogen om een regressielijn te kunnen maken waar een patroon/structuur uit is te halen. Een oplossing voor al deze problemen zou het gebruiken van een andere dataset (van een andere leverancier) kunnen zijn. Wellicht hebben andere energiebedrijven completere informatie in hun datasets opgenomen om het gedrag van de klant beter te kunnen observeren en voorspellen en dus de levering van stroom en gas beter op de klant te kunnen afstemmen.

6 Literatuurlijst

1. [Datasets van Liander]. (z.d.). Geraadpleegd op 26 juni 2018, van <https://www.liander.nl/partners/datadiensten/open-data/data>
2. [dataset stadsdelen]. (2017, 1 januari). Geraadpleegd op 20 juni 2018, van <https://www.ois.amsterdam.nl/download/01b-kerncijfers-22-gebieden-en-stadsdelen-1-januari-2017>
3. [hoe een huis geïsoleerd is]. (z.d.). Geraadpleegd op 26 juni 2018, van <https://www.energielabel.nl/>