mondaine Proof of Concept

Deliverable D3.2

Methodiek datakwaliteit

Dit project is mede gefinancierd door TKI-Energie uit de Toeslag voor Topconsortia voor Kennis en Innovatie (TKI's) van het ministerie van Economische Zaken en Klimaat.

Auteurs	Vincent Kamphuis (TNO), Joost Laarakkers (TNO), Edwin Matthijssen (TNO)
Versie	v1.0
Datum	26-02-2020
Reviewers	Roos de Kok (Quintel), Chael Kruip (Quintel), Arendjan van der Neut (Geodan), Ewoud Werkman (TNO)
Project	MOdels aNd DAta INterfaces for Energy – Proof of Concept
Referentie	1721401

QUINTEL INTELLIGENCE Geodan TNO

Inhoud

1. Inleiding

Goede beslissingen in de energietransitie vragen om betrouwbare inzichten, die onderbouwd worden door betrouwbare informatie. Een deel van deze informatie wordt afgeleid uit data. Deze wordt ofwel verzameld ofwel berekend door rekenmodellen. De data komt van veel verschillende aanbieders en bronnen en verschilt sterk in kwaliteit.

In het Mondaine Proof of Concept projectplan wordt benoemd dat er aandacht besteed moet worden aan het borgen van de kwaliteit van de groeiende hoeveelheid energie gerelateerde open data. Dit document geeft een aantal handvatten hoe dit gedaan kan worden.

Een kwalitatief goede dataset voldoet aan de volgende criteria:

- De data is goed beschreven
- De data is volledig
- De data beschrijft de werkelijkheid accuraat.

In de praktijk is data echter niet altijd even goed beschreven, soms onvolledig, of vertoont deze inconsistenties en/of bevat deze datapunten of statistische parameters die niet goed verklaarbaar of foutief zijn. Zo is energiedata op een laag schaalniveau niet beschikbaar bij onvoldoende aansluitingen binnen een gebied, worden aansluitingen op transportnetten niet altijd meegenomen in energieverbruiksgegevens, en zijn indelingen naar verbruikers categorieën niet altijd eenduidig. Zo vallen corporatiewoningen in Energie in Beeld onder zakelijk en niet onder particulier. Hierdoor kan de data door de gebruiker verkeerd geïnterpreteerd of gebruikt worden, wat kan leiden tot verkeerde conclusies en beslissingen binnen de energietransitie.

Het is dus wenselijk om meer inzicht te krijgen in de kwaliteit van deze energietransitie data. Naarmate modellen steeds meer data (automatisch) combineren en verrijken wordt de kwaliteit van deze data en het beheersen daarvan steeds belangrijker. Betere kwaliteit van inputdata kan ook de uitkomsten van modellen verbeteren, waardoor betere beslissingen kunnen worden genomen binnen de energietransitie. Alleen de vraag stellen “wat is de kwaliteit van deze data?” kan al helpen, maar biedt onvoldoende richting om de vraag ook daadwerkelijk te beantwoorden. Daarom is er in het Mondaine Proof of Concept project een methodiek ontwikkeld die inzicht biedt in de kwaliteit van data. De methodiek signaleert inconsistenties binnen een dataset en probeert ook zo veel mogelijk inconsistenties tussen datasets onderling te ontdekken. De methodiek geeft tevens inzicht in de statistische verdeling voor de indicatoren binnen een dataset. Randvoorwaarde voor toepassing van deze methodiek is dat de data op een goede manier beschreven is. De methodiek kan toegepast worden op alle soorten data. Zoals gemeten data (bijv. energieverbruiken), kengetallen (bijv. kosten van verschillende soorten warmtepompen) en berekende data (bijv. uitkomsten uit een model). Verder betreft het zowel open data als gesloten data. De methodiek dient gebruikt te kunnen worden door zowel de leveranciers van data als de gebruikers van de data. Voor leveranciers van data geeft het zekerheid over de kwaliteit voor publicatie van de data. Gebruikers kunnen het gebruiken als checklist om inzicht te krijgen in de kwaliteit van de data. Het verlaagt daarmee het risico dat er bijvoorbeeld fouten gemaakt worden door verkeerde interpretatie of verkeerd gebruik van de data.

Omdat er veel verschillende datasets en ook grote datasets bestaan, zou het wenselijk zijn als (een deel van) de methodiek op een bepaalde manier geautomatiseerd kan worden. Ook het feit dat sommige datasets periodiek geüpdatet worden, waardoor de kwaliteit iedere keer getoetst dient te worden, versterkt deze wens.

Zowel voor het eenduidig beschrijven als voor het automatiseren lijkt het gebruik van de Energy System Description Language (ESDL[1]) een goede keuze. ESDL maakt het mogelijk om datasets op een eenduidige manier te beschrijven en mede daardoor “automatisch” consistentie checks uit te voeren. Dit dient in de toekomst verder verkend te worden. Mogelijk geeft het TKI Mondaine project (de opvolger van dit project) hier mogelijkheden toe.

De methodiek wordt toegelicht in hoofdstuk 2. In hoofdstuk 3 wordt een Proof of Concept (PoC) beschreven aan de hand van drie voorbeelden: (1) een consistentiecheck tussen uitkomsten van het Energietransitiemodel (ETM) van Quintel en het Vesta model van PBL, (2) consistentie- en statistische checks op de uitkomsten van de Energiepotentieelscan voor Bedrijventerreinen van TNO en (3) consistentie checks op en tussen verschillende energieverbruik datasets. De conclusies en aanbevelingen worden gegeven in hoofdstuk 4.

[1] https://www.tno.nl/nl/aandachtsgebieden/informatie-communicatie-technologie/expertisegroepen/monitoring-control-services/grip-op-de-energietransitie-met-esdl/

2. Methodiek

De dataset kwaliteit methodiek signaleert inconsistenties binnen een dataset en inconsistenties tussen datasets. Tevens geeft de methodiek inzicht in de statistische verdeling voor de indicatoren binnen een dataset. Op basis hiervan kan, in de context van de toepassing van de dataset, de kwaliteit van een dataset bepaald worden. Randvoorwaarde voor succesvolle toepassing van de methodiek is een goede beschrijving van de datasets. Deze 3 aspecten worden in dit hoofdstuk beschreven. Vervolgens wordt de relatie tussen modelkwaliteit en datakwaliteit beschreven. De methodiek wordt in de appendix tevens als checklist beschreven. Deze kan gebruikt worden door de dataeigenaar of gebruiker om inzicht te krijgen in de kwaliteit van een dataset.

Datasets zijn goed beschreven

Een goede beschrijving van een dataset voldoet aan de volgende voorwaarden:

1. De data is goed gedocumenteerd; alles is helder en eenduidig gedefinieerd, eenheden worden gegeven. Dit geldt voor alle parameters en variabelen binnen een dataset.
2. De verzamelmethode voor ruwe data is duidelijk beschreven. Voor metingen wordt het tijdstip of de meetperiode gegeven.
3. De aannames die gedaan zijn voor verrijking van ruwe data zijn duidelijk beschreven. Denk aan interpolatie, interpretatie en andere bewerkingen.
4. Indien de data uitkomsten van een rekenmodel betreft, wordt verwezen naar een duidelijke beschrijving van de gevolgde rekenmethodiek. Tevens wordt de modelrun dusdanig beschreven dat deze reproduceerbaar is.
5. Publicatiedatum, eigendom, contactgegevens, gebruiksrecht en licentievoorwaarden worden gegeven.
6. Eventuele onvolledigheden worden expliciet toegelicht.
7. De scoping wordt aangegeven: voor welke toepassingen is de data wel en niet bruikbaar?
8. Periode van geldigheid wordt gegeven: tot wanneer is de data te gebruiken/bruikbaar.
9. Onderhoudscyclus wordt gegeven: wordt de data onderhouden? Met welke cyclus?

De verantwoordelijkheid voor een goede beschrijving van een dataset ligt bij de eigenaar van de dataset.

Datasets zijn (onderling) consistent

Het is van belang dat datasets (onderling) consistent zijn, zeker als modellen deze gaan gebruiken en combineren. Voorbeelden van consistenties zijn:

- Aggregatie
- Energiebalans
- Vergelijkbaarheid

Consistentie onder aggregatie betekent dat data op een lager aggregatieniveau optelt tot data op een hoger aggregatieniveau. Dit kan ruimtelijke (geografische) aggregatie zijn, maar ook aggregatie over (sub)sectoren, assetgroepen of energiedragers. Bijvoorbeeld:

- Ruimtelijk: van Postcode 6 (PC6) naar gemeente
- Sectoren: van subsector naar hoofdsector (van huizen tot gebouwde omgeving)
- Assets: van individuele warmtebronnen naar totaal warmte aanbod
- Energiedragers: van warmte, gas en elektriciteit naar totale energievraag

Consistentie in de energiebalans representeert de wet van behoud van energie. Deze geldt (uiteraard) zowel voor individuele assets als voor een gebied. Bijvoorbeeld:

- Voor een gebied: productie – verlies + import - export = energieverbruik + opslag + conversie
- Voor een elektrisch warmtepomp: warmteproductie = COP x elektriciteitsverbruik

Tot slot is het wenselijk dat de waardes, correlaties en spreidingen voor indicatoren die in verschillende datasets voorkomen vergelijkbaar zijn, of dat verschillen verklaard kunnen worden door een verschil in uitgangspunten of definities.

Om deze (in)consistenties op een "automatische" manier inzichtelijk te maken, kan ESDL gebruikt worden. Met deze taal kan energietransitie data op een eenduidige manier worden beschreven en onderdeel gemaakt worden van een energiesysteem (*"de data wordt in z'n context geplaatst"*). Hierdoor is het vervolgens gemakkelijker om op een ESDL-versie van een dataset bovengenoemde consistentie aspecten te controleren (zoals bijvoorbeeld aggregaties). Let op: de methodiek biedt een signaleringsfunctie en geeft richting aan mogelijke oplossingen voor inconsistenties. Het lost inconsistenties niet op.

Zowel de vertaling van variabelen en parameters in een dataset naar eenduidige definities in ESDL, als het maken van scripts om de conversie van grote datasets naar ESDL eenvoudiger te maken, kan gedaan worden door de eigenaren van data en modellen.

Statistische analyse van datasets

Een andere manier om meer inzicht te krijgen in een dataset, en daarmee ook in de kwaliteit van data, is een statistische analyse. Belangrijke indicatoren waarvoor een statistische analyse meer inzicht kan geven in de kwaliteit van een dataset zijn bijvoorbeeld:

- Energieverbruik per m2
- Energieverbruik per inwoner
- Energieverbruik per FTE (voor bedrijven)
- Energieverbruik per kilo product (voor bedrijven)
- Investeringskosten per m2
- CO2-uitstoot per m2

Het is belangrijk om de statische verdeling voor deze indicatoren te verklaren en waar mogelijk te vergelijken met beschikbare kengetallen. Deze verantwoordelijkheid ligt bij de eigenaar van de data, maar kan wel als sanity check door de modeleigenaar gebruikt worden voor de data te accepteren.

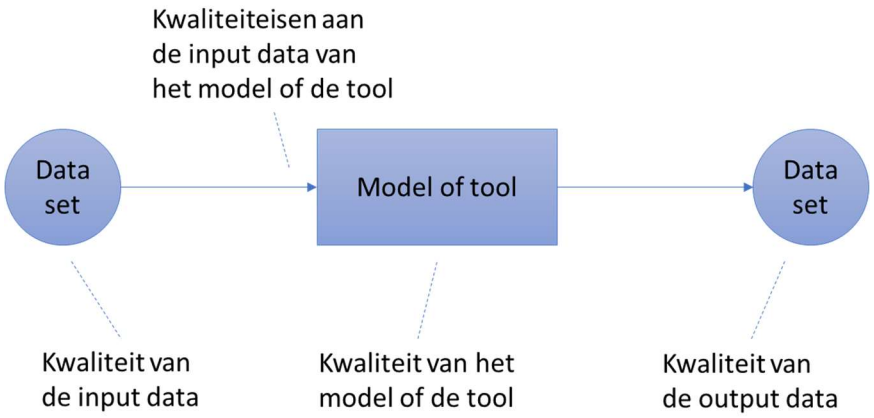
Het kan meerwaarde hebben om de statistische parameters ook in ESDL te definiëren, als onderdeel van de ESDL-beschrijving van een dataset. Hier zijn al de eerste experimenten mee gedaan; er zijn objecten waarmee bijv. gemiddelden en standaarddeviaties van data beschreven kunnen worden. Dit wordt op dit moment echter niet gebruikt. Indien er hier serieuze behoefte naar ontstaat, zou daar meer aandacht aan besteed moeten worden.

Relatie tussen datakwaliteit en modelkwaliteit

De kwaliteit van een dataset die output is van een rekenmodel hangt af van de kwaliteit van het rekenmodel én de kwaliteit van de gebruikte inputdata. De verschillende interfaces die hierbij een rol spelen zijn weergegeven in onderstaand figuur:

- Natuurlijk ten eerste de kwaliteit van de input datasets. De verantwoordelijkheid voor deze kwaliteit ligt bij de producent van deze data.
- Het model (of de modeleigenaar) moet daarna wel een basis kwaliteitscheck doen om de verifiëren dat deze data aan de vereiste input requirements van het model voldoet.

- De kwaliteit van het model zelf, is natuurlijk de verantwoordelijkheid van de modelleverancier.
- Net zoals de kwaliteit van de data die het model produceert. Die kwaliteit is een combinatie zijn van model kwaliteit en bron datakwaliteit.



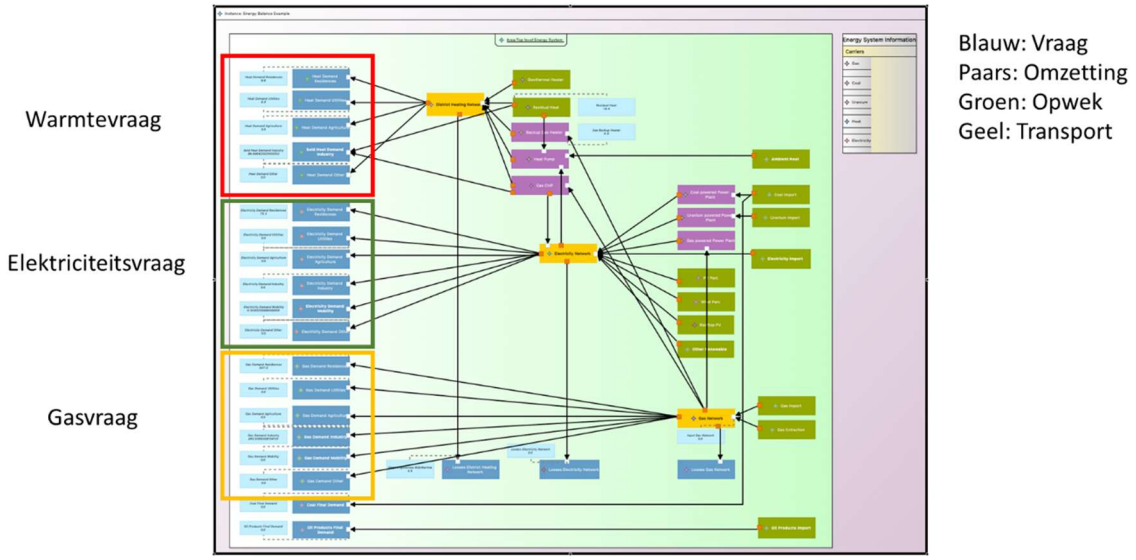
De geschetste methodiek kan gebruikt worden voor zowel de input data voor als de output data van een model of tool, en zo inzicht geven in de kwaliteit van deze data. De leverancier van de deze data kan de kwaliteit duiden in een breed toepassingsperspectief voor de data, de gebruiker kan de kwaliteit duiden binnen de scope van het model. Merk op dat de output data van een model ook weer als input gebruikt kan worden voor andere modellen en toepassingen.

3. Proof of concept

Het Proof of Concept (PoC) voor de methodiek is uitgewerkt aan de hand van drie voorbeelden: (1) een consistentiecheck tussen uitkomsten van het Energietransitiemodel (ETM) van Quintel en het Vesta model van PBL, (2) consistentie- en statische checks op de uitkomsten van de Energiepotentieelscan voor Bedrijventerreinen van TNO en (3) consistentie checks op en tussen verschillende energieverbruik datasets.

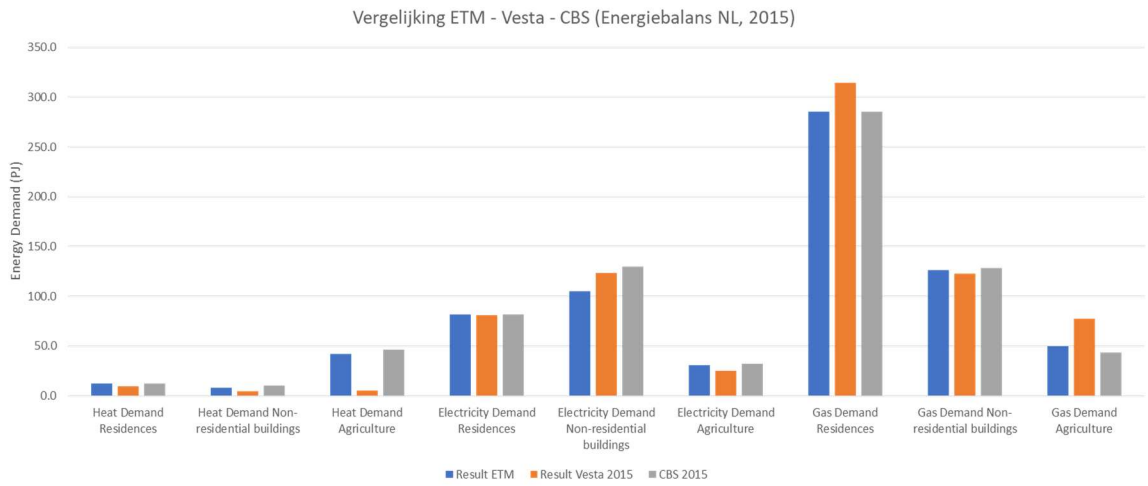
Energiebalans Nederland

Het ETM en Vesta bevatten beiden een energiebalans voor Nederland als uitgangssituatie. Om deze met elkaar te kunnen vergelijken zijn beide energiebalansen gemapped op een ESDL-configuratie voor de energiebalans van Nederland. Deze configuratie is weergegeven in Figuur 1. Tevens is de CBS energiebalans op deze configuratie gemapped.



Figuur 1. Visuele weergave van de ESDL-configuratie van de Nederlandse energiebalans.

De waardes voor de (directe) warmtevraag, de elektriciteitsvraag en de gasvraag in 2015 in ETM, Vesta en de CBS energiebalans zijn vergeleken voor woningen, utiliteitsbouw en landbouw in Figuur 2. De getallen zijn vergelijkbaar (maar niet gelijk) voor de meeste indicatoren, maar wijken vooral voor de (directe) warmtevraag en de gasvraag van de landbouw sterk af. Deze afwijkingen zijn significant en kunnen niet vanuit nauwkeurigheid verklaard worden. Een mogelijke verklaring voor dit verschil is dat de gasvraag van WKK's als (directe) warmtevraag is meegenomen in het ETM en de CBS energiebalans en als gasvraag in Vesta. Het is belangrijk deze verschillen te verklaren en eventuele onjuistheden op te laten lossen. Zo ontstaat meer inzicht in de betrouwbaarheid van de data en de modeluitkomsten. Het oplossen van eventuele onjuistheden kan de kwaliteit van de data verhogen.



Figuur 2. Vergelijking van de (directe) warmte-, elektriciteit- en gasvraag voor woningen, utiliteit en landbouw in de modellen ETM en Vesta en de CBS energiebalans in 2015.

Energiepotentieelscan voor Bedrijventerreinen (EPS)

De Energiepotentieelscan voor Bedrijventerreinen (EPS) geeft een eerste orde inschatting van de business case voor energiemaatregelen op bedrijventerreinen, zowel voor individuele ondernemers als voor een bedrijventerrein als geheel. Deze is inmiddels succesvol toegepast op ruim 50 bedrijventerreinen in Nederland. Daarmee zitten er inmiddels ruim 8000 bedrijven en ruim 3000 panden in de rekendatabase.

Op deze dataset zijn de volgende consistentie checks succesvol uitgevoerd:

- Gasverbruik gebouwgebonden + gasverbruik industriële processen = gasverbruik totaal
- Elektriciteitsgebruik niet-proces + elektriciteitsgebruik industriële processen = elektriciteitsgebruik totaal
- Gasverbruik + elektriciteitsverbruik + warmteverbruik = energieverbruik
- Totaal energieverbruik bedrijven = totaal energieverbruik bedrijventerreinen

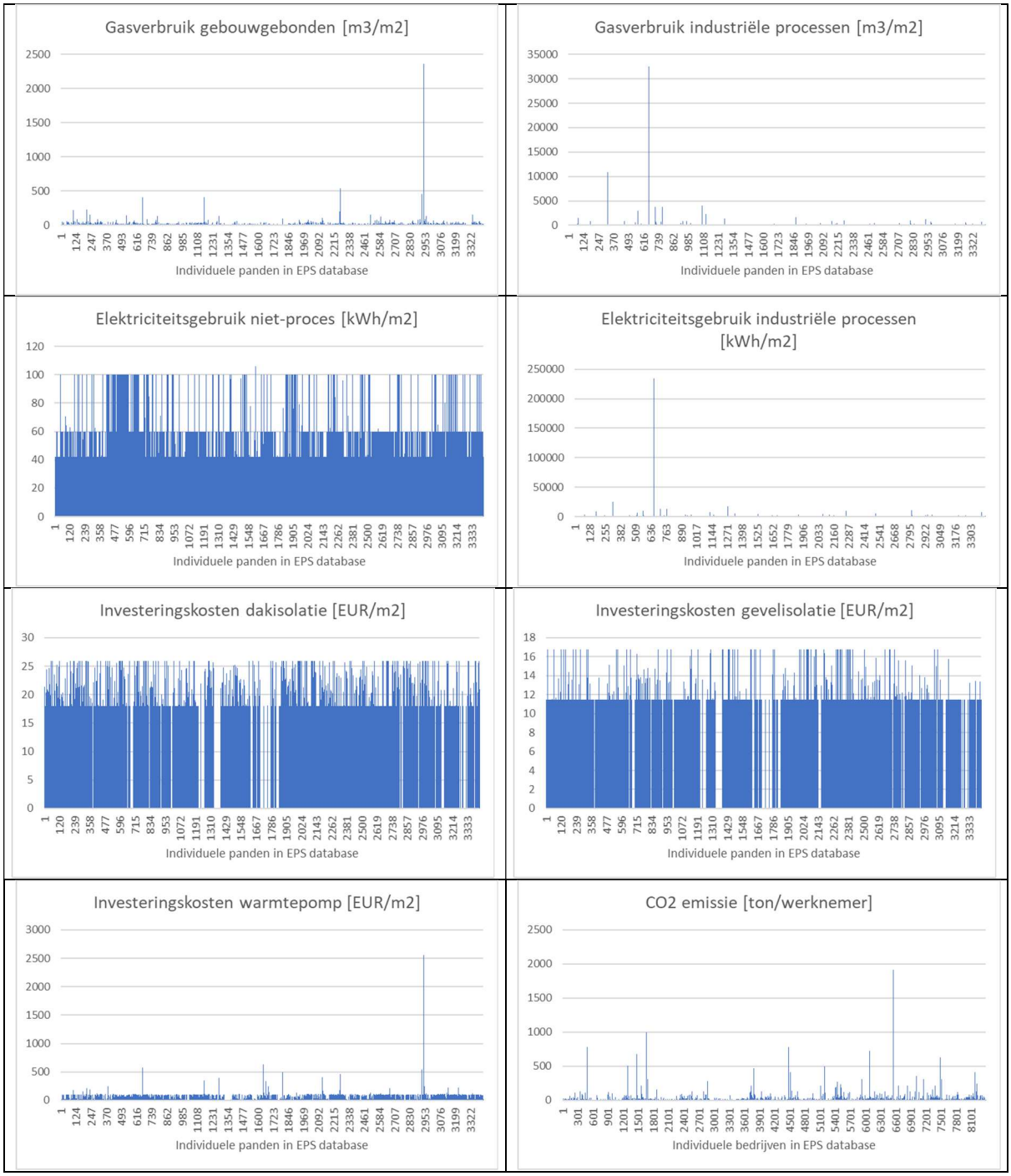
Om meer inzicht te krijgen in de kwaliteit van de resultaten zijn tevens de volgende indicatoren bepaald voor alle panden in de EPS rekendatabase:

- Gasverbruik gebouwgebonden [m3/m2]
- Gasverbruik industriële processen [m3/m2]
- Elektriciteitsgebruik niet-proces [kWh/m2]
- Elektriciteitsgebruik industriële processen [kWh/m2]
- Investeringskosten dakisolatie [EUR/m2]
- Investeringskosten gevelisolatie [EUR/m2]
- Investeringskosten warmtepomp [EUR/m2]

En de volgende indicator voor alle bedrijven:

- CO2-emissie [ton/werknemer]

De waardes voor deze indicatoren zijn weergegeven in onderstaande figuren:



Tabel 1. Gasverbruik gebouwgebonden [m3/m2], gasverbruik industriële processen [m3/m2], elektriciteitsgebruik niet-proces [kWh/m2], elektriciteitsgebruik industriële processen [kWh/m2], investeringskosten dakisolatie [EUR/m2], investeringskosten gevelisolatie [EUR/m2] en investeringskosten warmtepomp [EUR/m2] voor 3000+ panden en CO2-emissie [ton/werknemer] voor 8000+ bedrijven in de EPS database.

De verdeling van de waardes voor deze indicatoren maakt duidelijk dat er bedrijven zijn met grote uitschieters voor Gasverbruik gebouwgebonden (m3/m2), Gasverbruik industriële processen (m3/m2), Elektriciteitsverbruik industriële processen (kWh/m2), Investeringskosten warmtepomp (EUR/m2) en CO2-emissie (ton/werknemer). Het verklaren van deze uitschieters en het oplossen van eventuele onjuistheden in data en rekenmethodiek die hieruit volgen, kan leiden tot verbetering van de kwaliteit van de uitkomsten van de EPS, of duidelijk maken dat voor deze uitschieters het model niet direct toepasbaar is. De verschillen tussen de waardes voor de andere 3 indicatoren representeren de verschillen in kentallen voor verschillende functies van verblijfsobjecten en

bouwjaren van panden. Deze zijn goed verklaarbaar en geven dus de goede kwaliteit van de dataset aan. Een uitzondering (die niet op basis van verschillen in kengetallen verklaard kan worden) is het pand dat meer dan 100 kWh/m2 aan elektriciteit gebruikt voor niet-processen. Dit datapunt dient verder uitgezocht worden.

Energieverbruiksdata

Er zijn verschillende bronnen voor (geografische) energieverbruiksdata:

- Open data netbeheerders
- Energie in Beeld
- CBS
- Klimaatmonitor
- Energieatlas (2014)

Een lastigheid bij de vergelijking tussen deze datasets is dat ze verschillende definities gebruiken. Zo betreft de open data van netbeheerders enkel KVB (Klein Verbruikers) aansluitingen en bevat Energie in Beeld geen data over hogedruk gasnet aansluitingen. Daarnaast maakt Energie in Beeld onderscheid tussen zakelijk (waaronder corporatiewoningen) en particulier, terwijl CBS-data (tevens gebruikt voor de Klimaatmonitor en de Energieatlas) onderscheid maakt tussen woningen en bedrijven. Mogelijk dat het VIVET (Verbetering Informatievoorziening voor de Energietransitie) traject een rol kan spelen in het afstemmen van deze definities.

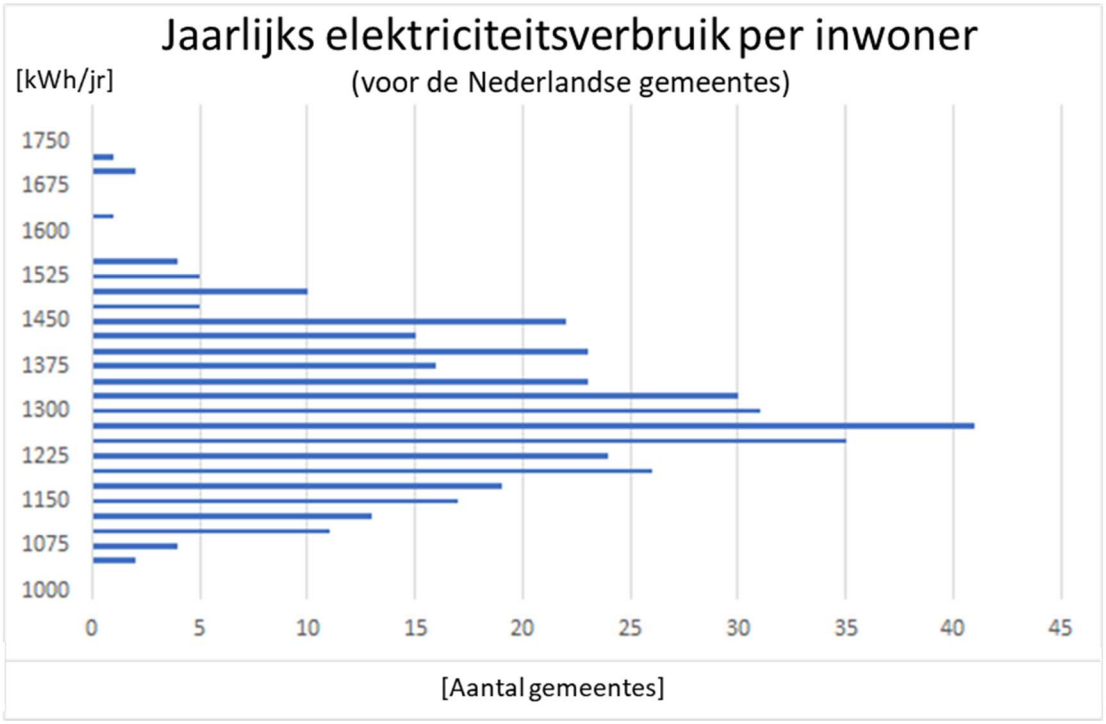
In dit proof of concept hebben we gekeken naar volledigheid en statistiek binnen de Klimaatmonitor dataset.

De Klimaatmonitor combineert en ontsluit veel verschillende datasets. Dit geeft erg nuttige inzichten maar maakt ook duidelijk dat sommige datasets niet compleet zijn, en daardoor niet aggregeerbaar zijn. Een voorbeeld is de CO2-uitstoot, die niet voor elke provincie beschikbaar is:

	Totaal bekende CO2-uitstoot	CO2-uitstoot Gebouwde Omgeving (gas, elektr. en warmte, tier 3/tier 2)	CO2-uitstoot Verkeer en vervoer incl. auto(snel) wegen, excl. elektr. railverkeer (scope 1, tier 1)	CO2-uitstoot Industrie, Energie, Afval en Water (gas en elektr., tier 3)	CO2-uitstoot Landbouw, bosbouw en visserij, SBI A (gas, elektr., tier 3)
Groningen	?	1.849.519	?	3.557.400	106.313
Friesland	?	1.852.321	?	783.565	210.679
Drenthe	3.605.600	1.492.153	1.139.215	?	190.627
Overijssel	7.507.100	3.278.622	2.267.281	1.718.214	242.996
Flevoland	2.642.300	1.016.272	1.072.337	209.556	344.138
Gelderland	14.726.600	5.988.424	5.272.290	2.692.579	773.308
Utrecht	7.772.400	3.614.594	3.313.064	699.898	144.808
Noord-Holla...	?	8.695.635	?	3.661.672	1.372.442
Zuid-Holland	?	10.138.003	?	9.509.673	4.249.681
Zeeland	?	1.168.030	?	5.391.924	184.253
Noord-Brabant	19.848.400	7.463.042	5.949.344	5.220.734	1.215.303
Limburg	11.726.800	3.518.600	2.394.455	5.037.112	776.628

De Klimaatmonitor-dataset van gas- en elektriciteitsverbruik per gemeente kan gebruikt worden om het verbruik per inwoner of per woning te vergelijken. Dan blijkt dat gas meer van het aantal woningen en elektriciteit meer van het aantal inwoners afhangt. Deze verdeling kan voor validatie van andere datasets gebruikt worden.

Als voorbeeld hieronder een figuur met de statistische spreiding van het elektriciteitsverbruik per inwoner voor de verschillende gemeentes; Minimum elektriciteitsverbruik is 1033 kWh per inwoner, maximum is 1704 kWh en het gemiddelde is 1286 kWh per inwoner. Dit kan gebruikt worden voor een eerste validatie van elektriciteitsdatasets.



4. Conclusies en aanbevelingen

In de 3 proof of concept voorbeelden zijn:

1. Inconsistenties gesignaleerd tussen de energiebalans in Vesta, in het Energietransitiemodel (ETM) en die van het CBS.
2. De uitkomsten van de Energiepotentieelscan voor Bedrijventerreinen consistent bevonden onder geografische aggregatie en aggregatie over energiedragers. Tevens is de statistiek voor deze resultaten inzichtelijk gemaakt.
3. Onvolledigheden gesignaleerd voor provinciale CO2 data in de Klimaatmonitor. Tevens is de statistiek voor gemeentelijke energieverbruiksdata in de Klimaatmonitor inzichtelijk gemaakt.

De voorbeelden tonen aan dat consistentie checks en een statische analyse inzicht geven in de kwaliteit van een dataset en bijdragen aan de interpretatie van datasets en verschillen tussen datasets. Essentieel hierbij is een goede beschrijving van de datasets. Dit vormen de 3 kernpunten van de datakwaliteit methodiek:

1. Beschrijf de dataset op een goede manier
2. Voer consistentie checks uit, zowel op de dataset als in vergelijking met andere datasets
3. Voer een statistische analyse uit

Deze methodiek kan gevolgd worden door zowel de data-eigenaar als de datagebruiker. De data-eigenaar kan de kwaliteit van een dataset inzichtelijk maken binnen de scope van een brede toepassing, de gebruiker van een dataset vanuit het perspectief van zijn of haar specifieke toepassing. De methodiek is geldig voor zowel brondata als model output data. In de appendix wordt een checklist gegeven waarmee de methodiek door de data-eigenaar of -gebruiker kan worden doorlopen.

Het heeft voordelen om de consistentie checks te automatiseren, vanwege de complexiteit en vereiste inspanning om de controles uit te voeren. Het handmatig uitvoeren op basis van verifiëren en vertalen van definities, kan hiermee tot een minimum beperkt worden. ESDL kan hier een bijdrage aan leveren door de samenhang tussen datasets via ESDL-systeem beschrijvingen/configuraties te laten lopen. ESDL systeembeschrijvingen gaan dus dienen als de methode om consistentie van een dataset en consistentie tussen datasets aan te tonen.

Hiertoe kan een ESDL consistentie-module ontwikkeld worden, die aantoont of de datasets kloppen met elkaar (tot op zekere hoogte, met een zekere marge) of niet. Deze module signaleert inconsistenties. Deze inconsistenties kunnen vervolgens ofwel opgelost worden, of enkel geduid binnen de scope van de toepassing van de dataset(s).

Gebruik van de methodiek brengt inzicht in de betrouwbaarheid en de kwaliteit van datasets, voor eigenaren en gebruikers van datasets. Dit inzicht is essentieel bij het gebruik van datasets binnen de energietransitie, zodat data en modeluitkomsten op een goede manier geïnterpreteerd en gebruikt worden bij het maken van beslissingen binnen de energietransitie. Deze verantwoordelijkheid wordt gedeeld tussen de eigenaar en de gebruiker van data. Zij kunnen beiden de kwaliteit van de dataset beoordelen en een (subjectief) oordeel vellen over of de dataset betrouwbaar genoeg is om te gebruiken. Beiden kunnen de checklist in de appendix gebruiken om de beschreven datakwaliteit methodiek te doorlopen.

Appendix: Checklist Data Kwaliteit

Het volgen van deze checklist geeft de eigenaar of gebruiker van een dataset inzicht in de kwaliteit van een dataset.

Beschrijving

☐ Is de data helder en eenduidig gedefinieerd? Worden eenheden weergegeven?

☐ Is de verzamelmethode voor ruwe data duidelijk beschreven? Is voor metingen het tijdstip of de meetperiode gegeven?

☐ Zijn de aannames die gedaan zijn voor ruwe data duidelijk beschreven? Denk aan interpolatie, interpretatie en andere bewerkingen.

☐ Indien de data uitkomsten van een rekenmodel betreft: wordt er verwezen naar een duidelijke beschrijving van de gevolgde rekenmethodiek?

☐ Metadata: worden publicatiedatum, eigendom, contactgegevens, gebruiksrecht en licentievoorwaarden gegeven?

☐ Worden eventuele onvolledigheden expliciet toegelicht?

☐ Wordt de scoping aangegeven: voor welke toepassingen is de data wel/niet bruikbaar?

☐ Wordt er een periode van geldigheid gegeven? Tot wanneer is de data te gebruiken/bruikbaar?

☐ Is er een onderhoudscyclus? Wordt de data onderhouden? Met welke cyclus?

☐ Is de data goed gedocumenteerd? Is duidelijk welke documentatie beschikbaar is? En wat de inhoud van deze documentatie is?

Consistentie

Controle:

☐ Staat er überhaupt iets in het data-veld? Of is het leeg?

☐ Is het een getal? Of een string?

☐ Is het niet-negatief? (Als het over energie/potentie etc. gaat verwacht je alleen positieve getallen)

☐ Is het conform de gekozen decimale punt/komma conventie?

Aggregatie: tellen de subtotalen op tot totalen?

☐ Ruimtelijk

☐ Sectoren

☐ Assets

☐ Energiedragers

Energiebalans: voldoet de data aan de wet van energiebehoud?

☐ Voor assets

☐ Voor een gebied

☐ Voor een energiesysteem

☐ Is de data consistent met kengetallen en andere datasets en kunnen verschillen verklaard worden? E.g.

- Voorbeeldwoningen
- Cijfers en tabellen
- Energieverbruiksgegevens
- Eerdere jaren
- Etc.

Statistiek

☐ Is de dataset volledig? Of ontbreken er veel datapunten?

☐ Is de statistische spreiding van de variabelen binnen een dataset verklaarbaar?

Bepaling van de volgende variabelen kan hierbij helpen:

☐ Verbruik, e.g.

- Per m2
- Per FTE
- Per omzet (in EUR of product)
- Per huishouden

☐ Kosten, e.g.

- Per m2
- Per W
- Per kWh

☐ CO2, e.g.

- Per m2
- Per FTE
- Per omzet (in EUR of product)
- Per huishouden