



**Gaussian-Mixture Typegenerator in
Gaia LV Network Design**

22-061 AIS

14 december 2022

Building 026
Koningstraat 27-1d
6811 DG Arnhem

Postbus 2013
2800 BD Gouda

026 352 37 00
www.phasetophase.nl



Documentversies

Versie	Datum	Auteur	Commentaar	Review
1.0	14-12-2022	AIS	Initiële versie	CWI

© Phase to Phase BV, Arnhem, Nederland. Alle rechten voorbehouden.

Dit document bevat vertrouwelijke informatie. Overdracht van de informatie aan derden zonder schriftelijke toestemming van of namens Phase to Phase BV is verboden. Hetzelfde geldt voor het kopiëren van het document of een gedeelte daarvan.

Phase to Phase BV is niet aansprakelijk voor enige directe, indirecte, toekomstige of gevolgschade ontstaan door of bij het gebruik van de informatie of gegevens uit dit document, of door de onmogelijkheid die informatie of gegevens te gebruiken.



INHOUD

1	INLEIDING	1
1.1	Overzicht GM-belastingtypegenerator in Gaia	1
2	GAUSSIAN MIXTURE MODELLEN	3
2.1	GM-model wiskundige formule.....	4
2.2	Expectation-Maximization algoritme	5
3	FUNCTIONALITEITEN EN INSTELLINGEN	6
3.1	Apparaten.....	6
3.2	Basisbelasting	6
3.3	Werkdag/weekend	7
3.4	Trend verwijderen	7
3.5	Correlatie.....	7
3.6	Overzicht.....	8
3.7	Van factoren naar gewichten	8
4	EXPECTATION-MAXIMIZATION ALGORITME	9
4.1	Bewerkingsstap.....	9
4.1.1	Data	9
4.1.2	Aantal tijdstippen bepalen.....	9
4.1.3	Datanormalisatie	9
4.1.4	Detrend	10
4.2	Invoeringsstap	10
4.3	Uitvoeringsstap.....	11
4.3.1	Startoplossing	12
4.3.2	Iteratieproces	13
4.3.3	Toevoegingen.....	14
4.4	Overzicht.....	14
5	POST-PROCESSEN.....	15
5.1	Waarderingsstap.....	15
5.1.1	Startoplossing	15
5.1.2	Iteratie	15
5.1.3	Interpolatie	16
5.2	Factorenstap	16
5.2.1	Maandfactoren.....	16
5.2.2	Trendfactoren.....	17
5.2.3	Kwartierfactoren.....	17
5.2.4	Apparaten en weekendfilter	17



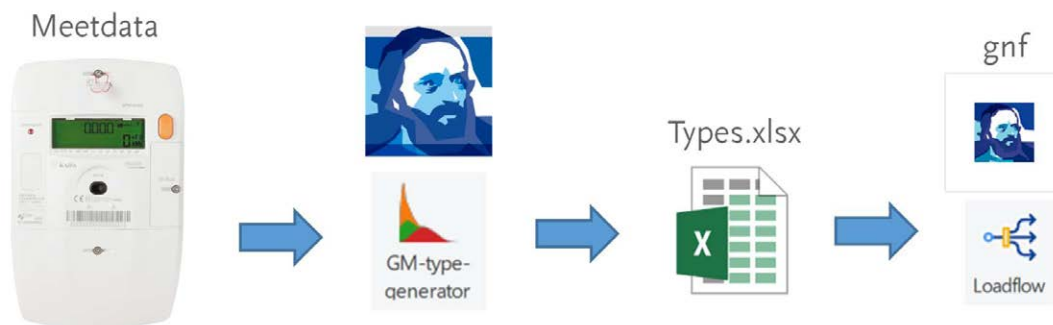
1 INLEIDING

Gaia LV Network Design is een gereedschap specifiek ontwikkeld voor de laagspanningsdistributie, enkele punten specifiek voor de LS-distributie zijn: asymmetrische kabels, asymmetrische belasting, niet-lineair net en onzekere belasting. Dit laatste punt is onderwerp in dit rapport; de onzekerheid in netbelasting. Individuele aangeslotenen vertonen in het elektriciteitsverbruik een zeker groepsgedrag, maar gedragen zich momentaan gezien als onafhankelijke individuen. Immers, niet bij iedereen draait de wasmachine op hetzelfde moment. Dit gegeven wordt in Gaia meegenomen door elke aansluiting op een bepaald moment in de tijd een kans mee te geven op een bepaalde belasting. Dit in de vorm van een, of meerdere, normaalverdeling(en).

De GM-typegenerator, zoals behandeld in dit rapport, is een tool voor het bepalen van parameters van het wiskundige Gaussian Mixture model (GM-model). De GM-typegenerator genereert op basis van data de GM-modellen die bestaan uit verschillende normaalverdelingen. De parameters van deze normaalverdelingen worden bepaald vanuit het Expectation-Maximization algoritme (EM-algoritme) en het post-processen. De kansverdeling die voortkomt uit de GM-typegenerator, wordt toegepast bij de stochastische loadflowberekeningen aan laagspanningsnetwerken.

1.1 Overzicht GM-belastingtypegenerator in Gaia

Het werkproces van het genereren van een nieuw GM-type is schematisch in Figuur 1 afgebeeld. De eerste stap is het verzamelen van meetdata (van slimme meters of metingen van specifieke apparaten). De meetdata moet eerst gecategoriseerd worden (bijvoorbeeld, gebaseerd op jaarverbruik en/of type woning/apparaat of andere criteria). Dit is de verantwoordelijkheid van de gebruiker. Na de categorisatie en omzetten van meetdata naar het door de GM-typegenerator geaccepteerd tekstbestand (csv-formaat) kan het bijhorende stochastische GM-type gegenereerd worden in Gaia. Daarna is het mogelijk om de parameters van GM-type naar het typebestand te kopiëren en te plakken. GM-types uit het typebestand kunnen verder in Gaia-netwerkbestanden (gnf-bestanden) bij de gegevens van aansluitingen toegepast worden. Kansverdelingen die bij de GM-types horen, worden uiteindelijk bij de stochastische loadflow en daarmee ook de netbelastingberekening gebruikt.



Figuur 1: Schematisch werkproces van het genereren van een nieuw GM-type

De GM-typegenerator gebruikt csv-formaat (txt-bestand) als input waarvan de eerste kolom de datum-tijd bevat en de andere kolommen de bijhorende vermogens in kW zoals gemeten (slimme meters/apparaten). Het is bij de GM-typegenerator aangenomen dat slimme meterdata al gecategoriseerd is en, indien wenselijk, gedisaggregeerd (wat verbruik van grote apparaten, zoals bijvoorbeeld PV en warmtepompen, betreft). De meetdata moet minstens 1 jaar representeren (meerdere jaren zijn ook toegestaan), per kwartier gesampled en zo veel mogelijk kolommen (lees: metingen) bevatten (zie Figuur 2). Hoe meer meetdata beschikbaar, hoe beter de statistische benadering van het resulterende GM-type.



Vermogenswaarden per kwartier



	A	B	C	D	E	F	G	H	I
1	1-1-2021 00:00	1,389	0,420	0,375	0,468	0,398	0,915	0,646	1,708
2	1-1-2021 00:15	1,200	0,354	0,374	0,396	0,390	0,929	0,638	1,671
3	1-1-2021 00:30	1,011	0,289	0,372	0,323	0,394	0,943	0,631	1,634
4	1-1-2021 00:45	1,062	0,270	0,356	0,217	0,404	0,954	0,603	1,658
5	1-1-2021 01:00	1,113	0,251	0,341	0,111	0,414	0,965	0,576	1,682
6	1-1-2021 01:15	1,385	0,349	0,272	0,118	0,525	0,934	0,508	1,707
7	1-1-2021 01:30	1,656	0,446	0,203	0,124	0,636	0,903	0,441	1,732
8	1-1-2021 01:45	1,222	0,425	0,186	0,147	0,604	0,880	0,417	1,691
9	1-1-2021 02:00	0,788	0,403	0,169	0,169	0,573	0,858	0,393	1,650
10	1-1-2021 02:15	0,846	0,284	0,165	0,183	0,446	0,831	0,392	1,058
11	1-1-2021 02:30	0,904	0,165	0,161	0,196	0,319	0,804	0,391	0,465

N slimme meters
... (hoe meer, hoe beter)
van dezelfde categorie

↓ 1 of meerdere jaren

Geen privacy issues, alleen maar tijden en waarden zijn van belang!

Figuur 2: Voorbeeld van slimme meterdata (als input voor GM-typegenerator)

De GM-typegenerator bepaalt op basis van de metingen gemiddelden, standaarddeviaties en gewichten van het GM-type. Deze parameters kunnen eenvoudig gekopieerd/geplakt worden naar het typebestand (tabblad GM) zoals in Figuur 3 weergegeven. Gemiddelden en standaarddeviaties variëren niet met de tijd, tijdsafhankelijkheid wordt door middel van gewichten gemodelleerd. In de gewichten zit dus impliciet de stochastische jaarprofiel van een belasting/apparaat.

GM-typegenerator

Invoer: Databestand C:\Gaia\GM-types\Tm GM-types\London\SVJ 5000.txt

Opties: Apparaat Werkdag/weekend Correlate Trend verwijderen

Berekening voltooit

n	Gemiddelde	Standaarddeviatie	Weging
1	0,21	0,09	0,46
2	0,62	0,23	0,22
3	1,21	0,45	0,21
4	2,32	0,92	0,11

Kwarterfactoren

	00:00	00:15	00:30	00:45	01:00	01:15	01:30	01:45	02:00	02:15
1 - Werkdag	0,55	0,59	0,63	0,67	0,70	0,73	0,75	0,78	0,79	0,81
2 - Werkdag	0,37	0,33	0,30	0,27	0,25	0,23	0,21	0,19	0,17	0,16
3 - Werkdag	0,07	0,06	0,05	0,04	0,04	0,03	0,03	0,02	0,02	0,02
4 - Werkdag	0,02	0,02	0,02	0,02	0,02	0,01	0,01	0,01	0,01	0,01
1 - Weekend	0,57	0,61	0,65	0,69	0,72	0,75	0,77	0,79	0,81	0,82
2 - Weekend	0,35	0,32	0,28	0,26	0,23	0,21	0,19	0,17	0,16	0,15
3 - Weekend	0,06	0,05	0,05	0,04	0,03	0,03	0,02	0,02	0,02	0,02
4 - Weekend	0,02	0,02	0,02	0,01	0,01	0,01	0,01	0,01	0,01	0,01

Maandfactoren

n	1	2	3	4	5	6	7	8	9	10	11	12
1	1,00	1,04	1,25	1,43	1,54	1,58	1,54	1,64	1,50	1,31	1,13	1,00
2	1,00	1,00	0,89	0,93	0,96	0,98	1,04	0,99	0,99	1,02	1,01	0,96
3	1,00	0,94	0,83	0,64	0,52	0,46	0,41	0,41	0,54	0,69	0,87	1,05
4	1,00	0,99	1,34	0,85	0,52	0,37	0,26	0,26	0,40	0,55	0,84	1,11

Typenaam: GM-type

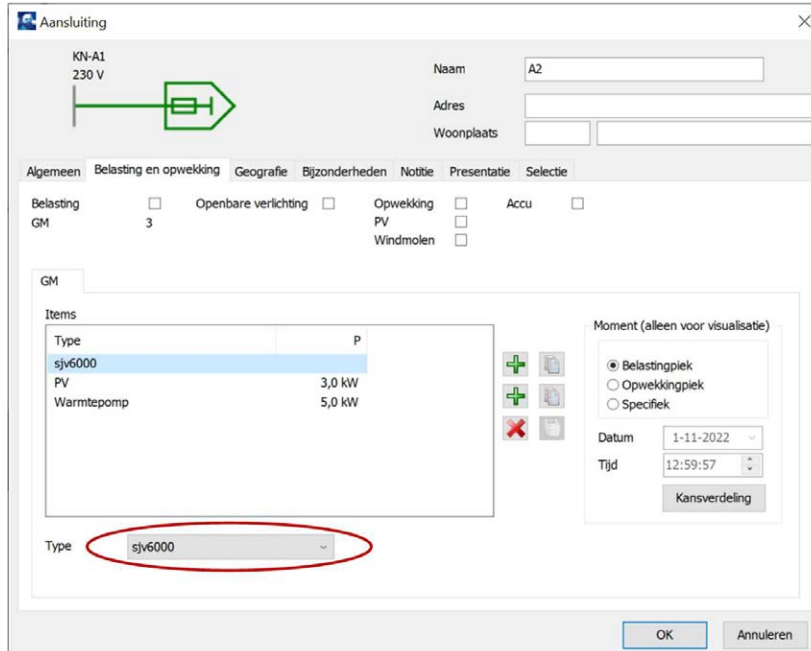
Types.xlsx

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Name	Cos Indicator	Nummer	Correlation	Average[1]	Average[2]	Average[3]	Average[4]	Deviation[1]	Deviation[2]	Deviation[3]	Deviation[4]	
9	Hoog	0,95	Belasting	4	0	1,43	2,98	0,89	0,61	0,894	1,158	0,656	0,346
10	Laag	0,95	Belasting	4	0	0,42	0,3	0,08	0,04	0,548	0,447	0,224	0,141
11	Elektrisch koken	0,95	Apparaat	1	0	29,9609				18,573			
12	Airco	0,95	Apparaat	1	0	36,122				9,749			
13	sjv500	0,98	Belasting	4	0	0,056	0,160	0,017	0,802	0,022	0,070	0,012	0,334
14	sjv1000	0,98	Belasting	4	0	0,112	0,320	0,035	1,204	0,045	0,140	0,024	0,666
15	sjv1500	0,98	Belasting	4	0	0,171	0,390	0,204	0,051	0,071	0,157	1,461	0,030
16	sjv2000	0,98	Belasting	4	0	0,231	0,459	2,864	0,087	0,098	0,174	2,254	0,036
17	sjv2500	0,98	Belasting	4	0	0,057	2,579	0,180	0,447	0,022	1,922	0,078	0,185
18	sjv3000	0,98	Belasting	4	0	0,047	2,294	0,129	0,435	0,008	1,590	0,059	0,190
19	sjv3500	0,98	Belasting	4	0	0,195	2,671	0,078	0,515	0,074	1,990	0,028	0,194
20	sjv4000	0,98	Belasting	4	0	0,262	3,047	0,109	0,596	0,090	2,371	0,048	0,192
21	sjv4500	0,98	Belasting	4	0	0,614	0,083	0,262	3,253	0,240	0,048	0,081	2,532
22	sjv5000	0,98	Belasting	4	0	0,632	0,057	0,262	3,459	0,287	0,048	0,072	2,694
23	sjvRNN	0,98	Belasting	4	0	0,624	0,074	0,146	3,272	0,236	0,077	0,082	2,388

Figuur 3: GM-typegenerator en koppeling met Gaia-typebestand

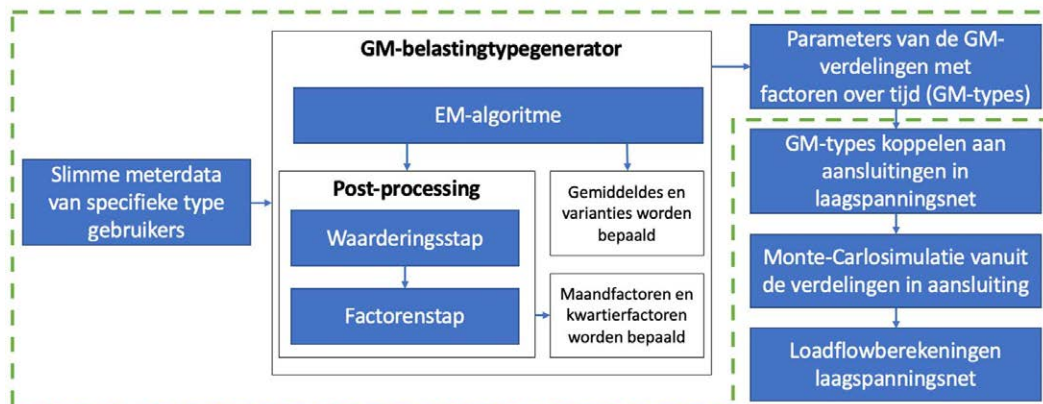


In Gaia kunnen de stochastische GM-types via de parameters van een aansluiting toegewezen worden (zie Figuur 4). GM-types van basisbelasting(en) en apparaten kunnen binnen één aansluiting gestapeld worden. Bij het openen van oudere netten (met Strand-Axelsson belastingen formaat tot Gaia 7.15) in Gaia 8.x worden ze automatisch geconverteerd naar de nieuwe GM-types. Informatie over deze conversie is in een afzonderlijk rapport beschikbaar op onze website.



Figuur 4: Toewijzing van een stochastische GM-type in de parameters van een aansluiting

Het stroomschema van de stochastische modellering, van meting tot loadflow, in Gaia is op een meer gedetailleerde manier weergegeven in Figuur 5. Verder in dit rapport wordt de focus gelegd op de processen die binnen het groene kader vallen.



Figuur 5: Stroomschema GM-belastingtypegenerator

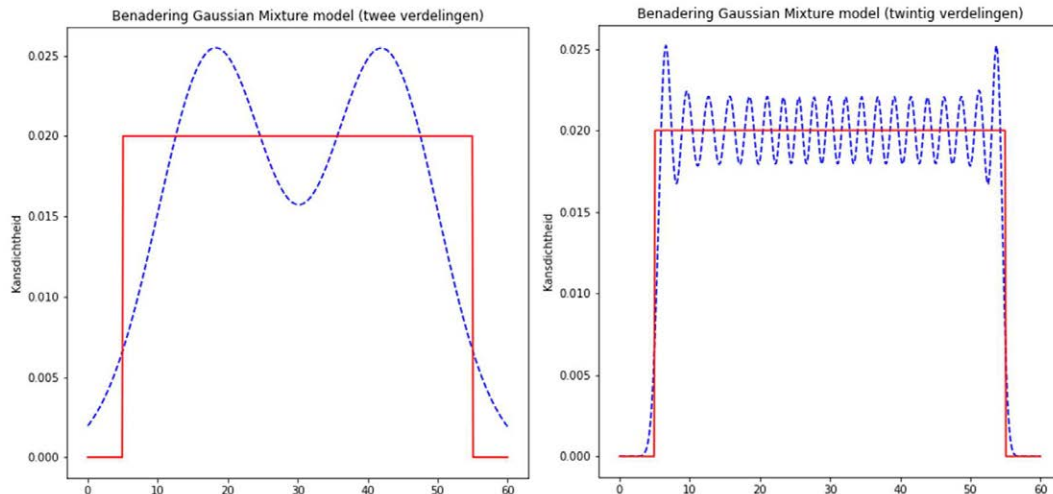
2 GAUSSIAN MIXTURE MODELLEN

Tot kort bestond de belasting op laagspanningsnetten vooral uit gloeilampen, wasmachines, drogers, televisies en koelkasten. Tegenwoordig is de basisbelasting veranderd door ledverlichting, vaatwasmachines, computers en is er een opkomst te zien in apparaten zoals warmtepompen,



zonnepanelen en laadpalen voor elektrisch vervoer. De onzekerheid van de belasting van aansluitingen begint toe te nemen mede door deze apparaten van een andere orde.

Het oude Strand-Axelsson model zoals geïmplementeerd in Gaia t/m versie 7.15, dat werkt met één normaalverdeling, voldoet niet meer bij de toepassing van meerdere apparaten in een huishouden. Dit is de reden dat in Gaia het Gaussian-Mixture model is geïntroduceerd, bestaande uit meerdere normaalverdelingen. Deze normaalverdelingen worden gebruikt om de waargenomen empirische verdeling te benaderen. Het hoeft dus niet zo te zijn dat de onderlinge datapunten de normaalverdeling volgen. Een Gaussian Mixture model kan worden gebruikt om elke continue en differentieerbare kansverdeling te benaderen, zolang het aantal componenten maar hoog genoeg is.



Figuur 6: Benadering continue uniforme distributie met een Gaussian Mixture model

In Figuur 6 wordt een benadering van een continue uniforme distributie met behulp van een Gaussian Mixture model weergegeven. Hoewel één normaalverdeling deze data niet nauwkeurig zou kunnen vertegenwoordigen, lukt het een Gaussian Mixture model al een stuk beter. Hier komt de flexibiliteit en inzetbaarheid van GM-modellen duidelijk naar voren. Men hoeft zich niet meer af te vragen of de gekozen verdeling wel correct is. De focus kan zo worden verlegd naar het kiezen van een optimaal aantal componenten rekening houdend met de rekentijd en opslagcapaciteit.

2.1 **GM-model wiskundige formule**

Een Gaussian Mixture model werkt dus met meerdere normaalverdelingen. Deze verschillende normaalverdelingen hebben allemaal een bepaald gewicht. Het gewicht kan beschreven worden als het oppervlak dat de normaalverdeling inneemt in de totale kansverdeling van het GM-model. Het gewicht ligt altijd tussen de nul en één. Als een gewicht van een normaalverdeling groter wordt, dan stijgen de y-waardes, en dus de kans, van die specifieke normaalverdeling. Wel moet de som van gewichten altijd gelijk zijn aan 1, anders is er geen sprake van een correcte kansverdeling¹. De kansdichtheidsfunctie van het 1D GM-model kan beschreven worden aan de hand van de volgende vergelijking:

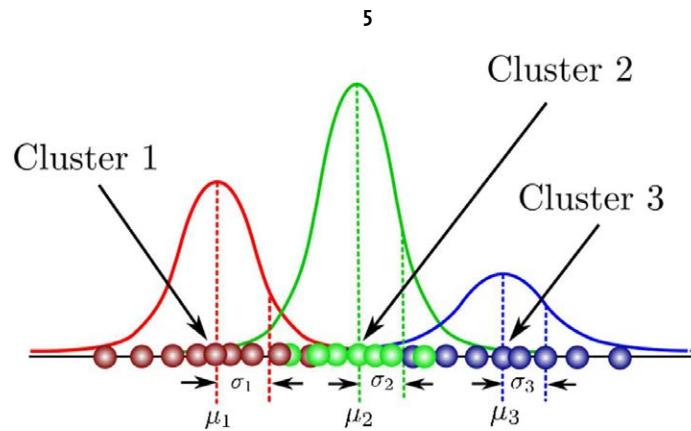
$$f(x) = \sum_{j=1}^N \omega_j \cdot \frac{1}{\sqrt{2\pi}\sigma_j} \cdot e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}}$$

waarbij geldt dat:

$$0 \leq \omega_j \leq 1 \text{ en } \sum \omega_j = 1.$$

Hierbij staat N voor het aantal normaalverdelingen dat gebruikt wordt, ω_j voor het gewicht van normaalverdeling j , σ_j staat voor de standaarddeviatie van normaalverdeling j en μ_j voor het gemiddelde van normaalverdeling j .

¹ Indien de som van gewichten van 1 afwijkt, kunnen de gewichten genormaliseerd worden zodat de som gelijk aan 1 is en de oorspronkelijke verhoudingen onveranderd blijven. Dit wordt in Gaia toegepast omdat de som van tijdsfactoren (kwartierfactor maal maandfactor) op een willekeurig tijdstip niet altijd aan 1 gelijk is.



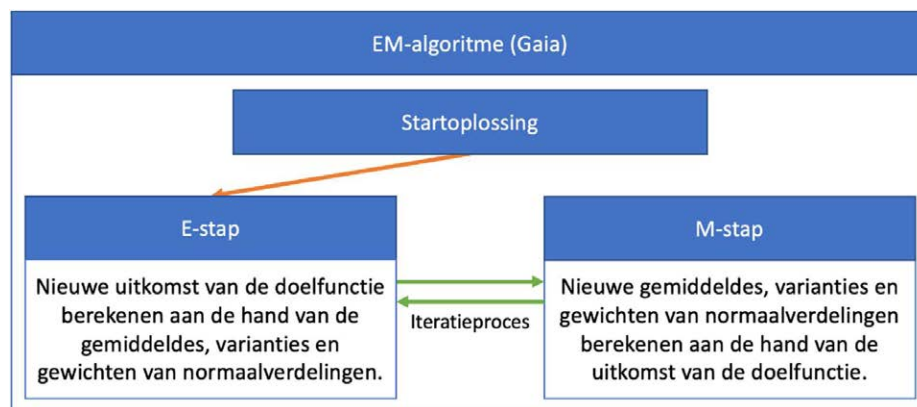
Figuur 7: Voorbeeld van een GM-model

In Figuur 7 wordt een voorbeeld van een één-dimensionaal Gaussian Mixture model weergegeven. Dit GM-model bestaat uit drie normaalverdelingen. De bolletjes stellen de individuele metingen voor waaruit de normaalverdelingen zijn opgebouwd. In de werkelijkheid is het vaak zo dat één meting kan bijdragen aan het opbouwen van meerdere normaalverdelingen, dit principe wordt soft clustering genoemd. Het figuur is dus een versimpelde weergave.

2.2 Expectation-Maximization algoritme

Om uit data één normaalverdeling op te stellen is tamelijk eenvoudig. Het gemiddelde en de standaarddeviatie kunnen berekend worden via een formule. Het is echter een stuk lastiger om de verschillende gemiddelden, varianties en gewichten van een GM-model te bepalen. Hier zijn geen deterministische vergelijkingen voor.

Om een zo goed mogelijke benadering te behalen wordt er gebruik gemaakt van het EM-algoritme. Het EM-algoritme is een iteratief algoritme dat bestaat uit twee stappen: de E-stap en M-stap. In Figuur 8 wordt het EM-algoritme schematisch weergegeven.



Figuur 8: Versimpelde schematische weergave EM-algoritme

Bij elke iteratie wordt de doelfunctie (die gebaseerd is op de kansdichtheidsfunctie van de normaalverdeling) steeds hoger, tot dat er uiteindelijk een (lokaal) maximum wordt bereikt. Later in het rapport wordt de gekozen implementatie in GM-typegenerator van Gaia verder uitgewerkt.

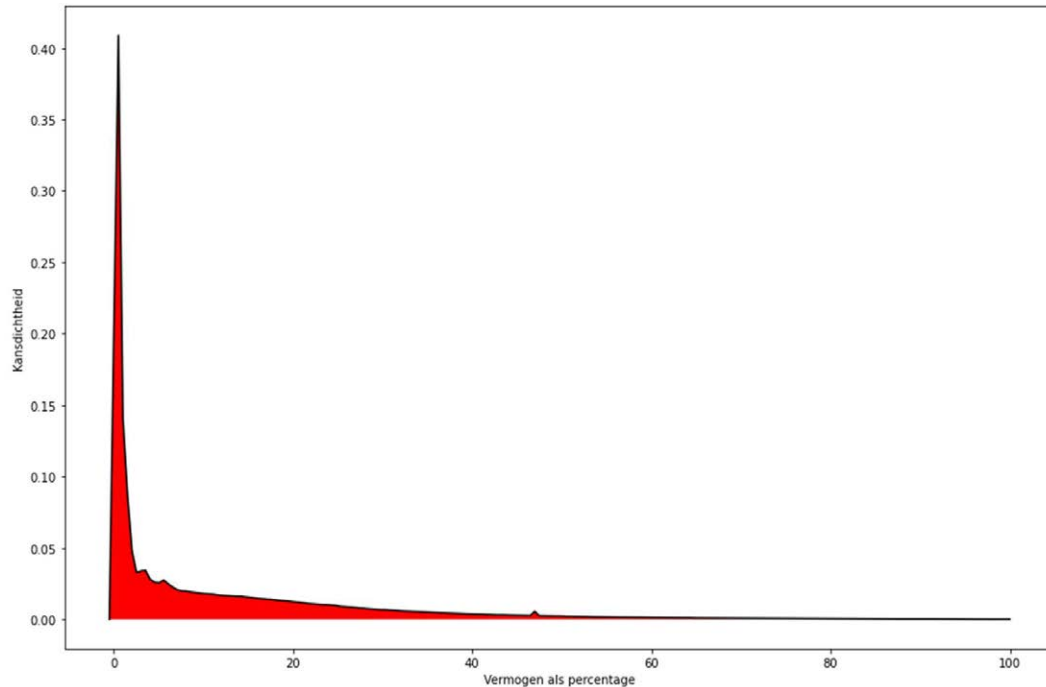


3 FUNCTIONALITEITEN EN INSTELLINGEN

Dit hoofdstuk beschrijft eigenschappen en instellingen van de GM-typegenerator voor het bepalen van parameters van GM-types van apparaten en basisbelastingen.

3.1 Apparaten

In de GM-typegenerator is er de keuze gemaakt om een belastingtype toe te kunnen wijzen aan zelfstandige apparaten, zoals warmtepompen en zonnepanelen. De uitvoer van deze apparaten wordt uitgedrukt in percentages, waarna de gebruiker zelf een specifiek vermogen kan instellen bij verschillende aansluitingen in het LS-net in Gaia. In Figuur 9 is de kansverdeling van de genormaliseerde gemeten vermogens van de verschillende warmtepompen in kaart gebracht.



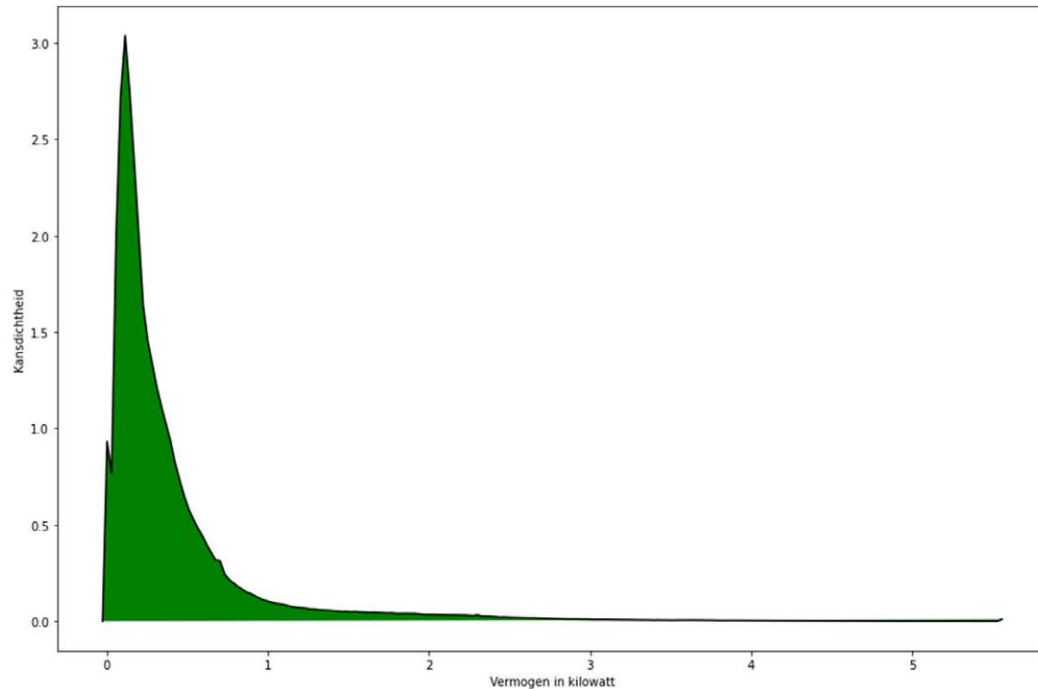
Figuur 9: Kansverdeling van de genormaliseerde warmtepompvermogens (gebaseerd op meetdata)

Om deze kansverdeling te benaderen is er gekozen om twee normaalverdelingen te gebruiken. Deze keuze is gemaakt zodat datapunten die dicht bij het vermogenpercentage van nul liggen gerepresenteerd kunnen worden door de eerste normaalverdeling zodat de (tweede) normaalverdeling dan kan worden opgesteld zonder dat apparaten die in de sluimerstand staan veel invloed hierop uitoefenen. Het idee is dat het gemiddelde en de variantie van de tweede normaalverdeling dus wordt gebaseerd op data van wanneer de apparaten daadwerkelijk worden gebruikt. Standaard staat de eerste normaalverdeling op een gemiddelde van 0 met een variantie van 5%. Deze waarden worden geforceerd in het GM-model om zo vervolgens nog één normaalverdeling te vinden die past op de rest van de dataset.

3.2 Basisbelasting

In de GM-typegenerator is er de keuze gemaakt om een GM-type toe te kunnen wijzen aan de basisbelasting van bijvoorbeeld een huisaansluiting. De basisbelasting is de belasting van de verzameling van apparaten die gekoppeld zijn op één aansluiting in het LS-net, maar niet apart gemodelleerd worden. In Figuur 10 is de kansverdeling van vermogens in kW van verschillende huishoudens in kaart gebracht.

Veel apparaten staan standaard op stand-by denk aan televisies, wasmachines, enzovoort. Hierdoor is de grote piek dicht bij nul te verklaren. Er is voorlopig gekozen om de basisbelasting te benaderen met vier normaalverdelingen, omdat dit in meeste situaties tot voldoende nauwkeurige benaderingen leidt.



Figuur 10: Elektriciteitsgebruik van huishoudens omgezet tot kansverdeling

3.3 **Werkdag/weekend**

Werkdag/weekend is een filter die de data opsplijst in twee categorieën: het werkdagegebruik en weekenddagegebruik. Dit resulteert uiteindelijk in een stuk meer opslagdata, maar wel een nauwkeurigere benadering van het werkdagegebruik en weekendgebruik.

3.4 **Trend verwijderen**

Als een apparaat terugkerende (dag- of maand)patronen van elektriciteitsgebruik volgt, dan kan dit worden beschreven aan de hand van een bepaalde trend. Een zonnepaneel is hier een voorbeeld van. Vaak is er in de ochtend en avond weinig zonlicht en in de middag veel zonlicht. Daarnaast is er op een willekeurige ochtend in januari vaak minder zonlicht dan op een willekeurige ochtend in juli. Er is hier een terugkerend patroon te vinden. Als in dit geval het standaardrekenproces toegepast wordt, kunnen de gemiddelde en de standaarddeviatie onjuist bepaald worden (in acht nemend dat ze in de huidige modellering voor het hele jaar vastgesteld worden). Eerst moet de trend van de meetdata verwijderd worden, vervolgens worden de gemiddelde en de standaarddeviatie door GM-typegenerator geschat. De trend (percentages per kwartier van een dag en maandfactoren) wordt in dit geval apart bewaard. Tijdens de stochastische loadflow berekening op een bepaald tijdstip wordt de gemiddelde van een apparaat uit typebestand opgeteld met de trendwaarde (percentage per kwartier maal maandfactor). Op deze manier kan de actuele gemiddelde (geüpdatet met trend) van de aan toestand van het apparaat variëren met tijd.

3.5 **Correlatie**

De optie correlatie is op dit moment alleen maar door de gebruiker in te vullen in de GM-typegenerator. Het idee achter de optie is dat er in de LS-netten apparaten zijn waarvan het verloop van elektriciteitsgebruik vergelijkbaar is met elkaar. Een voorbeeld hiervan zijn zonnepanelen. Wanneer het bekend is dat een zonnepaneel elektriciteit opwekt is de kans groot dat ook de andere zonnepanelen in het betreffende netwerk elektriciteit opwekken. Hierbij heeft de externe factor zonintensiteit invloed op andere zonnepanelen in het net en is er hierdoor een correlatie tussen deze zonnepanelen. Met behulp van deze optie kan er een bepaalde correlatiefactor (die door de gebruiker zelf bepaald moet worden) worden toegevoegd aan het GM-type die vervolgens toegepast kan worden bij de Monte-Carlosimulatie van de kansverdelingen.



3.6 **Overzicht**

In Figuur 11 zijn de verschillende opties te zien. Deze opties worden aangezet door ze aan te vinken.

Figuur 11: Invoermogelijkheden GM-typegenerator

Sommige opties kunnen worden gecombineerd en sommige opties moeten zelfs worden gecombineerd. De optie *Trend verwijderen* kan alleen worden uitgevoerd als de optie *Apparaat* is aangevinkt, verder kan de optie *Trend verwijderen* niet worden gecombineerd met een werkdag/weekend filter. De optie *Correlatie* kan alleen worden uitgevoerd als *Apparaat* ook is aangevinkt. Als *Apparaat* niet is aangevinkt dan wordt de basisbelasting (verzameld elektriciteitsgebruik van bijvoorbeeld een huishouden) aangehouden. In Tabel 1 zijn de functionaliteiten en instellingen kort samengevat. Het is belangrijk te vermelden dat de geforceerde parameters in deze tabel betrekking hebben op de uitvoering van het EM-algoritme.

Tabel 1: Overzicht van verschillende opties

	BASISBELASTING	APPARAAT
Aantal normaalverdelingen	Vier	Twee (waarvan één weergegeven wordt)
Geforceerde parameters	Geen	Eén onzichtbare vastgestelde normaalverdeling ($\mu: 0, \sigma^2: 5\%$)
Aantal per aansluiting	Eén	Meerdere
Trend verwijderen	Niet mogelijk	Wel mogelijk
Correlatie meenemen	Niet mogelijk	Niet geïmplementeerd
Weekend/werkdag	Wel mogelijk	Wel mogelijk

De verschillen in instellingen en functionaliteiten zijn ook terug te zien in de output van de GM-typegenerator. In Tabel 2 is dat weergegeven.

Tabel 2: Uitvoer van de GM-typegenerator (GM-type parameters)

Basisbelasting	Apparaat
4 gemiddelden	1 gemiddelde
4 standaarddeviaties	1 standaarddeviatie
384 (4 x 96) kwartierfactoren	96 kwartierfactoren
48 (4 x 12) maandfactoren	12 maandfactoren

Dit zijn de parameters die worden opgeslagen wanneer er een GM-type wordt aangemaakt. Het getal 96 is afkomstig van het aantal kwartieren op een dag en twaalf voor het aantal maanden in het jaar. Daarbij worden er nog twee keer zoveel kwartierfactoren opgeslagen als het weekend/werkdag filter aanstaat en 204 ($2 \times 96 + 12$) trendfactoren opgeslagen als ‘trend verwijderen’ is geselecteerd. In de maximale situatie worden er 824 parameters opgeslagen door de GM-belastingtypegenerator.

3.7 **Van factoren naar gewichten**

Om per kwartier een verdeling op te stellen wordt er gebruik gemaakt van de maand- en kwartierfactoren. Deze factoren worden omgezet naar gewichten. Bij de basisbelasting kan het gewicht ω_j van normaalverdeling j in een GM-model worden opgesteld aan de hand van de volgende vergelijking:

$$\omega_j = \frac{kf_{k,j} \cdot mf_{m,j}}{\sum_i^N kf_{k,i} \cdot mf_{m,i}}$$



De variabele $kf_{k,j}$ staat voor de kwartierfactor en $mf_{m,j}$ staat voor de maandfactor. Hierin staat k voor het kwartiernummer (1 t/m 96), m voor het maandnummer (1 t/m 12) en j voor de specifieke normaalverdeling in het GM-model.

4 EXPECTATION-MAXIMIZATION ALGORITME

In dit hoofdstuk wordt de implementatie van het EM-algoritme in de GM-belastingtypegenerator behandeld. Voordat het EM-algoritme kan worden uitgevoerd, wordt de data bewerkt. Vervolgens wordt er een keuze gemaakt welke data wordt ingevoerd in het algoritme. In Tabel 3 wordt het verschil van deze processen weergegeven bij de verschillende instellingen.

Tabel 3: Overzicht van de bewerking van data bij verschillende instellingen

	Basisbelasting	Apparaat	Apparaat + trend verwijderen
Ruwe databewerking	geen aanpassingen	datanormalisatie	eerst datanormalisatie, daarna detrend
Invoer datapunten	±1000 piekwaarden & ±1000 dalwaarden	±1000 piekwaarden	±1000 piekwaarden

4.1 **Bewerkingsstap**

De bewerkingsstap is de stap waarin de data van apparaten wordt aangepast voordat de data als invoer worden gebruikt voor het EM-algoritme. Dit is nodig omdat de vermogens moeten worden omgezet naar percentages en omdat de trend mogelijk verwijderd moet worden.

4.1.1 DATA

Om de algoritmes en processen te kunnen beschrijven is het noodzakelijk om eerst te beschrijven welke data er verwerkt moet worden. De data die in de algoritmes wordt verwerkt zijn elektriciteitsvermogens die gekoppeld zijn aan tijdstippen. Vaak zijn deze vermogens uitgedrukt in kW of W. Belangrijk is dat de eenheden in de dataset niet verschillen. De tijdstippen worden afgerond naar het dichtstbijzijnde kwartier. De data is afkomstig uit een tekstbestand. Hierin staan de gemeten vermogens van alle kwartieren onder elkaar weergegeven.

4.1.2 AANTAL TIJDSTIPPEN BEPALEN

Slechts een beperkt aantal tijdstippen dienen als invoer voor het EM-algoritme. Dit m -aantal tijdstippen wordt berekend aan de hand van de volgende vergelijking:

$$m = \left\lceil \frac{1000}{\text{aantal vermogens per tijdstip}} \right\rceil$$

Om uiteindelijk de 1000 piekwaarden of dalwaarden te bepalen worden er een m -aantal tijdstippen gekozen volgens de bovenstaande vergelijking (naar boven afgerond). De gemeten vermogens bij deze tijdstippen worden allemaal meegenomen in het EM-algoritme. Het aantal datapunten is dus altijd 1000 of hoger, mits er voldoende data beschikbaar zijn.

4.1.3 DATANORMALISATIE

Zoals eerder aangegeven wordt de datanormalisatie alleen uitgevoerd op apparaten. De datanormalisatie vindt plaats voordat het EM-algoritme wordt uitgevoerd. Het werkt als volgt: eerst wordt er voor elk gemeten apparaat (per kolom) de maximale waarde bepaald. Vervolgens wordt elke waarde van het apparaat (in de kolom) gedeeld door de maximale waarde en vermenigvuldigd met 100 om percentuele waarde te krijgen. Als de waarde uiteindelijk lager blijkt te zijn dan 5%, dan wordt de waarde naar nul gezet. De waarden onder de vijf procent vertegenwoordigen de uitgeschakelde toestand van het apparaat. In Tabel 4 wordt dit aan de hand van een voorbeeld uitgewerkt.

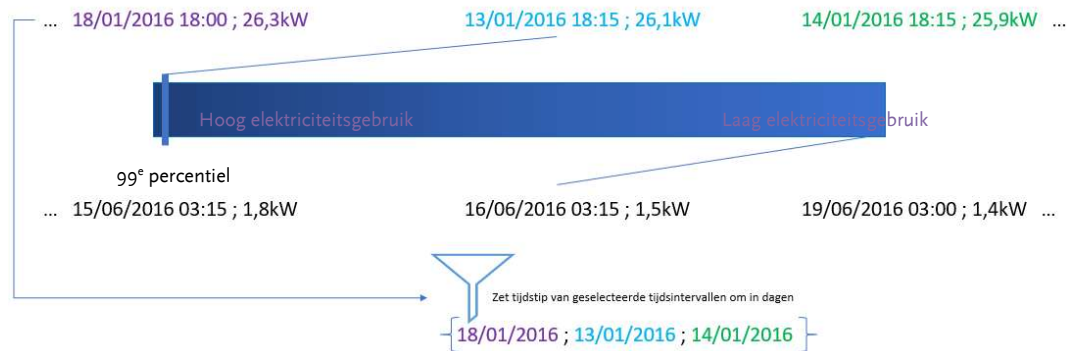


Tabel 4: Voorbeeld van datanormalisatie

Tijd	Apparaat 1	Apparaat 2	Normaliseer →	Tijd	Apparaat 1	Apparaat 2
00:00	2,5kW	6,2kW		00:00	67,6%	66,0%
00:15	2,7kW	0,4kW		00:15	73,0%	4,3% → 0,0%
00:30	3,1kW	8,6kW		00:30	83,8%	91,5%
00:45	3,4kW	9,4kW		00:45	91,9%	100,0%
01:00	3,7kW	8,4kW		01:00	100,0%	89,4%
...

4.1.4 DETREND

Detrend is de techniek die de trend uit de data verwijdert. Het werkt als volgt: eerst worden de tijdstippen gesorteerd op basis van het totale vermogen (van hoog naar laag). Vervolgens wordt er *m*-aantal tijdstippen geselecteerd rond het 0,99-kwantiel. Dit houdt in dat 1% van de tijdstippen een hoger totaal gemeten vermogen heeft. Het werken met deze kwantielen is een robuuste methode om uitschieters te vermijden. De datums van deze tijdstippen worden vervolgens opgeslagen in een lijst. In Figuur 12 wordt dit proces weergegeven.



Figuur 12: Selectie trenddagen

Als de dagen bepaald zijn, worden vervolgens alle vermogens van de tijdstippen die binnen de dagen vallen geselecteerd. De vermogens van tijdstippen die in hetzelfde kwartier vallen worden bij elkaar opgeteld en in een lijst bijgehouden. Deze lijst heeft een lengte van 96 (96 kwartieren op een dag). Vervolgens worden de waarden in de lijst gedeeld door het aantal dagen en het aantal gemeten vermogens per tijdstip. Het resultaat is een gemiddeld vermogen voor elk kwartier.

Deze lijst wordt de trend genoemd. De laatste stap die wordt uitgevoerd is om de trend van de data af te trekken. Er wordt over elke individuele waarde in de genormaliseerde datalist gelopen en vervolgens wordt er de corresponderende kwartierwaarde van de trend afgetrokken. De trend (percentages per kwartier van een dag en maandfactoren) wordt in het GM-type apart bewaard. Ze worden later toegepast bij de stochastische loadflowberekening.

4.2 Invoeringsstap

Van de beschikbare data wordt er maar een beperkt aantal datapunten geselecteerd die als invoer dienen voor het EM-algoritme. Deze datapunten horen bij piekmomenten (apparaat zowel basisbelasting) en dalmomenten (basisbelasting). Op deze tijdstippen wordt het GM-type getraind om betere probabilistische benadering te krijgen op piek- en dalmomenten. Het precieze aantal datapunten dat wordt meegenomen kan worden berekend aan de hand van de volgende vergelijkingen:

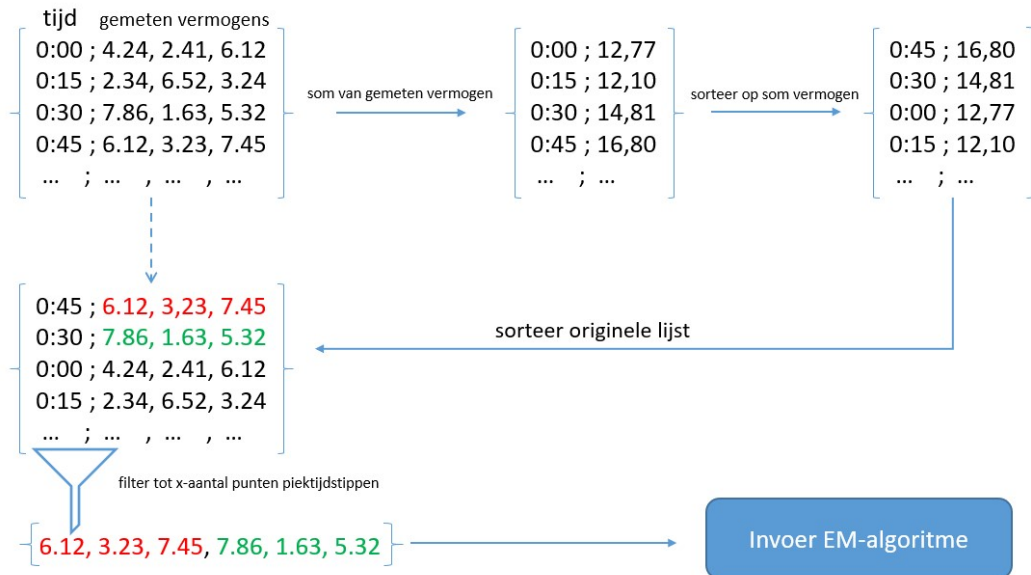
Piekmomenten (apparaat): $aantal\ datapunten = m \times aantal\ vermogens\ per\ tijdstip;$
 Piek- en dalmomenten (basisbelasting): $aantal\ datapunten = 2 \times m \times aantal\ vermogens\ per\ tijdstip.$

Van het *m*-aantal tijdstippen (zie paragraaf 4.1.2 voor de formule) met het hoogste totale vermogen, worden alle individuele vermogens meegenomen in het EM-algoritme. Dit noemt men de piekwaarden.



Bij de basisbelasting worden ook de dalwaarden meegenomen. Dit zijn het m -aantal tijdstippen met het laagste totale vermogen. Hiervan worden ook alle individuele vermogens meegenomen in het EM-algoritme.

Een voorbeeld van hoe piekmomenten worden geselecteerd is in Figuur 13 visueel te zien.



Figuur 13: Visualisatie datasetselectie procedure

4.3 Uitvoeringsstap

Het doel van het EM-algoritme is om de parameters van het GM-model zo te bepalen (gewicht, gemiddelde, variantie) dat de dataset optimaal wordt gerepresenteerd door het GM-model. Dit houdt in dat de vorm van de data zo goed mogelijk overeenkomt met de vorm van de verdelingen in het GM-model. Dit gebeurt middels een iteratief proces waarin de onderstaande doelfunctie f langzaam convergeert naar een (lokaal) maximum. Hieronder wordt de doelfunctie f weergegeven:

$$\max f = \sum_i \sum_j S(i, j)$$

waarbij $S(i, j)$ als een score kan worden gezien die berekend kan worden aan de hand van de volgende vergelijking:

$$S(i, j) = P(x_i | N_j) \cdot P(N_j) .$$

In de vergelijking van $S(i, j)$ staat x_i voor het individuele datapunt i en N_j de normaalverdeling j . Eerder in het verslag werd het gewicht beschreven als ω_j . Het gewicht ω_j kan in het EM-algoritme worden gezien als de kans op een specifieke normaalverdeling bij een willekeurige trekking uit de totale verdeling. Er geldt dus $P(N_j) = \omega_j$. De term $P(x_i | N_j)$ is de voorwaardelijke kansdichtheidsfunctie van de desbetreffende normaalverdeling. De vergelijking wordt hieronder weergegeven:

$$P(x_i | N_j) = \frac{1}{\sqrt{2\pi}\sigma_j} \cdot e^{-\frac{(x_i - \mu_j)^2}{2\sigma_j^2}}$$

Het iteratieproces wordt beëindigd zodra de stopcondities zijn bereikt. In dit geval is dat wanneer er 300 iteraties zijn uitgevoerd of wanneer het verschil in waarde van de doelfunctie f per iteratiestap minder dan 0,01 is geworden. De voorwaardelijke kansdichtheidsfunctie $P(x_i | N_j)$ oefent invloed uit op



de aannemelijkheid $P(N_j | x_i)$. De aannemelijkheid oefent weer invloed uit op het gemiddelde, de variantie en de gewichten. Hieronder is de functie van de aannemelijkheid beschreven:

$$P(N_j | x_i) = \frac{P(x_i | N_j) \cdot P(N_j)}{P(x_i)} = \frac{P(x_i | N_j) \cdot P(N_j)}{\sum_j P(x_i | N_j) \cdot P(N_j)}$$

De aannemelijkheidsformule $P(N_j | x_i)$ volgt de stelling van Bayes. Hierin staat $P(x_i)$ voor de kans dat datapunt x_i wordt waargenomen.

Het 1D-EM-algoritme dat in Gaia gebruikt wordt is terug te leiden naar de onderstaande formulereeks:

Expectation-stap

$$P(x_i | N_j) = \frac{1}{\sqrt{2\pi}\sigma_j} \cdot e^{-\frac{(x_i - \mu_j)^2}{2\sigma_j^2}}$$

$$P(N_j | x_i) = \frac{P(x_i | N_j) \cdot P(N_j)}{P(x_i)} = \frac{P(x_i | N_j) \cdot P(N_j)}{\sum_j P(x_i | N_j) \cdot P(N_j)}$$

Maximization-stap

$$\mu_j = \frac{\sum_i P(N_j | x_i) \cdot x_i}{\sum_i P(N_j | x_i)}$$

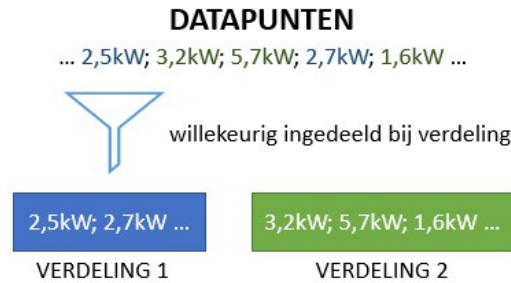
$$\sigma_j^2 = \frac{\sum_i P(N_j | x_i) \cdot (x_i - \mu_j)^2}{\sum_i P(N_j | x_i)}$$

$$\omega_j = \frac{\sum_i P(N_j | x_i)}{\sum_i \sum_j P(N_j | x_i)} = P(N_j)$$

In het EM-algoritme wordt er vaak een niet optimale startoplossing ingevoerd. Dit is ook het geval in de implementatie in Gaia. De startoplossing wordt berekend aan de hand van hard clustering. Dit houdt in dat de individuele datapunten alleen bijdragen aan het opstellen van een enkele normaalverdeling. Het EM-algoritme werkt met soft clustering. Dit houdt in dat individuele datapunten kunnen bijdragen aan het opstellen van meerdere normaalverdelingen. In de eerste iteratie veranderen de gemiddelden, varianties en gewichten daardoor drastisch. Dit zorgt er echter ook voor dat in de volgende iteratie de aannemelijkheden weer veranderen door de aangepaste gemiddelden en varianties. Dit heeft opnieuw effect op de gemiddelden, varianties en gewichten. In dit heen en weer rekenen van gemiddelden, varianties, gewichten en aannemelijkheden beginnen de datapunten zo steeds beter benaderd te worden door de normaalverdelingen en convergeert de doelfunctie f naar een (lokaal) maximum.

4.3.1 STARTOPLOSSING

De vergelijkingen die gebruikt worden in het EM-algoritme zijn afhankelijk van een eerder berekend gemiddelde, variantie en gewicht. Er is dus een startoplossing nodig om het algoritme te kunnen laten beginnen. De startoplossing wordt bepaald door de datapunten willekeurig toe te wijzen aan clusters. Van deze clusters worden vervolgens de variantie en het gemiddelde berekend. Het gewicht wordt ook berekend door het aantal toegewezen datapunten per cluster te delen door het totale aantal datapunten. Per cluster worden dus, met behulp van de data in het clusters zelf, de parameters van de normaalverdelingen berekend. Het indelen van de datapunten in de verschillende clusters wordt gevisualiseerd in Figuur 14.



Figuur 14: Voorbeeld van het hard clusteren van de datapunten in een willekeurige verdeling

Voor de basisbelasting worden er vier clusters gemaakt. Voor apparaten worden er twee clusters gemaakt en wordt het gemiddelde en de variantie van de normaalverdeling met het laagste gemiddelde later respectievelijk aangepast naar 0,00001 en 5.

Vervolgens wordt er ook voor elk datapunt x_i bij elke normaalverdeling N_j een score $S(i, j)$ berekend aan de hand van de eerder vermeldde vergelijking $S(i, j) = P(x_i | N_j) \cdot P(N_j)$.

4.3.2 ITERATIEPROCES

Om het verschil per iteratie te toetsen wordt de doelfunctie f uitgerekend, daarna worden de gewichten op nul gezet $\sum_j P(N_j) = 0$. Dit heeft geen gevolgen voor de al eerder berekende waarde $S(i, j)$.

Vervolgens wordt de aannemelijkheid $P(N_j | x_i)$ dat datapunt x_i afkomstig is van verdeling N_j berekend aan de hand van de volgende vergelijking:

$$P(N_j | x_i) = \frac{P(x_i | N_j) \cdot P(N_j)}{\sum_j P(x_i | N_j) \cdot P(N_j)} = \frac{S(i, j)}{\sum_j S(i, j)}$$

Deze berekening wordt uitgevoerd voor elk datapunt bij elke normaalverdeling. Daaropvolgend wordt de totale aannemelijkheid horende bij de verschillende normaalverdelingen berekend aan de hand van de vergelijking a_j :

$$a_j = \sum_i P(N_j | x_i)$$

De totale aannemelijkheid a_j wordt gebruikt om de nieuwe gewichten van de normaalverdelingen te bepalen volgens vergelijking ω_j :

$$\omega_j = P(N_j) = \frac{\sum_i P(N_j | x_i)}{\sum_i \sum_j P(N_j | x_i)} = \frac{a_j}{\sum_j a_j}$$

Dan worden de rijwaarden $R(i, j)$ berekend van de normaalverdeling. De aannemelijkheden worden gedeeld door de totale aannemelijkheid van de bijbehorende normaalverdeling. Dit gebeurt aan de hand van de volgende vergelijking:

$$R(i, j) = \frac{P(N_j | x_i)}{\sum_i P(N_j | x_i)} = \frac{P(N_j | x_i)}{a_j}$$

Met behulp van $R(i, j)$ kan nu het gemiddelde en de variantie berekend worden aan de hand van de onderstaande vergelijkingen:

$$\mu_j = \frac{\sum_i P(N_j | x_i) \cdot x_i}{\sum_i P(N_j | x_i)} = \sum_i R(i, j) \cdot x_i$$

$$\sigma_j^2 = \frac{\sum_i P(N_j | x_i) \cdot (x_i - \mu_j)^2}{\sum_i P(N_j | x_i)} = \sum_i R(i, j) \cdot (x_i - \mu_j)^2$$

Voordat de iteratie stopt wordt er voor elk punt opnieuw een score $S(i, j)$ berekend met behulp van de vergelijking $S(i, j) = P(x_i | N_j) \cdot P(N_j)$. Dit wordt gebruikt als invoer voor de volgende iteratiestap en dit iteratieproces herhaalt zich tot de stopcondities zijn bereikt.



4.3.3 TOEVOEGINGEN

De bovenstaande vergelijkingen kunnen omschreven worden als de kern van het EM-algoritme. In Gaia zijn er echter een aantal toevoegingen/wijzigingen aangebracht aan de vergelijkingen. Zo wordt er bij het berekenen van de score $S(i, j)$ rekening gehouden met een marge van error $m(N_j)$. Deze marge van error $m(N_j)$ wordt berekend aan de hand van de onderstaande vergelijking:

$$m(N_j) = 1/2 - 1/2 \times \operatorname{erf}\left(\frac{-\mu_j}{\sqrt{2\sigma_j^2}}\right).$$

Hierin refereert erf naar de wiskundige errorfunctie. De score $S(i, j)$ in de doelfunctie van Gaia wordt feitelijk berekend aan de hand van de volgende vergelijking:

$$S(i, j) = \frac{P(x_i|N_j) \cdot P(N_j)}{m(N_j)}$$

De convergentie van het EM-algoritme verloopt asymptotisch. Hierdoor kan de error marge $m(N_j)$ worden toegepast om het convergeren sneller te laten verlopen.

Als de stopcondities van het iteratieproces zijn bereikt dan worden er nog een aantal wiskundige operaties uitgevoerd. De nieuw berekende aannemelijkheden zijn nog niet omgezet in nieuwe gemiddelden, varianties en gewichten. Deze worden berekend aan de hand van lineaire interpolatie met de volgende vergelijkingen:

$$\begin{aligned}\mu_{j \text{ nieuw}} &= \mu_{j \text{ huidige iteratie}} + 1/2 \cdot (\mu_{j \text{ huidige iteratie}} - \mu_{j \text{ twee iteraties geleden}}) \\ \sigma_{j \text{ nieuw}}^2 &= \sigma_{j \text{ huidige iteratie}}^2 + 1/2 \cdot (\sigma_{j \text{ huidige iteratie}}^2 - \sigma_{j \text{ twee iteraties geleden}}^2) \\ \omega_{j \text{ nieuw}} &= \omega_{j \text{ huidige iteratie}} + 1/2 \cdot (\omega_{j \text{ huidige iteratie}} - \omega_{j \text{ twee iteraties geleden}})\end{aligned}$$

Dit is een eenvoudigere methode dan het gebruikelijke terugrekenen van de nieuwe gemiddelden, varianties en gewichten.

Als de optie 'Apparaat' geselecteerd is dan wordt het gemiddelde van de normaalverdeling met het laagste gemiddelde veranderd naar 0,00001 en de variantie naar 5. Als varianties van normaalverdelingen bij de basisbelasting en bij apparaten lager zijn dan 0,00001, dan wordt deze aangepast naar 0,00001. Ook is er een mechanisme dat afstand bewaart tussen de gemiddeldes van de normaalverdelingen. Dit gebeurt alleen bij de basisbelastingen en dit werkt als volgt:

- Het gemiddelde verschil ($\Delta\mu$) van de μ van de normaalverdelingen wordt berekend.
- Er wordt gekeken of het verschil tussen deze aparte gemiddelden lager is dan $0,1 \cdot \Delta\mu$. Als dit zo is dan wordt er bij het hoogste gemiddelde $0,05 \cdot \Delta\mu$ toegevoegd en het laagste gemiddelde $0,05 \cdot \Delta\mu$ afgetrokken, mits het nieuwe gemiddelde niet lager wordt dan 0.

Dit mechanisme wordt uitgevoerd om te voorkomen dat de normaalverdelingen in elkaar vallen (gelijke gemiddelden). De extra spreiding tussen de gemiddelden van de gekozen normaalverdelingen komt goed uit voor het post-processen. Hierin wordt het vermogen bij verschillende tijdstippen benaderd met de normaalverdelingen die zijn vastgezet in het EM-algoritme. Dat wordt lastig als sommige van de normaalverdelingen dicht op elkaar liggen en daardoor niet veel toegevoegde waarde hebben.

4.4 Overzicht

Eerst worden de piek- en dalmomenten geselecteerd. Deze tijdstippen worden gebruikt voor de training van GM-type. Vervolgens worden de vermogens bij deze tijdstippen als invoer gebruikt voor het EM-algoritme. Het EM-algoritme is een iteratief algoritme dat per iteratie telkens een beetje aan het convergeren is naar een betere score, deze score moet uiteindelijk gemaximaliseerd worden.

Het berekenen van de scores en aannemelijkheden noemt men de E-stap. Vanuit de aannemelijkheden worden de gewichten, gemiddelden en varianties opgesteld. Dit noemt men de M-stap. De kansdichtheidsfunctie van de normaalverdeling is te vinden in de aannemelijkheidsfunctie, dit zorgt ervoor dat de normaalverdelingen de data uiteindelijk gaan benaderen.



De gemiddeldes en varianties van de verschillende normaalverdelingen in de GM-belastingtypegenerator zijn afkomstig uit dit EM-algoritme. Het gewicht wordt later (opnieuw) per kwartier bepaald tijdens het post-processen.

5 **POST-PROCESSEN**

In dit hoofdstuk wordt de wiskundige werking achter het post-processen behandeld. Het post-processen bestaat uit de waarderingsstap en de factorenstap. Met behulp van de maand- en kwartierfactoren, die worden berekend in het post-processen, worden uiteindelijk de gewichten van de verschillende normaalverdelingen per kwartier bepaald.

5.1 **Waarderingsstap**

In de waarderingsstap worden er voor alle individuele tijdstippen gewichten bepaald. Deze gewichten dienen als invoer voor de factorenstap. Ook voor de waarderingsstap geldt dat er eerst een startoplossing nodig is. De waarderingsstap is in feite een kort EM-algoritme (twee iteraties) en maakt dus gebruik van dezelfde vergelijkingen met een paar bijzonderheden.

5.1.1 **STARTOPLOSSING**

De parameters die worden gehanteerd bij de startoplossing zijn:

$$\begin{aligned} \mu_j &= \mu_{j \text{ EM-algoritme}} \text{ ,} \\ \sigma_j^2 &= \sigma_{j \text{ EM-algoritme}}^2 \text{ .} \end{aligned}$$

De gemiddeldes μ_j en varianties σ_j^2 worden dus overgenomen uit het EM-algoritme. De gewichten $P(N_j)$ van de normaalverdelingen worden gelijkgesteld, via de volgende vergelijking:

$$P(N_j) = \frac{1}{J}$$

waarin J staat voor het totaal aantal normaalverdelingen. Er wordt een marge van error $m(N_j)$ berekend voor elke verdeling via de volgende vergelijking:

$$m(N_j) = 1/2 - 1/2 \times \operatorname{erf} \left(\frac{-\mu_j}{\sqrt{2\sigma_j^2}} \right)$$

5.1.2 **ITERATIE**

Over elk tijdstip wordt apart geïtereerd. Elk tijdstip krijgt uiteindelijk een gewicht toegewezen. Hierin wordt alleen de data van het specifieke tijdstip meegenomen om de gewichten te bepalen. Hiervoor wordt eerst de score $S(i, j)$ berekend aan de hand van de volgende vergelijking:

$$S(i, j) = \frac{P(x_i | N_j) \cdot P(N_j)}{m(N_j)}$$

De score $S(i, j)$ wordt gebruikt als invoer voor de aannemelijkheidsfunctie $P(N_j | x_i)$, die berekend wordt met de volgende vergelijking:

$$P(N_j | x_i) = \frac{P(x_i | N_j) \cdot P(N_j)}{\sum_j P(x_i | N_j) \cdot P(N_j)} = \frac{S(i, j)}{\sum_j S(i, j)}$$

Vervolgens wordt de totale aannemelijkheid a_j per verdeling uitgerekend aan de hand van de onderstaande vergelijking:

$$a_j = \sum_i P(N_j | x_i) \text{ .}$$



De totale aannemelijkheid a_j van een normaalverdeling wordt gebruikt om het gewicht $P(N_j)$ te berekenen. Dit gebeurt aan de hand van de volgende vergelijking:

$$P(N_j) = \omega_j = \frac{\sum_i P(N_j|x_i)}{\sum_i \sum_j P(N_j|x_i)} = \frac{a_j}{\sum_j a_j}$$

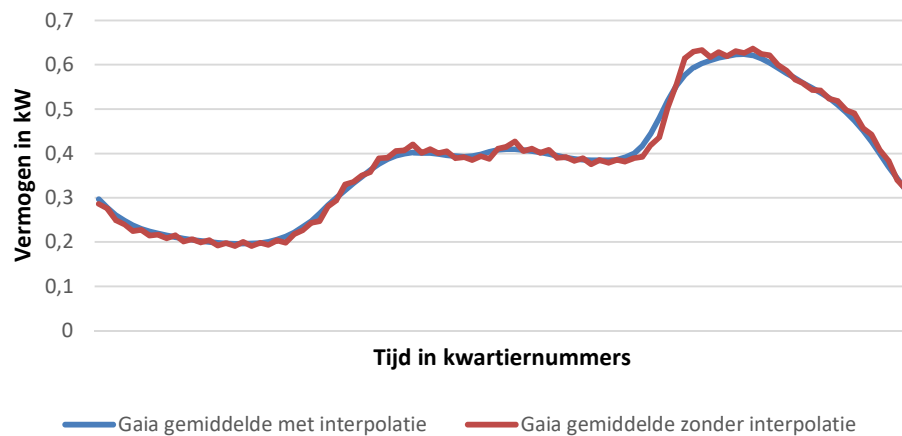
Om een betere benadering te maken wordt deze iteratie nog een keer uitgevoerd. De startoplossing van het volgende tijdstip verandert nu naar het uitgerekende gewicht van dit tijdstip.

5.1.3 INTERPOLATIE

Als alle gewichten zijn berekend dan worden de gewichten geïnterpoleerd over zeven tijdstippen. Dit gebeurt voor elk gewicht volgens de volgende vergelijking:

$$\omega_{n,j} = \frac{1}{7} \cdot \sum_{t=n-3}^{n+3} \omega_{t,j}$$

Hierin zijn t en n de indexatie van een specifiek tijdstip. Dankzij deze interpolatie worden gewichten die veel uitschieters gedempt. Dit is toegevoegd omdat deze afzonderlijke tijdstippen weinig data bevatten, enkele uitschieters kunnen dan grote verschuivingen in het gewicht veroorzaken. In Figuur 15 is het effect van de interpolatie op het gemiddelde van de GM-verdelingen per kwartier te zien. Deze is berekend met de volgende vergelijking: $\mu(x) = \sum_{j=1}^J \omega_j \cdot \mu_j$.



Figuur 15: Interpolatie effect op gemiddelde GM-modellen maand januari

5.2 Factorenstap

In de factorenstap worden de maandfactoren, de weekendfactoren en de werkdagfactoren berekend. Dit gebeurt met behulp van de eerder berekende gewichten in de waarderingsstap.

5.2.1 MAANDFACTOREN

De factorenstap begint met het berekenen van de maandfactoren. Eerst worden alle gewichten $\omega_{m,j}$ van het maandnummer m (1 t/m 12) en verdeling j berekend aan de hand van de volgende vergelijking:

$$\omega_{m,j} = \frac{1}{T_m} \cdot \sum_i^{T_m} \omega_{i,j}$$

waarin T_m de tijdstippen zijn die in het maandnummer m vallen. Het gewicht $\omega_{i,j}$ is afkomstig van tijdstip i en verdeling j uit de waarderingsstap.

Mochten er bij bepaalde maanden gewichten ontbreken, dan worden die gewichten bepaald via interpolatie aan de hand van de volgende vergelijking:



$$\omega_{m,j} = \frac{1}{T_{m-1} + T_{m+1} + T_{13-m}} \cdot \left(\sum_i^{T_{m-1}} \omega_{i,j} + \sum_i^{T_{m+1}} \omega_{i,j} + \sum_i^{T_{13-m}} \omega_{i,j} \right)$$

De maandfactor $mf_{m,j}$ van maandnummer m en verdeling j wordt berekend aan de hand van de volgende vergelijking:

$$mf_{m,j} = \frac{\omega_{m,j}}{\omega_{m_{argmax},j}}$$

Hierin verwijst m_{argmax} naar het maandnummer dat de meeste tijdstippen bevat. Deze draagt immers het meeste bij aan de parameters van de verschillende normaalverdelingen. Bij maandnummers met een gelijk aantal tijdstippen wordt het laagste maandnummer gekozen.

5.2.2 TRENDFACTOREN

De trendfactor v_m wordt bepaald via de volgende vergelijking:

$$v_m = \frac{1}{T_m} \cdot \sum_i^{T_m} \sum_j^t x_j$$

waarin $\sum_j^t x_j$ alle datapunten zijn die in een tijdstip van T_m vallen. Mocht er bij bepaalde maanden data missen dan worden deze waarden geïnterpoleerd aan de hand van de volgende vergelijking:

$$v_m = \frac{1}{3} \cdot (v_{m-1} + v_{m+1} + v_{13-m})$$

Vervolgens wordt de trend genormaliseerd met behulp van de volgende vergelijking:

$$v_{m,pu} = \frac{v_m}{v_{max}}$$

Hierin is max het maandnummer waarbij v_m maximaal is.

5.2.3 KWARTIERFACTOREN

Als de maandfactoren bekend zijn kunnen ook de kwartierfactoren worden uitgerekend. Dat gebeurt eerst door de parameter $b_{k,j}$ te bepalen voor verdeling j . Hierin staat T_k voor de tijdstippen die in het kwartiernummer k (1 t/m 96) vallen. De parameter $b_{k,j}$ wordt berekend aan de hand van de volgende vergelijking:

$$b_{k,j} = \frac{1}{T_k} \cdot \sum_i^{T_k} \omega_{i,j} \cdot mf_{m,j}$$

Mochten er bij bepaalde kwartiernummers waarden onbekend zijn, dan worden deze waarden bepaald middels interpolatie met behulp van de volgende vergelijking:

$$b_{k,j} = \frac{1}{2} (b_{k-1,j} + b_{k+1,j})$$

Daarna kan de kwartierfactor $kf_{k,j}$ worden berekend voor verdeling j en kwartiernummer k aan de hand van de volgende vergelijking:

$$kf_{k,j} = \frac{b_{k,j}}{\sum_j b_{k,j}}$$

5.2.4 APPARATEN EN WEEKENDFILTER

Voor apparaten worden er parameters van maar één verdeling berekend. Voor het weekendfilter wordt er onderscheid gemaakt tussen tijdstippen die in het weekend vallen en op werkdagen. Het proces van kwartierfactoren opstellen wordt afzonderlijk van elkaar uitgevoerd met deze data. Daardoor ontstaan er dus twee keer zoveel kwartierfactoren.