

Inleiding

Om bij de Investeringsstatistiek TopDown-gaafmaken mogelijk te maken, is een deel van het topdown-package zoals uitgerold door Agile Team Generiek (versie 0.2.4) geïmplementeerd in de huidige gaafmaakttool. Met de scores die hierbij in de tool beschikbaar komen, kunnen de meest afwijkende en invloedrijke records het eerst bekeken en eventueel worden aangepast. In deze documentatie is meer informatie te vinden over de werking van het TopDown-gaafmaken en over in te stellen variabelen.

Korte inhoudsopgave

1. Algemene info
2. Opbouw van de tool
3. Regels
4. Datapreparatie in SQL Server
5. Module in R.
6. Aftrappen scoreberekening en logging

1. Algemene info

In december 2021 is bij de investeringsstatistiek TopDown gaafmaken geïmplementeerd. Bij TopDown gaafmaken worden voor elk record scores berekend waarbij de scores de kwaliteit van een record proberen weer te geven. In het kort komt het erop neer dat per record, per variabele wordt bekeken of deze een invloedrijke afwijking heeft t.o.v. een bepaalde referentie. De referentie is hierbij veelal dezelfde variabele voor hetzelfde record, maar dan in jaar T-1 en de invloedrijkheid wordt bepaald aan de hand van hoeveel de waarde van deze variabele in het record bijdraagt aan een grotere aggregatie (groep van meerdere records, bijvoorbeeld een SBI). De combinatie van deze twee factoren wordt vastgelegd in een score voor een variabele (lokale score) en meerdere scores vormen voor dat record een globale score. Het is uiteindelijk de bedoeling dat records met de hoogste score (meest verdachte records; grootste afwijking i.c.m. grootste invloed) geprioriteerd worden om gaaf te maken dan wel te herbekijken. Deze manier van hoogste scores eerst bekijken wordt dan ook TopDown gaafmaken genoemd.

Over hoe de scorefuncties precies geïnterpreteerd moeten worden, is meer te vinden in *bijlage 1*, waar een kopie is opgenomen van het bestand *scorefuncties_voor_dashboards_v8.docx*.

Bij deze implementatie is gebruik gemaakt van het TopDown-package van Agile Team Generiek (versie 0.2.4). Dit package zorgt voor een eenduidige manier van scores berekenen, waarbij voor elke score een doelvariabele, referentievariabele, deler en aggregatie aangewezen wordt.

2. Opbouw van de tool

De 4 eenheden die nodig zijn om de scores te berekenen worden voorbereid middels datapreparatie in SQL Server Views op de database van de investeringsstatistiek:

4.1.4

De data wordt voorbereid in de volgende views:

```
sco.vw_data_waarde
sco.vw_data_referentie
sco.vw_data_totaal
```

Deze views worden ingelezen in een R-script

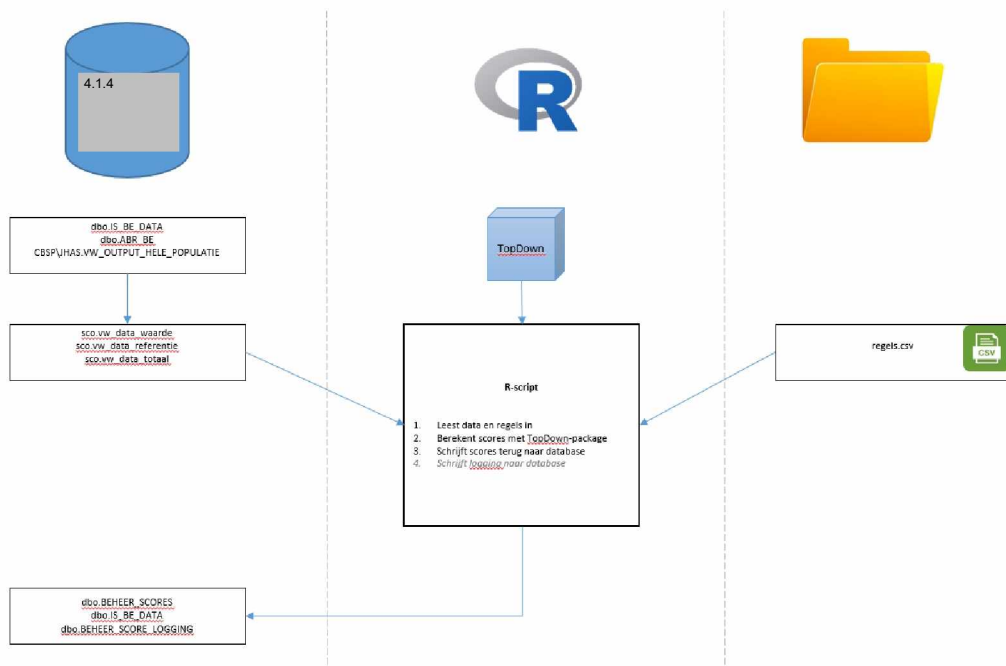
4.1.4

verder) waarin ook het TopDown package en een regelbestand worden ingeladen. Hiermee worden de scores berekend. Deze scores worden vervolgens terugschreven naar de volgende tabellen:

<code>dbo.BEHEER_SCORES</code>	(inclusief alle lokale scores)
<code>dbo.BEHEER_SCORES_LOKAAL</code>	(inclusief alle lokale scores, diepte-opslag)
<code>dbo.IS_BE_DATA</code>	(alleen de globale score waarop TopDown gaafmaken wordt gebaseerd)

De laatste tabel wordt uiteindelijk door de statistiek gebruikt om TopDown gaaf te kunnen maken. De tweede tabel wordt gebruikt om te kunnen achterhalen welke lokale score een eventuele hoge score veroorzaakt.

In onderstaande afbeelding wordt weergegeven hoe de verschillende onderdelen met elkaar samenhangen.



Afbeelding 1: *samenhang tussen de verschillende onderdelen voor TopDown bij investeringen.*

In de volgende paragrafen worden de losse onderdelen een voor een uitgelegd, waarbij bij ieder onderdeel ook duidelijk wordt welke eventuele aanpassingen gemaakt moeten worden bij het implementeren van meer scores of het wijzigen van scores.

3. Regels

In het TopDown package wordt een set regels ingelezen die de lokale scores van een eenheid bepalen. Deze regels zijn door de afdeling investeringsstatistiek (5.1.2.e) en methodologie (5.1.2.e) gezamenlijk opgesteld en zien er als volgt uit.

Lokale scores

Alle scores behalve score6 en score10 hebben de volgende vorm:

$$\text{score} = \frac{|\text{waarde} - \text{referentiewaarde}|}{\text{totaal}},$$

waarbij totaal een aggregaat op SBI 2-digit-niveau is.

Een score van, zeg, 0,05 komt overeen met een afwijking tussen waarde en referentiewaarde die gelijk is aan 5% van een totaalwaarde op SBI 2-digit-niveau waaraan de betreffende eenheid bijdraagt.

De volgende tabel geeft aan hoe deze formule is ingevuld voor de verschillende scores. De nummering van de scores komt overeen met de prioritering uit het Excelbestand "20211006 EBN2x Brainstorm regels investeringen". Variabelen hebben betrekking op het huidige jaar (T) tenzij anders aangegeven. Een score die gebruikmaakt van individuele waarden uit vorig jaar (T-1) blijft leeg als geen (T-1)-waarde bekend is.

score	waarde	referentiewaarde	totaal	opmerking
1_MVA_Tm1	C014004	0	C014004(T-1)	
1_lmVA_Tm1	C017004	0	C017004(T-1)	
2_MVA	C014004(T-1)	0	C014004(T-1)	
2_lmVA	C017004(T-1)	0	C017004(T-1)	
3_MVA	C014004	C014004(T-1)	C014004(T-1)	
3_lmVA	C017004	C017004(T-1)	C017004(T-1)	
4	E001001	0	E001001(T-1)	
5	F001001	F002001(T-1)	F002001(T-1)	
6	is gelijk aan 0.5 als $GK \geq 7$ en $C014004 + C017004 = 0$, anders 0 (en blijft leeg als SBI begint met 78)			de waarden 0.5 en 7 zijn instelbare parameters
8	C014004	I001001	C014004(T-1)	

Tabel 1: uitwerking score 1 t/m 8 – variant 1

Update: extra regel toegevoegd op 18/11/2022:

score	waarde	referentiewaarde	totaal	opmerking
13_F002	F002001	0	C014004(T-1)	

Tabel 1: uitwerking score 13

Deze eerste regels zijn allen van de vorm zoals (A1) in *bijlage 1*. Bij regel 9 is vervolgens het idee om voor de deelscores 9.1 t/m 9.13 gebruik te maken van variant 2 uit het topdown-package in plaats van variant 1. Dit is hetgeen dat staat omschreven als formule (A1*) in *bijlage 1*. Hierbij krijgt de variabele C014004 (oij) een rol als variabele waar steeds aan gerelateerd wordt.

score	waarde	referentiewaarde	oij	totaal	opmerking
9.1	C001004	C001004(T-1) / C014004(T-1)	C014004	C014004(T-1)	deelscore
9.2	C002004	C002004(T-1) / C014004(T-1)	C014004	C014004(T-1)	deelscore
...	...	...	...	...	...
9.13	C013004	C013004(T-1) / C014004(T-1)	C014004	C014004(T-1)	deelscore

Tabel 2: uitwerking score 9 – variant 2 met blokscore

De deelscores score9.1 t/m score9.13 worden vervolgens samengevat in score9 door het maximum te nemen.

Verder is score10 gelijk aan het maximum van de deelscores score10.1 t/m score10.4, mits minimaal één van deze deelscores niet leeg is (en anders zelf leeg). Deze deelscores hebben de vorm

$$\text{score} = \begin{cases} \frac{\max\{\text{referentiewaarde} - \text{waarde}, 0\}}{\text{totaal}} & \text{als referentiewaarde} > 0 \\ \text{leeg} & \text{anders} \end{cases}$$

waarbij geldt:

Score	waarde	referentiewaarde	totaal	opmerking
10.1	C002004	MvA01001 – MvA01002	C002004(T-1)	deelscore
10.2	C003004	MvA02001 – MvA02002	C003004(T-1)	deelscore
10.3	C004004	MvA03001 – MvA03002	C004004(T-1)	deelscore
10.4	C014004 – (C002004 + C003004 + C004004)	MvA04001 – MvA04002	C014004(T-1) – (C002004(T-1) + C003004(T-1) + C004004(T-1))	deelscore

Tabel 3: uitwerking score 10 met blockscore

Waar relevant is de absolute waarde $|\text{waarde} - \text{referentiewaarde}|$ die gebruikt is in de berekening van een score ook opgeslagen in het Excelbestand onder de naam “diff”. Bij diff9 (horend bij score9) en diff10 (horend bij score10) betreft dit het maximum van de diff-waarde voor alle deelscores.

Globale score

Per eenheid wordt een globale score berekend die de verschillende lokale scores samenvat in één getal. Er zijn twee varianten bepaald:

- globale_score_gem is het (ongewogen) gemiddelde van de volgende lokale scores: score1_MvA_Tm1, score1_lmVA_Tm1, score2_MVA, score2_lmVA, score3_MVA, score3_lmVA, score4, score5, score6, score7, score8, score9 en score10. Per eenheid worden lokale scores die leeg zijn overgeslagen.
- globale_score_max is het maximum van dezelfde scores.

In elke score is telkens de waarde, referentiewaarde en totaal van belang. Deze zijn telkens in het regelbestand opgenomen en worden in de datapreparatie voorbereid. Om het gebruik van de tool overzichtelijk te maken en niet het TopDown package als een black box te laten fungeren, is de samenhang tussen de datapreparatie en de variabelen die in de regels genoemd worden dan ook heel belangrijk; de variabelen die in de regels worden genoemd moeten in de afzonderlijke views samengesteld worden. Het regelbestand is te vinden op de volgende locatie:

4.1.4

Hierin bevatten de eerste 9 kolommen (**kolom A t/m I**) vooral meta-informatie (o.a. naam van de regel en eventuele aanpassingsdatum), behalve kolom E.

Kolom E bevat het gewicht van een de individuele regel. Hoe hoger dit gewicht, hoe meer dit meetelt in de globale score.

Kolom J bevat het gewicht van het record. Gezien dit bij investeringen (nog) niet aan de orde is, is er voor gekozen om een hier een kolom DummyWeight voor aan te maken waarin altijd de waarde 1 is opgenomen. Deze variabele wordt in de R-module aangemaakt in de data. Mocht een gewicht van

het record in de toekomst wel belangrijk worden, dan kan de variabelenaam van dat gewicht hier gewoon ingevuld worden en moet er bij de view `vw_data_waarde` voor gezorgd worden dat deze variabele onder diezelfde naam meegegeven wordt.

Kolom K (goal) bevat de doelvariabele, zoals in bovenstaande tabellen omschreven met *waarde*. In de datapreparatie (`vw_data_waarde`) komt de suffix `_T` of `_Tm1` nog niet voor, maar in het regelbestand moet deze wel toegevoegd worden, afhankelijk van of je een doelvariabele uit het uitgangsjaar T of uit T-1 neemt.

Kolom L bevat de referentiewaarde, die in bovenstaande tabellen per regel gedefinieerd is. De referentiewaarde wordt in `vw_data_referentie` aangemaakt en krijgt in het regelbestand altijd de suffix `_ref_T` of `_ref_Tm1`, afhankelijk van of een waarde uit jaar T of jaar T-1 de referentiewaarde bevat. Deze suffixes zijn altijd nodig om in de data (binnen de R-module) altijd duidelijk het onderscheid te kunnen maken tussen een variabele die als doelvariabele, referentievariabele of totaalvariabele gebruikt wordt. Dit heeft te maken met het aan elkaar plakken (LEFT_JOINEN) van de data. Als daar twee keer exact dezelfde variabelenaam wordt gebruikt, dan kunnen daarbij problemen ontstaan.

Als er geen referentiewaarde is, kan hier gewoon None worden ingevuld. Leeg laten zorgt voor errors in de berekening. Het TopDown package verwacht hier namelijk een kolomnaam (/variabelenaam) en als die niet wordt ingevuld dan gaat de module op zoek naar een kolom zonder naam, die niet bestaat. De kolomnaam *None* wordt in de R-module automatisch aan de data toegevoegd met voor alle records waarde 0.

Kolom M bevat de variabele waar telkens aan gerelateerd wordt. Deze hoeft alleen gevuld te zijn als in kolom P is gekozen voor variant 2. Er wordt in dat geval bij een score naar structuur gekeken (t.o.v. *Oij*) in plaats van naar de absolute waarden. De waarde in kolom M krijgt altijd een suffix *_T* of *_Tm1*.

Kolom N bevat de deler. Deze variabele geeft aan welke variabele het aggregaat bevat waardoor gedeeld moet worden. Omdat bij investeringen de totale populatie uit een andere tabel komt (CBSP\JHAS.VW_OUTPUT_HELE_POPULATIE) dan de gaaf te maken records (dbo.IS_BE_DATA), is dit aggregaat telkens voorberekend. Dit wordt gedaan in *vw_data_totaal*. Het totaal wordt momenteel voorbereid op 2-cijferig SBI-nummer. Belangrijk is dat in deze kolom altijd de suffix *_totaal* gebruikt moet worden.

Kolom O bevat de variabele waarop de deler van kolom N betrekking heeft. Gezien het totaal in N moest worden voorberekend, is deze *group_by* niet van toepassing en geldt deze dus eigenlijk per record. Vandaar dat hier altijd een variabele moet staan die een record individueel kan identificeren. We gebruiken daarvoor de *Beid*.

In **kolom P** staat vermeld welke variant van de scorefunctie gebruikt moet worden voor de regel. Variant 1 staat voor formule A1 in *bijlage 1*. Variant 2 staat voor formule A1* in *bijlage 1*. Als variant 2 wordt gekozen, moet ook altijd een variabele in kolom M worden ingevuld.

In de uitwerking van de regels hierboven moet van regel 9 en 10 het maximum genomen worden van alle deelscores. Hiervoor wordt in **kolom R** het block gevuld. Alle deelscores die bij elkaar horen, krijgen dezelfde block-naam. In het R-script is het zo geregeld dat alleen het maximum van deze blockscores mee wordt genomen in de bepaling van de globale score.

In **kolom S** kan aan deze blockscores nog een gewicht worden meegegeven; er kan worden aangegeven of de ene score belangrijker is dan de andere. Dit heeft enkel nut als een blockscore wordt bepaald door het gemiddelde van de deelscores te nemen. Aangezien dat hier niet het geval is (het maximum wordt genomen) is deze score hier niet van belang. Er moet wel een gewicht worden ingevuld om de module juist te laten lopen.

4. Datapreparatie in SQL Server

De datapreparatie vindt in SQL Server plaats. Hier is bewust voor gekozen om de volgende redenen:

- datapreparatie duidelijk gescheiden van het R-script en het TopDown package → geen black box;
- duidelijke samenhang tussen regelbestand en datapreparatie; de variabelen die in het regelbestand worden genoemd, moeten beschikbaar worden gemaakt in de datapreparatie (exclusief de suffixes zoals in paragraaf 3 benoemd: *_T*, *_Tm1*, *_ref_T*, *_ref_Tm1*, *_totaal*);
- gebruikers van de investeringenstatistiek zijn zeer bekend met SQL Server, waardoor bewerkingen hierin makkelijk uitvoerbaar zijn door de business zelf.

De data wordt in 3 views op de database voorbereid:

- *vw_data_waarde*
- *vw_data_referentie*
- *vw_data_totaal*

vw_data_waarde

Hierin worden de variabelen voorbereid die in het regelbestand zijn ingevuld in kolom K en die in tabel 1 t/m 3 te vinden zijn in de kolom *waarde*. De view is een samenvoeging van de tabellen *dbo.IS_BE_DATA* en *dbo.ABR_BE*, waarin de eerste de variabelen bevat die in de scores gebruikt worden en de tweede gebruikt wordt om SBI en grootteklasse aan een record te koppelen. De SBI-

informatie is nodig om het record door in de score aan de juiste deler te kunnen koppelen. Een record met SBI 29 moet bijvoorbeeld gedeeld worden door de groep records met SBI 29.

Voor de meeste variabelen spreekt de view voor zich. Er zijn twee uitzonderingen:

1. Regel 6 voldoet niet aan de regels zoals het TopDown package die kan interpreteren en is als volgt gedefinieerd:
is gelijk aan 0.5 als $GK \geq 7$ en $C014004 + C017004 = 0$, anders 0 (en blijft leeg als SBI begint met 78)
Dit krijgen we voor elkaar door deze voorwaarde hier al volledig uit te werken, de referentiewaarde altijd 0 te maken (zie *vw_data_referentie*) en de deler altijd 1 (zie *vw_data_totaal*).
2. Regel 10 voldoet niet aan de regels zoals die door het TopDown package kunnen worden geïnterpreteerd (referentiewaarde – waarde i.p.v. waarde – referentiewaarde), daardoor wordt de teller hier al volledig voorbereid.

vw_data_referentie

Hierin worden dezelfde twee basistabellen gebruikt als in *vw_data_waarde*, maar deze view bereidt de data voor zoals deze in de kolom *referentiewaarde* van de tabellen 1 t/m 3 in dit document staan. Bij regel 9 bestaat de referentiewaarde uit een deelsom. Omdat zowel teller als noemer leeg kunnen zijn, is het volgende bepaald:

- als er helemaal geen (T-1)-waarden kunnen worden aangekoppeld, dan wordt de nieuwe score 9 NA;
- als er wel (T-1)-waarden kunnen worden aangekoppeld, maar $C014004(T-1) = NA$ of 0 (ingevuld), dan wordt de nieuwe score 9 ook NA, omdat er geen structuurinformatie beschikbaar is uit T-1;
- als er wel (T-1)-waarden kunnen worden aangekoppeld, $C014004(T-1)$ is ingevuld en ongelijk aan 0, en een van de variabelen $C001004(T-1)$ t/m $C013004(T-1) = NA$, vervang die NA dan in de berekening van de betreffende deelscore door 0.

vw_data_totaal

Omdat de data voor de aggregaten waardoor gedeeld moet worden (totale populatie) een andere basisset is dan de data die gaafgemaakt moet worden, moet ook de *denominator* vooraf berekend worden. Dat gebeurt in de view *vw_data_totaal*. Deze view heeft de tabel [CBSP\JHAS].[VW_OUTPUT_HELE_POPULATIE] als bron en sommeert alle relevante variabelen op SBI-2-digit-niveau.

5. Module in R

In het R-script vindt de berekening van de scores plaats met behulp van het generieke TopDown package. De packages die het R-script nodig heeft (naast TopDown nog een aantal andere) zijn allemaal geïnstalleerd in dezelfde map als waar het R-script staat en worden ook van daaruit aangeroepen. Dit heeft als voordeel dat niemand de packages lokaal hoeft te installeren en dat iedereen (ook JOBS) dezelfde versie gebruikt. Het is dus bij het uitrollen van een nieuwe versie van een van de packages dat ook van belang dat op die plek te doen en goed te testen.

Het R-script is verdeeld in een aantal stappen, die hieronder kort worden beschreven.

1. Het inladen van de data.
De data uit de eerder beschreven datapreparatie (drie views) wordt opgehaald voor de jaren T en T-1).
2. Het koppelen van de data.
De data uit de drie views wordt gekoppeld (left-joins op de *vw_data_waarde* van jaar T) en er worden eventuele additionele variabelen aangemaakt (zoals `None = 0` en `DummyWeight = 1`).

3. Het regelbestand wordt ingelezen.
4. De scores worden berekend
Hierbij worden de blockscores berekend uit lokale (deel)scores en deze blockscores gaan dan weer mee in de berekening voor de globale score. Voor een uitleg van de globale score, zie *bijlage 1*.
5. De scores worden teruggeschreven naar de tabellen op de database.
Eerst worden alle lokale scores en de globale score weggeschreven naar respectievelijk de tabel *dbo.BEHEER_SCORES* en *dbo.BEHEER_SCORES_LOKAAL*. De bedoeling van deze tabellen is om bij het zien van een hoge score eventueel in te kunnen zoomen op de lokale scores die daar onder liggen.
Vervolgens wordt de globale score automatisch weggeschreven naar de tabel *dbo.IS_BE_DATA*.

Om te bepalen of het totaal waardoor gedeeld moet worden (denominator) jaar T-1 of T moet zijn, wordt automatisch het grootste (maximum) jaartal uit *vw_data_totaal* genomen. Het is dus aan de gebruiker om daar de gegevens van het juiste jaartal in te zetten.

6. Aftrappen scoreberekening en logging

De berekening van de scores kan gestart worden middels een dubbelklik op het volgende bestand:

4.1.4

Het uitvoeren van de berekeningen en wegschrijven naar de database duren bij schrijven van deze documentatie ongeveer 90 seconden.

Bij een succesvolle run wordt datum en tijd van de run en de term *script succesvol doorlopen* weggeschreven naar de tabel *dbo.BEHEER_SCORE_LOGGING*. Als er een fout is opgetreden is de melding die beschrijft waar in het script er iets is fout gegaan. Die melding kan in het script worden opgezocht om zo de fout te kunnen herstellen. Let op: aangezien het bestand afgetrapt gaat worden vanuit JOBS, is dit de enige indicator of de module al dan niet succesvol is doorlopen.

In bovengenoemd CMD-bestand is het uitgangsjaar als parameter opgenomen. De scores worden dus voor dat jaar berekend. Dit jaar kan veranderd worden door met de rechtermuisknop op het bestand te klikken, dan op bewerken te klikken en vervolgens het jaartal te veranderen.

Bijlage 1: Scorefuncties in een generiek dashboard voor top-down-gaafmaken

5.1.2.e met input 5.1.2.e
5.1.2.e 5.1.2.e en 5.1.2.e (16-9-2021)

Dit document geeft een overzicht van de formules van de scorefuncties voor top-down-gaafmaken die zijn ingebouwd in het prototype-dashboard dat is ontwikkeld tijdens de POC EBN 2.x Cluster 1 – Fase 2. Deze scorefuncties zijn een nadere uitwerking van een deel van de theorie uit het document “Kwaliteitsmaten voor selectief gaafmaken (scores)” uit Fase 1 van de POC. Voor meer toelichting over de methodologie achter de scorefuncties wordt verwezen naar dat document.

De beschrijving is opgedeeld in vier hoofdstukken:

- A. Scorefuncties voor gebruik binnen één statistiek, per eenheid ten opzichte van een vast aggregatieniveau
- B. Samenvattende scorefuncties per aggregaat, voor gebruik binnen één statistiek
- C. Scorefuncties voor confrontatie tussen statistieken, per eenheid ten opzichte van een vast aggregatieniveau
- D. Samenvattende scorefuncties per aggregaat, voor confrontatie tussen statistieken

A. Scorefuncties voor gebruik binnen één statistiek, per eenheid t.o.v. een vast aggregatieniveau

Lokale scores bij eenheid i :

De meest basale lokale scorefunctie heeft de volgende vorm:

$$s_{ij,G} = \frac{|v_i * (y_{ij} - \tilde{y}_{ij})|}{|Y_{j,G}|}, \quad (j = 1, \dots, J). \quad (A1)$$

Hierbij is v_i het ophooggewicht van eenheid i , y_{ij} de ruwe waarde van doelvariabele y_j voor eenheid i en $\tilde{y}_{ij}$ de referentiewaarde van variabele y_j voor eenheid i . De aggregaten $Y_{j,G} = \sum_{i \in G} v_i * y_{ij}$ hebben alle betrekking op hetzelfde aggregatieniveau: het zijn bijvoorbeeld allemaal kernceltotalen, of allemaal bedrijfstakttotalen. In de noemer van (A1) staat steeds het betreffende aggregaat (kerncel, bedrijfstak, regkol, ...) waaraan eenheid i zelf bijdraagt.

Afhankelijk van de specifieke toepassing kunnen ook andere lokale scorefuncties worden gedefinieerd. Bijvoorbeeld bij structuurvariabelen in de Productiestatistieken (PS):

$$s_{ij,G} = \frac{|v_i * (y_{ij} - \tilde{q}_{ij} * o_{ij})|}{|Z_{j,G}|}, \quad (j = 1, \dots, J), \quad (A1^*)$$

waarbij o_{ij} de waarde van eenheid i is voor een variabele o_j waarmee de doelvariabele y_j wordt vergeleken (bij de PS meestal totale bedrijfsopbrengsten). Verder is $\tilde{q}_{ij}$ een referentiewaarde voor de verhouding $q_{ij} = y_{ij}/o_{ij}$. In de noemer staat ofwel $Z_{j,G} = Y_{j,G}$ (bij doelvariabelen met strikt positieve totalen), ofwel $Z_{j,G} = O_{j,G}$ (bij doelvariabelen waarvoor een totaal ≤ 0 kan zijn).

N.B. De index $j = 1, \dots, J$ loopt over de lokale scores, niet over de doelvariabelen. In het algemeen kan dezelfde doelvariabele voorkomen in meerdere lokale scores. Ook bestaan er geavanceerdere scorefuncties dan (A1) en (A1*) waarin twee of meer doelvariabelen tegelijk voorkomen.

Globale deelscore per blok bij eenheid i :

$$S_{i,G}(\alpha, D_v) = \left\{ \frac{\sum_{j \in D_v} \left(w_{j,G} * \frac{s_{ij,G}}{M_{j,G}} \right)^\alpha}{\sum_{j \in D_v} w_{j,G}^\alpha} \right\}^{1/\alpha}, \quad (v = 1, \dots, V). \quad (A2)$$

(Voor de notatie $M_{j,G}$, $w_{j,G}$, α en D_v , zie de uitleg onder 'Instelbare parameters' verderop.) In het speciale geval $\alpha = \infty$ gaat deze score over in:

$$S_{i,G}(\infty, D_v) = \frac{\max_{j \in D_v} \left(w_{j,G} * \frac{s_{ij,G}}{M_{j,G}} \right)}{\max_{j \in D_v} w_{j,G}}, \quad (v = 1, \dots, V). \quad (A2^*)$$

Een globale deelscore van de vorm (A2) kan in de praktijk ook gebruikt worden om de lokale scores te vertalen naar een score per doelvariabele, die bijvoorbeeld in een dashboard kan worden gebruikt om de waarden van de doelvariabelen te kleuren afhankelijk van het verwachte risico op een invloedrijke fout. De score per doelvariabele is dan gelijk aan $S_{i,G}(\alpha, D_j)$, waarbij het blok D_j alle lokale scores $s_{ij,G}$ bevat waarin de betreffende variabele voorkomt als doelvariabele (y_j). In het speciale geval dat een variabele voorkomt in precies één lokale score is $S_{i,G}(\alpha, D_j)$ gelijk aan $s_{ij,G}/M_{j,G}$ voor de betreffende lokale score, ongeacht de keuze van α .

Globale score bij eenheid i :

$$S_{i,G}(\alpha) = \left\{ \frac{\sum_{v=1}^V [W_G(\alpha, D_v) * S_{i,G}(\alpha, D_v)]^\alpha}{\sum_{v=1}^V [W_G(\alpha, D_v)]^\alpha} \right\}^{1/\alpha}$$

$$= \left\{ \frac{\sum_{j=1}^J \left(w_{j,G} * \frac{S_{ij,G}}{M_{j,G}} \right)^\alpha}{\sum_{j=1}^J w_{j,G}^\alpha} \right\}^{1/\alpha}, \quad (A3)$$

waarbij¹ $W_G(\alpha, D_v) = \left(\sum_{j \in D_v} w_{j,G}^\alpha \right)^{1/\alpha}$ voor $v = 1, \dots, V$.

In het speciale geval $\alpha = \infty$ gaat deze score over in:

$$S_{i,G}(\infty) = \frac{\max_{v=1, \dots, V} [W_G(\infty, D_v) * S_{i,G}(\infty, D_v)]}{\max_{v=1, \dots, V} [W_G(\infty, D_v)]}$$

$$= \frac{\max_{j=1, \dots, J} \left(w_{j,G} * \frac{S_{ij,G}}{M_{j,G}} \right)}{\max_{j=1, \dots, J} w_{j,G}}, \quad (A3^*)$$

waarbij $W_G(\infty, D_v) = \max_{j \in D_v} w_{j,G}$.

Instelbare parameters:

$M_{j,G}$: (optioneel) ‘maximaal acceptabele’ relatieve invloed per eenheid van een mogelijke fout in de doelvariabele(n) van lokale score $s_{ij,G}$ op het aggregaat in de noemer. Deze normering is mede bepalend voor de invloed van de lokale score op een globale (deel-)score [zie (A2) en (A3)] én voor de kleuring van de lokale score in het dashboard (zie hieronder bij grenswaarden). Default-keuze: $M_{j,G} = 1$ voor alle j (d.w.z. maak hierbij geen onderscheid tussen verschillende lokale scores).

Voorwaarde: $M_{j,G} > 0$ voor alle j .

$w_{j,G}$: (optioneel) gewicht voor het belang van lokale score $s_{ij,G}$ in een globale score (gegeven de normering t.o.v. $M_{j,G}$). Default-keuze: $w_{j,G} = 1$. Voorwaarden: $w_{j,G} \geq 0$ voor alle j en er is minimaal één j met $w_{j,G} > 0$.

α : bepaalt de vorm van de globale scorefunctie. Voorwaarde: $\alpha \geq 1$. Meestal wordt $\alpha = 1$ (gewogen gemiddelde), $\alpha = 2$ (gewogen Euclidische norm) of $\alpha = \infty$ (gewogen maximum) gekozen.

$D_1, \dots, D_V$: (optioneel) indeling van de lokale scores $j = 1, \dots, J$ in niet-overlappende blokken. Default-keuze: $V = 1$ (d.w.z. alleen één globale score).

$\tau_G^{(0)}, \tau_G^{(1)}, \tau_G^{(2)}$: grenswaarden (met $0 \leq \tau_G^{(0)} \leq \tau_G^{(1)} \leq \tau_G^{(2)}$) die bepalen welke kleur een lokale of globale (deel)score krijgt op het dashboard. Voor een lokale score $s_{ij,G}$ geldt:

- als $s_{ij,G}/M_{j,G} \leq \tau_G^{(1)}$ dan is de kleur groen;
- als $\tau_G^{(1)} < s_{ij,G}/M_{j,G} < \tau_G^{(2)}$ dan is de kleur oranje;
- als $s_{ij,G}/M_{j,G} \geq \tau_G^{(2)}$ dan is de kleur rood.

De kleuring van globale (deel)scores werkt op dezelfde manier:

- als $S_{i,G}(\alpha, D_v) \leq \tau_G^{(1)}$ [resp. $S_{i,G}(\alpha) \leq \tau_G^{(1)}$] dan is de kleur groen;
- als $\tau_G^{(1)} < S_{i,G}(\alpha, D_v) < \tau_G^{(2)}$ [resp. $\tau_G^{(1)} < S_{i,G}(\alpha) < \tau_G^{(2)}$] dan is de kleur oranje;
- als $S_{i,G}(\alpha, D_v) \geq \tau_G^{(2)}$ [resp. $S_{i,G}(\alpha) \geq \tau_G^{(2)}$] dan is de kleur rood.

¹ Deze keuze zorgt ervoor dat de twee manieren in (A3) om de globale score te berekenen – vanuit de globale deelscores of direct vanuit de lokale scores – equivalent zijn:
$$\left\{ \frac{\sum_{v=1}^V [W_G(\alpha, D_v) * S_{i,G}(\alpha, D_v)]^\alpha}{\sum_{v=1}^V [W_G(\alpha, D_v)]^\alpha} \right\}^{1/\alpha} = \left\{ \frac{\sum_{v=1}^V \sum_{j \in D_v} \left(w_{j,G} * \frac{S_{ij,G}}{M_{j,G}} \right)^\alpha}{\sum_{v=1}^V \sum_{j \in D_v} w_{j,G}^\alpha} \right\}^{1/\alpha} = \left\{ \frac{\sum_{j=1}^J \left(w_{j,G} * \frac{S_{ij,G}}{M_{j,G}} \right)^\alpha}{\sum_{j=1}^J w_{j,G}^\alpha} \right\}^{1/\alpha},$$
 aangezien elke lokale score meetelt in precies één van de blokken $D_1, \dots, D_V$.

De laagste grenswaarde $\tau_G^{(0)}$ speelt geen rol in de kleuring, maar wordt gebruikt in de samenvattende scores die hieronder worden gedefinieerd in hoofdstuk B.

s_{max} : (optioneel) maximum waarde voor een lokale score. Elke score $s_{ij,G}$ uit formule (A1) of (A1*) die een waarde groter dan s_{max} oplevert wordt afgekapt en in verdere berekeningen vervangen door s_{max} . Op deze manier kan worden voorkomen dat lokale scores extreme waarden aannemen (bijvoorbeeld omdat er gedeeld wordt door een aggregaat-waarde in de buurt van 0) en daardoor de bijbehorende globale (deel)scores domineren. Default-keuze: $s_{max} = \infty$ (d.w.z. niet afkappen). Voorwaarde: $s_{max} > 0$.

B. Samenvattende scorefuncties per aggregaat, voor gebruik binnen één statistiek

Samenvattende score per aggregaat voor een afzonderlijke lokale score:

$$S_{j,G} = \sum_{i \in G} \frac{S_{ij,G}}{M_{j,G}} * I \left\{ \frac{S_{ij,G}}{M_{j,G}} \geq \tau_G^{(0)} \right\}, \quad (j = 1, \dots, J), \quad (B1)$$

waarbij $I\{\cdot\} = 1$ als het argument waar is en anders $I\{\cdot\} = 0$ en $S_{ij,G}$ afkomstig is uit (A1) of (A1*).

De grenswaarde $\tau_G^{(0)}$ is in hoofdstuk A gedefinieerd en zo gekozen dat in (B1) alle oranje en rode lokale scores sowieso meetellen in de som en daarnaast mogelijk (als $\tau_G^{(0)} < \tau_G^{(1)}$) ook een deel van de groene lokale scores.

Samenvattende score per aggregaat voor een blok:

$$S_G(\alpha, D_v) = \sum_{i \in G} S_{i,G}(\alpha, D_v) * I \{ S_{i,G}(\alpha, D_v) \geq \tau_G^{(0)} \}, \quad (v = 1, \dots, V). \quad (B2)$$

waarbij $S_{i,G}(\alpha, D_v)$ afkomstig is uit (A2).

Samenvattende globale score per aggregaat:

$$S_G(\alpha) = \sum_{i \in G} S_{i,G}(\alpha) * I \{ S_{i,G}(\alpha) \geq \tau_G^{(0)} \}. \quad (B3)$$

waarbij $S_{i,G}(\alpha)$ afkomstig is uit (A3).

Om de kleuring van deze samenvattende scores per aggregaat op het dashboard te bepalen stellen we de volgende aanpak voor:

- Als minimaal één van de onderliggende scores die bijdraagt aan de som in (B1), (B2) of (B3) rood wordt gekleurd volgens de bijbehorende grenswaarde $\tau_G^{(2)}$, dan wordt de samenvattende score rood gekleurd.
- Zo niet, maar als wel minimaal één van de onderliggende scores die bijdraagt aan de som in (B1), (B2) of (B3) oranje wordt gekleurd volgens de bijbehorende grenswaarde $\tau_G^{(1)}$, dan wordt de samenvattende score oranje gekleurd.
- Zo niet, dan wordt de samenvattende score groen gekleurd.

Merk op: indien $\tau_G^{(0)} = \tau_G^{(1)}$ wordt gekozen kleurt de samenvattende score groen dan en slechts dan als deze gelijk is aan 0. Als $\tau_G^{(0)} < \tau_G^{(1)}$ kunnen er ook samenvattende scores ongelijk aan 0 voorkomen die groen worden gekleurd.

C. Scorefuncties voor confrontatie tussen statistieken, per eenheid t.o.v. een vast aggregatieniveau

Lokale scores bij eenheid i :

Bij elke conceptuele doelvariabele y_l ($l = 1, \dots, L$) waarvoor een confrontatie tussen statistieken plaatsvindt (confrontatievariabele) noteren we de waargenomen waarden voor eenheid i in de verschillende bronnen/statistieken als $y_{il}^{(1)}, \dots, y_{il}^{(P)}$. De lokale scorefunctie voor de confrontatie van statistiek p met statistiek q voor deze variabele heeft de volgende vorm [vergelijk (A1)]:

$$s_{il,G}^{(p,q)} = \frac{|v_i^{(p)} * (y_{il}^{(p)} - y_{il}^{(q)})|}{|Y_{l,G}|}, \quad (l = 1, \dots, L; p = 1, \dots, P; q = 1, \dots, P; q \neq p). \quad (C1)$$

Hierbij is $v_i^{(p)}$ het ophooggewicht van eenheid i voor statistiek p . Het aggregaat $Y_{l,G}$ is een robuuste schatting gebaseerd op de verschillende aggregaten voor de betreffende doelvariabele uit de afzonderlijke statistieken, $Y_{l,G}^{(p)} = \sum_{i \in G} v_i^{(p)} * y_{il}^{(p)}$ voor $p = 1, \dots, P$. Als een van de statistieken al eerder is gaafgemaakt, zou deze als robuuste schatting kunnen worden gebruikt. In andere situaties zou bijvoorbeeld de mediaan van de aggregaten $Y_{l,G}^{(1)}, \dots, Y_{l,G}^{(P)}$ kunnen worden gekozen, of een aangepaste schatting waarin mogelijke uitbijters een verlaagd gewicht hebben gekregen. Voor confrontaties tussen statistieken zal vaak de regkol als aggregatieniveau worden gekozen.

Bij een doelvariabele die (voor een bepaalde eenheid) waargenomen is bij P statistieken worden $P(P-1)$ lokale scores van de vorm (C1) berekend. Merk op: deze lokale scores zijn niet symmetrisch in p en q , dus in het algemeen zal gelden dat $s_{il,G}^{(p,q)} \neq s_{il,G}^{(q,p)}$. Verder kan P afhankelijk zijn van l en i , maar dit wordt hier niet expliciet aangegeven om de notatie niet onnodig ingewikkeld te maken.

In sommige toepassingen is het wenselijk om statistieken over een doelvariabele ook te confronteren met een of meer andere bronnen waar geen statistiek bij hoort (en dus ook geen ophooggewicht). Een dergelijke bron kan in (C1) wel gebruikt worden als index q maar niet als index p . Het aantal paarsgewijze lokale scores is in deze situatie gelijk aan $P_0(P-1)$, waarbij $P_0 \leq P$ het aantal bronnen is waar wel een statistiek bij hoort. In het vervolg nemen we gemakshalve aan dat $y_{il}^{(1)}, \dots, y_{il}^{(P_0)}$ afkomstig zijn uit bronnen met een statistiek en $y_{il}^{(P_0+1)}, \dots, y_{il}^{(P)}$ uit bronnen zonder statistiek.

Globale deelscore per confrontatievariabele bij eenheid i :

Op basis van de lokale scores uit (C1) kan een globale deelscore bepaald worden per eenheid i en confrontatievariabele y_l :

$$S_{il,G}(\alpha) = \left\{ \sum_{p=1}^{P_0} \sum_{\substack{q=1 \\ q \neq p}}^P \left(\frac{s_{il,G}^{(p,q)}}{M_{l,G}^{(p,q)}} \right)^\alpha \right\}^{1/\alpha}. \quad (C2)$$

(Voor de notatie $M_{l,G}^{(p,q)}$ en α , zie de uitleg onder 'Instelbare parameters' verderop.) In het speciale geval $\alpha = \infty$ gaat deze score over in:

$$S_{il,G}(\infty) = \max_{\substack{p=1, \dots, P_0, \\ q=1, \dots, P, q \neq p}} \left(\frac{s_{il,G}^{(p,q)}}{M_{l,G}^{(p,q)}} \right). \quad (C2^*)$$

Globale score bij eenheid i :

Indien gewenst kan ook een globale score per eenheid worden gedefinieerd over alle confrontatievariabelen heen, analoog aan (A3) of (A3*):

$$S_{i,G}(\alpha) = \left\{ \frac{\sum_{l=1}^L [w_{l,G} * S_{il,G}(\alpha)]^\alpha}{\sum_{l=1}^L w_{l,G}^\alpha} \right\}^{1/\alpha}. \quad (C3)$$

(Voor de notatie $w_{l,G}$, zie de uitleg onder ‘Instelbare parameters’ verderop.) In het speciale geval $\alpha = \infty$ gaat deze score over in:

$$S_{i,G}(\infty) = \frac{\max_{l=1,\dots,L} [w_{l,G} * S_{il,G}(\infty)]}{\max_{l=1,\dots,L} w_{l,G}}. \quad (C3^*)$$

Globale deelscore per paar statistieken bij eenheid i :

Ten slotte kan per paar statistieken een globale deelscore worden bepaald over alle variabelen heen. De vorm van deze globale deelscore voor alle confrontaties van statistiek p met bron q is:

$$S_{i,G}^{(p,q)}(\alpha) = \left\{ \frac{\sum_{l=1}^L \delta_l^{(p,q)} * \left[w_{l,G} * \frac{S_{il,G}^{(p,q)}}{M_{l,G}^{(p,q)}} \right]^\alpha}{\sum_{l=1}^L \delta_l^{(p,q)} * w_{l,G}^\alpha} \right\}^{1/\alpha}, \quad (C4)$$

waarbij $\delta_l^{(p,q)} = 1$ als confrontatievariabele y_l beschikbaar is uit statistiek p en bron q , en anders $\delta_l^{(p,q)} = 0$. In het speciale geval $\alpha = \infty$ gaat deze score over in (merk op dat $[\delta_l^{(p,q)}]^\alpha = \delta_l^{(p,q)}$):

$$S_{i,G}^{(p,q)}(\infty) = \frac{\max_{l=1,\dots,L} \left[\delta_l^{(p,q)} * w_{l,G} * \frac{S_{il,G}^{(p,q)}}{M_{l,G}^{(p,q)}} \right]}{\max_{l=1,\dots,L} [\delta_l^{(p,q)} * w_{l,G}]}. \quad (C4^*)$$

Instelbare parameters:

$M_{l,G}^{(p,q)}$: (optioneel) factor per paar statistieken/bronnen die aangeeft wat de ‘maximaal acceptabele’ relatieve afwijking is in de (voor statistiek p opgehoogde) waarden van een eenheid ten opzichte van het aggregaat $Y_{l,G}$. Bij de keuze van $M_{l,G}^{(p,q)}$ is relevant of verwacht wordt dat $y_{il}^{(q)}$ op microniveau een goede referentiewaarde is voor $y_{il}^{(p)}$. De default-keuze is $M_{l,G}^{(p,q)} = 1$ voor alle paren en alle doelvariabelen (d.w.z. maak hierbij geen onderscheid tussen verschillende lokale scores). Voorwaarde: $M_{l,G}^{(p,q)} > 0$ voor alle l, p en q . Om het aantal te specificeren parameters beperkt te houden zou men in de praktijk aparte factoren $M_{l,G}^{(p)}$ ($p = 1, \dots, P$) per bron kunnen kiezen en vervolgens definiëren: $M_{l,G}^{(p,q)} = \max \{M_{l,G}^{(p)}, M_{l,G}^{(q)}\}$.

$w_{l,G}$: (optioneel) gewicht voor het relatieve belang van confrontatievariabele y_l in een globale score over alle confrontatievariabelen heen. Default-keuze: $w_{l,G} = 1$. Voorwaarden: $w_{l,G} \geq 0$ voor alle l en er is minimaal één l met $w_{l,G} > 0$.

α : bepaalt de vorm van de globale scorefunctie. Voorwaarde: $\alpha \geq 1$. Meestal wordt $\alpha = 1$ (gewogen gemiddelde), $\alpha = 2$ (gewogen Euclidische norm) of $\alpha = \infty$ (gewogen maximum) gekozen.

$\tau_G^{(0)}, \tau_G^{(1)}, \tau_G^{(2)}$: grenswaarden (met $0 \leq \tau_G^{(0)} \leq \tau_G^{(1)} \leq \tau_G^{(2)}$) die bepalen welke kleur een lokale of globale (deel)score krijgt op het dashboard, op dezelfde manier als in hoofdstuk A. Dat wil zeggen, er geldt:

- als $S(.) \leq \tau_G^{(1)}$ dan is de kleur groen;
- als $\tau_G^{(1)} < S(.) < \tau_G^{(2)}$ dan is de kleur oranje;
- als $S(.) \geq \tau_G^{(2)}$ dan is de kleur rood.

Hierbij mag voor $S(.)$ een van de volgende scores worden ingevuld: de *genormeerde* lokale score $s_{il,G}^{(p,q)} / M_{l,G}^{(p,q)}$ gebaseerd op (C1), de globale deelscore $S_{il,G}(\alpha)$ uit (C2), de globale score $S_{i,G}(\alpha)$ uit (C3) of de globale deelscore $S_{i,G}^{(p,q)}(\alpha)$ uit (C4).

De laagste grenswaarde $\tau_G^{(0)}$ speelt wederom geen rol in de kleuring, maar wordt gebruikt in de samenvattende scores die hieronder worden gedefinieerd in hoofdstuk D.

s_{max} : (optioneel) maximum waarde voor een lokale score. Elke score $s_{i,G}^{(p,q)}$ uit formule (C1) die een waarde groter dan s_{max} oplevert wordt afgekapt en in verdere berekeningen vervangen door s_{max} , net als in hoofdstuk A. Default-keuze: $s_{max} = \infty$ (d.w.z. niet afkappen). Voorwaarde: $s_{max} > 0$.

D. Samenvattende scorefuncties per aggregaat, voor confrontatie tussen statistieken

Samenvattende score per aggregaat voor een afzonderlijke lokale score voor confrontaties:

$$S_{l,G}^{(p,q)} = \sum_{i \in G} \frac{S_{il,G}^{(p,q)}}{M_{l,G}^{(p,q)}} * I \left\{ \frac{S_{il,G}^{(p,q)}}{M_{l,G}^{(p,q)}} \geq \tau_G^{(0)} \right\}, \quad (l = 1, \dots, L; p = 1, \dots, P_0; q = 1, \dots, P; q \neq p), \quad (D1)$$

waarbij $I\{\cdot\} = 1$ als het argument waar is en anders $I\{\cdot\} = 0$ en $S_{il,G}^{(p,q)}$ afkomstig is uit (C1). De grenswaarde $\tau_G^{(0)}$ is in hoofdstuk C gedefinieerd en zo gekozen dat in (D1) alle oranje en rode lokale scores sowieso meetellen in de som en daarnaast mogelijk (als $\tau_G^{(0)} < \tau_G^{(1)}$) ook een deel van de groene lokale scores.

Samenvattende score per aggregaat voor een globale deelscore over één confrontatievariabele:

$$S_{l,G}(\alpha) = \sum_{i \in G} S_{il,G}(\alpha) * I \{S_{il,G}(\alpha) \geq \tau_G^{(0)}\}, \quad (l = 1, \dots, L), \quad (D2)$$

waarbij $S_{il,G}(\alpha)$ afkomstig is uit (C2).

Samenvattende score per aggregaat voor een globale score over alle confrontaties heen:

$$S_G(\alpha) = \sum_{i \in G} S_{il,G}(\alpha) * I \{S_{il,G}(\alpha) \geq \tau_G^{(0)}\}, \quad (D3)$$

waarbij $S_{il,G}(\alpha)$ afkomstig is uit (C3).

Samenvattende score per aggregaat voor een globale deelscore over één paar statistieken:

$$S_G^{(p,q)}(\alpha) = \sum_{i \in G} S_{il,G}^{(p,q)}(\alpha) * I \{S_{il,G}^{(p,q)}(\alpha) \geq \tau_G^{(0)}\}, \quad (D4)$$

waarbij $S_{il,G}^{(p,q)}(\alpha)$ afkomstig is uit (C4).

Om de kleuring van deze samenvattende scores per aggregaat op het dashboard te bepalen stellen we dezelfde aanpak voor als beschreven in hoofdstuk B.