

Nieuwe verwerking e-commercevariabelen (september 2025)

Hier volgt een voorstel voor een nieuwe verwerking van de e-commercevariabelen. Hierin neem ik aan dat:

- De DRT als hoofdbron voor de omzet gebruikt wordt,
- Er alleen enquêtegegevens voor de omzet beschikbaar zijn voor bedrijven in bedrijfstakken die niet tot de DRT-populatie behoren,
- De omzet uit verkoop via website/apps en EDI beschikbaar zijn in aantallen (2025)

Aankoppelen externe bronnen

Voor zover mogelijk gebruiken we de omzet uit de DRT (**omzet_drt**) en de BTW (**omzet_netto_btw**) voor de totale omzet. Als eerst koppelen we **omzet_drt** en **omzet_netto_btw** aan de enquêtedata. Na het koppelen met de DRT hebben we al voor zo'n 87% (n=17 000) van de bedrijven in de respons omzetgegevens uit de DRT. Er ontbreekt dus nog 13% (n=2000). Hiervan:

- Zit een deel wel in de DRT-populatie, maar ontbreekt er data (n=100) → Wijst erop dat het bedrijf niet actief was in het afgelopen jaar, dus vul een 0 in voor de omzet.
- Zit een deel niet in de DRT-populatie, maar is wel BTW-data beschikbaar (n=750) → Gebruik BTW-data
- Zit een deel niet in de DRT-populatie, maar is enquêtedata beschikbaar (n=750) → Gebruik enquêtedata
- Zit een deel niet in de DRT-populatie, en is geen enquêtedata beschikbaar (n=350) → Imputeer

Gaafmaken van de omzet uit de enquête

De omzet uit de enquête moeten we nog wel gaafmaken, bijv.:

- **Omzet van 0 eruit halen**
- **Duizendfouten corrigeren**

Om de duizendfouten te corrigeren, moeten we **omzet_enq** vergelijken met een andere bron. Echter, omdat we voor de bedrijven waar we **omzet_enq** willen gebruiken, géén DRT en BTW-omzet hebben, gaan we eerst een tijdelijke schatting (**omzet_totaal_duizendfout**) maken voor hun verwachte omzet op basis van de *gemiddelde omzet* van bedrijven uit de hele populatie met dezelfde **gk_sbs** en dezelfde **sbi_digit2**.

We stellen dat er een duizendfout plaatsvindt wanneer:

- De geschatte omzet tussen 50% en 150% ligt van **omzet_enq** * 1000
- De geschatte omzet tussen 50% en 150% ligt van **omzet_enq** / 1000

Indien nodig, corrigeren we de duizendfouten die we op deze manier geïdentificeerd hebben.

Maken perc_verkoop_web en perc_verkoop_edl

Voordat de totale omzet helemaal gaafgemaakt is, beginnen we ook al met het maken van **omzet_web** en **perc_verkoop_web**. Een belangrijke aanname is dat hoewel **omzet_enq** en **omzet_web** in euro's niet hoeven te kloppen (want bijv. **omzet_enq** komt niet overeen met **omzet_drt** of er zijn duizendfouten in beide waarden), bedrijven **de ratio tussen de twee** wel goed opgeven.

We maken hier dus al **perc_verkoop_web** aan, met de verhouding **omzet_web/omzet_enq**. Een uitzondering is:

- Als bedrijven verkopen via EDI én website en $(\text{omzet_edi} + \text{omzet_web}) > \text{omzet_totaal}$
- Als bedrijven hebben ingevuld dat $\text{omzet_enq} == 0$ (geloven we niet)
- Als bedrijven hebben ingevuld dat $\text{omzet_web} == 0$ (geloven we niet)

In deze gevallen laten we **perc_verkoop_web** leeg. Hetzelfde geldt voor **perc_verkoop_edi**.

De resterende ontbrekende waarden voor **perc_verkoop_web** en **perc_verkoop_edi** worden later geïmputeerd in **ImputeGaps**.

- **Uitbijters verwijderen**

Tot slot verwijderen we de uitbijters uit **omzet_enq** door te kijken naar de *mediaan omzet* van bedrijven met dezelfde **gk_sbs** en dezelfde **sbi_digit2**. (Als dat niet mogelijk is, dan kijken we naar de *mediaan omzet* van bedrijven met dezelfde **gk_sbs** en dezelfde **sbi_digit1**). Als een waarde 100 keer zo groot, of 100 keer zo klein, is als de mediaan, dan beschouwen we het als een uitbijter, en verwijderen we de waarde.

Imputeren nog ontbrekende omzetten

Tot slot gaan we alle omzetwaarden die nu nog ontbreken, imputeren. Dit doen we niet alleen voor de bedrijven in de enquête, maar voor de hele populatie. Dit is nodig omdat we de totalen later nodig hebben voor de weging. We maken een schatting voor de omzet van een bedrijf op basis van de *gemiddelde omzet* van bedrijven met dezelfde **gk_sbs** en dezelfde **sbi_digit2**. De gaafgemaakte omzet slaan we op in de variabele **omzet_totaal**.

Gaafmaken negatieve omzet

In de DRT komen negatieve omzetten soms voor. Eurostat accepteert echter geen negatieve omzetten. Zet alle negatieve omzetten in **omzet_totaal** daarom op 0. Zet in deze gevallen ook **omzet_web**, **perc_verkoop_web** en **omzet_edi**, **perc_verkoop_edi** op 0, waar van toepassing.

Als de variabelen **perc_verkoop_web** en **perc_verkoop_edi** nog ontbrekende waarden hebben, gaan we die imputeren (**ImputeGaps**). Na het imputeren kan het nog wel zo zijn dat $\text{perc_verkoop_web} + \text{perc_verkoop_edi} > 100$. Als dat zo is, nemen we **perc_verkoop_web** als vast, en zetten de **perc_verkoop_edi** zo dat de optelsom van beide 100 is. Hierna zijn **perc_verkoop_web** en **perc_verkoop_edi** helemaal klaar.

Aanmaken perc_verkoop_ecom

Nu kunnen we ook **perc_verkoop_ecom** berekenen door simpelweg de som te nemen van **perc_verkoop_web** + **perc_verkoop_edi**.

Berekenen omzet_web en omzet_edi

Tot slot gaan we **omzet_web** en **omzet_edi** (her)berekenen. Aangezien we **perc_verkoop_web** en **perc_verkoop_edi** helemaal klaar zijn, kunnen we **omzet_web** berekenen met: $\text{omzet_web} = (\text{perc_verkoop_web} / 100) * \text{omzet_totaal}$. Hetzelfde geldt voor **omzet_edi**.

Overige variabelen berekenen

Tot slot kunnen we alle andere afgeleide variabelen (zoals **verk_alle_klasseLT_eeen** etc.) + de Eurostat-variabelen berekenen.