Methodologie

De vorming van categorieën van gemeenten berust in de eerste plaats op de samenstelling van een sociaaleconomische database op het gemeentelijk echelon (2.1.) en de vorming van indicatoren (2.2.) op basis van de brutogegevens. Daarna volgt een doorgedreven statistische verwerking in twee fasen: een analyse in belangrijke bestanddelen afgeleid uit de beginvariabelen (2.3.) met daarna een opklimmende hiërarchische classificatie op basis van de factorscores waaruit de categorieën van gemeenten (clusters) (2.4.) kunnen worden afgeleid.

1. De keuze van de beginvariabelen

Bij een statistische classificatiemethode rijst vooral het probleem van de selectie en de beschikbaarheid van de gewenste gegevens.
Om de diversiteit van de gemeenten statistisch te benaderen, moet men zich baseren op variabelen die niet alleen representatief zijn voor de morfologische verschillen (fysiek waarneembare) maar ook voor de sociaaleconomische verscheidenheid van de gemeenten (bevolkingsstructuur, economische activiteit enz.).

De variabelen (waarvan de lijst zich in bijlage 1 bevindt) werden geselecteerd om voor elke gemeente een beeld van de vijf volgende dimensies te verkrijgen:

- bestemming van de bodem en van de gebouwen, kenmerken en uitrustingsgraad van de woningen;
- niveau van de inkomsten (van de gezinnen en uit de grondbelasting);
- economische activiteit en structuur van de beroepsbevolking;
- structuur en evolutie van de bevolking;
- voorzieningen van algemeen nut, externaliteiten en aantrekkingskracht.

Bovendien werden voor elk van deze dimensies zowel **statische** (toestand op een bepaald moment) als **dynamische indicatoren** (evolutiepercentage) geselecteerd. Daardoor vermijden we een classificatie die uitsluitend een momentopname weergeeft. Aan de vijf voornoemde dimensies wordt dus de dimensie tijd toegevoegd.

De analyse gebruikt variabelen die hoofdzakelijk afkomstig zijn van het Nationaal Instituut voor de Statistiek (NIS). Meer precies gaat het om de resultaten van de laatste enquête (2001). Een dergelijke telling blijft een opmerkelijke en onvervangbare bron voor statistische informatie over de bevolking, haar activiteiten en haar leefomgeving.

Naast de periodieke publicaties van het NIS en van ECODATA over de bevolkingscijfers en het inkomen, werd onze gegevensbasis ook aangevuld met andere statistische bronnen die meer specifieke informatie verschaffen over onder meer de werkloosheid (RVA), het aantal begunstigden van het leefloon (FOD Volksgezondheid), de uitsplitsing van de werkgelegenheid per activiteitssector (RSZ, RIZIV), alsook over de waarde en de samenstelling van het kadastraal inkomen (FOD Financiën).

Ook werden sommige indicatoren opgevraagd bij de gewest- of gemeenschapsadministraties zoals bijvoorbeeld de cijfers over het onderwijs en de toeristische activiteit. Ten slotte stelden universitaire onderzoekscentra[25] bepaalde zeer specifieke gegevens ter beschikking.

2. De samenstelling van indicatoren

In plaats van op basis van de absolute waarden te werken, bepaalden wij systematisch een indicator waarmee het brutobegingegeven kon worden gerelativeerd. De gegevens die we in de analyse invoerden werden dus allemaal in relatieve termen uitgedrukt (index, percentage, inwoners per km^2 ...). Via deze optie kan tijdens de statistische verwerking een

[25] Uitrustingsgraad (K.U.Leuven-ISEG) en toegevoegde waarde per gemeente (ULB-IGEAT).

systematische afwijking, veroorzaakt door een grootte-effect, worden vermeden. De grootste entiteiten vertonen immers systematisch de hoogste absolute waarden, ongeacht de beginvariabele.

Om ondanks alles de omvang van de gemeente in aanmerking te nemen, hebben wij een transformatie doorgevoerd van de variabele bevolking. Dankzij deze variabele kan rekening worden gehouden met het feit dat eenzelfde verschil in de bevolkingsvariabele (bijvoorbeeld een verschil van 5 000 inwoners) niet dezelfde betekenis heeft voor de kleinste gemeenten (bijvoorbeeld voor een gemeente van 5 000 inwoners en een andere van 10 000 inwoners) als voor grote steden (bijvoorbeeld een entiteit van 70 000 inwoners en van 75 000 inwoners).

3. De factoranalyse: een voorafgaande syntheseoefening

Factoranalyse is een techniek om gegevens te herleiden. De bedoeling is een nieuwe set van variabelen te vinden die kleiner is in aantal dan de aanvankelijke set en weergeeft wat gemeenschappelijk is onder de beginvariabelen.

Schematisch bestaat de factoranalyse erin om, op basis van de systematische analyse van de relaties tussen de **beginvariabelen** (de correlaties), de informatie over een beperkt aantal relevante **nieuwe variabelen** (**factoren** genoemd) te verzamelen. Deze nieuwe variabelen, lineaire combinaties van de beginvariabelen, bevatten zoveel mogelijk informatie met een minimum aan redundantie.

De analyse genereert in principe evenveel factoren als er beginvariabelen zijn. In tegenstelling tot deze laatste hebben de nieuwe variabelen echter het voordeel onderling niet gecorreleerd te zijn en stuk voor stuk een maximum aan niet redundante informatie te verschaffen die wordt gesorteerd volgens een hiërarchie. De eerste component verklaart namelijk het best de veranderlijkheid van de begingegevens, de tweede verklaart het best de veranderlijkheid van de informatie die niet verklaard wordt (het residu) door de eerste enz.
Op die manier is het dus mogelijk om het aantal beginvariabelen aanzienlijk te verminderen door enkel de eerste vastgestelde factoren in aanmerking te nemen en het verlies aan informatie die in de begingegevens vervat zit, te minimaliseren.

Wanneer de factoren vastgesteld zijn, moet men vervolgens de betekenis ervan vastleggen en ze een benaming geven. De interpretatie van een factor berust voornamelijk op zijn **saturaties** (correlatiecoëfficiënt factor-variabele), die de link vormen tussen de beginvariabelen, waarvan de betekenis bekend is, en de nieuwe creaties, nl. de te interpreteren factoren. De betekenis wordt vastgelegd op basis van de beginvariabelen die de hoogste saturaties vertonen (positieve of negatieve).

In het voorbeeld in *tabel 2* bevinden zich beginvariabelen die zowel onderling een verband hebben als met de factor. De interpretatie die men zeer duidelijk kan maken, is dat de factor representatief is voor de levensstandaard van de bevolking.

Het **factorcijfer** (of de factorscore) is de waarde van een observatie (gemeente) voor een bepaalde factor. Aangezien de factoren gestandaardiseerd zijn, is het gemiddelde van de cijfers nul en de variantie gelijk aan 1.
Hoe sterker een factorscore voor een gemeente afwijkt van 0 (positief of negatief), hoe sterker de gemeente een uitgesproken karakter voor deze factor vertoont. Ten opzichte van het voorbeeld van de factor vermeld in *tabel 2* zal een gemeente met een meer dan gemiddelde levensstandaard een hoge positieve factorscore vertonen, terwijl een gemeente met een minder

Tabel 2: Voorbeeld van het verband tussen de beginvariabelen en een factor

Negatieve saturaties (< -0,75)	Positieve saturaties (> 0,75)
Aantal begunstigden van het leefloon	Gemiddeld inkomen per inwoner
% belastingaangiften < 7 500 EUR	% woningen met "groot comfort"
	% gezinnen met drie voertuigen
	% universitaire diploma's
	% gezinnen met pc en internet

begunstigde bevolking een negatieve score zal hebben. Ten slotte zal een gemeente met een levensstandaard die dicht bij het regionale gemiddelde ligt, een score vertonen die in de buurt van nul uitkomt.

4. De clustermethode: de samenstelling van klassen

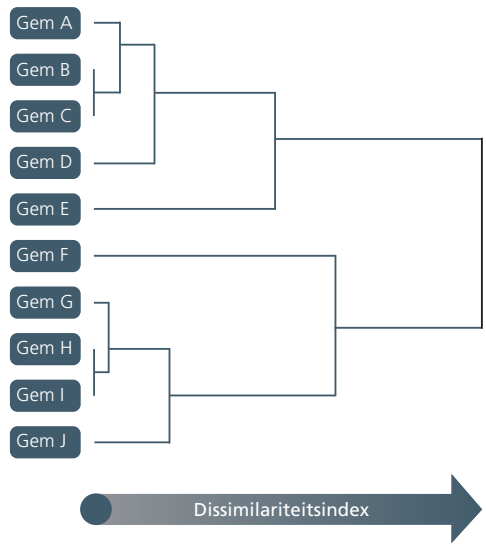
De gebruikte classificatiemethode is een hiërarchische procedure die gebruik maakt van aggregatie.
Het algoritme zoekt via een iteratief proces in de ruimte van n dimensies (volgens het aantal in aanmerking genomen factoren) de twee gemeenten die de kleinste "afstand" vertonen, d.w.z. die de dichtste combinatie van factorscores hebben (en die dus de meest gelijkende sociaaleconomische context bezitten). Deze twee gemeenten worden samengevoegd en vormen tijdens de volgende fase een nieuwe observatie. De classificatieprocedure groepeert beetje bij beetje de observaties in steeds minder klassen tot wanneer één enkele groep wordt verkregen, gevormd door het geheel van de waarnemingen.
Men spreekt van hiërarchische classificatie omdat elke klasse van een partitie ingesloten zit in een klasse van de volgende partitie[26].

De gebruikte methode groepeert dus in de eerste plaats de gemeenten met de kleinste verschillen (d.w.z. dicht bij het gemiddelde) en brengt uiteindelijk de gemeenten met een atypisch profiel samen.

Grafiek 3 illustreert het aggregatieproces in de vorm van een hiërarchische boomdiagram.

Op basis van dit boomdiagram, vormt elke gemeente in het beginstadium een klasse (de inertie tussen de klassen is maximaal en stemt overeen met de totale inertie[27]).
Aan de top van de boom, d.w.z. het ultieme stadium van het iteratief proces, beschikt men over een partitie van alle gemeenten gegroepeerd in eenzelfde klasse (de inertie binnen de klasse is in dat geval gelijk aan de totale inertie).
Op de tussentijdse niveaus bestaat deze totale inertie uit een som van de inertie binnen de klassen en de inertie tussen de klassen. Met de hergroeperingen neemt de inertie binnen de klasse toe en daalt de inertie tussen de klassen.
Clustering is dus het zoeken naar een compromis tussen het streven naar de vaststelling van een beperkt aantal categorieën enerzijds en de samenstelling van zo homogeen mogelijke groepen anderzijds.

Grafiek 3: Boomdiagram van de verschillende aggregaties

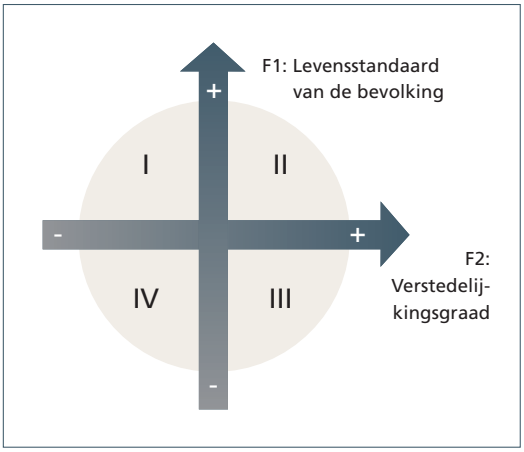


Clusters analyseren op basis van factorscores en niet op grond van de basisgegevens biedt de volgende voordelen:
- de redundantie van de begininformatie is geëlimineerd;
- het aantal variabelen wordt gevoelig verminderd (van 150 variabelen naar een tiental), wat de interpretatie van de clusters grotendeels vergemakkelijkt;
- de factoren worden gestandaardiseerd (gemiddelde = 0 en variantie = 1), zodat de analyse niet kan worden beïnvloed door meeteenheden.

Om evenwel rekening te houden met het zeer veranderlijk verklarend vermogen van de verschillende factoren (cf. punt 3.), hebben wij de gebruikte factoren gewogen (op basis van de inertiegraad).

In de illustratie van deze benaderingswijze ziet men intuïtief dat een classificatie die slechts op twee criteria gebaseerd is (bijvoorbeeld de levensstandaard en de verstedelijkingsgraad), in de volgende categorieën kan resulteren:
- "landelijke" gemeenten met hoge levensstandaard (kwadrant I);
- "verstedelijkte" gemeenten met hoge levensstandaard (kwadrant II);

[26] De Wasseige Y., Laffut M., Ruyters Ch., Schleiper P. en Vanden Dooren L., "Bassins d'emploi et régions fonctionnelles", Discussion papers, Service des Etudes et de la Statistique de la Région Wallonne, mei 2002.
[27] De totale inertie van een puntenwolk meet de spreiding van deze punten rond het zwaartepunt van de wolk. Volgens het principe van Huygens kan de totale inertie (I t) worden uitgesplitst in inertie binnen klassen (I w) en inertie tussen klassen (I b) volgens de formule: I t = I w + I b.



Grafiek 4: Statistische verwerking van de gegevens – algemeen overzicht

a. Beginmatrix

	Beginvariabelen												
	V1	V2	V3	V4	V5	V6	…	…	V25	…	V50	…	**Vn**
Gemeenten													
G1													
G2													
G3													
G4													
G5													
…													
G25													
…													
Gn													

b. Factoranalyse

	Factoren					
	F1	F2	F3	…	…	**Fn**
Gemeenten						
G1						
G2						
G3						
G4						
G5						
…						
G25						
…						
Gn						

c. Clusteranalyse

	Factoren					
	F1	F2	F3	…	…	**Fn**
Clusters						
CL1						
CL2						
…						
CLn						

- "verstedelijkte" gemeenten met lage levensstandaard (kwadrant III);
- "landelijke" gemeenten met lage levensstandaard (kwadrant IV);

De analyse in clusters veralgemeent de voorgaande aanpak en gebeurt op basis van een algoritme dat de overeenkomst tussen gemeenten zoekt in de waarden van de diverse gebruikte factoren.

Kortom, de typologie van de gemeenten wordt verkregen na een dubbele statistische verwerking van de begininformatie. Deze laatste heeft de vorm van een grote matrix met twee dimensies (aantal gemeenten, aantal beginvariabelen) (cf. *grafiek 4*: Statistische verwerking van de gegevens – a. Beginmatrix).

Dankzij de **factoranalyse** kan het aantal beginvariabelen (Vn) beperkt worden door de redundante informatie te elimineren en de in aanmerking genomen informatie toe te spitsen op een beperkt aantal nieuwe variabelen (Fn), factoren genoemd. Door deze eerste statistische verwerking kan de horizontale dimensie van de beginmatrix dus gereduceerd worden. Die telt nu immers minder dan tien "nieuwe" variabelen (Fn) in plaats van de zowat honderd beginvariabelen (Vn) (cf. *grafiek 4*: Statistische verwerking van de gegevens – b. Factoranalyse).

De **clusteranalyse** bestaat er vervolgens in om in een multidimensionele ruimte de "dicht bij elkaar liggende" observaties te zoeken. De dichtst bij elkaar liggende gemeenten, d.w.z. die voor de diverse factoren vergelijkbare waarden hebben (en die dus eenzelfde sociaaleconomische omgeving als kenmerk hebben) worden ondergebracht in klassen of "clusters".
Deze tweede statistische verwerking draagt er dus toe bij dat de verticale dimensie van onze beginmatrix wordt gereduceerd. Het aantal gemeenten (Gn) wordt immers ondergebracht in een beperkt aantal clusters (CLn) (cf. *grafiek 4*: Statistische verwerking van de gegevens – c. Clusteranalyse).