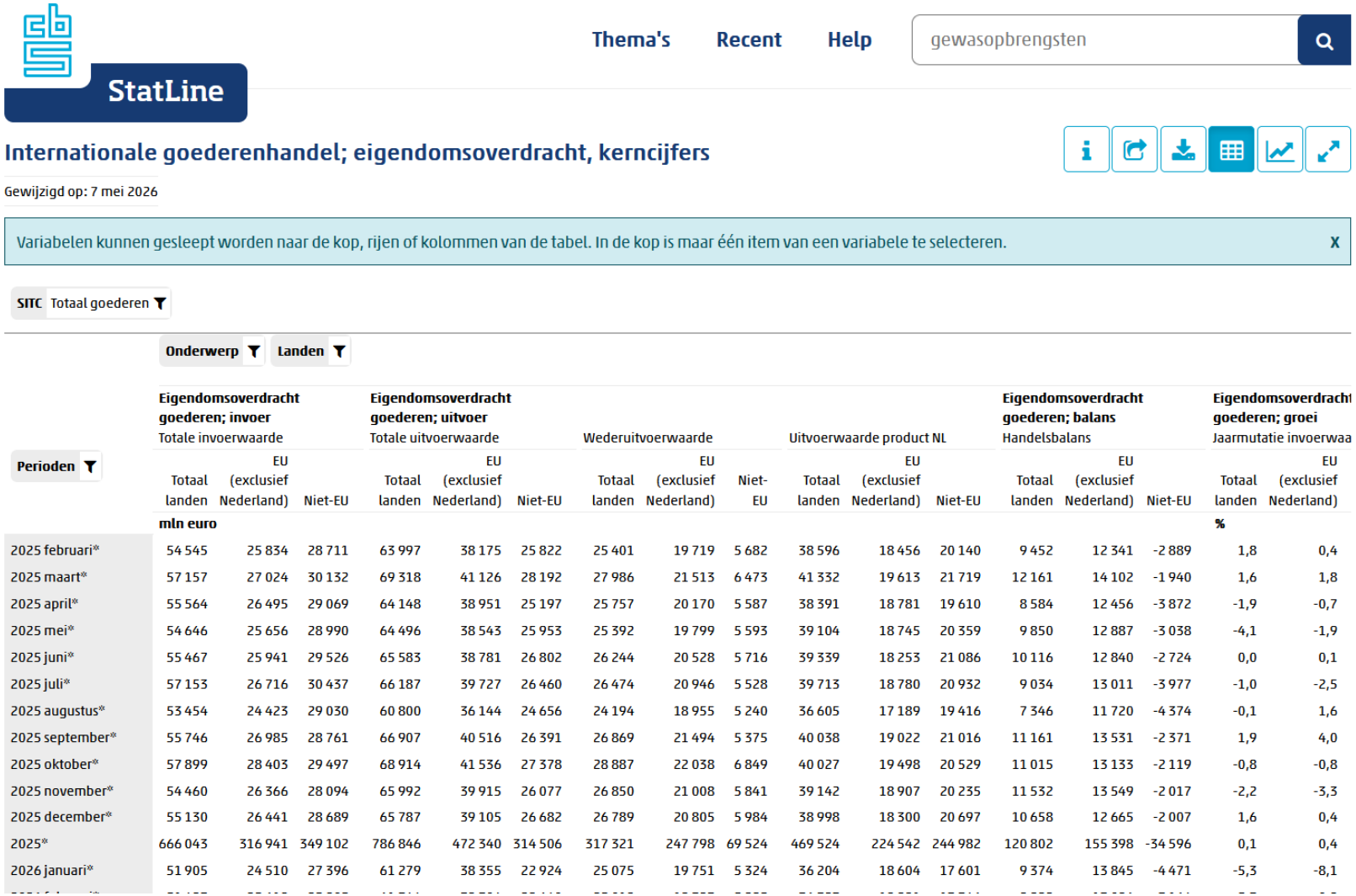
From Pixels to Statistics

EO- and in-situ data integration

Workshop on integration of spatial data sources

Our ‘bread and butter’ as statisticians



StatLine

Thema's Recent Help

gewasopbrengsten

Internationale goederenhandel; eigendomsoverdracht, kerncijfers

Gewijzigd op: 7 mei 2026

Variabelen kunnen gesleept worden naar de kop, rijen of kolommen van de tabel. In de kop is maar één item van een variabele te selecteren. X

SITC Totaal goederen

Onderwerp Landen

Perioden	Eigendomsoverdracht goederen; invoer Totale invoerwaarde			Eigendomsoverdracht goederen; uitvoer Totale uitvoerwaarde			Wederuitvoerwaarde			Uitvoerwaarde product NL			Eigendomsoverdracht goederen; balans Handelsbalans			Eigendomsoverdracht goederen; groei Jaarmutatie invoerwaa	
	Totaal landen	EU (exclusief Nederland)	Niet-EU	Totaal landen	EU (exclusief Nederland)	Niet-EU	Totaal landen	EU (exclusief Nederland)	Niet-EU	Totaal landen	EU (exclusief Nederland)	Niet-EU	Totaal landen	EU (exclusief Nederland)	Niet-EU	Totaal landen	EU (exclusief Nederland)
	mln euro															%	
2025 februari*	54 545	25 834	28 711	63 997	38 175	25 822	25 401	19 719	5 682	38 596	18 456	20 140	9 452	12 341	-2 889	1,8	0,4
2025 maart*	57 157	27 024	30 132	69 318	41 126	28 192	27 986	21 513	6 473	41 332	19 613	21 719	12 161	14 102	-1 940	1,6	1,8
2025 april*	55 564	26 495	29 069	64 148	38 951	25 197	25 757	20 170	5 587	38 391	18 781	19 610	8 584	12 456	-3 872	-1,9	-0,7
2025 mei*	54 646	25 656	28 990	64 496	38 543	25 953	25 392	19 799	5 593	39 104	18 745	20 359	9 850	12 887	-3 038	-4,1	-1,9
2025 juni*	55 467	25 941	29 526	65 583	38 781	26 802	26 244	20 528	5 716	39 339	18 253	21 086	10 116	12 840	-2 724	0,0	0,1
2025 juli*	57 153	26 716	30 437	66 187	39 727	26 460	26 474	20 946	5 528	39 713	18 780	20 932	9 034	13 011	-3 977	-1,0	-2,5
2025 augustus*	53 454	24 423	29 030	60 800	36 144	24 656	24 194	18 955	5 240	36 605	17 189	19 416	7 346	11 720	-4 374	-0,1	1,6
2025 september*	55 746	26 985	28 761	66 907	40 516	26 391	26 869	21 494	5 375	40 038	19 022	21 016	11 161	13 531	-2 371	1,9	4,0
2025 oktober*	57 899	28 403	29 497	68 914	41 536	27 378	28 887	22 038	6 849	40 027	19 498	20 529	11 015	13 133	-2 119	-0,8	-0,8
2025 november*	54 460	26 366	28 094	65 992	39 915	26 077	26 850	21 008	5 841	39 142	18 907	20 235	11 532	13 549	-2 017	-2,2	-3,3
2025 december*	55 130	26 441	28 689	65 787	39 105	26 682	26 789	20 805	5 984	38 998	18 300	20 697	10 658	12 665	-2 007	1,6	0,4
2025*	666 043	316 941	349 102	786 846	472 340	314 506	317 321	247 798	69 524	469 524	224 542	244 982	120 802	155 398	-34 596	0,1	0,4
2026 januari*	51 905	24 510	27 396	61 279	38 355	22 924	25 075	19 751	5 324	36 204	18 604	17 601	9 374	13 845	-4 471	-5,3	-8,1

As statisticians, what do we do with geospatial data?

- Estimate values of:
  - amount or acerage or percentage of…
  - variables like biomass, crop yield, crop types, urban area, …
- Per:
  - country, Nuts 2 or 3, city, parcel,
  - Year, month, quarter, day(?)
- So that we may:
  - create a map, or use a geocoded data
  - that classifies areas, e.g. parcels, grid cells, adresses
  - determine extent/ shape
  - determine class, amount
- By:
  - manual coding
  - Inferences/ heuristics
  - statistical modelling,
  - physical modelling
  - machine learning
  - computer vision models

Examples of spatial datasources

- EO data (specifically satellite data)
- Digital (topographical or other functional) maps
- Administrative systems
- Measuring stations
- Fieldwork-based data sources
- Survey data
- (Regional) statistics
- Models …
- 3d Models
- Graphs (e.g. roads, waterways, public transport)

But… The data rarely matches the statistics we want

Data source	Transformation	Desired outcome
Earth observation pixels	?	Crop yield per NUTS 2
Measuring stations hourly values	?	Air quality per city district
Containerscounts in major harbor	?	Trade volume per year per country
Species observations for PQ	?	Biodiversity indicator per habitat
Regional energy consumption	?	Energy demand per industry

As statisticians, we are often forced to estimate quantities on supports where they were never directly observed.

Data sources:

May have

- different supports
  - (spatial extents like district borders)
- different concepts
  - (not specifically designed for target statistic)
- different timing
- different uncertainty structures
- different quality demands

Every integration step is a transformation

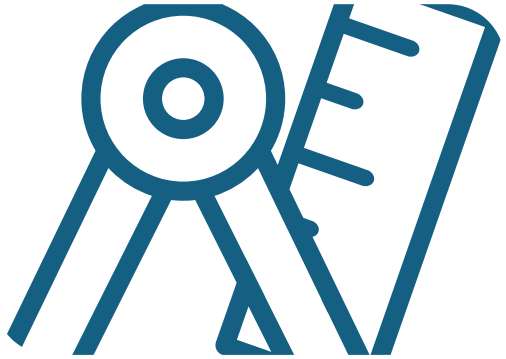
From/ To	Point	Grid/ raster	Polygon	Network
Point			Aggregation	Route calculation
Grid	Sampling	Resampling		
Polygon			Areal weighting	
Network		graph propagation		

Each transformation implies assumption about (the level of homogeneity) of space, time and the phenomenon itself

Typical assumptions

Assumption	Example
Homogeneity	Yield similar within field
Representativity	Stations represent surroundings
Proxy validity	NDVI relates to biomass
Temporal stability	Measurements compare over time

Assumptions

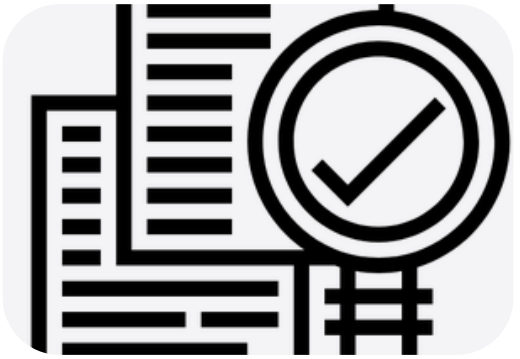


Geometry

Nearby things are similar

Timing

Measurements represent the same period



Proxy validity

Observed variable represents target phenomenon

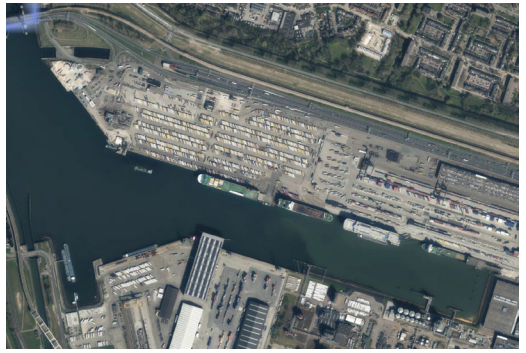
Scale

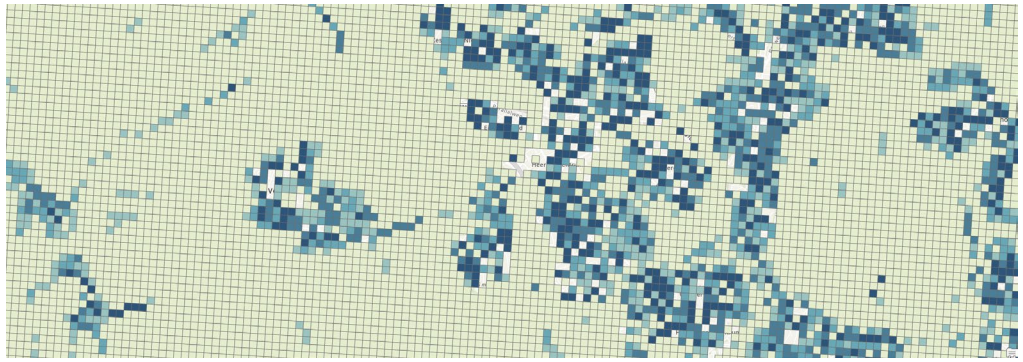
Patterns remain valid after aggregation/disaggregation

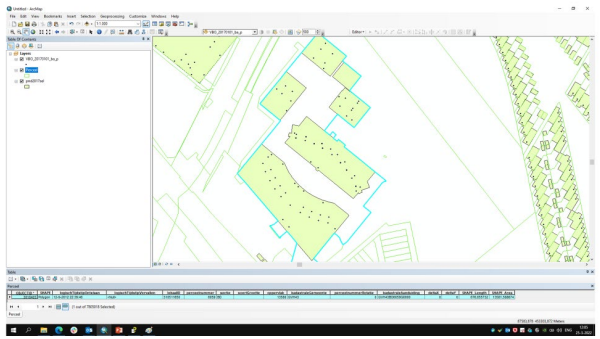
Existing methods for transformations

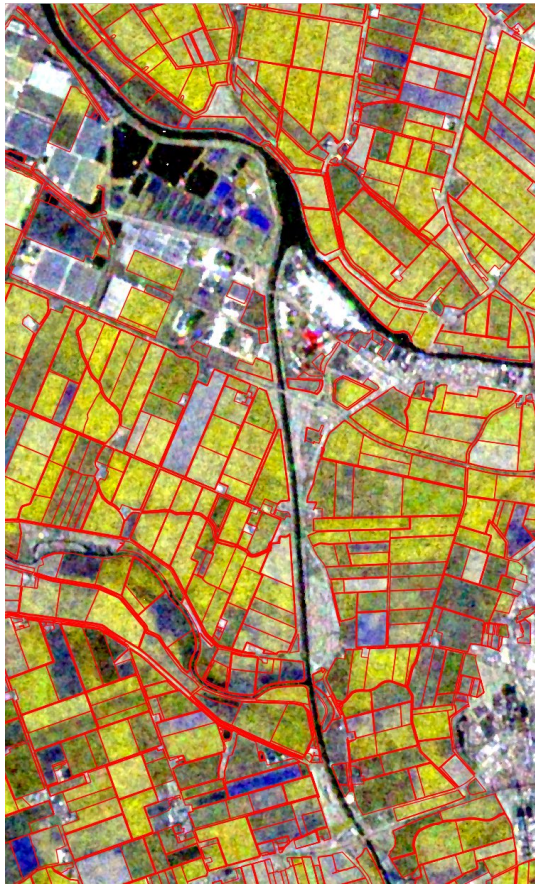
Challenge	Typical methods
Aggregation	Areal weighting, dasyemtric methods
Spatial interpolation	Kriging, spatial smoothing
Combining proxies	Regression, ML, DL
Preserving totals	Benchmarking, constrained estimation
Sparse observations	Small area estimators
…	

The same integration problem appears in many domains









From data integration to statistical integration:

- clear target variable
- known spatial and temporal support
- explicit assumptions
- validation strategy
- uncertainty awareness

What we ask from you today:

1. Discuss the spatial characteristics of the data sources and target statistic
2. Identify the transformations
3. Identify assumptions and discuss what works and fails (identify and diagnose risks)
4. Think about validation
5. Think about what is still missing

Per case template

- **Step 1: Describe the data**
- Spatial support (point / raster / polygon / network)
- Temporal support (instant / daily / yearly) (not to be discussed in depth)
- Variable type (measurement / proxy / modelled)
- **Step 2: Identify transformations**
  Between:
- In-situ with EO
- Combination with statistical output
- **Step 3 : Diagnose the risks**
- Misalignment (spatial / temporal)
- Concept mismatch
- Scale effects
- Heterogeneity
- **Step 4: What helps?**
- Auxiliary data?
- Models?
- Constraints (mass preservation, benchmarking)
- **Step 5: Practical judgement/ validation**
- What validation is possible?
- When would you trust this?
- When not?

CASE 1 — Crop type statistics (LPIS and Sentinel2 → NUTS2)

Policy/statistical question:

What is the distribution of crop types at NUTS2 level?

Data sources:

Administrative

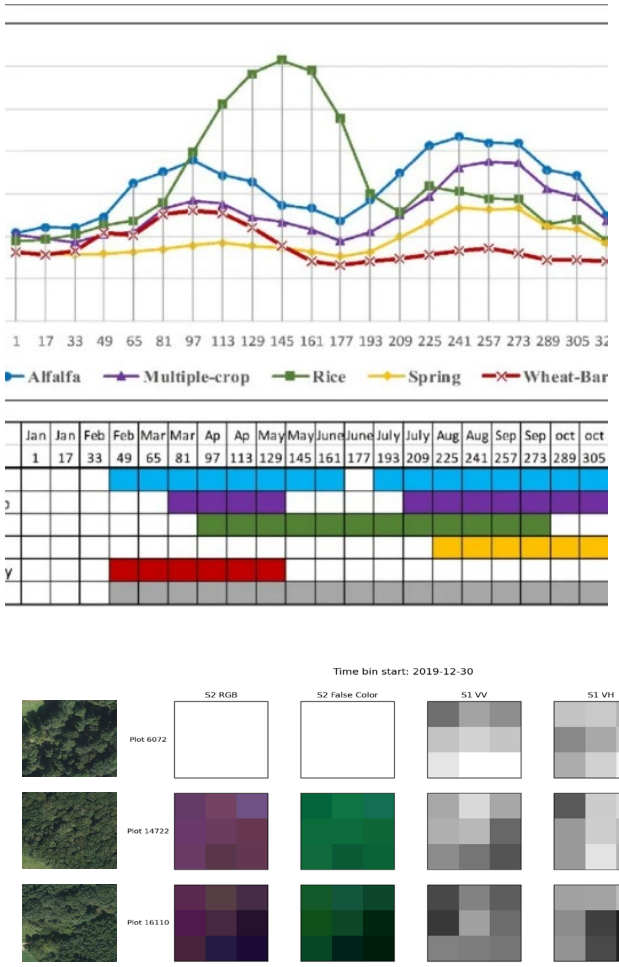
- LPIS parcels
- Geometry: polygons (fields)
- Variable: declared crop type (categorical)
- Timing: annual declaration

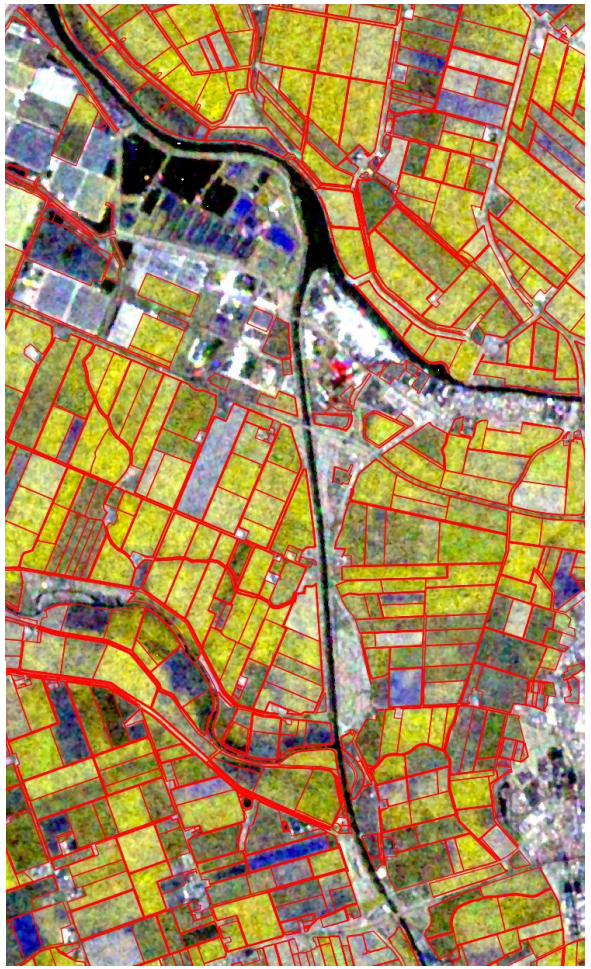
Earth Observation

- Sentinel2 imagery
- Geometry: raster (10–20m pixels)
- Variable: spectral reflectance / derived indices
- Timing: multi-temporal (growing season)

Target statistic

- Crop shares per NUTS2 region
- Geometry: polygons
- Timing: annual





Step 1: Data description (example)

Input/output data	Spatial dimensions	Comments/issues
LPIS admininstrative data	Polygon	May contain several subparcels
…		

Step 2: Transformations (example)

Inputs for transformation	Type of transformation	Output	Comments/issues/ decisions
LPIS – EO raster	Selection of raster cells based on parcel polygon	Parcel boundaries	Border grid cells partly overlap
…			

Step 3: Risks/ problems (example)

Type of transformation	Type of transformation	Comments	Risks/ issues
LPIS/ EO	From raster to parcel	Parcel border and grid cells only partly overlap	Assigning values to cross sections may be biased
…			

Step 4: Diagnose risks/ problems (example)

Comments	Risks/ issues	Diagnosis	What could help?
Parcel border and grid cells only partly overlap	Assigning values to cross sections may be biased	If only inner cells are included: how much area is missing? Are cross section cells homogenuous?	If homogenuous: standard area weighting If not: auxiliary info available? -> dasymetric mapping
…			

Step 5: Validate (example)

Comments	Risks/ issues	What could help?	Validate/ corroborate
Parcel border and grid cells only partly overlap	Assigning values to cross sections may be biased	If homogenuous: standard area weighting If not: auxiliary info available? -> dasymetric mapping	Sampled field studies, variance analysis, cross-validate with farmer declarations, …
…			