Spatial representativeness of NO_2 monitoring stations with respect to Sentinel-5P satellite based estimates

Patryk Grzybowski, Institute of Geophysics, Faculty of Physics, University of Warsaw
Krzysztof Markowicz, Institute of Geophysics, Faculty of Physics, University of Warsaw
Jan Musiał, CloudFerro S.A.

- Introduction
- Represenativeness – what is it?
- Methods: correlation, semivariograms and CSF
- Workflow – how it works?
- Results – how many stations are required… and if?
- Summary

Objectives

- Establish what area of Poland is covered by spatial representative NO_2 surface concentrations measurement network
- Find how many people live over the areas which are not covered by spatial representativeness measurements
- Propose where new stations should be located in to cover Poland by a fully representative NO_2 measurement network

Represenativeness

But… how to define it?

We need to dicuss it…
United States Environmental Protection Agency (1978)

Representativeness is the extent to which a set of measurements taken in a space-time domain reflects the actual conditions *in the same or different space-time domain taken on a scale appropriate for a specific application* (Nappo et al., 1982)

Measurement representativeness - an extent to which a measurement reflects the actual conditions being measured;

Point-to-point representativeness – the possibility of applying measurements taken at one location for use at another location

Point-to-volume representativeness - that establishes a relationship between measurement taken at one location over horizontal surfaces

Represenativeness… in Europe

Stations should be divided according to three criteria:

- type of station (traffic, industrial, background)
- type of zone (urban, suburban, traffic)
- characterization of zone (residential, commercial, industrial, agricultural, natural and combinations of these)

(EEA, Larssen et al., 1999)

The following range of values as a radius of area was proposed:

- Traffic stations - not applicable
- Industrial stations - 10–100 m
- Background stations:
  - Urban background stations 100 m-1 km
  - Near-city background stations 1–5 km;
  - Regional stations 25–150 km

  (EEA, Larssen et al., 1999)

These are leads
Conditions over specific locations should be evaluated

Represenativeness… in Europe

- Representativeness
  - a ±7.5 μg/m^3 threshold for background rural areas
  - a ±2.5 μg/m^3 threshold for background urban areas
  - concentration differs up to ±20 % of those recorded on the station

  (Report for EC, Spangl et al., 2007)

Maximum spatially representative (SR) distance from a measuring site:

- 100 km (2007) - Report for EC, Spangl et al., 2007
- 200 km (2008) – EC, 2008
- 100 km (2011) – EC, 2011

Then...

Representativeness is the extent to which a set of measurements taken in a space-time domain reflects the actual conditions *in the same or different space-time domain taken on a scale appropriate for a specific application* (Nappo et al., 1982)

Point-to-volume representativeness - that establishes a relationship between measurement taken at one location over horizontal surfaces (Nappo et al., 1982)

Stations should be divided according to three criteria:

- type of station (traffic, industrial, background),
- type of zone (urban, suburban, traffic) (EEA, Larssen et al., 1999)

Maximum spatially representative (SR) distance from a measuring site – 100km (EC, 2011)

Methods

- Methods for assessment of spatial representativeness area
  - Global Moran I – general verification
  - Method 1 - Correlation of NO_2 surface mass concentration between stations with respect to distance
  - Method 2 – Semivariograms
  - Method 3 - Similarity threshold

Correlation of NO_2 surface mass concentration between stations with respect to distance

Calculate distance-based correlations: For each station, correlate NO_2 values with surrounding points in distance intervals (3.5–100 km) to see how correlation decreases with distance

Average results across stations/images: Combine correlations from all stations to get a representative distance–correlation curve for urban and rural areas.

Apply Fisher r-to-z transformation: Convert correlations to z-scores to statistically compare correlations at distance vs at zero distance.

Test significance ($z \geq 1.96$): Identify distances where correlation becomes significantly lower than the baseline, using the 95% confidence threshold.

Define SR area: The last distance with still-significant correlation is taken as the spatial relationship (SR) range.

Semivariograms

Compute semi-variances between station and surrounding points: Semi-variance measures how NO_2 differences grow with distance, allowing detection of spatial structure around each station.

Build semivariograms (semi-variance vs distance): Semivariograms show how spatial dependence changes with distance and reveal irregular SR shapes that simple correlation-distance methods cannot detect.

Identify nugget, sill, and range: Nugget shows small-scale variability or measurement noise; sill indicates total variance; range defines the distance beyond which NO_2 values are no longer spatially related (the SR boundary).

Fit multiple semivariogram models: Test exponential, Gaussian, circular, wave, spherical, and linear models, selecting the best-fitting one using R^2.

Extract SR range separately for urban and rural areas: Using the best-fit model, determine the range parameter for each environment to quantify the spatial relationship area.

Similarity threshold

Compare concentration time series using CSF: For each station, NO_2 time series at the station are compared with time series at every grid point within 100 km to assess similarity in temporal behaviour.

Compute percentage differences at each time step: For every t, calculate the relative difference between concentrations at the station and each surrounding grid point.

Select a threshold via sensitivity analysis: Test multiple percentage thresholds and choose the one that best balances representativeness—based on population share deviations—separately for urban and rural areas.

Calculate the frequency function fsite(x,y): For each grid point, compute how often the percentage difference stays below the chosen threshold; this frequency quantifies temporal similarity.

Define SR area based on similarity frequency: Points with **fsite(x,y) > 0.9** are considered part of the station's spatial representativeness (SR) area.

Determining urban and rural areas

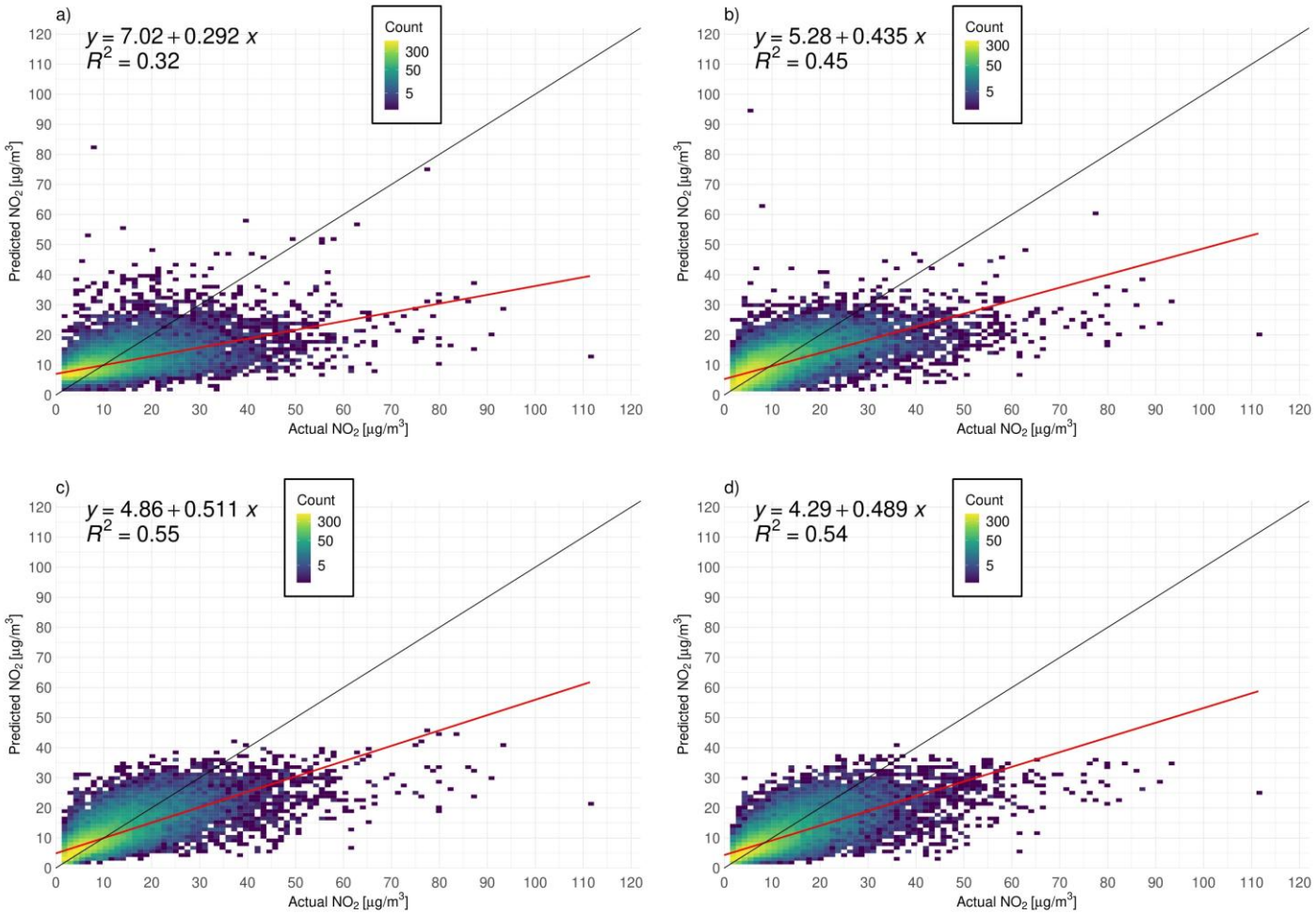
- Urban areas
- Rural areas (non-urban areas)
- Based on Corine Land Cover

- Excluding ‚Transport' and ‚Industrial' areas

Surface NO_2 mass concentration – conversion model overview

Time period analysed: July 2018 – June 2021.
Data availability: Years 2019 and 2020.
Data processing: Values were linearly interpolated to match the S-5P acquisition times.
Spatial resampling: Data were rescaled to a 3.5 km × 5.5 km spatial resolution using interpolation.

Surface NO2 mass concentration estimations							
		MEAN	**MIN**	**MAX**	**SD**	**1st Q**	**3rd Q**
Testing dataset (n=27241)		10.07	0.01	111.41	8.89	4.40	13.12
Training dataset (n=22678)		10.44	0.01	100.59	8.61	4.44	12.93
Hourly measurements		**R2**	**MSE**	**RMSE**	**Bias [ug/m3]**	**MAE [ug/m3]**	**MAPE [%]**
	LM - S5P	0.32	53.22	7.30	0.11	4.87	48.38
	MLM	0.45	43.11	6.57	0.42	4.24	42.08
	RF	**0.55**	**34.83**	**5.90**	**-0.04**	**3.74**	**37.15**
	SVM	0.54	37.67	6.09	1.04	3.72	36.94



Grzybowski, P. T., Markowicz, K. M., & Musiał, J. P. (2023). Estimations of the ground-level NO_2 concentrations based on the sentinel-5P NO2 tropospheric column number density product. *Remote Sensing*, *15*(2), 378.

Surface NO_2 mass concentration – conversion model overview

Linear regression with one undependable variable (LM)	Multiple linear regression with several undependable variables (MLM)	Random forest with several undependable variables (RF)	Radial kernel support vector machine with several undependable variables (SVM)
NO_2 TVCD	NO_2 TVCD	NO_2 TVCD	NO_2 TVCD
	Air temperature (T)	Air temperature (T)	Air temperaturę (T)
	Pressure (P)	Pressure (P)	Pressure (P)
	Solar radiation (RADNET)	Solar radiation (RADNET)	Solar radiation (RADNET)
	Wind speed (WS)	Wind speed (WS)	Wind speed (WS)
	Planetary boundary layer height (PBLH)	Planetary boundary layer height (PBLH)	Planetary boundary layer height (PBLH)
	Nightlights (NIGHTLIGHT)	Nightlights (NIGHTLIGHT)	Nightlights (NIGHTLIGHT)
	Population (POP)	Population (POP)	Population (POP)
	Road density (ROADS)	Road density (ROADS)	Road density (ROADS)
	Elevation (ELEVATION);	Elevation (ELEVATION);	Elevation (ELEVATION);
		Circulation type (CT)	Circulation type (CT)

Study area

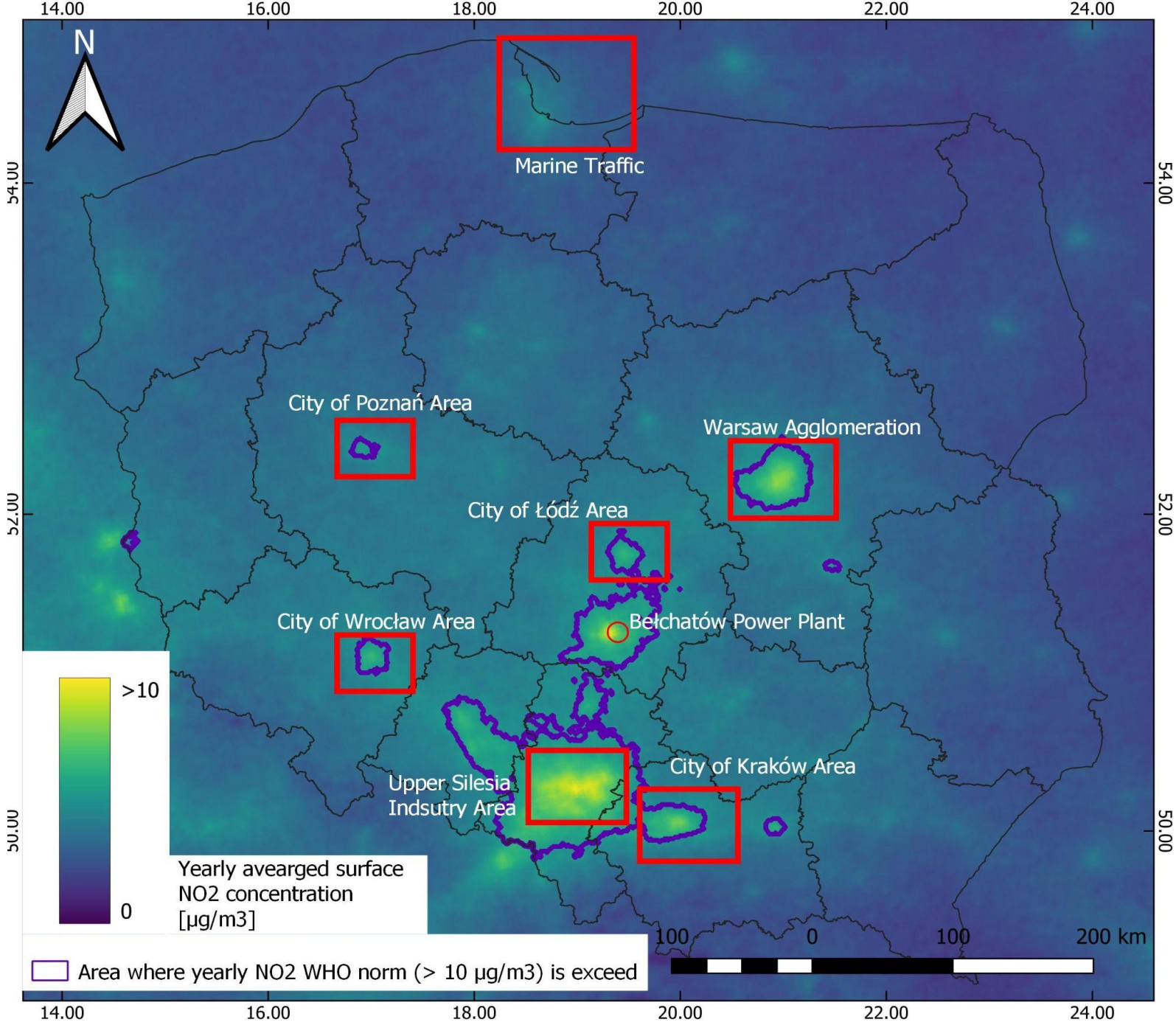
Study area was limited to homogeneous areas, such as urban and rural areas, which were defined with use of Corine Land Cover data

"Transport" and "Industrial" areas have been excluded from the study

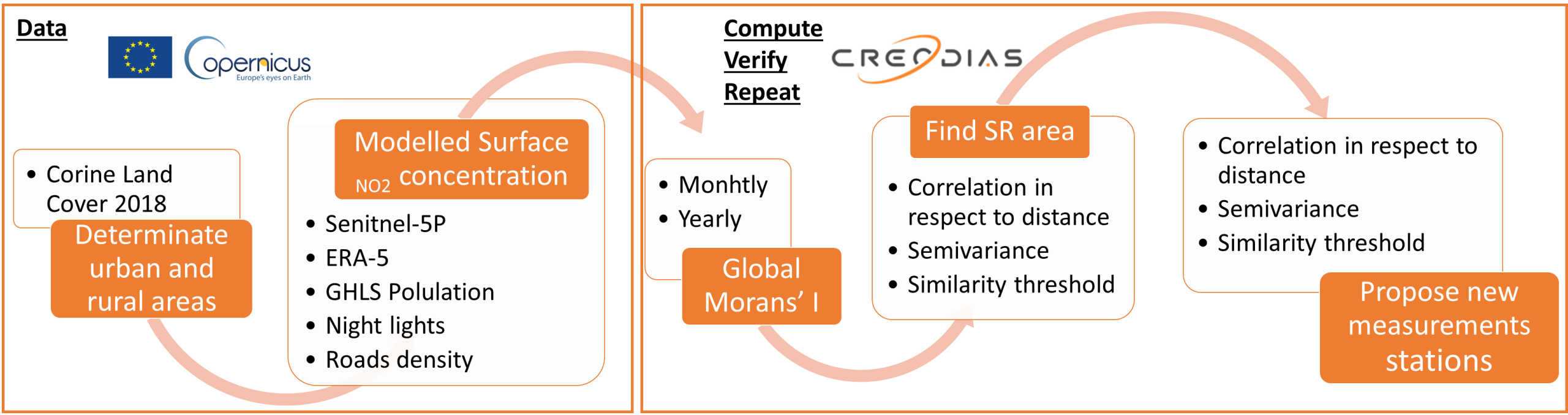
1000 points which included c.a. 170 real existing stations and c.a. 870 virtual stations.

Virtual rural stations were randomly placed at c.a. 400 sites, while virtual urban stations were randomly distributed across c.a. 470 sites.

Areas where the annual surface NO_2 concentration equaled or exceeded the WHO annual recommendation of 10 $\mu g/m_3$


N
Marine Traffic
City of Poznań Area
Warsaw Agglomeration
City of Łódź Area
City of Wrocław Area
Bełchatów Power Plant
Upper Silesia
Indsutry Area
City of Kraków Area
>10
0
Yearly avearged surface
NO2 concentration
[µg/m3]
Area where yearly NO2 WHO norm (> 10 µg/m3) is exceed
100
0
100
200 km
14.00
16.00
18.00
20.00
22.00
24.00
54.00
52.00
50.00

Workflow


Data
opernicus
Europe's eyes on Earth
• Corine Land Cover 2018
Determinate urban and rural areas
Modelled Surface NO2 concentration
• Senitnel-5P
• ERA-5
• GHLS Polulation
• Night lights
• Roads density
Compute
Verify
Repeat
CREODIAS
• Monhtly
• Yearly
Global Morans' I
Find SR area
• Correlation in respect to distance
• Semivariance
• Similarity threshold
• Correlation in respect to distance
• Semivariance
• Similarity threshold
Propose new measurements stations

Results

Global Moran I

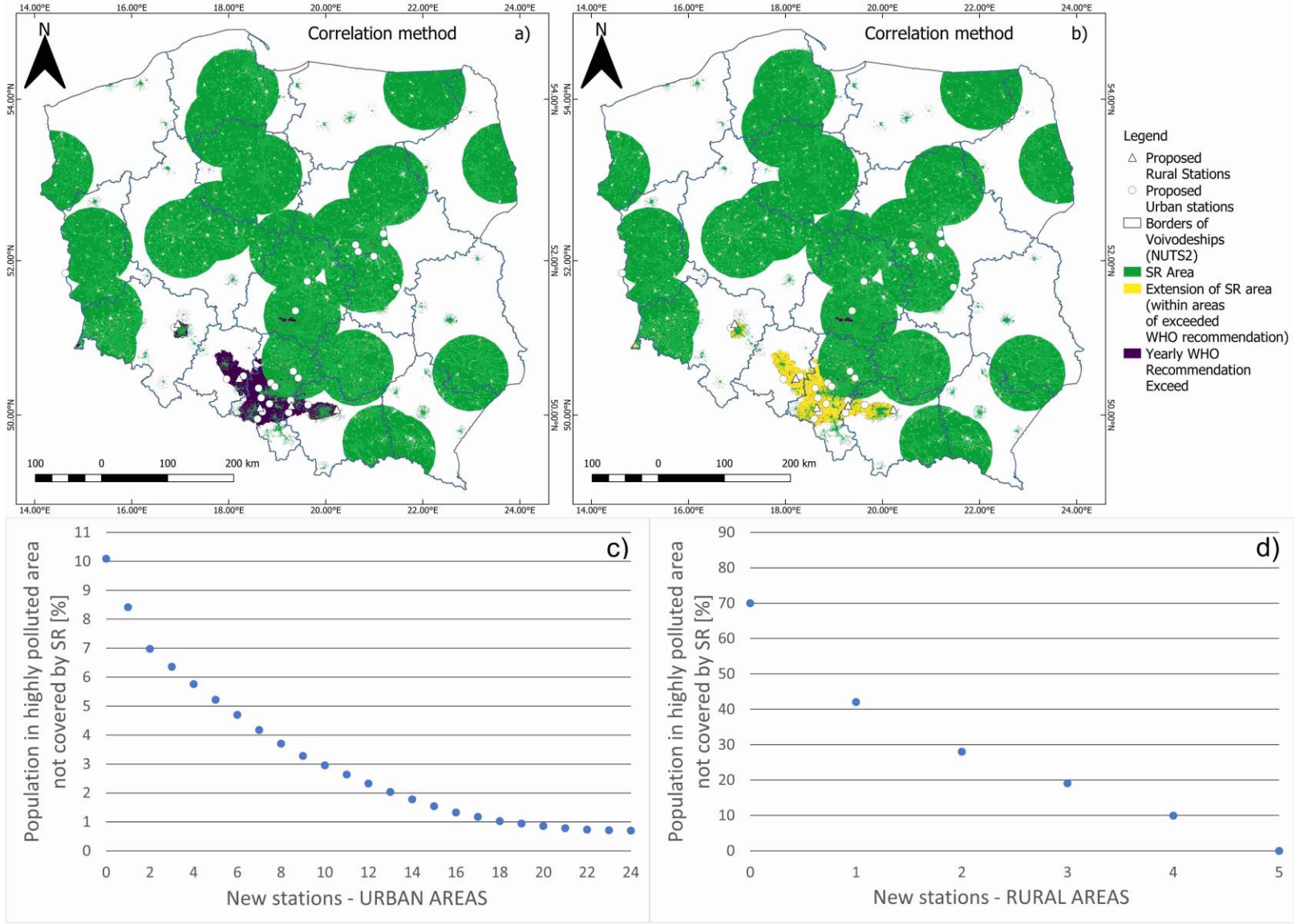
- **Global Moran's I (0.90–0.97)** confirms very strong and significant spatial clustering of surface NO_2 across Poland
- Spatial autocorrelation remains high year-round, with a **slight decrease in winter** when pollution levels are elevated
- **Lower value in May** is linked to temperature variability affecting pollutant variabilities, but overall spatial dependence stays consistently strong

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year
Global Moran's I	0.90	0.92	0.95	0.96	0.90	0.95	0.97	0.94	0.95	0.95	0.91	0.92	0.94

Correlation of NO_2 surface mass concentration between stations with respect to distance

SR Area buffer zone
Urban areas: 10.5 km
Rural areas 56,0 km

10.09% of population lives outside SR area



Advantages and Limitations

Advantages	**Limitations**
Easiest to implement; fastest data preparation and coding	Produces highly generalized results
Lowest memory and computational requirements	Cannot capture anisotropy or directional effects
Very easy to interpret	Limited detail compared to pixel-based or variogram methods
Identifies similar SR locations as more advanced methods	

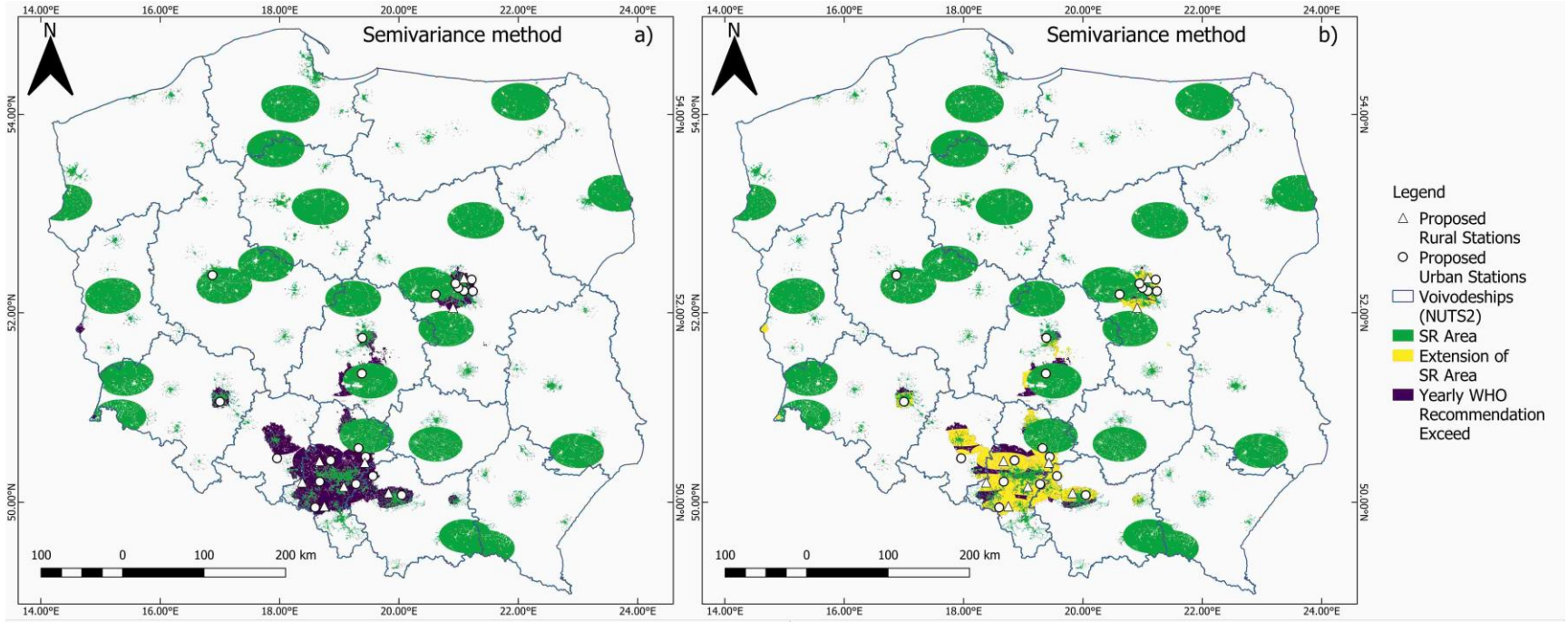
Semivariograms

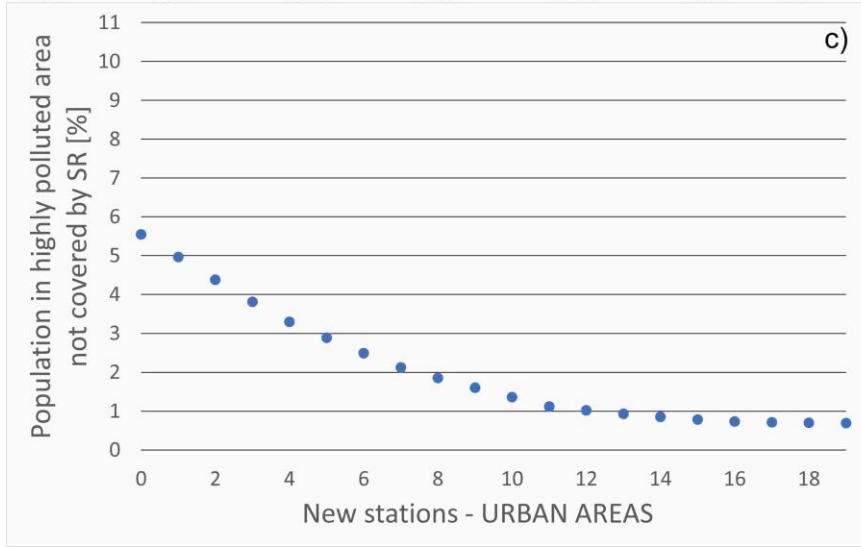
SR Area buffer zone

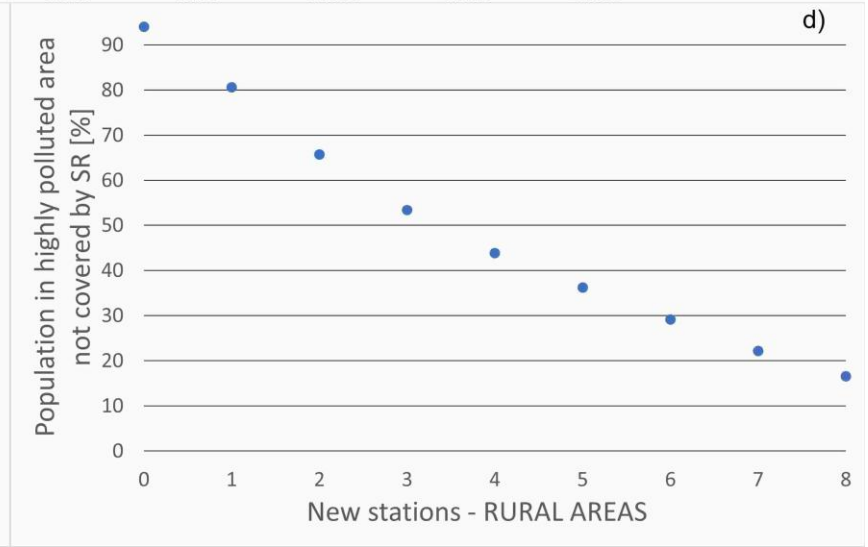
Urban areas: 26 km latitudinally/21 km meridianaly

Rural areas: 64 km latitudinally/41 km meridianaly

5.55% of population lives outside SR area







Advantages and Limitations

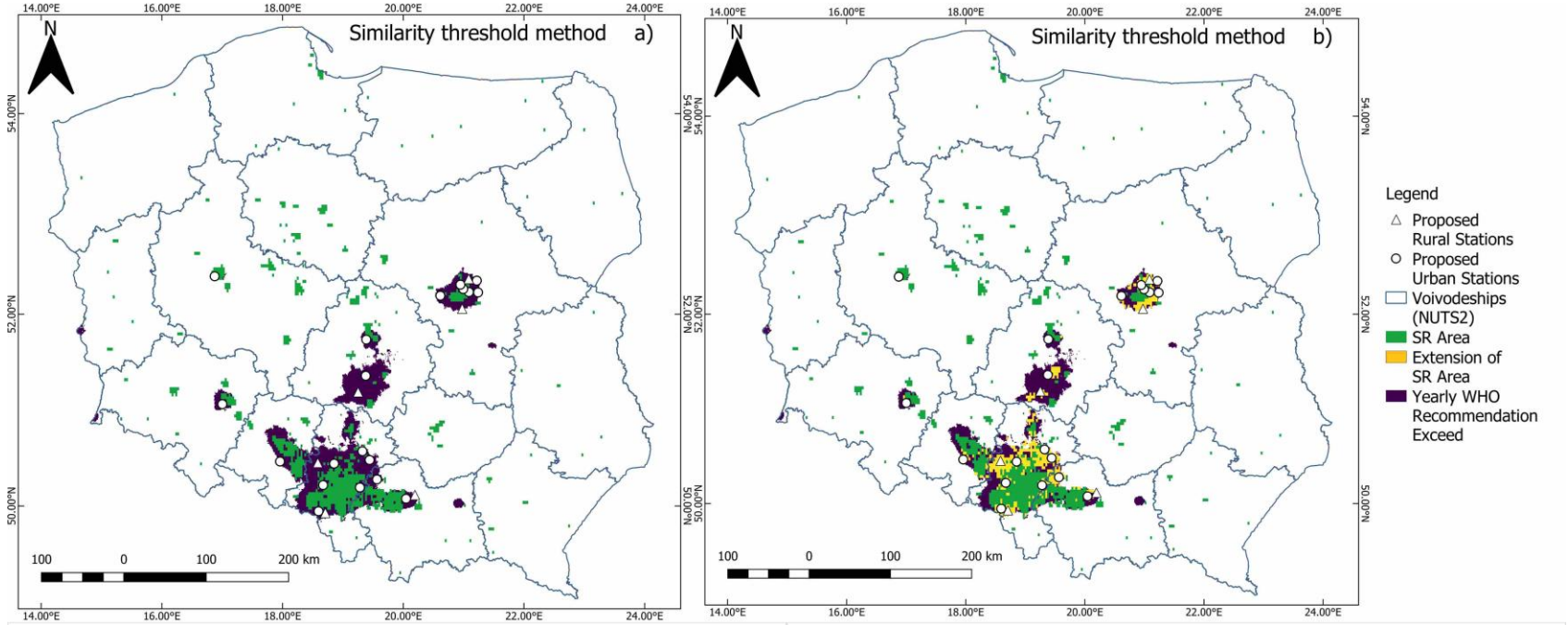
Advantages	**Limitations**
Captures anisotropy and wind-driven directional patterns	More complex to implement and interpret
Allows use of different variogram models tailored to local conditions	Requires careful model selection and fitting
Strong theoretical foundation in spatial autocorrelation research	More computationally demanding than correlation
Produces SR shapes that reflect ‚more' real spatial structure	

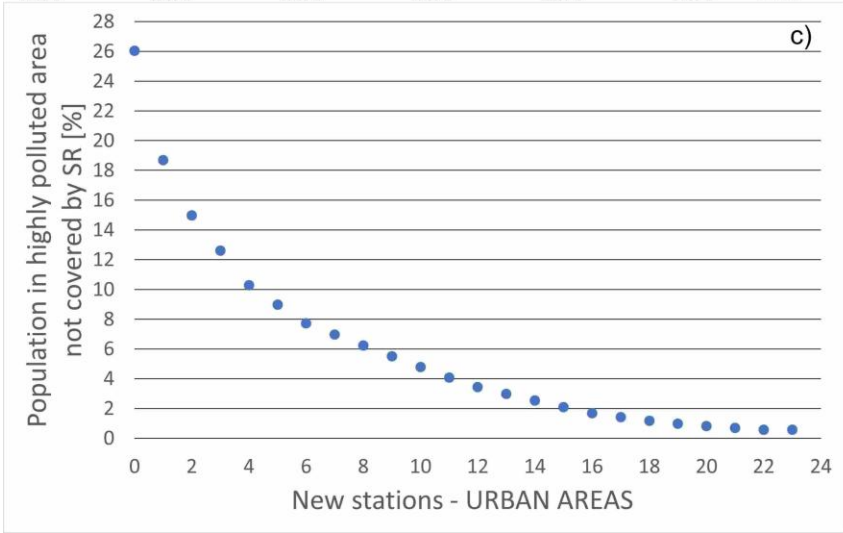
Similarity threshold

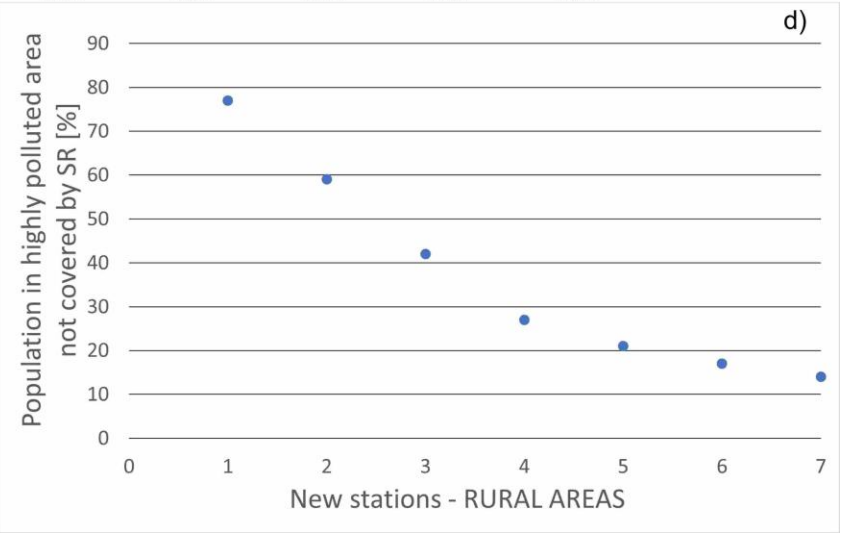
Pixel based

Similiarity within 100 km buffer zone for each station

26% of population lives outside SR area







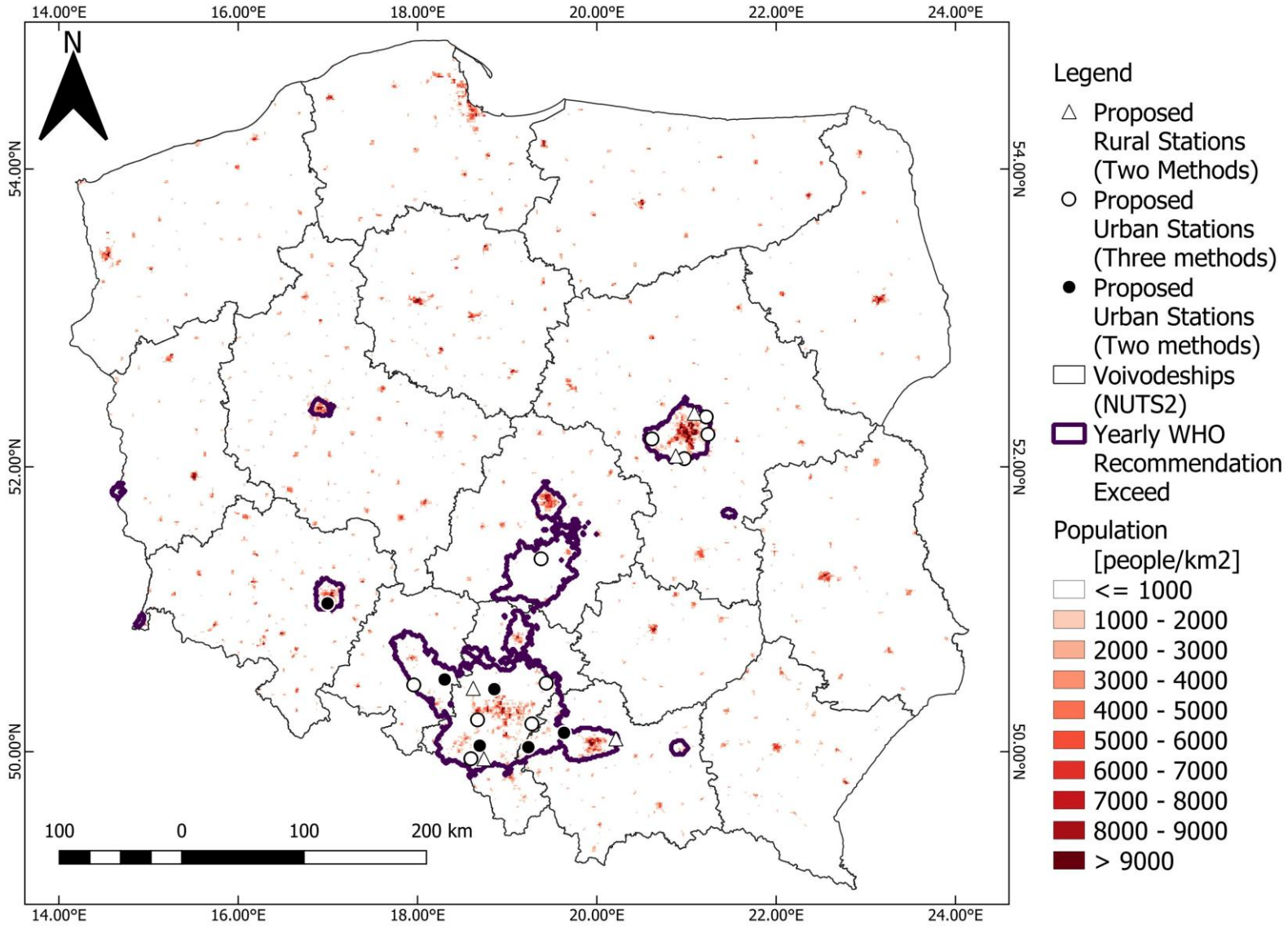
Advantages and Limitations

Advantages	**Limitations**
Produces the most detailed, pixel-level SR zones	Ineffective in low-NO_2 areas (relative percentage-based)
Highly sensitive to small spatial changes	Requires high spatial resolution (<1 km)
Strongly aligned with EU-level representativeness methodology	Most time-, power- and memory-intensive method
Ideal for complex urban environments with high variability	Computational load may limit large-domain applications

Proposed stations

Locations *proposed* by three methods
Jastrzębie-Zdrój (~90,000)
Jaworzno (~85,000)
Zawiercie (~50,000)
Bełchatów (~55,000)
Piaseczno (~45,000)
Eastern Warsaw (~5,000)
Kobyłka (~20,000)
Błonie (~12,000)
Knurów (~40,000)
Krapkowice (~15,000)

Locations *proposed* by two methods
Tarnowskie Góry (~60,000)
Żory (~60,000)
Myszków (~30,000)
Oświęcim (~40,000)
Olkusz (~35,000)
Strzelce Opolskie (~20,000)
Southern Wrocław (~20,000)



Advantages and Limitations

Advantages	**Limitations**
SR areas verified using continuous spatial data	Reliance on Sentinel-5P optical data limited by cloud cover
Consistent proposed station locations across methods	Seasonal gaps: autumn ~24%, winter ~18% usable data
New station locations identified for improved network design	Synthetic filling may reduce model accuracy and decrease the realism of the underlying atmospheric conditions

Summary

- NO_2 air pollution is a **highly spatially auto-correlated phenomenon**
- According to correlation approach, semivariograms and similarity threshold, the results revealed that c.a. **88 %, c.a. 94 % and 74 %** of urban population where yearly NO_2 recommendation was exceeded, is covered by representative NO_2 measurements network
- According to correlation approach, semivariograms and similarity threshold, the results revealed that c.a. **30 %, c.a. 10 % and 10 %** of rural population where yearly NO_2 recommendation was exceeded, is covered by representative NO_2 measurements network
- Results based on the semivariogram approach indicate that the spatial representativeness is influenced by anisotropy. **The SR area is much larger in the latitudinal dimension than in the longitudinal dimension**, for both urban and rural areas
- Anisotropy **influences more significantly on homogeneous areas, such as agriculture or forests**
- The **similarity threshold** method is particularly suited for **highly polluted areas**
- **10 urban and no rural** stations were proposed in the same location **by all methods**
- **17 urban and 5 rural** stations were proposed in the same location by at **least two methods**

Thank you!

Patryk Grzybowski
pgrzybowski@cloudferro.com