RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INERIS
maîtriser le risque pour un développement durable

APPLICATION OF SESAM (data fusion with SEnSors for Air quality Mapping)

WG6 sensors and data fusion

FAIRMODE technical meeting 7-9/10

Alicia Gressent

SESAM (data fusion with SEnSors for AQ Mapping)

METHODOLOGY 1/4

Geostatistical approach → universal kriging with an external drift (KED)


Environment International
Volume 143, October 2020, 105965

Data fusion for air quality mapping using low-cost sensor observations: Feasibility and added-value

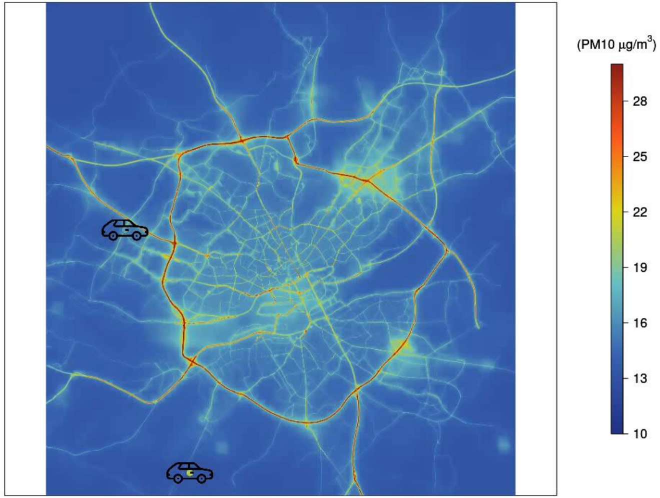
Alicia Gressent [a], Laure Malherbe [a], Augustin Colette [a], Hugo Rollin [a], Romain Scimia [b]

LCSQA Laboratoire Central de Surveillance de la Qualité de l'Air

https://github.com/AliciaGressent/SESAM

Initial development in the framework of the French central laboratory for air quality monitoring:

- Merging fixed and mobile sensor data with model outputs at the urban scale
- Take into account uncertainty and variability of sensor observations
- Application: French city /Nantes (modelling data provided by Air Pays de la Loire – a regional AQ monitoring association / PM sensor data provided by AtmoTrack)



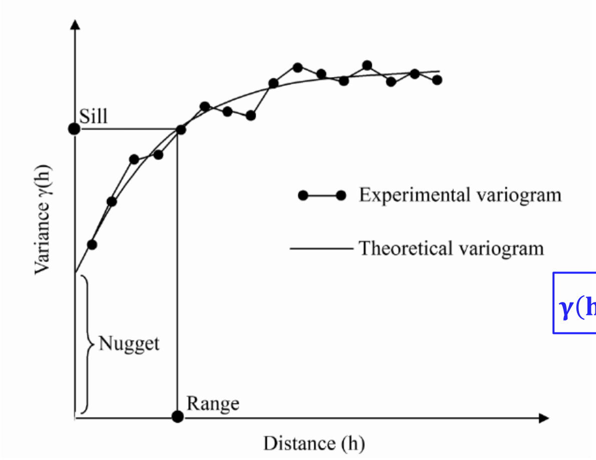
Application in Nantes for PM_{10} based on AtmoTrack sensors and ADMS-Urban simulations.

SESAM (data fusion with SEnSors for AQ Mapping)

METHODOLOGY 2/4

KRIGING

- Estimate that considers observed values and the information on the position
- Response on the spatial regularity and anisotropy of the regionalized phenomenon
- Spatial continuity: 2 observations located close to each other should on average be more similar than 2 distant observations



- Pollutant concentration fields to be estimated ⇔ random process with Z(x): realization of the random variable in x described by the variogram: $\gamma(h)$

$$\boxed{\boldsymbol{\gamma(h) = \frac{1}{2}\,\mathrm{var}[Z(x) - Z(x+h)]}}$$

SESAM (data fusion with SEnSors for AQ Mapping)

METHODOLOGY 3/4

KRIGING

- Non-stationary framework:
  Z(x) expectation is unknown, and it varies in space

  At a location s_0, the concentration $y(s_0)$:

$$\boldsymbol{y}(\boldsymbol{s_0}) = \boldsymbol{m}(\boldsymbol{s_0}) + \varepsilon(s_0)$$

Random process

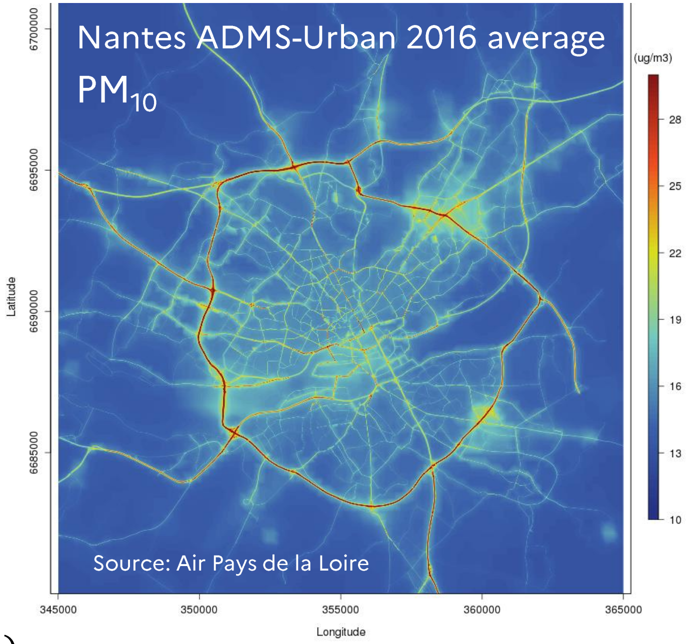
with: $$\boldsymbol{m}(\boldsymbol{s_0}) = \boldsymbol{b_0} + \boldsymbol{b_1}\boldsymbol{x_1}(\boldsymbol{s_0}) + \boldsymbol{b_2}\boldsymbol{x_2}(\boldsymbol{s_0}) + \cdots + \boldsymbol{b_p}\boldsymbol{x_p}(\boldsymbol{s_0})$$

Expectation drift **Coefficients of the linear relation** **Auxiliary variables**

At a location s_0, the estimated concentration $\widehat{y(s_0)}$:

$$\widehat{y(s_0)} = \sum_{i=1}^{N} \lambda_i y(s_i) \quad \text{and} \quad \forall x_p : x_p(s_0) = \sum_{i=1}^{N} \lambda_i x_p(s_i)$$

Dispersion model simulation⇔ drift



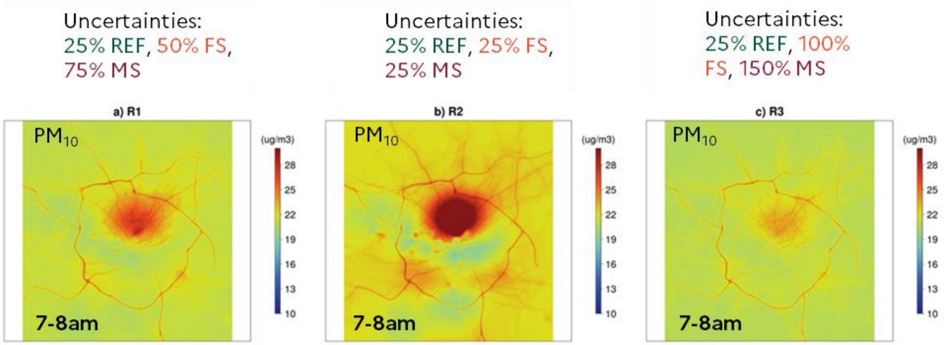
SESAM (data fusion with SEnSors for AQ Mapping)

METHODOLOGY 4/4

- Weight the importance of the sensor data in the spatial interpolation based on data reliability and dispersion

Variance of Measurement Error:

$$\mathbf{VME}= \left[\left(\frac{\sigma}{\sqrt{N}}\right)^2 + \frac{{v_r}^2}{N}\sum_{j=2}^{N}(C_j)^2\right]_i$$



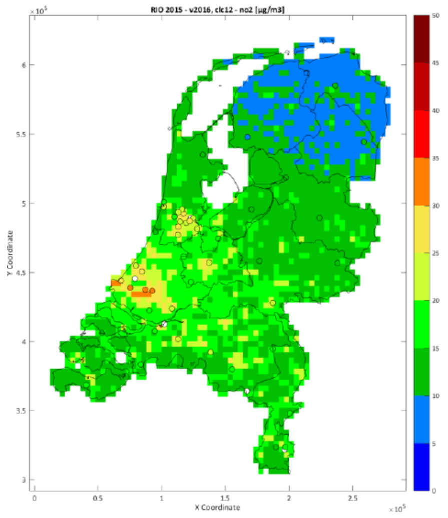
- σ is the standard deviation of the pollutant observations at the position i;
- N is the number of observations at the position i;
- v_r is the constant relative type uncertainty (which depends on the type of sensor → in the original version of the code: **50% fixed sensor observations** and **75% mobile sensor observations**);
- C_j is the jth pollutant concentration at the position i.

WG6 SENSORS AND DATA FUSION EXERCISE

SETUP

- Fairmode WG6 exercise: Netherlands, **$PM_{2.5}$** sensor and station data + **RIO model** estimations 1 x 1 km^2
- Data fusion for 3 runs: basic/standard, sparse and midOut ⇔ reference stations used in RIO model

Run	Official measurements	Sensor measurements	RIO map	Goal
Basic run	**All**	Raw and calibrated using **all** measurements	Based on **all** official measurements	Compare results from different data fusion methods
Option 1a	Leave out 1-2 official measurements in each province	Raw and calibrated using **all** measurements	Leave out the same 1-2 measurements in each province	Can **good** sensors compensate for less official measurements?
~~Option 1b~~	Leave out 1-2official measurements in each province	Leave out the same 1-2 measurements in each province in the calibration	Leave out the same 1-2 measurements in each province	Can **not so good** sensors compensate for less official measurements?
Option 2a	Leave out **all** official measurements in **one** province	Raw and calibrated using **all** measurements in **all** provinces	Based on **all** official measurements in the **other** provinces	Can **good** sensors compensate for a gap in the official measurements?
Option 2b	Leave out **all** official measurements in **one** province	Raw and calibrated using **all** measurements in the **other** provinces	Based on **all** official measurements in the **other** provinces	Can **not so good** sensors compensate for a gap in the official measurements?



https://vito.be/en/news/vito-model-local-air-quality-international-context

- Periods: **January 16-21, 2024, and August 25-30, 2024**

WG6 SENSORS AND DATA FUSION EXERCISE

DATASET 1/2

- Observations from sensors and reference stations on **January 16-21, 2024**

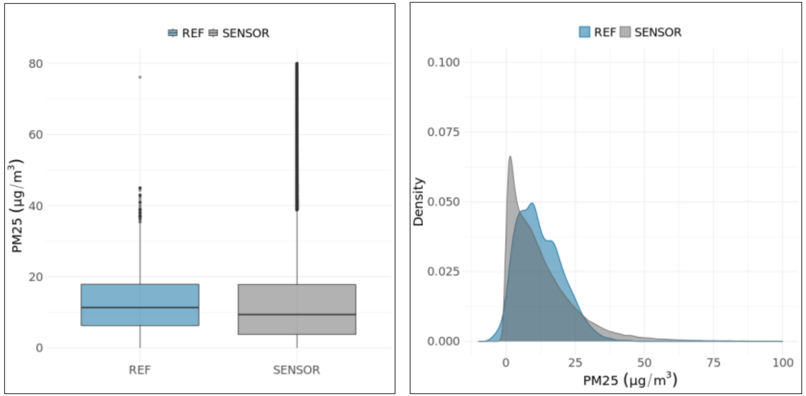
Reference data: 56 reference stations

```
     ID                    lon               lat                pol
Length:8815        Min.   : 46954   Min.   :317396   Min.   :-999.00
Class :character   1st Qu.: 99415   1st Qu.:400665   1st Qu.:   5.45
Mode  :character   Median :119510   Median :439310   Median :  10.76
                   Mean   :134562   Mean   :438255   Mean   : -17.11
                   3rd Qu.:177004   3rd Qu.:488302   3rd Qu.:  17.50
                   Max.   :259669   Max.   :594143   Max.   :  76.10
```

Sensor data: 1761 sensors

```
     ID                    lon               lat                pol
Length:214173      Min.   : 29266   Min.   :309377   Min.   : -458.2
Class :character   1st Qu.:138699   1st Qu.:425482   1st Qu.:    3.8
Mode  :character   Median :169468   Median :459471   Median :    9.7
                   Mean   :171350   Mean   :472241   Mean   :   36.3
                   3rd Qu.:203619   3rd Qu.:528124   3rd Qu.:   18.5
                   Max.   :275881   Max.   :610575   Max.   :13149.2
```

Data quick look



- Negative and > 500 µg/m³ $PM_{2.5}$ values have been removed from the dataset for data fusion

- Only stations that are not used in the RIO estimation for each run type are considered:
Basic: 34 stations, midOut: 31 stations and sparse: 8 stations

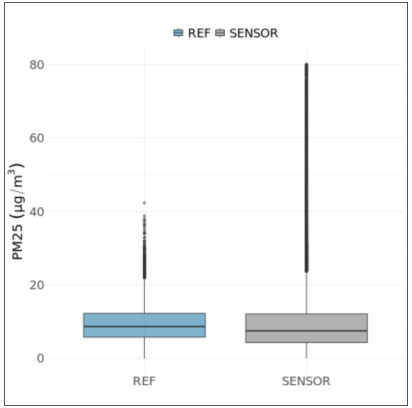
DATASET 2/2

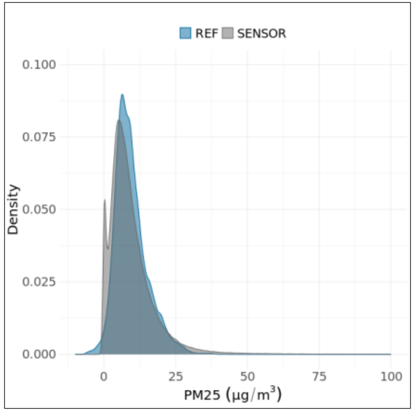
- Observations from sensors and reference stations on **August 25-30, 2024**

Reference data: 54 reference stations

```
      ID                  lon               lat                  pol
 Length:8431         Min.   : 46954   Min.   :317396   Min.   :-999.00
 Class :character    1st Qu.: 99415   1st Qu.:400665   1st Qu.:   5.30
 Mode  :character    Median :119806   Median :441664   Median :   8.21
                     Mean   :135922   Mean   :438436   Mean   : -27.84
                     3rd Qu.:178070   3rd Qu.:489238   3rd Qu.:  12.00
                     Max.   :259669   Max.   :594143   Max.   :  42.30
```

Data quick look





Sensor data: 1662 sensors

```
      ID                  lon               lat                  pol
 Length:229235       Min.   : 29411   Min.   :315002   Min.   :-446.60
 Class :character    1st Qu.:138108   1st Qu.:426396   1st Qu.:   4.40
 Mode  :character    Median :168504   Median :458651   Median :   7.70
                     Mean   :170476   Mean   :471856   Mean   :  45.11
                     3rd Qu.:203926   3rd Qu.:527944   3rd Qu.:  12.80
                     Max.   :273238   Max.   :610796   Max.   :4160.10
```

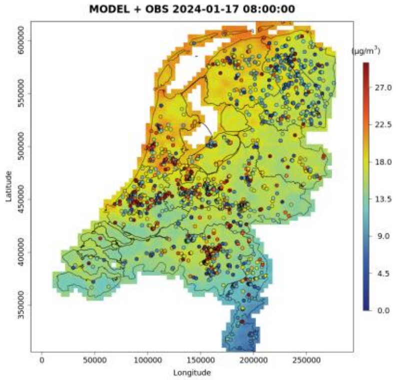
- Negative and > 500 µg/m³ $PM_{2.5}$ values have been removed from the dataset for data fusion
- Only stations that are not used in the RIO estimation for each run type are considered:

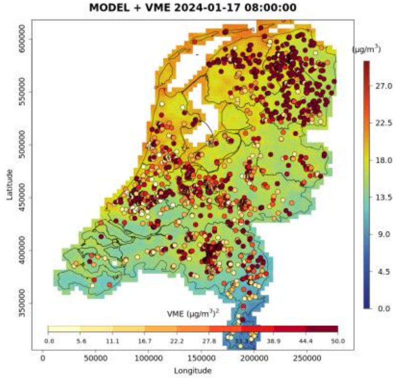
Basic: 34 stations, midOut: 31 stations and sparse: 8 stations

WG6 SENSORS AND DATA FUSION EXERCISE

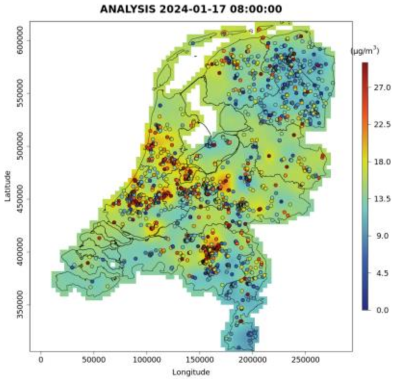
DATA FUSION 1/2

Example on January 17, 8 am

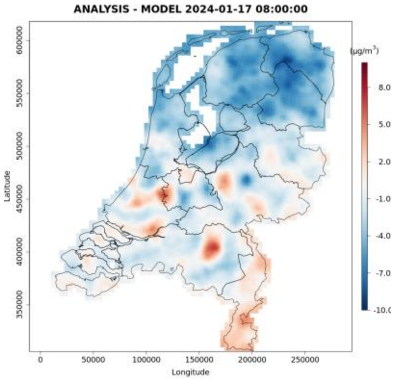




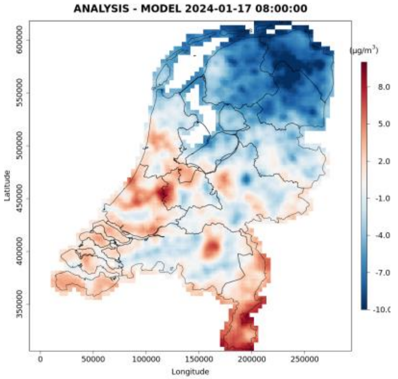
$$\text{VME} = \sqrt{U_{random}{}^2 + U_{calibration}{}^2}$$

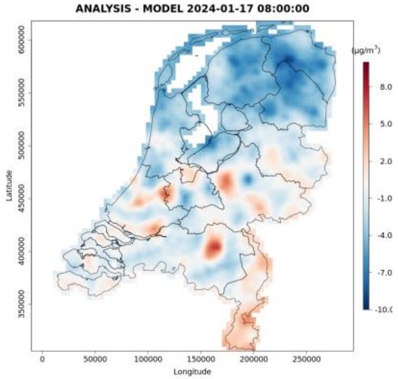


Basic run



Sparse run





MidOut run

WG6 SENSORS AND DATA FUSION EXERCISE

INTERIM CONCLUSIONS AND NEXT STEPS

Use of the SESAM tool:

- Adjustment of SESAM from the use at the urban scale to national scale
- Implementation of a more realistic VME into the kriging by using the calculated sensor uncertainties and reference station accuracy

→ SESAM seems to improve $PM_{2.5}$ concentration estimate at most stations but significant deterioration at some locations

To be implemented / tested with the SESAM tool:

- Take model uncertainty into account in the kriging approach → for instance error-in-variable KED model
- Test other approaches based on AI (Land use regression based on random forest for example and compare to kriging)

Average ratio RMSE over the two estimation periods

INERIS	STANDARD	MIDOUT	SPARSE
	91%	89%	84%
NL01485			110%
NL01487	107%	108%	104%
NL01488			
NL01489	95%	95%	90%
NL01491	111%	111%	122%
NL01493	113%	113%	114%
NL01494			
NL01495			96%
NL01496			
NL01913	91%	89%	85%
NL10131			62%
NL10136	80%	83%	80%
NL10138			
NL10230			
NL10240	106%	106%	87%
NL10241			73%
NL10247			57%
NL10248	75%	72%	69%
NL10404			77%
NL10418			97%
NL10444			83%
NL10449	89%	89%	86%
NL10450	94%	91%	78%
NL10538			74%
NL10636	83%	90%	85%
NL10641	83%	79%	79%
NL10643		90%	98%
NL10644		82%	
NL10738		74%	
NL10741	80%	79%	75%
NL10742			80%
NL10821			62%
NL10934			
NL10937	62%	64%	52%
NL10938			60%
NL49007	97%	94%	100%
NL49012	94%	94%	97%
NL49014			88%
NL49016			
NL49017	84%	82%	79%
NL49551			81%
NL49553			79%
NL49556			77%
NL49561			100%
NL49570			88%
NL49572			85%
NL49573			87%
NL49701			83%
NL49703			86%
NL49704			70%