Spatial correction of low-cost sensors observations for fusion of air quality measurements

Jean-Michel Poggi

Université de Paris

université PARIS-SACLAY
FACULTÉ DES SCIENCES D'ORSAY

MATHEMATIQUES ORSAY

Joint work with *Michel Bobbia (Atmo Normandie)* & *Bruno Portier (INSA Rouen)*

enbis

Outline

- *Context*
  - Pollution: micro-sensors
  - Statistical context: geostatistical assimilation

- *Methods and Algorithms*
  - Kriging method & Kriging residuals
  - Spatial correction algorithm

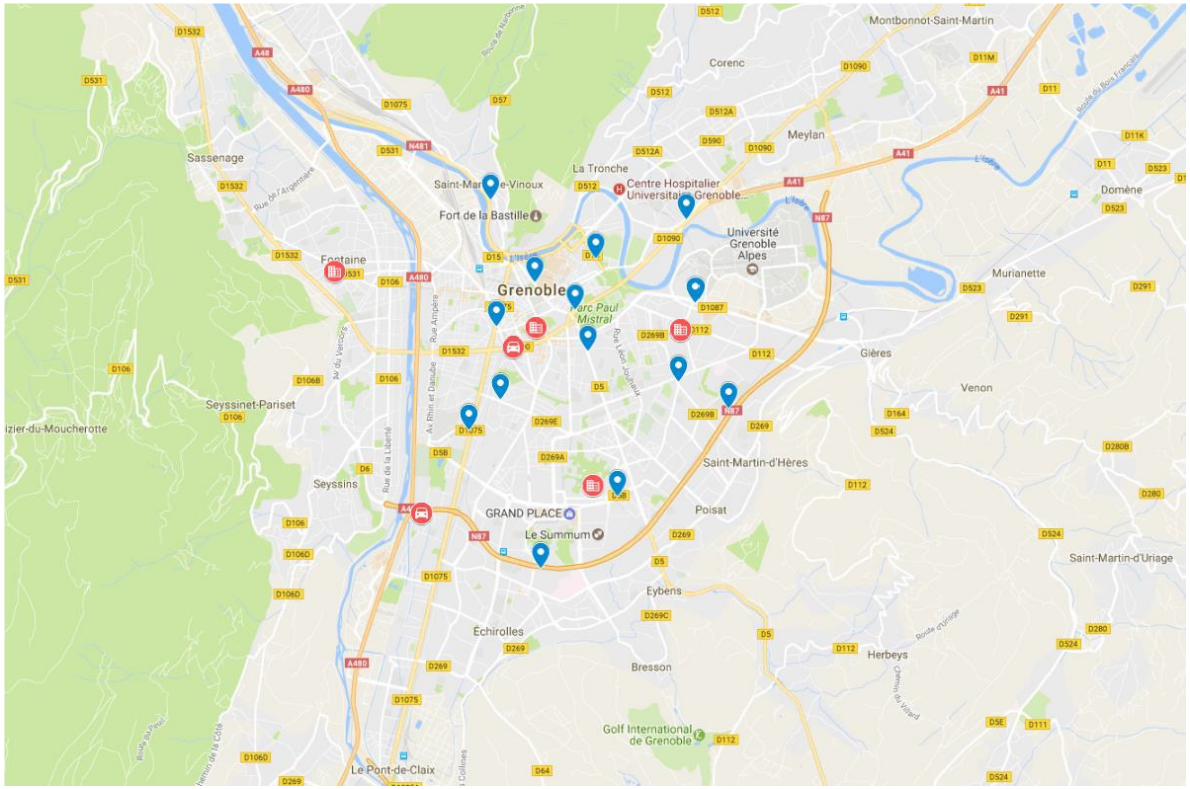
- *Numerical experiments*
  - Simulation study
  - Real data application

Motivation: use low-cost sensors to improve pollutant maps

- *Network 1 (fixed):* *reference* network, high quality but highly sparse
- *Network 2 (fixed or mobile):* *micro-sensors*, of less good quality but homogeneous and more dense
- *Network 3 (mobile):* of *micro-sensors* in connected objects, of medium (or unknown) quality and heterogeneous but potentially very dense

Illustration with 2 networks

The reference network 1 and the micro-sensors network 2

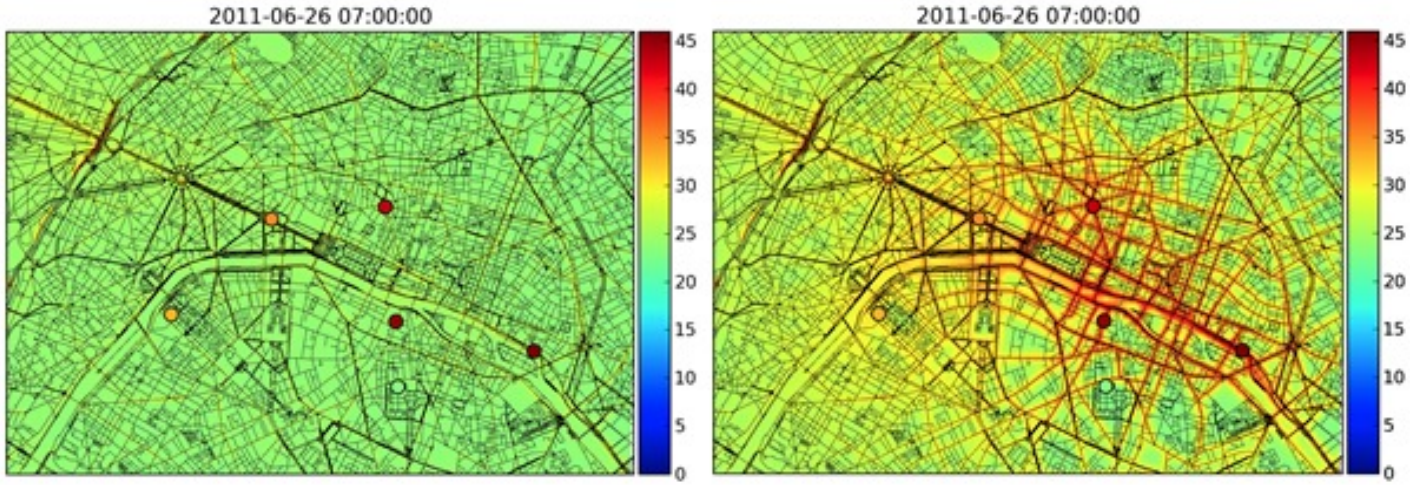


®Atmo Auvergne-Rhône-Alpes (2017)

City of Grenoble (France)

- Reference network 1
  quite sparse
- Micro-sensors network 2
  more dense

Data assimilation to combine numerical models and observations



Projet Votre Air : collaboration Airparif, Inria, Numtech

Model outputs (map) +
Pointwise observations

Mapping
assimilated

- A way to improve the quality of reconstructed map interest:
  to increase the density of sensors
- Availability of low-cost sensors in addition to reference stations
  measurements, opens a possibility without a prohibitive cost

Statistical context

- Geostatistical approach for the fusion of measurements

  Schneider et al. (2017) seminal study about NO2 in Oslo

  Gressent et al. (2020) a similar study for Nantes (France) including also mobile micro-sensors data

  Miskell et al. (2018), ***Weissert et al. (2020)*** hierarchical network design, low-cost sensor checking and correction by data fusion

- *1st step*: correct micro-sensors measures thanks to those given by the reference sensors. In general offline pre-processing, during a preliminary colocation study.

  Spinelle et al. (2015, 2017), ***Borrego et al. (2016)***

- However this calibration preprocessing could fail to adapt quickly to various changes (technology, preprocessing included by the sensors providers, inhomogeneous sources, …)

- ***We complement such approaches by a simple scheme merging the two steps by considering online spatial correction of micro-sensors***

Kriging method

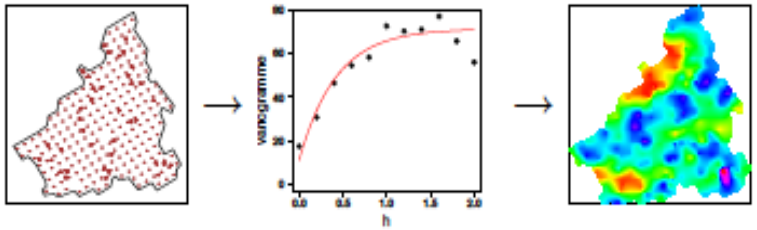
- See ***Cressie (1993)***
- Measures: $Z_t(s_k), 1 \leq k \leq K$

$$\hat{Z}(s_0) = \sum_k \lambda_k(s_0) Z(s_k)$$

- Underlying spatial interpolation model

$$Z(s) = \mu(s) + \varepsilon(s), s \in S$$

Measures → Model → Interpolation



μ(s) = m (ordinary kriging), ***m*** to be estimated

μ(s) = b₁f₁(s) + … + b_K f_K(s) (universal kriging) where the functions ***f_k(s)*** are given and the ***b_k*** are to be estimated

A special case is the kriging with external drift, modeling ***Z(s)*** as a linear function of a deterministic map typically a numerical model output

ε(s) is a zero mean stationary with a spatial dependence structure given by the variogram

$$\gamma(h) = \frac{1}{2}\text{var}\big(Z(s+h) - Z(s)\big) = C(0) - C(h)$$

where $C(h) = \text{cov}\big(Z(s), Z(s+h)\big)$

- *Remark*: weights depend on covariances and distances but not on *Z*

Kriging residuals

- Idea: correct predictions by measures, see ***de Fouquet et al. (2011)***

  Since the kriging requires some spatial stationarity, apply kriging to residuals (innovations) instead of concentrations

- Start from a predicted map given by a deterministic model like *Esmeralda* or *Chimere*: $(P_t(s), s \in S)$
- Define the pseudo-innovations by the prediction errors

$$\boldsymbol{E_t(s_k) = P_t(s_k) - Z_t(s_k), 1 \leq k \leq K}$$

- Kriging the innovation process to obtain estimates

$$\boldsymbol{(\widehat{E}_t(s), s \in S)}$$

and then deduce a corrected map

$$\boldsymbol{\widehat{P}_t(s) = P_t(s) - \widehat{E}_t(s), s \in S}$$

- *Remark*: here, kriging relates to the difference between observed concentration and model output, whereas kriging with external drift relates directly to the concentration at the measurement stations

Spatial iterative correction

Correction is first defined for an asymmetric situation: micro-sensor network Res_2 corrected by the reference network Res_1 supposed to be of high quality.

It consists in two steps performed iteratively:

1. Kriging a map $C\_Res_2(s), s \in S_2$ (typically a neighborhood or a city) using simply Res_2 measurements or by statistical adaptation of a map. This provides an estimate $\widehat{Res_1}(s)$ for all the points of network 1 included in S_2 (a few points)

2. Kriging the differences $Res_1(s) - \widehat{Res_1}(s)$ to get a correction $Corr(s)$ and then a new map from which corrected measurements of network 2 are extracted

$$Res_{2,corr}(s) = Res_2(s) + Corr(s)$$

This can then be ***iterated until stabilization*** of the quality of the maps obtained by successive corrections of Res_2 evaluated in the stations of Res_1, measuring the proximity of corrected microsensors measures to the reference measurements

Spatial iterative correction (2)

Spatial correction algorithm can also be useful in a more symmetric case, typically for measurements coming from two or more sub-networks of diverse quality. Iterative correction for the homogenization of micro-sensor data from 2 different populations (Res_2 and Res_3) can easily be defined:

1. Set $\boldsymbol{Res_a = Res_2}$ and $\boldsymbol{Res_b = Res_3}$
2. By kriging using Res_a measurements, build a map $C\_Res_a(s), s \in S_a$ to provide an estimate $\widehat{\boldsymbol{Res_b}}(\boldsymbol{s})$ at any point of the b network in S_a
3. By kriging the differences $\boldsymbol{Res_b(s) - \widehat{Res_b}(s)}$ get a correction $\boldsymbol{Corr(s)}$ and then deduce the corrected measurements of network a

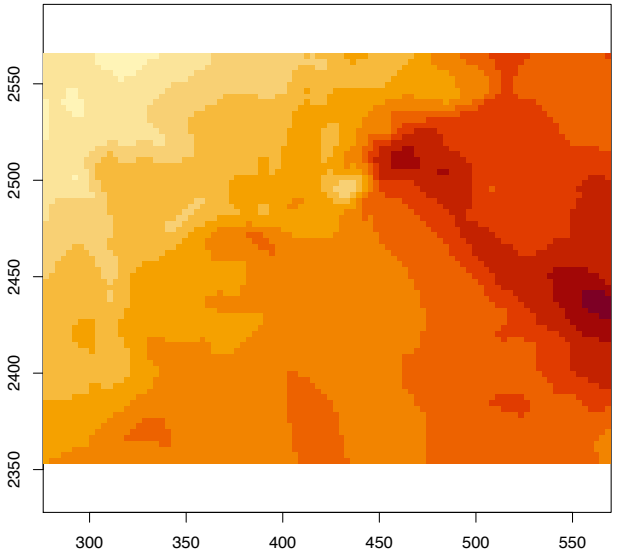
$$\boldsymbol{Res_{a,corr}(s) = Res_a(s) + Corr(s)}$$

4. Then we iterate such stages by exchanging the roles of the two networks. Swap $\boldsymbol{Res_a = Res_3}$ and $\boldsymbol{Res_b = Res_{a,corr}}$
5. Then one iterates such pairs of stages on the networks obtained by mutual correction, until stabilization of the difference of the two reconstructed maps

Numerical experiments

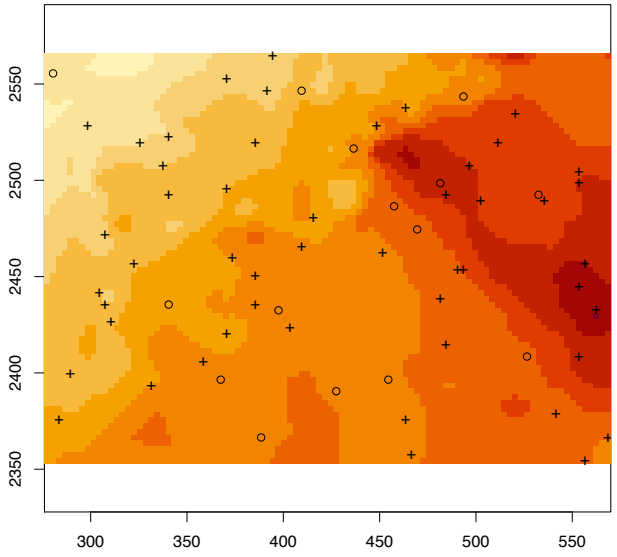
- Computations were performed using the `RGeostats` geostatistical package, developed with R language

- Simulated Data. Starting from the map of the daily maximum ozone concentration, August 26, 2019, output from Esmaralda



Simulated map obtained from concentration values and according to a cubic variogram of range 300

$$Y_{ms} = aY + b + eE$$

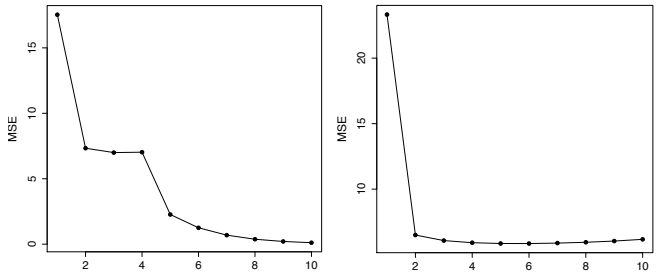


N_1 reference stations (network 1) o
N_1 = 5, 8, 15, 25, 50
N_2 micro-sensors (network 2) x
N_2 = 25, 50, 75, 100

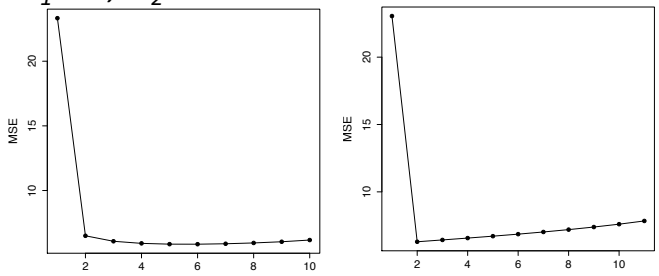
Simulation study

- How many iterations to converge?

N_1=8, N_2=25



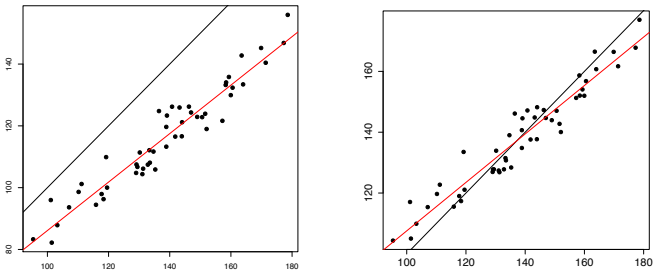
N_1=15, N_2=50



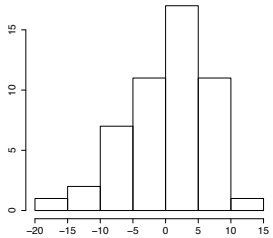
Errors associated with network 1 (left) and network 2 (right) along iterations

- Performance of the correction?

N_1=15, N_2=50



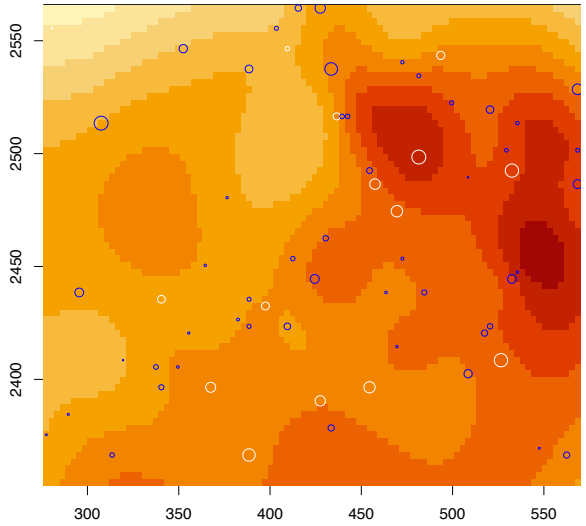
Micro-sensors measures (left) and corrected measures (right) versus observed concentrations



Distribution of errors after correction

Simulation study (2)

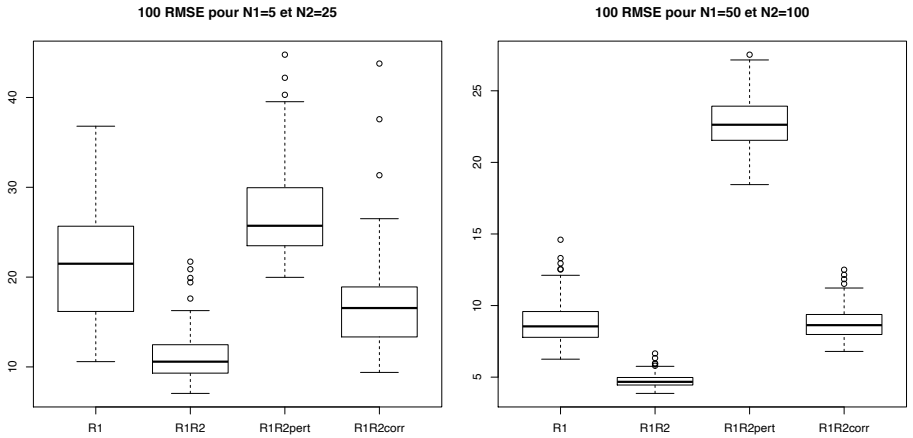
- Performance of the correction?



- Geographical location of errors represented by a proportional bubble.
- The smallest micro-sensor errors are often observed near fixed stations. They are also found at the lowest concentrations.

- Performance of the fusion?

Boxplots of the 100 RMSE values



N_1=5, N_2=25 (realistic)
the gain is significant

N_1=50, N_2=100
the correction is less interesting since the contribution of the micro-sensors is negligible

Application to real data

- In the next future, *Atmo Normandie* will, for the city of Rouen, collect in addition to their reference network of few stations, a second network of numerous micro-sensors
- But at this stage, no such data are available. We have used the real data from the Mobicit'air project of *Atmo Auvergne-Rhône-Alpes*, the pollution network of the Grenoble area, in partnership with *Grenoble-Alpes-Métropole*

Data

- Hourly average concentrations (in $\mu g/m^3$) of NO_2 pollutant
- 147 days, from January 5, 2017 to May 31, 2017
- N_1=6 reference stations located in the Grenoble area, close to traffic, in urban or suburban places
- N_2=15 micro-sensors

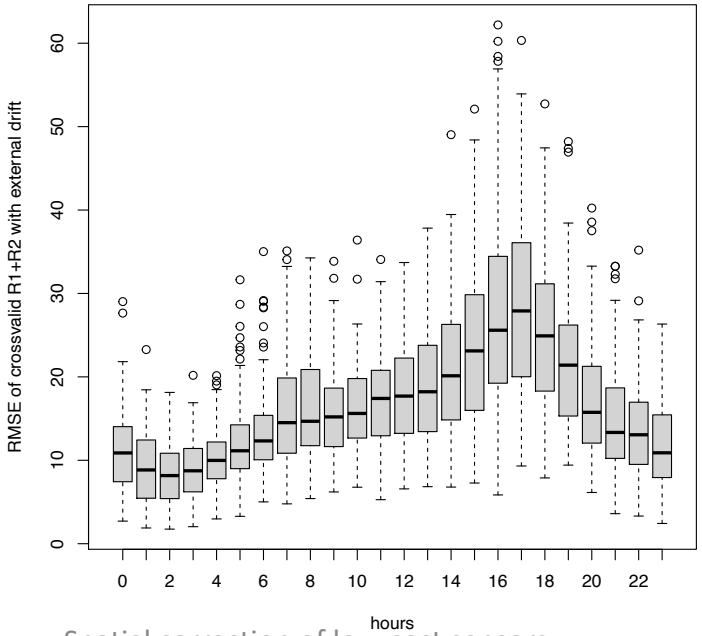
@Atmo Auvergne-Rhône-Alpes (2017) in report *''Assimilation de données de micro-capteurs dans les cartographies fines échelles''*

Global RMSE-CV

RMSE-CV to assess the correction method, "leave-one-out" resampling procedure, for each reference station, we can estimate the concentration at this point of *Res1* without using the actual measure

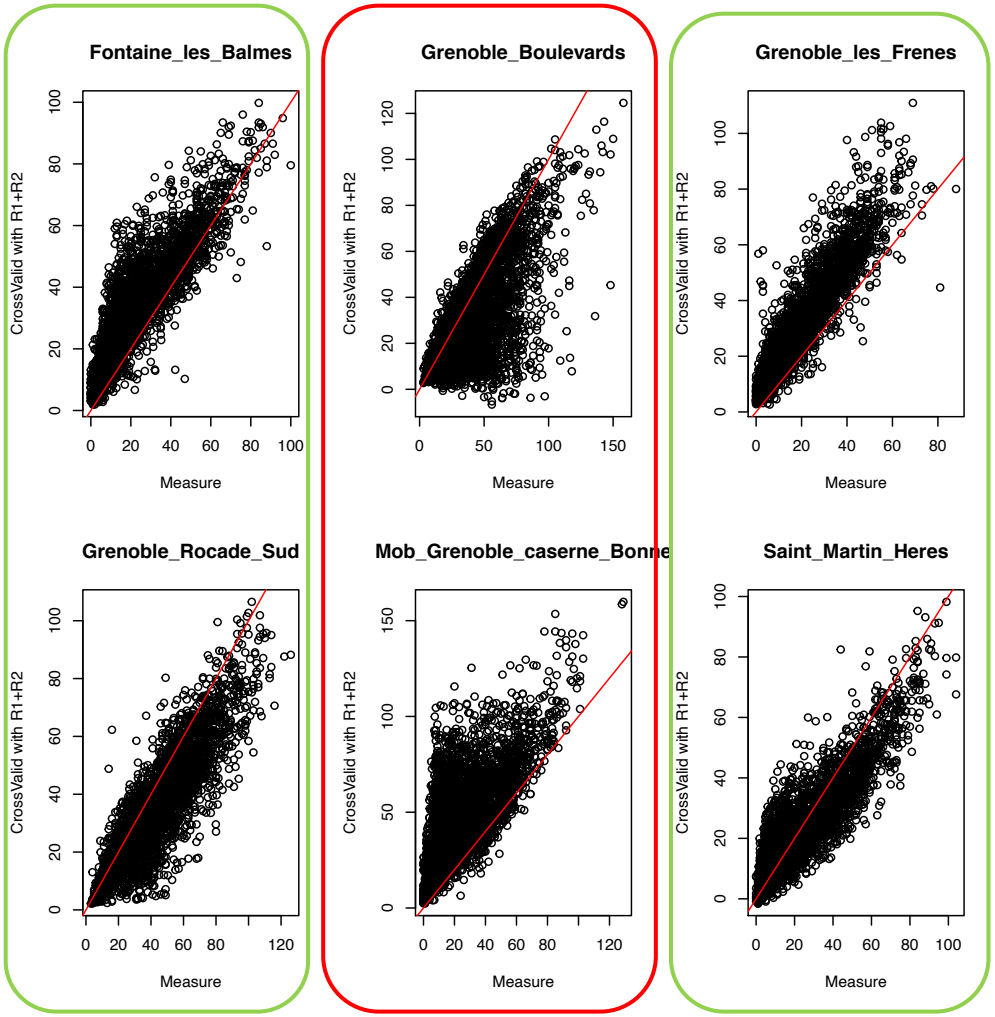
- For each point, apply the iterative correction procedure by kriging using only the measurements provided by all the other sensors and can deduce an estimation error at this point
- This process is repeated for all the sensors of network 1, and the quality is summarized by the average of the estimation errors

- Average RMSE over all the 2,980 available hours: 20,7 $\mu g/m^3$ for ordinary kriging 18,8 $\mu g/m^3$ using external drift



- The boxplots of the RMSE as a function of hour
- The quality depends on time instant
- The median as well the variability vary as expected, smaller during the night and higher during the end of the afternoon

Spatial RMSE-CV



Individual RMSE
10 to 14 for the 2nd group
30 to 23 for the 1st group

- Pretty large the two stations in the middle, located in Grenoble city-center
- Why? each station serves as a proxy for the other one in the CV scheme, -
  - one is a station near the *traffic* leading to *underestimation*
  - the other is a *background* urban station leading to *overestimation*

- Good or acceptable for the four other stations located on the border of Grenoble downtown (urban background stations, exhibiting more *balanced scatter plots*)

More details in the paper

RESEARCH ARTICLE

Spatial correction of low-cost sensors observations for fusion of air quality measurements

Michel Bobbia, Jean-Michel Poggi ✉, Bruno Portier

First published: 14 September 2022 | https://doi.org/10.1002/asmb.2713

Contact : Jean-Michel.Poggi@math.u-psud.fr

Jean-Michel Poggi

U Paris Cité & LMO, U Paris-Saclay