INLA/SPDE WORKSHOP

Introduction of the air pollution dataset and elements of comparison between space-time estimation methods applied to air quality forecasting

Maxime Beauchamp [1, 2*], Laure Malherbe[2]
Marta Valsania [3], Frédérik Meleux[2] and Anthony Ung[2]

[1] PSL University, Mines ParisTech

[2] Institut National de l'Environnement Industriel et des Risques (INERIS)

[3] University of Turin

November 8, 2018

Scientific context

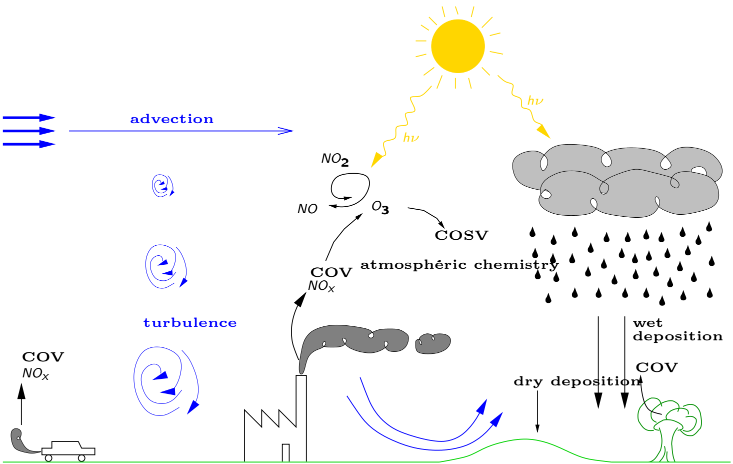
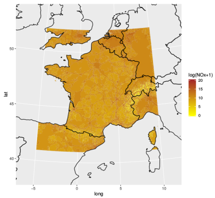
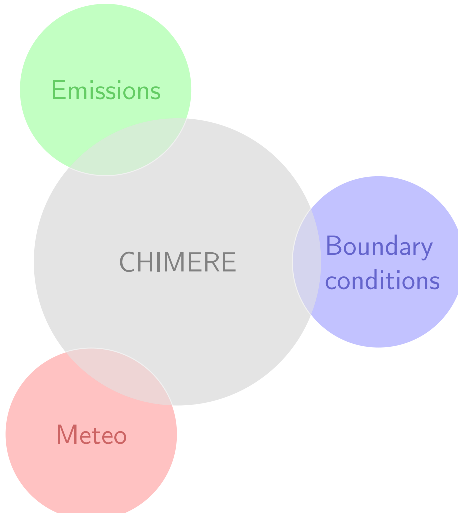


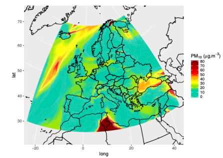
Figure 1 – Atmospheric dynamics of the pollutants

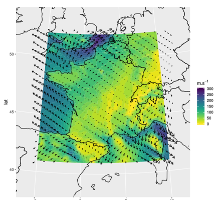
Scientific context

Implementation of mathematical models to describe the evolution processes of the chemical species (pollutant) in the troposphere









Air Pollution Dataset

CHIMERE

The RObject CHIM_2014.RData : data.frame with daily PM_{10} and $PM_{2.5}$ CHIMERE simulations in 2014.
CHIMERE is run over the AFM French simulation domain, with a 10 km resolution

Listing 1 – Reading the CHIMERE dataset

```
> load('CHIM_2014.RData')
> str(CHIM_daily2014)
'data.frame': 4092015 obs. of  5 variables:
 $ long     : num  -5 -4.85 -4.7 -4.55 -4.4 ...
 $ lat      : num  41 41 41 41 41 41 41 41 41 41 ...
 $ date     : POSIXct, format: "2013-12-31" "2013-12-31" "2013-12-31" "2013-12-31" ...
 $ PM10_CHIM: num  2.74 3.01 3.3 3.42 3.56 ...
 $ PM25_CHIM: num  2.13 2.32 2.53 2.6 2.71 ...
> date = as.POSIXct("20140315",format="%Y%m%d",tz="UTC")
> simu = CHIM_daily2014[CHIM_daily2014$date==date,]
> str(simu)
'data.frame': 11211 obs. of  5 variables:
 $ long     : num  -5 -4.85 -4.7 -4.55 -4.4 ...
 $ lat      : num  41 41 41 41 41 41 41 41 41 41 ...
 $ date     : POSIXct, format: "2014-03-15" "2014-03-15" "2014-03-15" "2014-03-15" ...
 $ PM10_CHIM: num  12.4 12.9 13.4 13.6 13.7 ...
 $ PM25_CHIM: num  9.47 9.98 10.42 10.6 10.57 ...
```

Air Pollution Dataset

CHIMERE

The RObject CHIM_2014.RData : data.frame with daily mean of the PM_{10} and $PM_{2.5}$ hourly CHIMERE simulations in 2014.
CHIMERE is run over the AFM french simulation domain, with a 10 km resolution

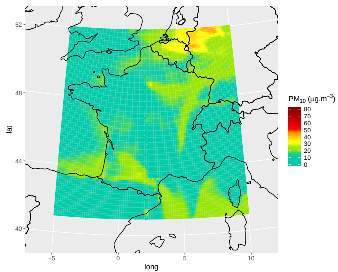
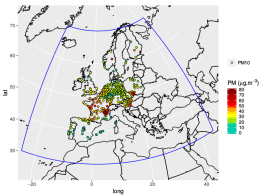


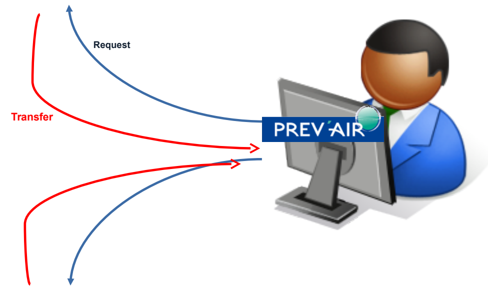
Figure 2 – CHIMERE simulation (15th of March 2014)

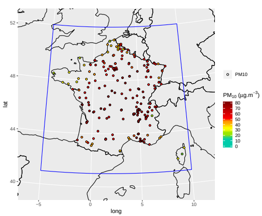
Collecting the data



Europe

European database :
1) Ozoneweb (real-time data)
2) Airbase (validated data)





National database :
GEOD'AIR
Real-time data
Validated data

France

Air Pollution Dataset

Observations

The RObject OBS_2014.Rdata : data.frame with daily mean of the PM_{10} and $PM_{2.5}$ hourly observations of the french GEOD'AIR database in 2014.

Listing 2 – Reading the observational dataset

```
> load('OBS_2014.RData')
> str(OBS_daily2014)
'data.frame': 185055 obs. of  7 variables:
 $ ID             : Factor w/ 507 levels "FR01001","FR01005",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ long           : num  5.8 5.8 5.8 5.8 5.8 ...
 $ lat            : num  49.5 49.5 49.5 49.5 49.5 ...
 $ type_of_station: Factor w/ 5 levels "","Background",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ date           : POSIXct, format: "2014-01-01" "2014-01-02" "2014-01-03" "2014-01-04" ...
 $ PM10           : num  6.8 4 10 10 7.9 5.8 8.7 8.8 6.1 9.2 ...
 $ PM25           : num  NA NA NA NA NA NA NA NA NA NA ...
```

Air Pollution Dataset

Observations

The RObject OBS_2014.Rdata : data.frame with daily mean of the PM_{10} and $PM_{2.5}$ hourly observations of the french GEOD'AIR database in 2014.

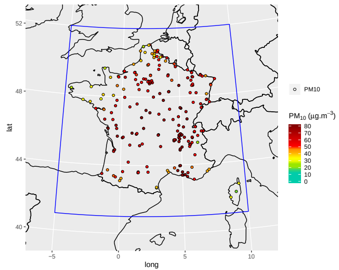


Figure 3 – Observations (15th of March 2014)

PREV'AIR

Operational system for air quality monitoring and forecasting over Europe and France, under the aegis of the Ministry in charge of the environment

- Partners : INERIS, Météo-France, CNRS, IPSL, LCSQA
- Set up in 2003 to deliver daily AQ forecasts and maps on France & Europe
- Based on deterministic chemistry-transport modelling and post-processing using in situ observation data
- During pollution episodes, alert procedures are mainly triggered according to the forecast situation for the previous day (D-1) and next days (D+0, D+1, D+2)

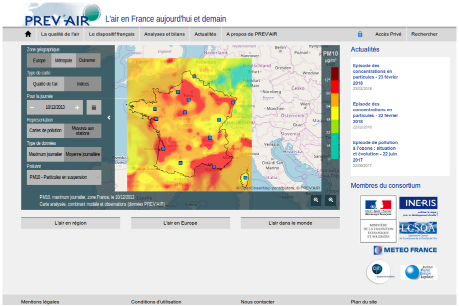


Figure 4 – Screenshot of the PREVAIR website http://www2.prevair.org/

Two products are delivered by the PREV'AIR system :

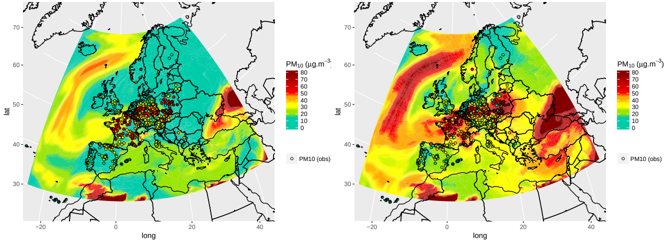
I) Analysis (Estimation problem)

Map of the previous day (D-1)

1) Meteorology, Emissions and Boundary conditions are used to run a CHIMERE simulation

2) Monitoring data are collected (France + Europe)

3) Kriging of **background** concentration measurements with CHIMERE data as external drift



(a) CHIMERE simulation

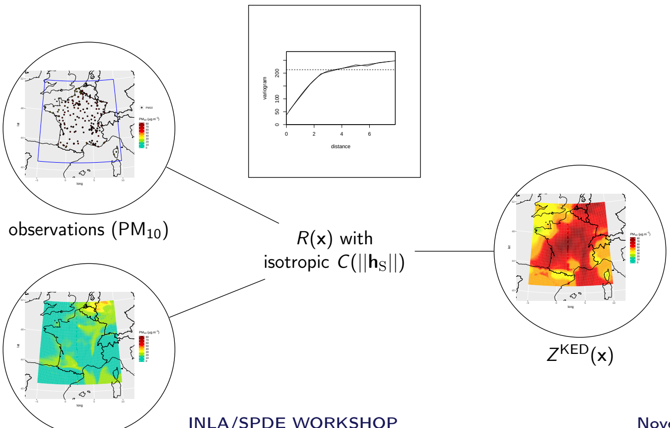
(b) Analysis (Kriging with external drift)

Figure 5 – CHIMERE daily simulation and analysis (11th of March 2014)

Kriging with external drift

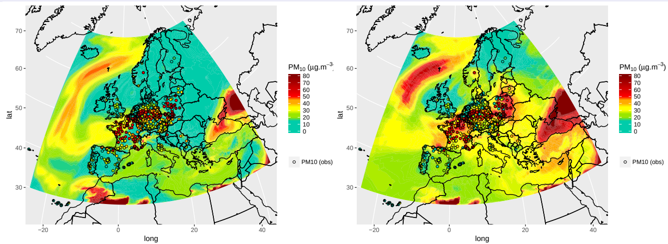
In the kriging with external drift model (Chiles and Delfiner, 2012), the relation between the explanatory variables $\varphi_l(\mathbf{x}_\alpha)$ (the model here) and the observations $Z(\mathbf{x}_\alpha)$ is assumed to be linear :

$$Z(\mathbf{x}) = \sum_l \beta_l \varphi_l(\mathbf{x}) + R(\mathbf{x})$$



II) Forecast (Prediction problem)

Forecast maps of the days D+0, D+1, D+2
1) Meteorology, Emissions and Boundary conditions are used to run a CHIMERE simulation
2) Local forecasting at the background monitoring sites by Multilinear regressions (CITEAIRII project, 2011) or Generalized additive models (Lavancier, 2016; Valsania, 2016)
3) Kriging of **background** concentration measurements with CHIMERE data as external drift



(a) CHIMERE simulation **(b)** Forecast (Kriging with external drift)

Figure 3 – CHIMERE daily simulation and forecast (11th of March 2014)

Advantages and Drawbacks

Advantages

- Methodology implemented and evaluated for several years
- Improvement of the forecasts, especially for D+0

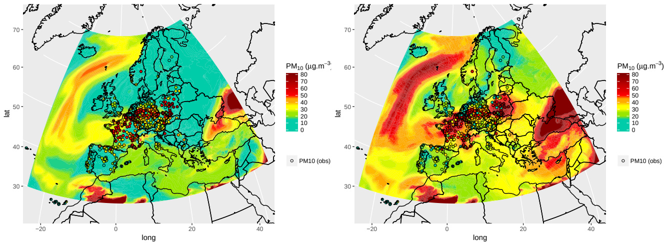
Predictors at the monitoring stations :
1) past observations (D-1 and first hours of D+0)
2) forecast meteorological variables
3) forecast concentrations

Drawbacks

- The statistical models at the stations have to be trained again each time the CHIMERE model is upgraded
- New monitoring stations cannot be introduced in the forecast before one year (in order to have enough data for the training)

Run the CHIMERE and meteorological models is very costly

The prediction problem



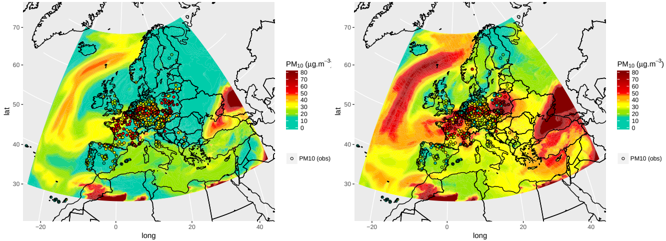
(a) CHIMERE simulation **(b)** Analysis (Kriging with external drift)

Figure 4 – CHIMERE daily simulation and analysis (11$^{\text{th}}$ of March 2014)

The **prediction problem** is usually solved by DA techniques (see e.g. Asch et al., 2016)

In AQ, impact(emissions) > impact(initial conditions) $\rightarrow$ space-time estimation techniques are very competitive (MACC project, 2015)

The prediction problem



(a) CHIMERE simulation **(b)** Analysis (Kriging with external drift)

Figure 4 – CHIMERE daily simulation and analysis (11^{th} of March 2014)

Idea

Consider the analysis (D-1) and the statistical adaptation (D+0, D+1, D+2) as a single product in a spatio-temporal kriging framework

Notations

Let $Z(\mathbf{x}_\alpha, t_k)$, $\alpha = 1, \cdots, N$, $k = 1, \cdots, M-1$ denote the space-time dataset of AQ concentrations observed at the monitoring sites $\mathbf{x}_\alpha$ between time t_1 and t_{M-1}

In the 2016 RESSTE workshop and its related publication (Allard et al., 2017), the kriging of the daily PM_{10} bias of CHIMERE is used to produce the analysis :

$$Z(\mathbf{x}, t) = \mu(\mathbf{x}, t) + R(\mathbf{x}, t) \tag{1}$$

with $\mu(\mathbf{x}, t)$, the local mean of the process is taken as the CHIMERE value and $R(\mathbf{x}, t)$ is the residual, here the bias of the model

Based on this previous work, a large dataset is used to compete the kriging predictions with the PREV'AIR system predictions

Data

- Type : Observations, CHIMERE and meteorological variables
- Pollutants : PM_{10} and O_3
- Time resolution : daily
- Domains : Europe (2014) & France (2013)

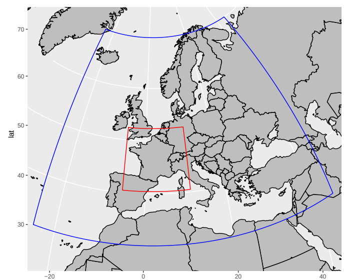


Figure 5 – MACC1e (blue) and FRA4k (red) domains

The intercomparison exercise

Operational context

- PREV'AIR has to provide the forecasts for D+0, D+1 and D+2 at D+0 09 :00
- Direct forecast of the daily mean concentration (**using only D-1 observations**)
- Daily mean concentration calculated as the average of the 24 hourly forecasts (**using D-1 observations & D+0 observations until 06 :00**)

Because big datasets are used, two options are used for the kriging :
(1) Usual **covariance-based kriging** with CHIMERE as external drift (identified better than kriging the bias), with (small) space-time moving neighbourhood

(2) **SPDE-based kriging** to deal with more data when solving the kriging system

Outline

- **Presentation of the methods**
- Covariance-based kriging performance
- Comparison with the statistical adaptation
- Contributions of the SPDE-based kriging

Statistical adaptation (PREV'AIR system)

step (1) : a Generalized Additive Model (**gam**) is built for each monitoring sites $\mathbf{x}_\alpha$:

$$Z(\mathbf{x}_\alpha, t_k) = \beta_0 + \sum_{i=1,\cdots,p} f_i\big(\varphi_i(\mathbf{x}_\alpha, t_k)\big) + \varepsilon \tag{2}$$

where $\varphi_i(.,.),\ i = 1, \cdots, p$ are explanatory variables of the process $Z(.,.)$. The training dataset has to be long, several years if possible

step (2) : the estimation at location $\mathbf{x}_0$ is given by a spatial kriging of the statistical forecasts obtained by these station-specific gam models

Covariance-based kriging

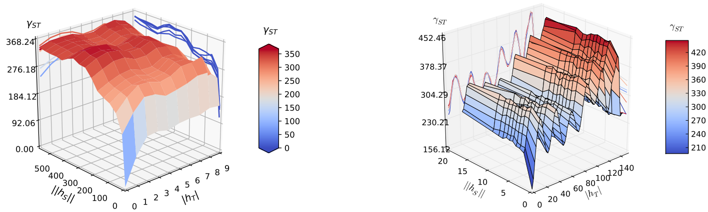
$Z(\mathbf{x}, t)$ is a random function with deterministic part $\mu(\mathbf{x}, t)$ and a residual $R(\mathbf{x}, t)$:

$$Z(\mathbf{x}, t) = \mu(\mathbf{x}, t) + R(\mathbf{x}, t) = \left[\beta_0 + \sum_{i=1}^{p} \beta_i \varphi_i(\mathbf{x}, t)\right] + R(\mathbf{x}, t) \tag{3}$$

with the coefficients β_0 and β_i unknown.
A space-time kriging $Z(\mathbf{x}, t) = \sum_{\alpha,\ k} \lambda_{\alpha,\ k} Z(\mathbf{x}_\alpha, t_k)$ is used for the estimation.
The weights $\lambda_{\alpha,\ k}$ are solution of the linear system (Chiles and Delfiner, 2012) :

$$\begin{cases} \sum_{\alpha=1}^{n} \lambda_\alpha \gamma(\mathbf{x}_\alpha - \mathbf{x}_\beta, t_k - t_l) + \mu_0 + \sum_{i=1}^{p} \mu_i \varphi_i(\mathbf{x}_\beta, t_l) & = \gamma(\mathbf{x}_\beta - \mathbf{x}_0, t_k - t_0) \quad \forall \beta \\ \sum_{\alpha=1}^{n} \lambda_\alpha & = 1 \\ \sum_{\alpha=1}^{n} \lambda_\alpha \varphi_i(\mathbf{x}_\alpha, t_k) & = \varphi_i(\mathbf{x}_0, t_0) \quad \forall i \end{cases} \tag{4}$$

where $\gamma(.,.)$ denotes a space-time authorized variogram model, (see e.g. Gneiting et al., 2007; Porcu et al., 2006; De Iaco et al., 2001)



(a) daily (PM_{10}) **(b)** hourly (O_3)

Figure 6 – Examples of daily and hourly variograms

Advantages

- Space-time moving neighbourhood → local fitting of β_i

Drawbacks

- The neighbourhood has to be small for reasonable inversion CPU time
- Small neighbourhood → using meteorological variables as predictors φ_i is useless (no variability)

SPDE-based kriging I

Model (Cameletti et al., 2012) :

$$Z(\mathbf{x},t) = \underbrace{\beta_0 + \sum \beta_i \varphi_i(\mathbf{x},t)}_{\text{local mean}} + \underbrace{\xi(\mathbf{x},t)}_{\text{latent field}} + \underbrace{\varepsilon(\mathbf{x},t)}_{\text{obs error}} \tag{5}$$

with $\varepsilon(\mathbf{x},t) \sim \mathcal{N}(0,\sigma_\varepsilon^2)$ and the latent field is an AR1 process :

$$\xi(\mathbf{x},t) = a\xi(\mathbf{x},t-1) + \omega(\mathbf{x},t) \tag{6}$$

with $\omega(\mathbf{x},t) \sim \mathcal{N}(0,\sigma_\omega^2 C(h))$, $C(\mathbf{h})$ a Mátern (spatial) covariance.

Coupled SPDE/INLA approach

(1) Rewrite the Model (5) based on the SPDE representation of the Gaussian field
(2) Estimation of the parameters in Model (5) with INLA, see e.g. Opitz (2016)

SPDE-based kriging II

A separable space-time covariance is built by approximating the Gaussian field by its Finite Elements representation :

$$\xi(\mathbf{x}, t) = \sum_k \psi_l(\mathbf{x}, t)\omega_k = \sum_k \psi_i^s(\mathbf{x})\psi_j^t(t)\omega_k \tag{7}$$

where the basis functions are seen as the product of purely spatial basis functions $\psi_i^s(\mathbf{s})$ and purely temporal basis functions $\psi_j^t(t)$, then the space-time stochastic PDE (Lindgren et al., 2011) defined by :

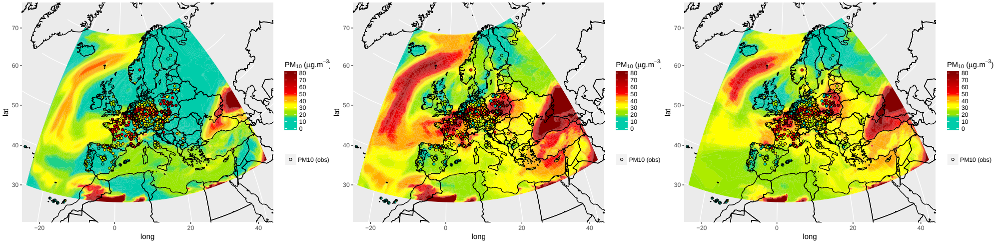
$$\frac{\partial}{\partial t}(\kappa(\mathbf{x})^2 - \Delta)^{\alpha/2}(\tau(\mathbf{x})\xi(\mathbf{x}, t)) = \mathcal{W}(\mathbf{x}, t), \quad (\mathbf{x}, t) \in \mathcal{D} \times \mathbb{R}$$

generates a precision matrix $\mathbf{Q}$ for the Gaussian weights ω_k so that :

$$\mathbf{Q} = \mathbf{Q_T} \otimes \mathbf{Q_S}$$

$\mathbf{Q_S}$ and $\mathbf{Q_T}$ are respectively the precision matrices of the purely spatial model and the Markovian random walk.

Outline



(a) CHIMERE **(b)** Analysis **(c)** Daily-based KED Prediction

Figure 7 – CHIMERE, analysis and daily KED predictions (11th of March 2014)

Direct kriging forecast of the daily mean concentration

- Most of the Western Europe patterns in the analysis are in the forecast...
- But still some strong differences
- Big differences in far-off spatial extrapolations

(I) CHIMERE

$C(\mathbf{x}_\beta, t_0)$, the daily outputs of CHIMERE interpolated at location $(\mathbf{x}_\beta, t_0)$

(II) ANALYSIS (LOOCV)

To estimate $Z(\mathbf{x}_\beta, t_0)$, the dataset is $\{Z(\mathbf{x}_\alpha, t_k)\},\ (\mathbf{x}_\alpha, t_k) \neq (\mathbf{x}_\beta, t_0)$ is used

How the spatial information brought by the neighbours at D+0 helps for the estimation of $Z(\mathbf{x}, t)$ at a location known in the past but not at the current time

(III) FORECAST (LOOCV)

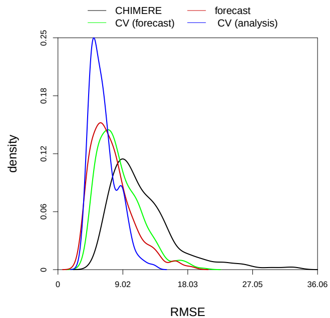
To estimate $Z(\mathbf{x}_\beta, t_0)$, the dataset is $\{Z(\mathbf{x}_\alpha, t_k)\},\ \alpha \neq \beta,\ k \neq 0$ is used : the time series in $(\mathbf{x}_\alpha, t_k)$ is removed

Assess the performance of the prediction without any information in space or time

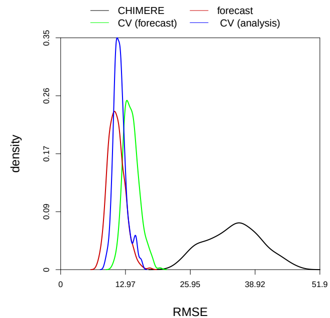
(IV) FORECAST

To estimate $Z(\mathbf{x}_\beta, t_0)$, the dataset is $\{Z(\mathbf{x}_\alpha, t_k)\},\ k \neq 0$ is used

The operational score



(a) Daily data (PM_{10})



(b) Daily data (O_3)

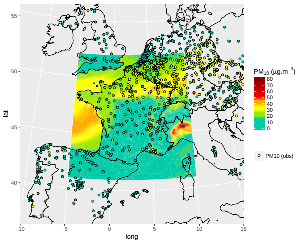
Figure 9 – RMSE

Logical order of performance :

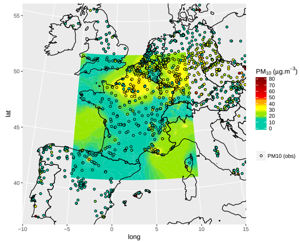
analysis > forecast > forecast in cross-validation > CHIMERE

Outline

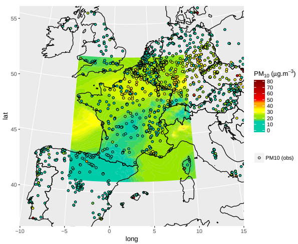
- Presentation of the methods
- Covariance-based kriging performance
- **Comparison with the statistical adaptation**
- Contributions of the SPDE-based kriging



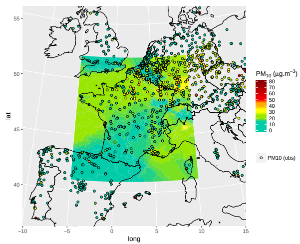
(a) CHIMERE



(b) Analysis



(c) Daily-based SA Prediction

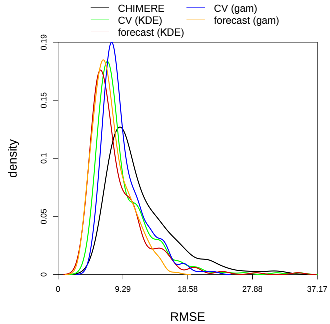


(d) Daily-based KED Prediction

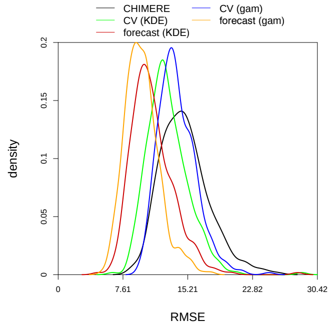
Figure 10 – CHIMERE, analysis and daily-based GAM & KED predictions (11th of March 2013)

Direct Forecast of the daily mean concentration

- Pollution plume over the North of France is better predicted by the statistical adaptation
- Maps very similar (consistent with the scores)



(a) Daily data (PM_{10})



(b) Daily data (O_3)

Figure 13 – RMSE

Forecast :

Performance very similar for PM_{10}
Statistical adaptation better for O_3...

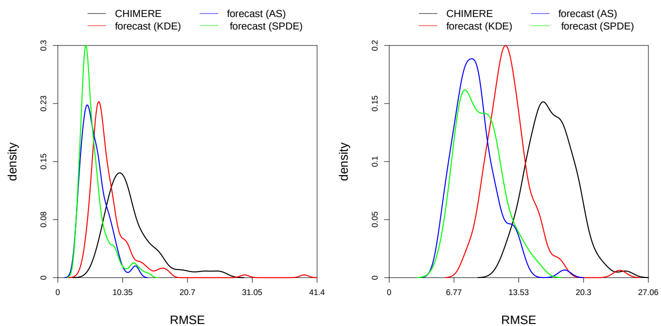
Cross-validation :

Close to the monitoring sites, SA is better
Elsewhere, the kriging is competitive

Outline

Remark

- Forecast scores only
- Reduced dataset (second semester of 2013 for PM_{10}, second quarter of 2013 for O_3)



(a) PM_{10} **(b)** O_3

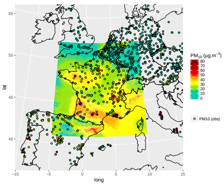
Figure 15 – RMSE

SPDE-based kriging better for PM_{10}
Statistical adaptation better for O_3 but SPDE-based kriging competitive

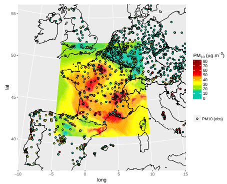
The predictive skills of the SPDE-based kriging approach is an important result

Case study

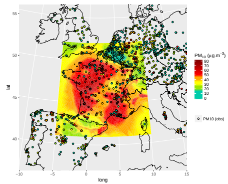
Pollution episode of December 2013, starting from the 9th and ending on the 14th



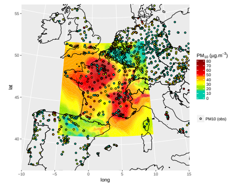
(a) 20131208



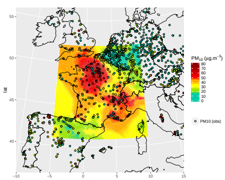
(b) 20131209



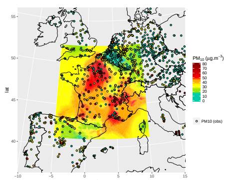
(c) 20131210



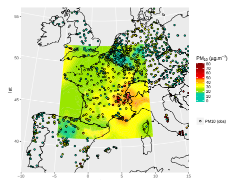
(d) 20131211



(e) 20131212



(f) 20131213



(g) 20131214

Figure 16 – Daily analyses during the pollution episode of December 2013

SPDE-based kriging better for the beginning and the end of pollution episodes

		Normalized Mean Bias		
		KED	GAM	SPDE
Day	20131207	17,26	2,76	-5,51
	20131208	12,46	4,32	-6,88
	20131209	-15,87	-12,98	-10,43
	20131210	-31,07	-14,61	-7,17
	20131211	-1,92	-8,57	-11,63
	20131212	-2,74	-12,65	-11,97
	20131213	-2,21	-8,47	-9,91
	20131214	21,36	9,24	-3,33

Table 1 – Normalized Mean Bias during the pollution episode of 2013

Why ?

Statistical adaptation uses the (D-1) observation as a predictor for the local mean $\mu(\mathbf{x}, t)$

Usual kriging approach : few data, only CHIMERE as covariate $\rightarrow$ poorly estimates the drift

SPDE/INLA approach : more data, more explanatory variables $\rightarrow$ $\mathrm{Var}[R(\mathbf{x}, t)] \searrow$

Conclusion

The main questions were :

1) **How does spatio-temporal kriging compare to the approach used in PREV'AIR to adjust CHIMERE forecasts ?**
Well. In addition, cross-validation results suggest good performance of spatio-temporal kriging in areas with sparse monitoring network.

2) **Does the coupled INLA-SPDE-based kriging approach bring any additional contribution to the performance of the usual covariance-based kriging ?**
Yes. Thanks to a bigger amount of data and explanatory variables, the INLA approach provides a better estimation of the local mean, which is a key point for the prediction of AQ pollution episodes.

Acknowledgements

This study was funded by the French Ministry in charge of Environment

Thank you for your attention

References I

D. Allard, M. Beauchamp, L. Bel, N. Desassis, E. Gabriel, G. Geniaux, L. Malherbe, D. Martinetti, T. Opitz, E. Parent, T. Romary, and N. Saby. Analyzing spatio-temporal data with R : Everything you always wanted to know – but were afraid to ask. *Journal de la Societe Française de Statistique*, page 33 p., 2017. URL `https://hal.archives-ouvertes.fr/hal-01606944`. Soumission.

M. Asch, M. Bocquet, and M. Nodet. *Data assimilation : methods, algorithms, and applications*. Fundamentals of Algorithms. SIAM, 2016. URL `https://hal.inria.fr/hal-01402885`.

M. Cameletti, F. Lindgren, D. Simpson, and H. Rue. Spatio-temporal modeling of particulate matter concentration through the spde approach. *AStA Advances in Statistical Analysis*, 97(2) :109–131, 2012. ISSN 1863-818X. doi : 10.1007/s10182-012-0196-3. URL `http://dx.doi.org/10.1007/s10182-012-0196-3`.

J. Chiles and P. Delfiner. *Geostatistics : modeling spatial uncertainty*. Wiley, New-York, second edition, 2012.

S. De Iaco, D. Myers, and D. Posa. Space-time analysis using a general product-sum model. *Statistics & Probability Letters, v. 52, no. 1*, pages 21–28, 2001.

References II

T. Gneiting, M. Genton, and P. Guttorp. *Geostatistical Space-Time Models, Stationarity, Separability, and Full Symmetry*, pages 151–175. C&H/CRC Monographs on Statistics & Applied Probability. Chapman and Hall/CRC, Oct 2007. ISBN 978-1-58488-593-1. doi : 10.1201/9781420011050.ch4. URL `http://dx.doi.org/10.1201/9781420011050.ch4`. 0.

F. Lavancier. Prévision journalière de l'Ozone et des PM_{10} á l'échelle nationale, 2016.

F. Lindgren, H. Rue, and J. Lindström. An explicit link between gaussian fields and gaussian markov random fields : the stochastic partial differential equation approach. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 73(4) :423–498, 2011. ISSN 1467-9868. doi : 10.1111/j.1467-9868.2011.00777.x. URL `http://dx.doi.org/10.1111/j.1467-9868.2011.00777.x`.

T. Opitz. Latent Gaussian modeling and INLA : a review with focus on space-time applications. soumis au Journal de la Société Francaise de Statistiques, 2016.

E. Porcu, P. Gregori, and J. Mateu. Nonseparable stationary anisotropic space–time covariance functions. *Stochastic Environmental Research and Risk Assessment*, 21(2) :113–122, 2006. ISSN 1436-3259. doi : 10.1007/s00477-006-0048-3. URL `http://dx.doi.org/10.1007/s00477-006-0048-3`.

M. Valsania. Generalized Additive Model per la previsione giornaliera di inquinamento da Ozono e da PM_{10} in Francia. Masters thesis, Universitá degli Studi di TorinoDipartimento di Economia e Statistica, November 2016.