Poll Res. 39 (3) : 577-592 (2020)

ISSN 0257–8050

SPATIAL AND TEMPORAL VARIABILTY IMPACT ON AIR POLLUTION INTERPOLATION: A CASE STUDY ON PM10 ESTIMATION IN NORTHEN-FRANCE

KHAOULA KARROUM* [1,2], ANTON SOKOLOV [2], YANN BEN MAISSA[3], MOHAMED EL HAZITI[1] AND HERVÉ DELBARRE[2]

[1] LRIT-CNRST, URAC 29, Faculty of Sciences Mohammed V University in Rabat, Rabat, Morocco
[2] Laboratoire de Physico-Chimie de l'Atmosphère, Université du Littoral Côte d'Opale, 59140 Dunkerque, France
[3]Telecommunication Systems, Networks and Services Lab National Institute of Posts and Telecommunications, Rabat Morocco

(Received 15 Janaury, 2020; accepted 24 March, 2020)

ABSTRACT

In this work, we compare the performance of a set of spatial interpolation techniques in estimating PM_{10} concentrations for Hauts-de-France region. Estimations of Coefficient of Determination, Root Mean Square Error and corresponding 95% confidence intervals show that classic and optimized version of Inverse Distance Weighting method and Gaussian Process Regression with two different kernels give comparable results. The spatial distribution of the error shows the high dependence on industry and coastal atmospheric phenomena. To assess the influence of the local meteorological effect on pollution dispersion, we estimated the Coefficient of Determination for the interpolation of time-averaged pollution data. It has a clear 24-hour maximum, corresponding to periodic atmospheric effects, such as the sea breeze. The sensitivity of interpolation techniques to the noise in measurements and to the data density shows that methods behave in the same way, leading to a bigger Root Mean Square Error following the magnitude of data perturbation and decreasing with the number of stations. In addition to that, regions with higher error are less sensitive to the perturbation.

KEY WORDS : Air pollution, PM_{10}, Interpolation, Local meteorology, Gaussian process regression, Inverse distance weighting

INTRODUCTION

Air pollution has a serious impact on the planet (e.g. it changes the albedo influencing the climate, Ramanathan and Feng 2008) and on human health (Künzli *et al.,* 2005; Kampa and Castanas, 2008 and Chen *et al.,* 2016), and studying this complex topic requires considering sources of emissions, meteorology, and chemistry. In order to control the air contamination level, the networks of stations are deployed measuring the pollution at certain sites. At unmonitored areas different interpolation techniques could be applied to estimate the air quality.

The objective of estimating air pollution differs from a work to another: it could be to analyze its effect on health (Son and Lee, 2010), to compare the rate of contribution of certain sources to this pollution (Hudda and Fruin, 2016) in an area, or to study the air pollution behavior according to meteorology or physicochemical phenomena. In this work, we focused on using the spatial interpolation techniques to study the Particulate Matter (PM_{10}) at a regional scale. PM_{10} is a mixture of solid and liquid species with a diameter of less than 10µm, suspended in the atmosphere of anthropogenic and natural sources.

Spatial interpolation is estimating a studied variable at unmonitored sites based on a limited set of known observations within the same area. Spatial

interpolation could be done either by modelling or by statistical techniques. The modelling approach requires inputs as meteorological, geographical and traffic information of the studied area, to describe the physical and chemical processes that are perceived to be the most important for defining the air pollution and the atmospheric dispersion and transformation of pollutants in the atmosphere. For instance, the work of Li, Lianfa *et al*. (2013), they use roads and traffic information to model on-road air pollution in southern California. Besides, statistical based spatial interpolation techniques exploit the statistical link between the measured data to estimate air pollution at unsampled locations, which is an alternative solution to the modelling approach (Deligiorgi and Philippopoulos, 2011).

Many statistical based spatial interpolation methods allow estimating the value of interest at unmonitored locations (Li and Heap 2014), among which the most frequent used ones (Li and Heap 2011) are Inverse Distance Weighting (IDW) and Gaussian Process Regression (GPR) (known as Kriging). These two techniques are still used as means to assess and analyze the pollutants to date (Ehrampoush, Mohammad Hassan, *et al.*, 2017; Zhang, *et al.*, 2018 and Qiao, Pengwei, *et al.*, 2019). But it is always difficult to select the convenient method for a specific dataset since many factors contribute in the spatial and temporal distribution of air pollution. Because the performance of each interpolator differs depending on the study area and also the measured pollutants, there is no way to expect the promising interpolator in a case other than by trying different interpolation techniques. Deligiorgi and Philippopoulos (2011) found that artificial neural network outperformed all the applied set of spatial interpolation techniques for NO_2 and O_3 except in one station. The interpolation methods behave differently with various monitoring density and distribution. Wong *et al.* (2004) tried four interpolation techniques on USA air pollution on five regions of the country: Northwest, Southern California, Southeast, Industrial Midwest, and Northeast. The authors noticed in the areas with high monitoring density, some methods tend to perform better than in low monitoring density areas. Kriging in California region succeeded in fitting a variogram to the air-monitoring data with the greatest correlation, in dense and regular monitoring stations' areas. The triangulation based methods are always used in spatial interpolation comparative studies like nearest neighbor technique (Wong *et al.*, 2004; Deligiorgi and Philippopoulos, 2011; Lee *et al.*, 2014), which attributes to the unmonitored location the nearest station measurements value (as in Miller *et al.*, 2007). In the current work we selected triangulation methods plus IDW and GPR to study the spatial distribution of PM_{10}.

In this paper, we intend to take into account the temporal behavior of air pollution to examine its influence on interpolation techniques. We compare the performance of interpolation methods in estimating PM_{10} concentrations, then we try to analyze this performance through the temporal variability by working with different time-averaged pollution measurements (e.g. 1-hour, 1-day, etc.,) rather than 15 minutes as provided in the observed data. Lastly, we verify the sensitivity of the interpolation techniques to data density and perturbation. To the best of our knowledge, this is the first study that works on Hauts-de-France region's PM_{10} concentrations estimation by coupling spatial interpolation with temporal variability.

The outline of this paper is as follows. Section 2 presents the study area and brief definitions of interpolation methods of nearest neighbor, linear interpolation, natural neighbor, spline, inverse distance weighting (IDW), IDW with optimized distance power, and Gaussian Process Regression with two different kernels. Section 3 displays and discusses the results of: the applied methods, the time-averaged pollution data and sensitivity of these methods to perturbation and data density. Section 4 provides the conclusions and perspectives of this paper.

Data and methods

The first part of this section is devoted to present the area of study and analyze some stations' behavior compared to the regional mean of PM_{10} concentrations, to be able to make some inferences about the spatial interpolation performance in the different parts of the whole area of interest (in section3). The second part is dedicated to the used interpolation methods and parameter of validation's definitions.

Study area and data

The study area is Hauts-de-France region (northern-France) which has approximately 6 million inhabitants and a population density of 180 inhabitants / km^2 [1], on 1st January 2016. It is the 3rd most populous region in France, and the 2nd most

densely populated region in metropolitan France, after Ile-de-France. Covering an area of 32,000 km², that represents 5.7% of the surface area of metropolitan France, the Hauts-de-France region is bordered on the North by the North Sea for a distance of 45 km, and on the West, by the Channel for a distance of 120 km (figure 1). The region is subject to a temperate, oceanic climate with cool, wet winters and mild summers. It is home to many industrial and agricultural activities, fishing ports and passenger transport, and significant road and sea traffic. Indeed, it is located in the center of northern Europe and the Paris-Brussels-London triangle (Gengembre, 2018).

Figure 1 illustrates the positions of the 37 measuring stations scattered over Hauts-de-France region. The northern part of the monitoring stations network is more dense in industrial and urban areas compared to the southern rural area. Some are in proximity to industrial activities in a part of Dunkerque city (e.g. DKI, DKC and DKG stations), others are next to a fishing port (BO stations). There arestations located in urban areas with a high traffic and population density as in Lille (as MN1 and MC7 stations) and Valenciennes (VA stations) cities, and stations that are along the Opal coast (CA and BO stations), etc. [1]. In a later step, we will see how these emissions information affects the air pollution behavior. The input data are quarter-hourly measurements of PM_{10} concentrations from the 1st of January 2016 to the 31st of December 2016, and they are provided by AtmoHauts-de-France [2]. The white frame in figure 1 corresponds to our region of interest, where we applied the spatial interpolation and represented it in the different maps later in this article.

[1]: https://hautsdefrance.cci.fr/wp-content/uploads/sites/6/2016/04/Atlas_NPCP_Edition2016.pdf

[2]: https://www.atmo-hdf.fr

The preprocessing of data consisted on first removing all the outliers (values that are bigger than the possible actual values of the PM_{10} concentration), then deleting all negative values and the invalid measurements represented by NAN (Not A Number) from the input database. In most cases, these outliers and negative values are a result

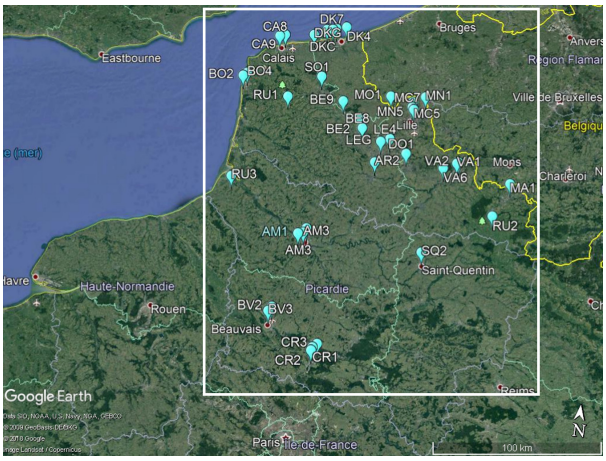

Fig. 1. The positions of measured PM_{10} by 37 stations in Hauts-de-France region

of a malfunction in the measuring sensor.

We first analyzed the PM_{10} behavior of some stations by comparing them to the regional mean (Table 1 and Figure 2 to 5), to study their variability at regional scale. We chose 12 stations among 37 to highlight the behavior of stations which we classified into four types corresponding to coastal, urban, rural and industrial exposure. For each category, we selected some stations and plotted their measurements along with the regional mean of one month (August) of the year 2016. The highest variability of PM_{10} takes place in summer where some specific meteorological phenomena like sea breeze occur (Miller *et al.*, 2003).

Table 1 displays the variance of some stations compared to the regional mean of PM_{10} concentrations.

In Table 1, the highest variance corresponds to DKG and DKC stations, which are exposed to the industrial activities of Dunkerque region and influenced by coastal meteorological phenomena, followed by MC7 station located in urban area. Right after, comes DK4 station which is a bit further from the industrial zone compared to DKC and DKG, but could be exposed to this source of emission. Next, comes VA1 and MN1 located in Valenciennes and Roubaix cities (which has a high population and human activities) and last, we have the rural and coastal stations (far from the industrial zone) with similar values and which have the

Table 1. Stations variance to the regional mean of PM_{10} concentrations

Station	RU1	RU2	RU3	CA9	CA8	BO4	MC7	VA1	MN1	DK4	DKC	DKG
$\mu g/m^3$	1.97	1.54	1.57	1.20	1.33	1.4	3.52	2.20	2.15	3.30	8.76	5.96

smallest values among the displayed stations. The upcoming figures (Figure 2 to 5) confirm these values (Table 1). In Figure 2, rural stations RU1 and RU3 follow quite well the regional average with some few peaks, while RU2 has stronger oscillations around this mean, probably due to the local meteorological phenomena with weak winds and temperature inversion (a hilly area). We plotted the same data for coastal stations in Figure 3, where we observe that three stations have a behavior that follows the regional pretty well more than the rural ones do. Knowing that these are coastal stations (land-sea area) where meteorological phenomena (like sea breeze) may take place and may cause a strong variation of PM_{10} concentrations around this regional mean, as in the period from 15 to 18 August in Figure 4, where CA8 and BO2 concentrations become much higher than the regional mean.

Figure 4 represents the PM_{10} concentrations of urban stations compared to the regional mean. The oscillations are stronger than in rural and coastal stations (confirming Table 1 results). MC7 and VA1 stations have similar behavior that follows the regional mean in most of August, while MN1 has a stronger variation around this mean during the whole month. It may be because MN1 is more exposed to a source of urban air pollution (traffic, combustion, heating, etc) than the two other stations MC7 and VA1). For stations exposed to industrial activities (Figure 5), we notice that DK and DKC stations' PM_{10} concentrations have the strongest oscillations around the regional mean among all the plotted, whereas DK4 tends more to a behavior that is stronger than the coastal stations and weaker than the industrial ones. The reasons behind these fluctuations are: first the meteorological phenomena (since they are coastal stations also), and second the industrial activities in this area (coastal part of

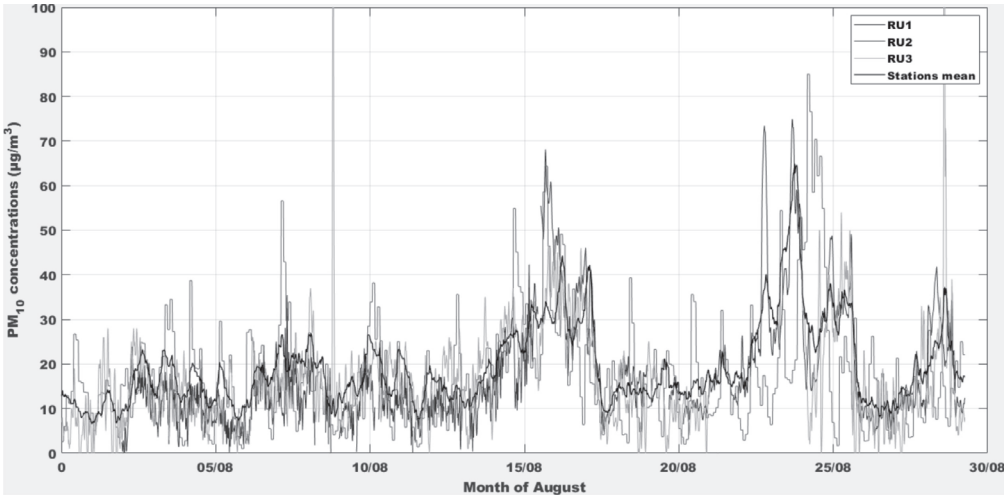

Fig. 2. PM_{10} concentrations of rural stations and the regional mean (August 2016)

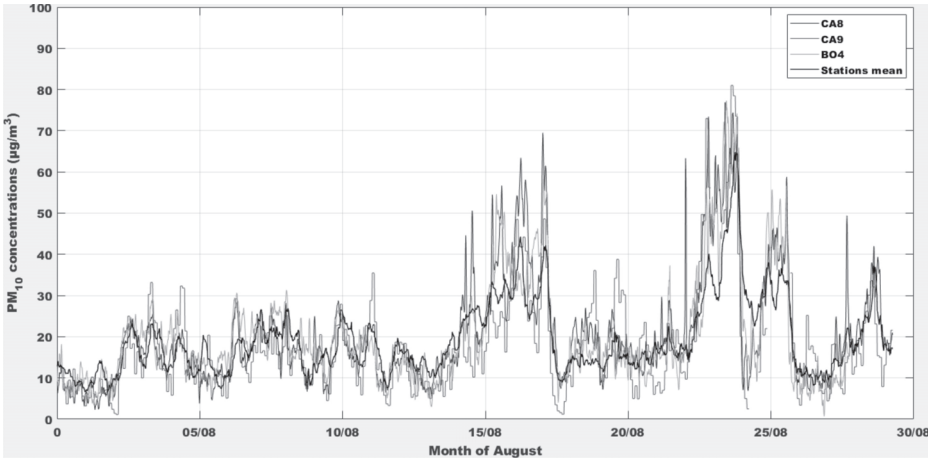

Fig. 3. PM_{10} concentrations of coastal stations along with the regional mean (August 2016)

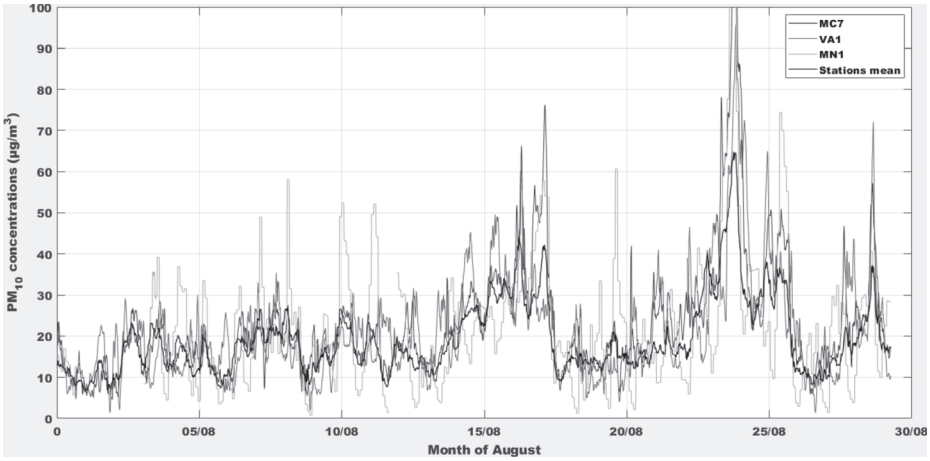

Fig. 4. PM_{10} concentrations of urban stations and the regional mean (August 2016)

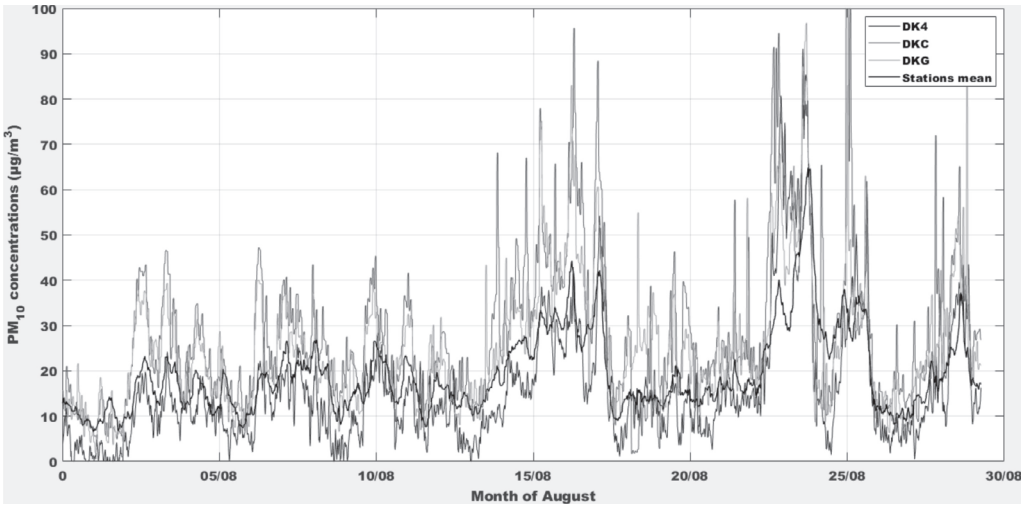

Fig. 5. PM_{10} concentrations of stations near industrial zone in Dunkerque and the regional mean (August 2016)

Dunkerque city).

Following the observations of the four figures above (figure 2-3-4-5) and Table 1, we infer that each station that behaves in a different way than the regional mean of PM_{10} concentrations in Hauts-de-France region, is exposed to local anthropogenic emissions and specific meteorological phenomena. PM_{10} peaks are expected to have a significant effect on interpolation, and this would be discussed in section 3.

Methods

We chose to use two groups of interpolation techniques in this work; triangulation-based methods, and IDW and GPR methods. In this section, we briefly present each method, plus the adopted procedure and parameters to assess the methods' accuracy.

Triangulation based methods

The four first methods we utilized are based on the triangulation technique. Delaunay triangulation consists on dividing a set P of points in a plane into triangles, in a way that no point in P is inside the circumcircle of any of the triangles of this triangulation (Lee and Schachter, 1980). The Voronoi diagram (called also tessellation) is the dual of Delaunay triangulation, and it is formed by perpendicularly bisecting the edges of the triangulation.

Nearest neighbor

It is the simplest technique in spatial interpolation methods, consisting on area partition to tessellation based on Voronoi diagram. Nearest neighbor selects the closest point to the site to estimate, without taking into consideration the other neighbors.

Nearest neighbor compares the distances between a point and its neighboring to choose by the end the value of the nearest one to be the query point's estimated value.

Linear interpolation

Given a set of points, linear interpolation method builds the Delaunay triangulation network by the three adjacent known points, so by the end we get a surface of triangular planes where no triangle edges are intersected by other triangles. The result is a patchwork of triangular faces over the extent of the grid. Each triangle defines a plane over the grid nodes lying within the triangle, with the tilt and elevation of the triangle determined by the three original data points defining the triangle. All grid nodes within a given triangle are defined by the triangular surface.

Natural neighbor

Natural neighbor is based on Voronoi diagram and consists on using the information of neighboring samples to estimate the query point. R. Sibson (Sibson, 1981) explained how to select and assign to each of the selected neighbors a weight corresponding to its influence. A Voronoi diagram of the input sampled data is constructed, then we add the query point to estimate. The algorithm selects the closest subset of input samples to this query point and assigns weights to them based on proportionate areas to interpolate a value. A detailed explanation is given in the work of Sibson (1981).

Spline

Spline interpolation in two dimensions is also based on Delaunay triangulation; it consists on creating a smooth surface that passes by the data points while minimizing the curvature as much as possible by a set of mathematical functions that passes by the input points (Longley, 2010). A commonly used example of spline is cubic splines, which consist of polynomials of third degree.

Inverse Distance Weighting (IDW)

Known for being one of the fastest interpolation techniques, Inverse Distance Weighting (IDW) method considers that it is possible to estimate the value y at an unmonitored site by the mean of the distance weighted average of the surrounding monitoring sites (Watson and Philip, 1985). IDW admits that the points closer to the site to estimate have a stronger influence than the farther ones, mathematically expressed by equations; that defines the weight of each of the sampled points in (2), and use this latter to estimate (3):

$$W_i(x) = \frac{1}{d\,(x,x_i)p} \quad .. (2)$$

$$y(x) = \frac{\Sigma_{i=0}^{N} w_i(x) y_i}{\Sigma_{i=0}^{N} w_i(x)} \quad .. (3)$$

where y is a function to estimate, y_i monitored (known)values of the function and their known locations x_i, d is the distance, N is the number of the influencing sites. and exponent p is the power ordistance exponent value.

p is the distance exponent value, which could be optimized to get a better estimation accuracy. As p becomes bigger, the assigned weight becomes smaller for the distant points, in other words it emphasizes the influence of the closest points compared to the distant ones. Thus, the IDW method with the infinite p corresponds to nearest neighbor interpolation. In the case of p tends to zero, all weights are equal to one in the expression (2), the denominator in (3) becomes the number of measurements, and the estimation (3) by IDW is an averaging of .

In this work, the p optimization consisted on finding the p leading to the smaller RMSE possible, by trying different interval values.Our numerical experiments showed that for our dataset the reasonable values of p are inside of the [0, 5] interval. This optimized by p version of IDW will be further abbreviated in this paper by IDW optimized.

Gaussian Process Regression (GPR)

We give a brief definition to Gaussian Process Regression (GPR), Rasmussen and Williams give a detailed explanation of GPR in (Ramussen and Williams 2006).

Let us consider a group of n predictors (inputs) $X = \{x_o,\ldots,x_n\}$, that we would like to learn its relationship with an output variable $Y = \{y_o,\ldots, y_n\}$, by a GPR model expressed by a regression function as follows :

$$y = f(x) + \varepsilon_i \quad .. (4)$$

with Gaussian noise. Assuming that inputs are well scaled and centered, and that the regression function has a Gaussian prior distribution with mean equal to zero, as follows:

$$y = [f(x_1), f(x_1), \ldots f(x_n)] \sim GP\,(0, K\,(x_i, x_j)), \quad ..(5)$$

where K is a covariance matrix (known also as kernel function) with x size, and GP means a Gaussian Process.

There exist a multitude of covariance functions, in this work we selected two kernels. The first one is the commonly used covariance function, which is the squared exponential that is written as:

$$K(x_i,x_j|\theta)= \sigma_f^2 exp\left[-\frac{1}{2}\frac{(x_i-x_j)^2}{\sigma_l^2}\right], \quad .. (6)$$

where $(c_i - x_j)$ is the distance between two input x_i and x_j, θ is the set of hyperparameters σ_f and σ_l; σ_f is the signal variance and is the lengthscale. We will use the abbreviation of GPR sqrexpo for GPR with this kernel in the remainder of this paper.

The second kernel is a squared exponential kernel with a separate length scale per predictor.

$$K(x_i,x_j|\theta)= \sigma_f^2 exp\left[-\frac{1}{2}\sum_{m=1}^{d}\frac{(x_{im}-x_{jm})^2}{\sigma_m^2}\right] \quad .. (7)$$

where σ_m, m∈[1,..., d] is the lengthscale that would be defined for each of the predictors with d dimension, unlike the previous kernel where we find one corresponding length scale for the set of predictors. We abbreviate GPR with this kernel by GPR ardexpo.

Accuracy assessment procedure and parameters

In order to compare the accuracy of the estimated measurements using different spatial interpolation methods, we utilized the following tools:

- *Leave One Out Cross Validation (LOOCV)*

To assess the performance of the applied techniques of spatial interpolation, we applied LOOCV to calculate RMSE and R^2. As the number of available data is limited, this technique allows generalizing estimations to a quasi-independent data set. The leave-one-out cross-validation is a special case of the cross-validation technique (Gong, 1986).

LOOCV consists on removing one of the observed data values from the set of input points, and try to estimate it by a method, then compute the difference between the estimated and the predicted values (error of estimation). Then, reproduce this process for each measurement in the set. We applied LOOCV to evaluate two statistical parameters: the Root Mean Square Error (RMSE, Willmott, 1982) and the coefficient of determination R^2 (Willmott and Matsuura, 2005). The former allows assessing the model precision, and the latter measures the performance by the percentage of explained variance of data.

We calculated RMSE for each station in the network (RMSE station), and for the whole data of all the stations (RMSE total), as follows:

$$\text{RMSE (station j)} = \sqrt{\frac{1}{n_j}\sum_{i=1}^{n}(X_{obs,i,j}-X_{estim,i,j})^2} \quad .. (8)$$

where $X_{obs,i,j}$ and $X_{estim,\,i,\,j}$ are the measured and estimated values of PM_{10} concentrations respectively at the station j at the time i, and n_j is the number of measurements available in j.

$$\text{RMSE (total)} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_j}\sum_{j=1}^{m}(X_{obs,ij}-X_{estim,ij})^2} \quad .. (9)$$

where m is the number of stations.

The coefficient of determination R^2 is defined as follows:

$$R^2 = 1-\frac{\sum_{i=1}^{n}(X_{obs,i}-X_{estim,i})^2}{\sum_{i=1}^{n}(X_{obs,i}-\overline{X_l})^2} \quad .. (10)$$

R^2 parameter shows the proportion of variance in the output variable, which could be explained by an estimator.

- *Bootstrap technique*

To assess the confidence intervals (Briggs *et al.*, 1997) of statistical estimates we obtained, we used bootstrap resampling method.

The bootstrap technique (Rao *et al.*, 1985) consists in resampling the original data we use to assess the performance of an algorithm, to compute the uncertainty of the parameter of validation (in our case it is RMSE). This technique selects repeatedly and with replacement random samples of the same size from the original dataset. So, each point of the original database could be selected zero, one or more times in every bootstrapped sample. Then, we calculate the evaluation parameter value for every bootstrapped sample for a number of times, we rank order these values and take the 95% confidence interval (CI) (Briggs *et al.*, 1997). Efron (1982) and Rao *et al.* (1985) give further explanation of bootstrapping.

RESULTS AND DISCUSSION

Before presenting the results, we plot the PM_{10} level in the studied area with giving some interpretations. The method we select to examine the time-averaged data effect on the interpolation methods' performance (section 3.2), and to generate all the maps is Inverse Distance Weighting (IDW), because

it has small RMSE and high R^2 (Table 1). In addition, IDW is fast and easy to compute, and it is physically more accurate than the triangulation techniques (Figure 14 in appendix). In all the maps of this article, the bold black line connecting France and Belgium represents the coast that extends from France to Belgium, while the bold red line is the borders between the two countries.

Figure 6 displays the spatial distribution of the annual mean of PM_{10} concentrations in 2016 in Hauts-de-France region, to get an overall idea about the air pollution level in the area of interest. We notice in this figure that the isolated stations are described by a concentric area of the same value or what it is named "the bull's-eye effect". This is a limitation of IDW algorithm, when the data points are sparse with a spatial distribution that has isolated sites. These factors lead to a decrease in R^2 in some methods and an overlearning problem in other as discussed later in this paper. As mentioned before, some of the stations are located next to air pollution emission sources, as the case where we notice a high PM_{10} concentration in some of Dunkerque's stations (DKC, DKI, and DKG) due to the industrial activity in this part of the city, for Boulogne there is a fishing port , Lille and Roubaix area is known as an urban zone with high population density with heavy road traffic, same for Valenciennes while Beauvais has an international airport that may be the main source of the air pollution there.

Moreover, we plotted the spatial distribution of annual standard deviation of PM_{10} concentrations (Figure 7), to see how the PM_{10} vary all over the region and be able to interpret the results we will get by the spatial interpolation techniques. This map shows high variations of PM_{10} in the industrial part of Dunkerque city (DKG, DKC, and DK7), in the urban stations in Lille and Roubaix cities (MN and MC5 stations). A high variation of PM_{10} concentrations is also noticed in Creil (CR stations), which we suppose is due to the air pollution coming from Paris given the proximity of Creil city to the agglomeration of Paris.

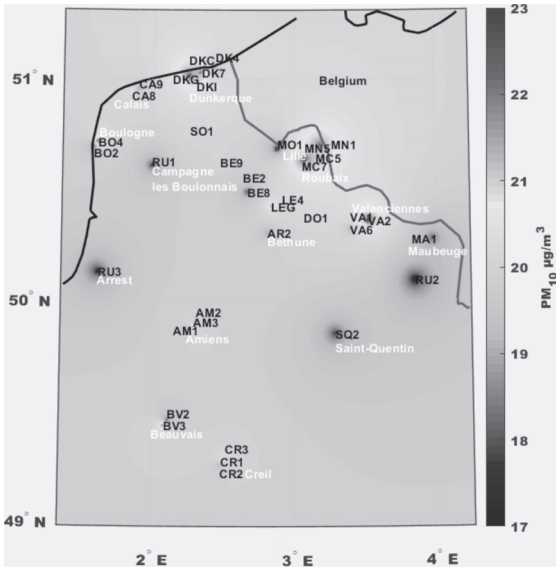

Fig. 6. 2016 annual mean of PM_{10} concentrations for Hauts-de-France region

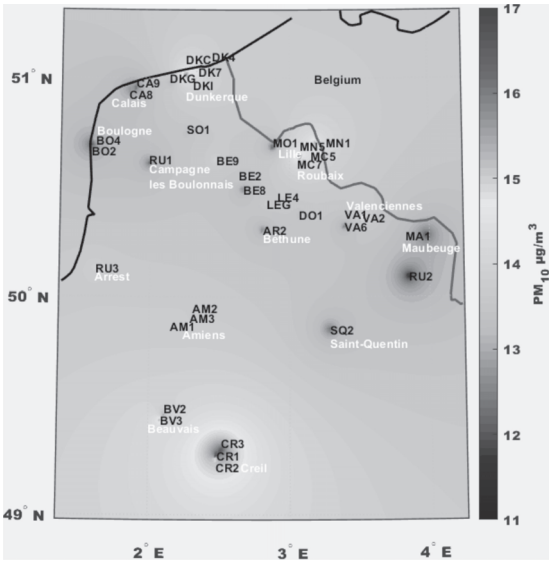

Fig. 7. 2016 annual standard deviation of PM_{10} concentrations for Hauts-de-France region

In what follows, we present the results of the interpolation methods, the time-averaged data, and last the sensitivity of methods to perturbation and data density.

Interpolation results

For the spatial interpolation methods, we wanted to see the efficiency of different kind of interpolation techniques in estimating air pollution, that is why we selected a set of methods which are different in terms of algorithm's complexity (methods figuring in Table 2).

After we remove the outliers represented by negative and extreme PM_{10} concentrations (which is in most cases because of a malfunction in the sensor measuring the concentrations) from the input database, we applied our set of interpolation techniques.

Results in table 1 show that IDW optimized

technique gives the smaller RMSE with 7.45 µg/m^3 and the highest R² with 70%, while the 95% confidence interval demonstrates that IDW, IDW optimized, GPR sqrexpo and GPR ardsqrexpo methods achieve comparable RMSE varying between 7.33 to 8.08 µg/m^3. All the triangulation-based methods (nearest neighbor, linear interpolation, natural neighbor and spline) resulted a 95% confidence interval RMSE of 8.59 to 9.77 µg/m^3. The difference in results between all these techniques is still slight. The small number of the measuring sites we have in the database plus the sensors positions overall the area of interest, makes it difficult to distinguish one interpolation method that would suit our case study region and be the only best interpolator. This proves the data density and geographical distribution impact on the interpolation methods.

To analyze the applied spatial interpolation behavior in the different areas of the studied region, we estimate the R² of each station by IDW, then interpolate these values by this same interpolation technique (Fig. 8). We observe that R² varies between 30% to more than 80%. The areas that are next to the industrial zone and exposed to local meteorological phenomena (as in Dunkerque city) have a high air pollution values and variance, and the lack of measuring stations in these same areas leads to high values in RMSE in spatial interpolation. For example, the spots where we have a small R² (like DK4 and DKC), are located next to an emission source of pollutants (industries), what produces a smaller R². Moreover, the lack of sensors in RU2 area may be the cause of having a smaller R², plus its geographical location that makes the air pollution coming from the neighboring area circulates in RU2 surroundings by dint of the mountains (mentioned in Section 2.1). For the other

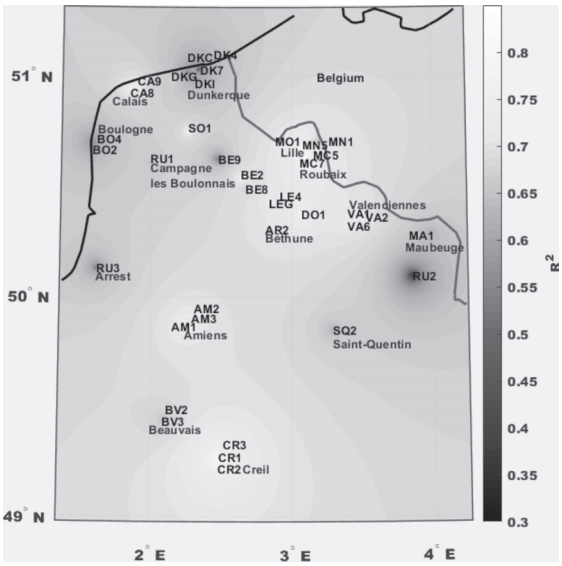

Fig. 8 R² of IDW interpolation of PM_{10} concentrations

station's interpretation of R², we still lack information about the overall region to be able to deduce the interpolation behavior there.

In figure 9 we calculated the R² of PM_{10} concentrations by linear interpolation technique in each station, then we interpolate these R² values by IDW. We notice that there is a big decrease in R² in almost all the map compared to the map of figure 8, with always having the smallest values of R² in the same areas that are the smallest in figure 8. This is due to the proximity to emission sources as industry or the exposition to the local meteorological phenomena's influence as wind or turbulence (as mentioned in section 2.1), which produces a high variation in the air pollution variance and consequently leads to a higher error in interpolation performance. We expected such a result since the linear interpolation is a simpler technique based on triangulation principle, which makes it having a

Table 2. Interpolation methods results

Interpolation method	RMSE µg/m^3	RMSE CI µg/m^3	R²
Nearest Neighbor	9.55	9.41-9.71	0.52
Linear interpolation*	8.84	8.70-8.99	0.58
Natural neighbor*	8.73	8.59-8.88	0.59
Spline*	9.61	9.44-9.77	0.51
IDW	7.74	7.63-7.84	0.68
IDW optimized	7.45	7.33-7.53	0.70
GPR sqrexpo	7.75	7.46-8.02	0.68
GPR ardsqrexpo	7.81	7.5-8.08	0.67

*: In addition to the used method for interpolation, we used nearest neighbor for extrapolation since it is the most stable technique among the set we have, to extrapolate.

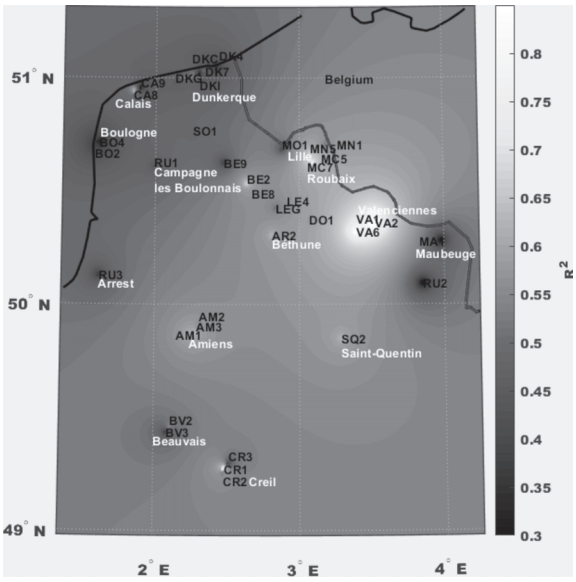

Fig. 9 R^2 of linear interpolation of PM_{10} concentrations

smaller R^2 than IDW, and even a smaller R^2 in areas next to emission sources of air pollution.

Time-averaged data results

The coastline area of Hauts-de-France region is exposed to effects of local meteorological phenomena leading to strong variation in PM_{10} concentrations. To show that high values of RMSE of interpolation are connected with local meteorology, short-term fluctuations of PM_{10} were smoothed. After working on enhancing the spatial interpolation of PM_{10} concentrations in Hauts-de-France region by optimizing parameters of the interpolation techniques using 15-minute data, we created a few datasets by calculating a simple moving average with a few time spans up to 3 months. Then, the previous interpolation methods were applied to estimate RMSE and R^2.

The results displayed in Table 3, show that R^2 calculated for IDW interpolation reaches the maximum value for the dataset corresponding to the time averaging period of 1-day. The 95% confidence intervals show that this 1-day R^2 value is significantly higher than R^2 for other time spans. This result demonstrates that the precision of spatial interpolation is sensitive to local meteorological effects and can be degraded by phenomena like the sea breeze. Thus, uncorrelated fluctuations observed in Figure 3 of the coastal stations around the mean could cause a degradation in the precision of spatial interpolation.

To see the influence of varying the time-averaging periods on the different stations of our network, we chose to illustrate the R^2 of IDW using data time-averaged data by 1-day period. As shown in Figure 10, we observe an increase of R^2 in all the region with respect to Figure 8. While we notice lower values of R^2 in the stations which are exposed to some air pollution emission sources like DK stations and RU2 (we kept the same scale of R^2, to be able to compare with the map in Figure 8). Therefore, the filtering of the local meteorological phenomena's effects resulted in high R^2 value in PM_{10} estimation, but with a different spatial distribution of these R^2 values over the studied region. Where we remark the lowest value of R^2 in RU2 area, this could be due to the exposition to an air pollution source of emission, plus having just one measuring station in this area. We actually went further in the mapping of other time averaging periods which we analyzed in the appendix of this paper (Figures from 15 to 18 in appendix).

Table 3. IDW interpolation results for different time averaging periods

Time averaging periods	R^2	Confidence interval
15 minutes	0.68	0.66-0.69
1 hour	0.69	0.68-0.7
3 hours	0.72	0.71-0.73
6 hours	0.74	0.73-0.75
1 day	0.80	0.79-0.81
1 week	0.77	0.76-0.78
2 weeks	0.72	0.71-0.73
1 month	0.62	0.61-0.63
3months	0.45	0.44-0.46

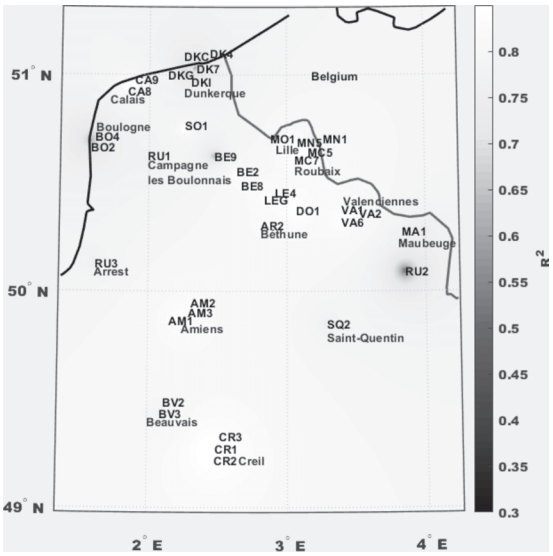

Fig. 10. R^2 of IDW for averaged data of 1-day period of PM_{10} concentrations

The sensitivity of methods to perturbation and input data density

In this part, we tried to examine the sensitivity of spatial interpolation methods to perturbation and analyze the role that data density plays in spatial interpolation.

Sensitivity to perturbations

The aim of perturbing PM_{10} concentrations measurements we have, is to see how the interpolation methods perform in real perturbation cases like when we have a malfunction in the monitored stations (which will be represented by a perturbation included in the real measured air pollution concentrations), and how this malfunction would affect the estimation (Figure 11). To verify the sensitivity of the applied interpolation techniques, we perturb our input database measurements by adding an uncorrelated Gaussian noise with a standard deviation of $\sigma = 5\ \mu g/m^3$. This value was chosen because it is close to the variance of real measurements in the original database. The increase of the interpolation's RMSE value by the noise in the data is estimated for each method (presented in Table 4).

We define the perturbed data by the equation (11), knowing that observed database is the provided database of PM_{10} concentrations:

Perturbed input database = observed database + Gaussiannoise ($\sigma = 5\ \mu g/m^3$) .. (11)

In the second column of Table 4, we display the RMSE already presented in Table 2 (column 1 of Table 2), then the RMSE of the perturbed database in the third column. The fourth column contains δ RMSE which is the difference between these two RMSE. Table 4 shows that δ RMSE is quite comparable between all the methods, where the lowest difference is for the optimized IDW with 1.45 $\mu g/m^3$, while the highest difference is for Spline by 2.21 $\mu g/m^3$. We notice that δ RMSE of IDW and GPR is smaller than the triangulation-based methods δ RMSE, and this due to the fact that these latter are based on only the three neighboring points to make the estimation of a query site. Thus, once one of these three points is perturbed this adversely affects the interpolation much more than in the case of IDW and GPR, which use all the network sites to perform the estimation. In addition, Spline is a technique that tries to pass through the given points exactly, which causes an overfitting problem and makes its δ RMSE the bigger among all the interpolation methods.

Table 4. Interpolation methods results on perturbed data by Gaussian noise.

Interpolation method	RMSE (observed data)	RMSE (perturbed data)	δ RMSE RMSE (difference)
Nearest Neighbor	9.55	11.70	2.15
Linear interpolation*	8.84	10.81	1.97
Natural neighbor*	8.73	10.67	1.94
Spline*	9.61	11.82	2.21
IDW	7.74	9.26	1.52
Optimized IDW	7.48	8.93	1.45
GPR sqrexpo	7.75	9.45	1.70
GPR ardsqrexpo	7.81	9.52	1.71

*: In addition to the used method for interpolation, we used nearest neighbor for extrapolation since it is the most stable technique among the set we have, to extrapolate.

Table 4 displays a little difference in δ RMSE between the different applied methods even though there are some that are more accurate than others, we suppose that it is because the database is not dense enough.

To get an idea about how this perturbation affects the different stations we have, we plot the calculated ä RMSE of perturbed data for every station by Inverse Distance Weighting method (Figure 11), and then interpolate it by this same technique. We remark that δ RMSE values are small in the stations that are exposed to emission sources of air pollution (stations discussed in section 3.1), unlikely to the stations that have a higher R^2 (figure 8) their δ RMSE

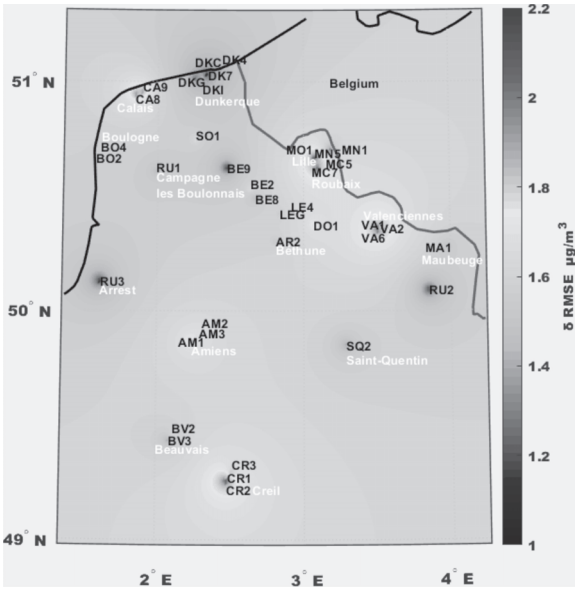

Fig. 11. The difference between RMSE of observed data and RMSE of perturbed data by IDW of PM_{10} concentrations

seems to be bigger. A possible interpretation of this behavior could be the stations near emission sources of air pollution (the industrial stations of Dunkerque city) are already uncorrelated due to the local air pollution emission sources. Perturbing the measurements of these stations do not have a big impact on their behavior, what is expressed by a small δ RMSE in these stations (like DK stations, RU2 and MC7). In contrast, the other stations that are far from any air pollution producing activities, they used to have a correlated behavior of PM_{10} concentrations, so once we perturb one of them it affects their behavior and leads to big error as shown in Figure 11.

Furthermore, in figure 12 we applied different values of perturbations on some of the used interpolation techniques. The elevation of the added perturbation (from 0 to 3) is accompanied by an increase in the RMSE values, which was expected. Moreover, we noticed that the IDW and GPR keep on behaving in a similar way having comparable RMSE values, while linear interpolation method had a larger RMSE than IDW and GPR.

Sensitivity to number of points

The last point we wanted to examine in this work, is to study the impact of data density on spatial interpolation techniques. To do so, we applied the previous interpolation techniques on the same PM_{10} concentrations database but with different number of measuring sites, each time we take out randomly one station of the available stations' set and see how this affects the method's RMSE. In Figure 13, we plot the RMSE of each method with a confidence interval of 95% as a function of the used number of stations in interpolation that decreases by a step of

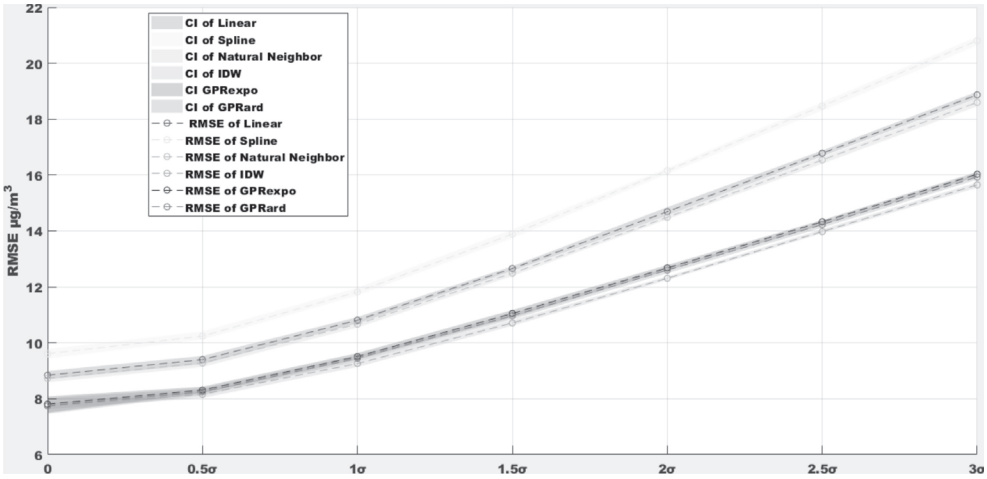

Fig. 12. Sensitivity of methods to perturbations

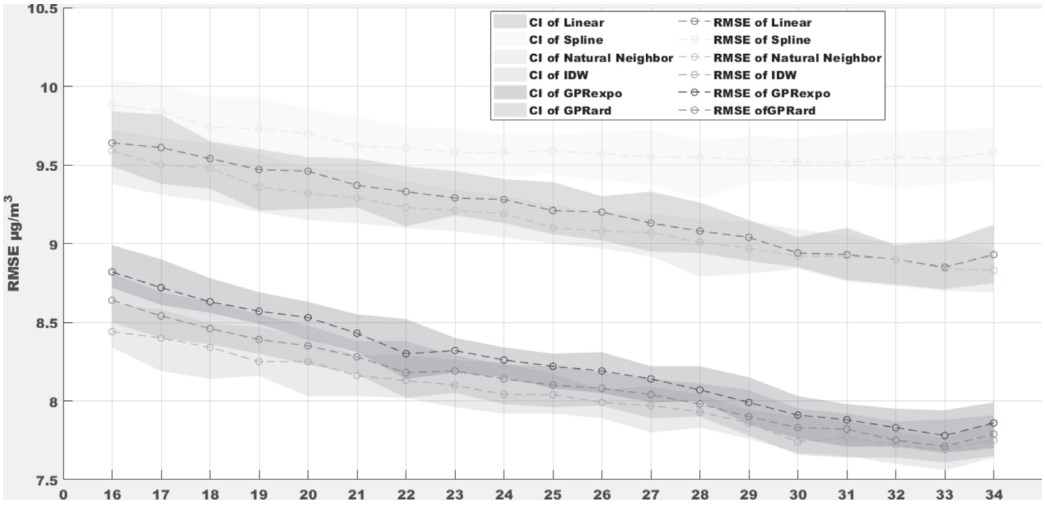

Fig. 13. Sensitivity of methods to data density (input stations number)

one. We selected six of the techniques we used before, namely linear interpolation, spline, natural neighbor, IDW and GPR with two kernels.

As expected, the diminution of the stations' number deteriorates the performance of the interpolation methods (Figure 13), what is expressed by the bigger RMSE we get each time we decrease the number of measuring stations in the different used interpolation methods. Thus, the data density plays a crucial role in making the methods estimation more accurate.

CONCLUSION

In this work, we estimated the PM_{10} concentrations in Hauts-de-France region by a few spatial interpolation techniques. The studied region has different locations subjected to air pollution of numerous industrial sources and by various atmospheric circulations. It allows estimating the performance of interpolation algorithms in different conditions.

It was shown that IDW in the classic and optimized versions and GPR with the two kernels gave similar results in the spatial interpolation of PM_{10} concentrations. Another group of methods as Linear Interpolation, 2D Splines, and Natural Neighbor is based on Delaunay triangulation. The RMSE and R^2 values show that the second group is less performant. These two groups of methods persist in the sensitivity study with respect to measurement's perturbation and density.

For all of the applied methods, the interpolation RMSE error was more important in areas near emission sources and in the zones subjected to intensive atmospheric phenomena (costal and hilly areas).

To verify the connection of interpolation precision with local atmospheric phenomena we filtered a high frequency pollution component by averaging of the input pollution data. We obtained that the maximum of R^2 corresponds to 1-day period. It shows the influence of the local periodic atmospheric phenomena like breezes on the pollution dispersion.

The estimation of the sensitivity of interpolation to the data perturbation demonstrates the diminishing of precision caused, by example, by malfunctioning of a sensor. It was shown that stations exposed to air pollution have an uncorrelated behavior making them less affected by this perturbation than other stations.

As expected, all the interpolation techniques showed the sensitivity to the number of available stations, the RMSE decreases with increase of the data points. The 95% confidence interval illustrates that the performance is comparable inside the first group.

In the perspective, it could be interesting to find a way to take into account meteorological parameters during the construction and optimization of statistical interpolators. Employing the temporal dimension of data could also bring more information on the local pollution dispersion and on precision of the interpolation.

Another possible continuation of the study is the optimization of measurement network for better retrieval of air pollution in the region.

ACKNOWLEDGEMENTS

This work was supported by the scholarship of Excellence from the National Center for Scientific and Technical Research (CNRST) of Morocco, and Université du Littoral Côte d'Opale of Dunkerque, France.

We would like to thank AtmoHauts-de-France for providing us the measurements used in this work.

The CaPPA project (Chemical and Physical Properties of the Atmosphere) is funded by the French National Research Agency (ANR) through the PIA (Programmed'Investissementd'Avenir) under contract "ANR-11-LABX-0005-01" and by the Regional Council " Nord-Pas de Calais » and the "European Funds for Regional Economic Development (FEDER)

Experiments presented in this paper were carried out using the CALCULCO computing platform, supported by SCoSI/ULCO (Service COmmun du Systèmed'Information de l'Université du Littoral Côte d'Opale).

Conflict of Interest: The authors declare that they have no conflict of interest.

REFERENCES

Briggs, A. H., Wonderling, D. E. and Mooney, C. Z. 1997. Pulling costeffectiveness analysis up by its bootstraps: a nonparametric approach to confidence interval estimation. *Health Economics.* 6(4) : 327-340.

Brunekreef, B. and Holgate, S. T. 2002. Air pollution and health. *The Lancet.* 360 (9341) : 1233-1242.

Chen, Hong, Jeffrey Kwong, Ray Copes, Karen Tu, Aaron

van Donkelaar, Perry Hystad, Paul Villeneuve, 2016.. Living near major roads and the incidence of dementia, Parkinson's disease and multiple sclerosis in Ontario, Canada: population-based study. In : *ISEE Conference Abstracts.* 2016.

Deligiorgi, D. and Philippopoulos, K. 2011. Spatial interpolation methodologies in urban air pollution modeling: application for the greater area of metropolitan Athens, Greece. *Advanced Air Pollution.* 341-362.

Efron, B. 1982. The jackknife, the bootstrap, and other resampling plans (Vol. 38). Siam.

Ehrampoush, M. H., Jamshidi, S., ZareSakhvidi, M. J. and Miri, M. 2017. A Comparison on Function of Kriging and Inverse Distance Weighting Models in PM10 Zoning in Urban Area. *Journal of Environmental Health and Sustainable Development*. 2(4) : 379-387.

Gengembre, C. 2018. Analyse dynamique, en champ proche et à résolution temporelle fine, de l'aérosol submicronique en situation urbaine sous influence industrielle. (Doctoral dissertation). Retrieved from http://www.theses.fr/ with ID 2018DUNK0489.

Gong, G. 1986. Cross-validation, the jackknife, and the bootstrap: Excess error estimation in forward logistic regression. *Journal of the American Statistical Association.* 81(393) : 108-113.

Green, P. J. and Sibson, R. 1978. Computing Dirichlet tessellations in the plane. *The Computer Journal*. 21(2) : 168-173.

Hudda, N. and Fruin, S. A. 2016. International airport impacts to air quality: size and related properties of large increases in ultrafine particle number concentrations. *Environmental Science & Technology.* 50(7) : 3362-3370.

Kampa, M. and Castanas, E. 2008. Human health effects of air pollution. *Environmental Pollution.* 151(2): 362-367.

Künzli, Nino, Michael Jerrett, Wendy J. Mack, Bernardo Beckerman, Laurie Labree, Frank Gillil, Duncan Thomas, John Peters, and Howard N. Hodis. 2005. *Ambient air pollution and atherosclerosis in Los Angeles. Environ Health Perspect*. 113 : 201-206.

Lee, D. T. and Schachter, B. J. 1980. Two algorithms for constructing a Delaunay triangulation. *International Journal of Computer & Information Sciences.* 9(3): 219-242.

Lee, Jui-Huan, 2014. Land use regression models for estimating individual NOx and NO_2 exposures in a metropolis with a high density of traffic roads and population. *Science of the Total Environment.* 472 : 1163-1171.

Li, J. and Heap, A. D. 2011. A review of comparative studies of spatial interpolation methods in environmental sciences: *Performance and impact factors. Ecological Informatics.* 6(3-4) : 228-241.

Li, J. and Heap, A. D. 2014. Spatial interpolation methods applied in the environmental sciences: A review. *Environmental Modelling & Software*. 53 : 173-189.

Li, L., Wu, J., Hudda, N., Sioutas, C., Fruin, S. A. and Delfino, R. J. 2013. Modeling the concentrations of on-road air pollutants in southern California. *Environmental Science & Technology.* 47(16): 9291-9299.

Longley, P. A., Goodchild, M. F., Maguire, D. J. and Rhind, D. W. 2005. *Geographic Information Systems and Science.* John Wiley & Sons.

Miller, K. A., Siscovick, D. S., Sheppard, L., Shepherd, K., Sullivan, J. H., Anderson, G. L. and Kaufman, J. D. 2007. Long-term exposure to air pollution and incidence of cardiovascular events in women. *New England Journal of Medicine.* 356(5) : 447-458.

Miller, S. T. K., Keim, B. D., Talbot, R. W. and Mao, H. 2003. Sea breeze: Structure, forecasting, and impacts. *Reviews of Geophysics.* 41(3).

Qiao, P., Li, P., Cheng, Y., Wei, W., Yang, S., Lei, M. and Chen, T. 2019. Comparison of common spatial interpolation methods for analyzing pollutant spatial distributions at contaminated sites. *Environmental Geochemistry and Health.* 1-22.

Ramanathan, V. and Feng, Y. 2009. Air pollution, greenhouse gases and climate change: Global and regional perspectives. *Atmospheric Environment.* 43(1) : 37-50.

Rao, S. T., Sistla, G., Pagnotti, V., Petersen, W. B., Irwin, J. S. and Turner, D. B. 1985. Resampling and extreme value statistics in air quality model performance evaluation. *Atmospheric Environment*. 19(9) : 1503-1518.

Seinfeld, J.H. and Pandis, S.N. 2012. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change.* Ed. Wiley.

Sibson, R. 1981. A brief description of natural neighbour interpolation. Interpreting multivariate data.

Son, J. Y., Bell, M. L. and Lee, J. T. 2010. Individual exposure to air pollution and lung function in Korea: spatial analysis using multiple exposure approaches. *Environmental Research.* 110(8) : 739-749.

Watson, D. F. 1985. A refinement of inverse distance weighted interpolation. *Geoprocessing.* 2 : 315-327.

Williams, C. K. and Rasmussen, C. E. 2006. *Gaussian Processes for Machine Learning* (Vol. 2, No. 3, p. 4). Cambridge, MA: MIT press.

Willmott, C. J. 1982. Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society.* 63(11) : 1309-1313.

Willmott, C. J. and Matsuura, K. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research.* 30(1) : 79-82.

Wong, D. W., Yuan, L. and Perlin, S. A. 2004. Comparison

of spatial interpolation methods for the estimation of air quality data. *Journal of Exposure Science and Environmental Epidemiology.* 14(5) : 404.

Zhang, X., Shi, R. and Chen, M. 2018, September). Comparative study of the spatial interpolation methods for the Shanghai regional air quality evaluation. In: *Remote Sensing and Modeling of Ecosystems for Sustainability XV* (Vol. 10767, p. 107670Q). International Society for Optics and Photonics.

[1]: https://hautsdefrance.cci.fr/wp-content/uploads/sites/6/2016/04/Atlas_NPCP_Edition2016.pdf

[2]: https://www.atmo-hdf.fr

APPENDIX

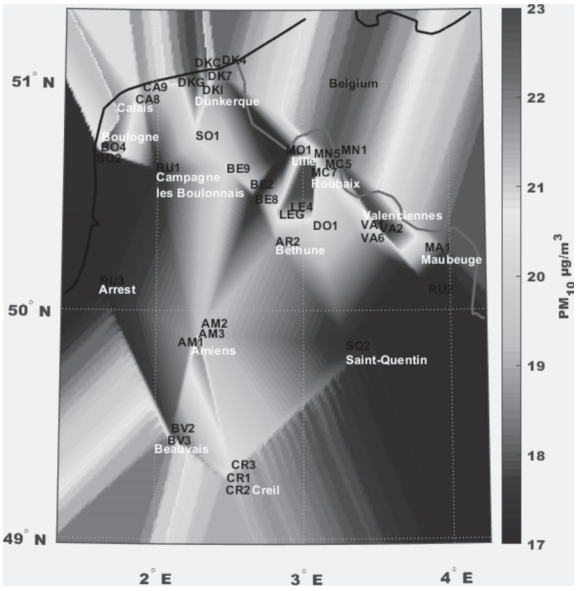

Fig. 14. 2016 annual mean of PM_{10} concentrations for Hauts-de-France region (interpolated by linear interpolation and nearest neighbor for extrapolation)

Figure 14 illustrates the annual mean of PM_{10} concentrations interpolated by the based triangulation method: linear interpolation. This map gives a physically wrong representation of the spatial distribution of air pollution, the reason why we chose IDW as the generator of all of the maps of this work.

The figures above (from 15 to 18) display the RMSE of PM_{10} concentrations in different time averaging periods. We applied the same interpolation method (which is IDW) in all these periods to be able to compare variation of RMSE corresponding to time averaging periods. In figure 15, which represents the interpolated RMSE of the original provided database with 15 minutes, we

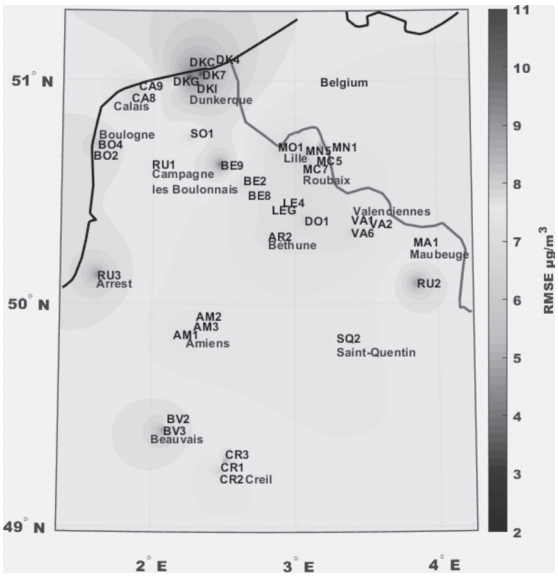

Fig. 15. RMSE of IDW interpolation for PM_{10} concentrations for Hauts-de-France region (for quarter hour data as observed)

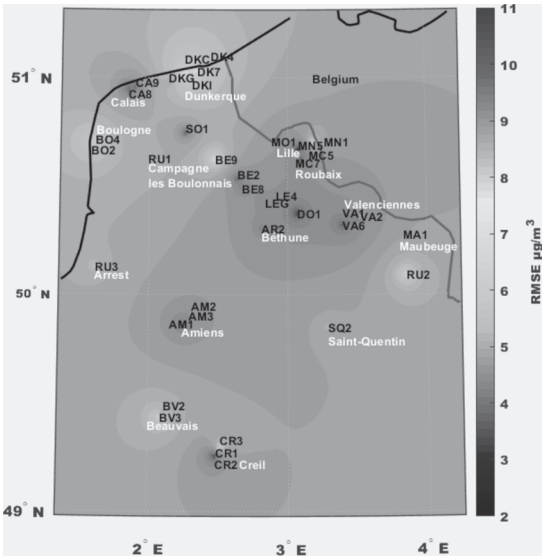

Fig. 16. RMSE of IDW interpolation for PM_{10} concentrations for Hauts-de-France region (for 1-day time averaged data)

have the biggest values of error among all the plotted maps, where we found (as indicated in section 3.1) that the areas exposed to emission sources of air pollution have the biggest error. Moving to a spatial interpolation with a time averaging period of 1 day (Figure 16), we observe a decrease of RMSE error values in the whole map with always being bigger in the same spots where it was bigger in the previous map (Figure 15). This is thanks to filtering the effect of meteorological

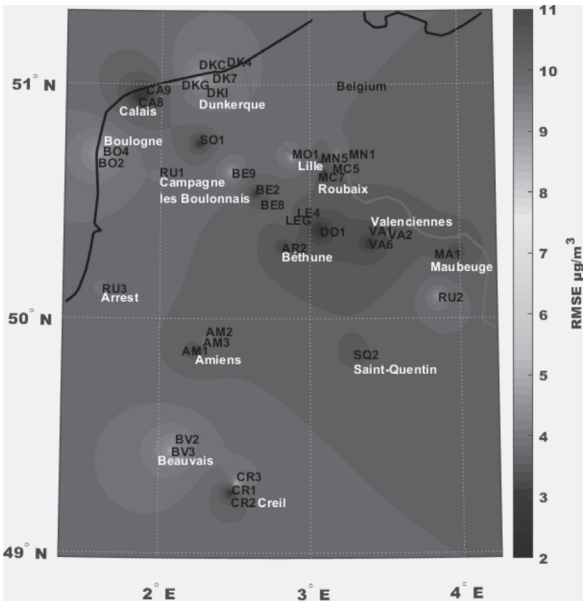

Fig. 17. RMSE of IDW interpolation for PM_{10} concentrations for Hauts-de-France region (for 1-week time averaged data)

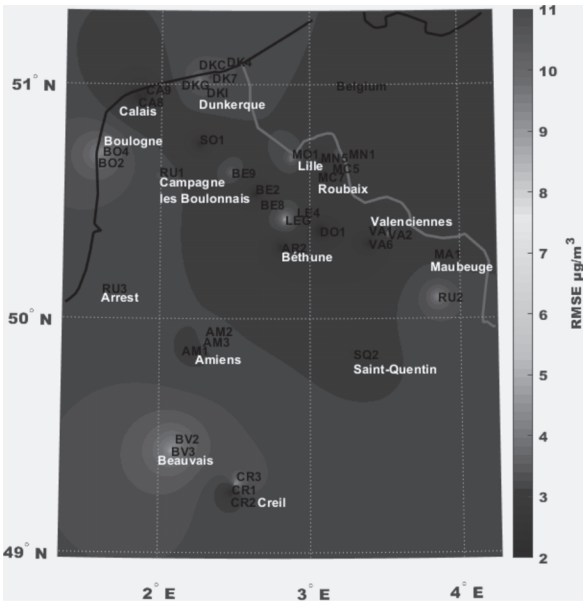

Fig. 18. RMSE of IDW interpolation for PM_{10} concentrations for Hauts-de-France region (for 1-month time averaged data)

phenomena with 1-day periodicity on PM_{10} concentrations. This RMSE keeps on decreasing as long as our time averaging period grows bigger, as we notice in Figures 17 and 18, where this time we filter the effect of 1-week to 1-month of the meteorological phenomena.