# Estimating indicators of river quality by geostatistics

C. Bernard-Michel, C. de Fouquet

Ecole des Mines de Paris, 35 rue Saint Honoré, 77305 Fontainebleau, France.

## 1 Introduction

In order to assess river water quality, nitrate concentrations are measured in different monitoring stations. The information contained in these measurements is summarized in a few synthetic quantitative indicators such as the 90% quantile of yearly concentrations or the annual mean making it possible to compare water quality in different stations, and its yearly evolution. The current French recommendations are based on the French water quality's evaluation system (SEQ EAU, http://www.rnde.tm.fr/) and the water framework directive in Europe, which aims at achieving good water status for all waters by 2015.

These calculations, however, use the classical statistical inference, essentially based on a hypothesis proved to be incorrect for nutrients (Bernard-Michel and de Fouquet 2003): time correlations are not taken into account. Moreover, the seasonal variations of concentrations and the monitoring strategy are ignored. For example, because of the run-off or the leaching of fertilizer, nitrate concentrations of Loire Bretagne basin are often high in winter and low in summer (Payne 1993). Thus, if sampling is increased in winter out of precaution, the annual mean and the quantile can be falsely increased. It is therefore necessary to develop methods that take into account both time correlations and sampling dates, especially in case of preferential sampling. We propose to assign kriging weights or segment of influence weights (Chilès 1999) to measurements for both indicators and to use a linear interpolation of the empirical quantile for the estimation of the 90$^{th}$ percentile.
In this paper, methods are presented and compared for nutriments on simulations.

## 2 Example

Figure 1 (left) shows an example of real nitrate concentrations measurements from the Loire River in 1985. The indicators have been estimated first with the totality of measurements (6 in summer, 12 in winter), then with an extracted sample of one regular measurement a month.

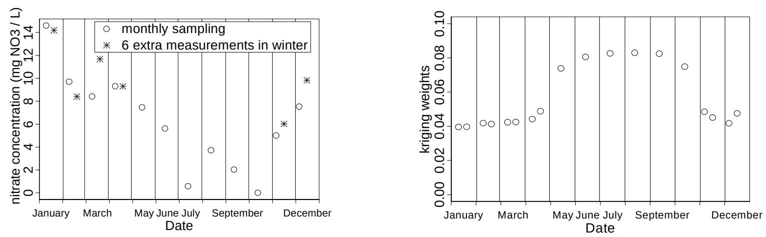


**Fig. 1.** Preferential sampling of nitrates concentration during one year at one monitoring station. Left: the measurements frequency is doubled in winter; Right: associated kriging weights.

**Table 1.** Statistical annual mean and quantile of nitrates concentrations.

| Sample size | Sample mean | 90% empirical quantile |
|---|---|---|
| 12 | 6.16 | 9.70 |
| 18 | 7.41 | 14.19 |

Classical estimations are presented in Table 1: usual indicators are obviously increased when sampling is reinforced in winter. It is a consequence of the preferential sampling and of the presence of time correlation showed for many nutriments (figure 4). It therefore appears necessary to develop methods to better assess yearly temporal mean.

## 3 The annual mean: statistical parameter or time average?

### 3.1 Methods

Experimental temporal variograms calculated on nitrates concentrations show for most of the monitoring stations the evidence of a time correlation. The sample mean (i.e. the arithmetic mean of experimental data) is an unbiased estimator for independent data or regularly spaced correlated data (with certain exception). In presence of time correlation, it is no longer the case when sampling is irregular or preferential. To correct this bias, two methods were studied:

- Kriging with unknown mean (OK) which takes into account correlation in the estimation of the annual mean and in the calculation of the estimation variance;
- A geometrical declustering whose objective is only to correct the irregularity of sampling.

These methods are presented below and compared later on simulations (2.2):

1. Classical statistical method (Saporta 1990): sample values $z_1, z_2, ..., z_n$ are interpreted as realizations of independent random variables $Z_1, Z_2, ...Z_n$ which all have the same distribution, with expectation $m$. The yearly mean corresponds to the estimation of this expectation, using the sample mean (1) denoted $m^*$. The estimation variance (2) is deduced from the experimental variance $\sigma^2$ :

$$m^* = \frac{1}{n}\sum_{i=1}^{n} Z_i \tag{1}$$

$$Var\left(\overline{Z}_n - m\right) = Var\left(\overline{Z}_n\right) = \frac{\sigma^2}{n} \text{ with } \sigma^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(Z_i - m^*\right)^2 \tag{2}$$

2. Temporal kriging (Chilès 1999): sample values are interpreted as a realization of a random correlated function $Z(t)$ at dates $t_1, t_2, ..., t_n$. We don't estimate anymore the parameter of a distribution, but the temporal mean $Z_T = \frac{1}{T}\int_T Z(t)dt$, still defined even in absence of stationarity. This quantity is estimated using ordinary "block" kriging, with constant but unknown mean :

$$Z_T^* = \sum_{i=1}^{n} \lambda_i Z_i \text{ where } \lambda_i \text{ are kriging weights} \tag{3}$$

$$Var(Z_T^* - Z_T) = \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} \lambda_\alpha \lambda_\beta C\left(t_\alpha - t_\beta\right) + \overline{C}\left(T,T\right) - 2\sum_{\alpha=1}^{n}\lambda_\alpha \int_T C\left(t_\alpha - t\right)dt \tag{4}$$

Analytical expressions are easy to calculate at 1D, without any discretization (Matheron 1970; Journel 1977). Figure 1 (right) gives an example of kriging weights, assigning lower weights to winter values, which avoids an estimation bias. The estimation variance and confidence interval, overestimated by classical statistics, are reduced by kriging taking into account the temporal correlation and the annual periodicity of the concentration.

3. Geometrical method (Chilès and Delfiner 1999) by segment declustering, corresponding to 1D polygonal declustering. This technique consists in weighting each data by the relative length of its segment of influence, in the linear combination (5). An example for 4 measurements is given in figure 2. Calculating the estimation variance necessitates the variogram.

$$Z_T^* = \sum_{i=1}^{n} \lambda_i Z_i \text{ where } \lambda_i \text{ are segment of influence weights} \tag{5}$$

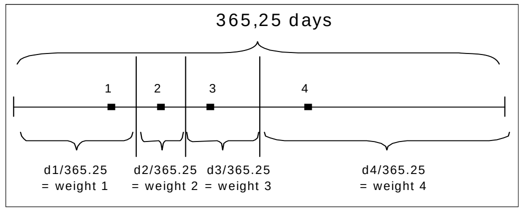


**Fig. 2.** Example of segment declustering for 4 measurements

## 3.2 Testing methods on simulations

### *3.2.1 Choice of the monitoring station*

In order to quantify the improvements of the new proposed methods, it would be necessary to have examples in which the annual mean is known. Because it is impossible, we propose to simulate 365 days of measurements based on a real dataset. We then admit that one measurement a day exactly determines the yearly mean.

The best sampled monitoring station available was on the Loire river in Orléans: in 1985, 1 measurement was taken every 2 days, in 1986, 1 measurement a week, and for other years 3 measurements a month. A Gaussian sequential conditional simulation of “daily” concentrations over ten years with respect to experimental data in Orleans and the fitted experimental variogram (figure 4) was constructed (figure 3, only the 1985 simulated values are presented). Then, samples were extracted to compare the annual means exactly known to the three estimation methods.

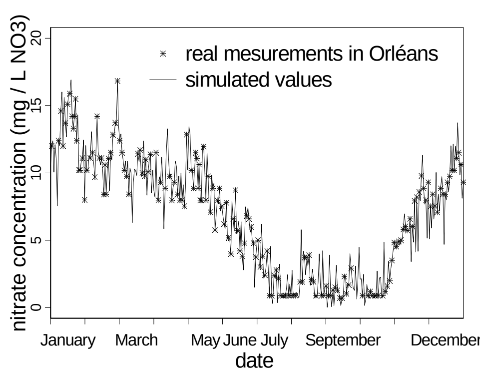


**Fig. 3.** Conditional simulation based on the real measurements. Loire river in Orléans.

### 3.2.2 Variogram model

The experimental variogram calculated over several years for Orléans (figure 4, right), reflects the annual periodicity of nitrate concentrations. The variogram calculated over one year with a lag of 2 days (figure 4) show the predominance of this periodical component. In these mean temporal variograms, winter or summer values are not distinguished. Variograms calculated for each season would differ, but as we are interested in the global annual statistics, the averaged variogram on one year is here sufficient (Matheron 1970).

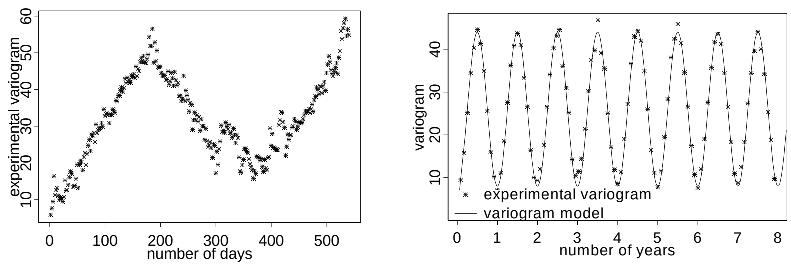


**Fig. 4.** Experimental variogram for nitrate at Orleans. Left, calculation for one year, lags of 2 days. Right, calculation for 8 years, lag of 30 days. This variogram has been fitted manually by the sum of a nugget effect, a spherical model and a cosinus model.

### 3.2.3 Simulations and results

In figure 5, a comparison between the estimation of the annual mean by statistics and geostatistics over 10 years is presented. The estimations are given with their 95% confidence intervals and compared to the real value of the annual mean estimated with 365 measurements. Samples are preferential (6 values in summer, 12 in winter) and have been extracted from the simulation. Figure 5 (left) confirms that the corresponding sample mean is often higher than yearly mean, and moreover it leads to a correspondingly large 95% confidence interval. The bias is well corrected by the geostatistical and geometrical methods for which weights are equivalent (figure 5, right). Nevertheless, kriging directly gives the estimation variance.

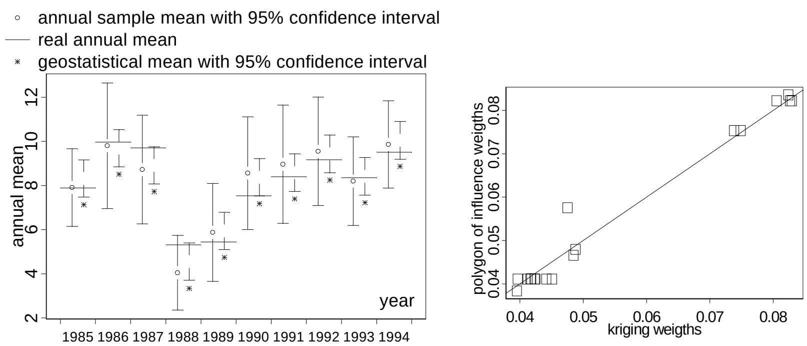


**Fig. 5.** On the left, estimations by sample mean and kriging are presented with their associated 95% confidence interval. They're compared to real annual mean. On the right, scatter diagram between weights, for kriging (abscissa) and for segments of influence (ordinates).

For most of the years, kriging gives better estimations than statistics and moreover 95% confidence intervals are about twice smaller than the ones given by statistics, and always include the true yearly mean. These results are confirmed on 1000 simulations (Table 2). Other examples for different stations, parameters and with different sampling strategy can be found in Bernard-Michel and de Fouquet, 2003.

As first conclusion, kriging corrected the bias in case of preferential sampling, better assessed yearly mean, and better predicted the precision of this estimation. However, if we are only interested in the value of annual mean, segment declustering can be used because of its simplicity. If the precision is needed, then kriging should be preferred.

**Table 2.** 1000 simulations : comparison of statistics and geostatistics estimations in average for a preferential sampling (12 measurements in winter, 6 in summer)-

| Average of the 1000 annual mean estimated with 365 measurements | 6.72 | |
|---|---|---|
| Preferential sampling | Statistics | Geostatistics |
| Average of the annual means | 7.62 | 6.72 |
| Average of the predicted standard deviations of estimation errors | 0.97 | 0.42 |
| Experimental standard deviation of error | 0.93 | 0.31 |
| Experimental 95% confidence interval | [7.10;8.12] | [6.07;7.37] |

## 4 Estimation of the 90% quantile

The 90% quantile is used by water agencies to characterize high concentrations, potentially the most dangerous for human health. However, today's recommendation to approach the 90% quantile is based on the empirical quantile. This statistical method is proved to be problematic for the following reasons:

– It is a biased estimator (Gaudoin 2002).
– As the sample mean, it does not take into account time correlation, and sample irregularity.

We first evaluated the bias of the empirical quantile in the case of independent variables, and proposed three methods to remedy. Then, we took into account the time correlation and the sampling irregularity by weighting the measurements.

### 4.1 Bias of the empirical 90% quantile of independent data

Generally, the estimation of percentiles is a part of extreme values theory (Coles 2001). However, this theory is based on asymptotic theorems which require many measurements. As we will only dispose of an average of 12 measurements a year, we propose to use a classical non-parametric estimator: the empirical quantile (Saporta 1990, Gaudoin 2002). But this estimator is proved to be biased (Gaudoin 2002). Moreover, this bias is a function of the sample size. This faces with a real problem when tracing yearly quality or comparing stations with different sampling sizes.

Several methods were studied to reduce the bias:

– Linear interpolation of the empirical quantile : when all the *n* experimental data are different, the quantile of order $i/n$ is $Z_{(i)} + \frac{Z_{(i+1)} - Z_{(i)}}{2}$, the one of order $0$ is the half of the minimum value, and the one of order 1 is the maximum experimental value. Linear interpolation is applied between these quantile values.
– Use of a Gaussian anamorphosis linearly interpolated (Rivoirard 1994);
– Use of a Gaussian anamorphosis fitted with an Hermite polynomial function (Rivoirard 1994).

The biases of the 90% quantile estimated by these methods are not theoretically calculable because the distribution of the concentrations is unknown. That’s why we propose to use simulations to evaluate them.

In case of a usual distribution, the expression of the bias is known theoretically but sometimes hard to calculate. We’ve calculated it for a uniform distribution in order to compare it with simulations results. Because of the similarity of results, we deduced that simulations are a good method to evaluate the bias.

Here we present results for 1000 realizations of an exponential distribution with expectation 1, samples sizes varying from 4 to 36. Because the variables are independent, it is not necessary to construct all the 365 daily values of a year to extract the samples of different sizes. Results are given in average for each different sample size. The evolution of the 90% quantile, the experimental estimation variance, the 95% confidence interval and the distribution of errors are presented figure 6.

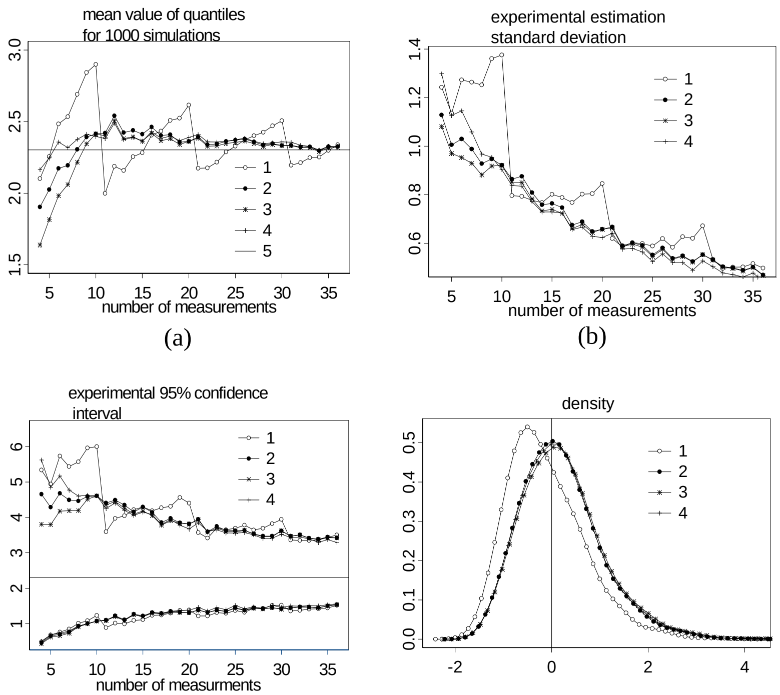


**Fig. 6.** Quantile estimation for independent variable, compared with theoretical value; results for 1000 simulations, as a function of sample size (a, b, c). Upper left figure (a): average of quantiles estimation. Upper right figure (b): experimental estimation standard error. Lower left figure (c): experimental 95% confidence interval. Lower right figure (d): histogram of quantile errors for samples of size 12. *Legend* : 1 represents the empirical quantile, 2 the linear interpolation of anamorphosis, 3 the hermitian interpolation of empirical quantile, 4 the hermitian interpolation of anamorphosis, 5 the real quantile.

The empirical quantile (figure 6 (a)) presents a bias, strongly reduced by the other methods. Moreover, strong discontinuities for sample sizes proportional to 10 make difficult the comparison between monitoring stations with different sampling strategy. The three proposed methods are quite similar for samples whose sizes are greater than 10. They don‘t show any more discontinuities, but converge quite regularly toward the theoretical value. For this distribution, with 12 measurements, the 90% quantile is overestimated using the three interpolations functions, and clearly underestimated using the empirical quantile.

Figure 6 (b) makes possible to evaluate committed errors in the quantile estimation. It gives the following experimental estimation standard error as a decreasing function of the sample size. However, precision is not really satisfying even with 36 measurements because it is still representing 20% of the real quantile.

Figure 6 (c) presents for each sample size the interval containing 95% of the 1000 quantile estimations, approximately symmetrical around the theoretical value.

For sample of size 12, figure 6 (d) shows that the distribution of the estimation errors is nearly Gaussian for the 3 interpolation functions, but not for the empirical quantile. In this last case, the errors are not centered and not symmetrical.

Other distribution examples (normal, lognormal, gamma and uniform distribution) have been tested leading to the same conclusions. Because of its equivalence to others methods, and its simplicity, the linear interpolation of quantile is advised for the estimation of 90% quantile and will always be used from now on.

## 4.2 Case of temporal correlation: weighted data

In presence of temporal correlation and in a limited field, we do not try any more to estimate the histogram or the quantile of the a priori distribution; this one corresponds, for an ergodic model, to the distribution of a realization in an infinite field. For a fixed realisation, the distribution to calculate is the one of a random point in the field. Because of the limited number of data per station for one year, we examine an approximate calculation of this “global” distribution.

### *4.2.1 Irregular sampling*

The bias of empirical quantile methods on independent variables can be resolved in practice with a linear interpolation of quantiles. As for the estimation of yearly mean, in presence of temporal correlation and irregular or preferential sampling, the weighting becomes necessary to avoid bias. The weights calculated for annual mean estimation (by kriging or segment of influence) are now used in the estimation of the experimental histogram. Then, the estimated quantiles are calculated are compared below on simulations for irregular but not preferential sampling.

The following example is based on real nitrate data of the Indre River. We have proceeded with 1000 conditional simulations of 365 days respecting real measurements, using the fitted variogram presented on figure 7. From each simulation, samples of different sizes have been extracted, from 4 to 36 measurements a year, irregularly spaced in time. Thus we obtain 1000 samples of size 4, 1000 of size 5 etc…. We estimate the 90% quantile for each sample and for each method. Results are given in average for each different sample size and shown in figure 8. They are compared to quantiles calculated in average on the 1000 simulations of 365 days. That means we consider that a 90% quantile is well determined with one measurement a day.

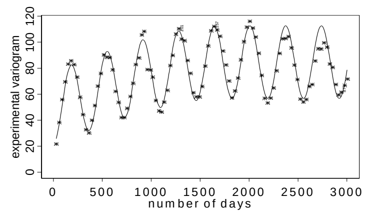


**Fig. 7.** Mean experimental variogram ($mg^2$ / $L^2$ NO3) calculated with monthly sampling and fitted model for the monitoring station on the Indre River. The model is composed of nugget effect (21), cosinus model (period 365.25, amplitude 56) and spherical model (range 1795, sill 35)

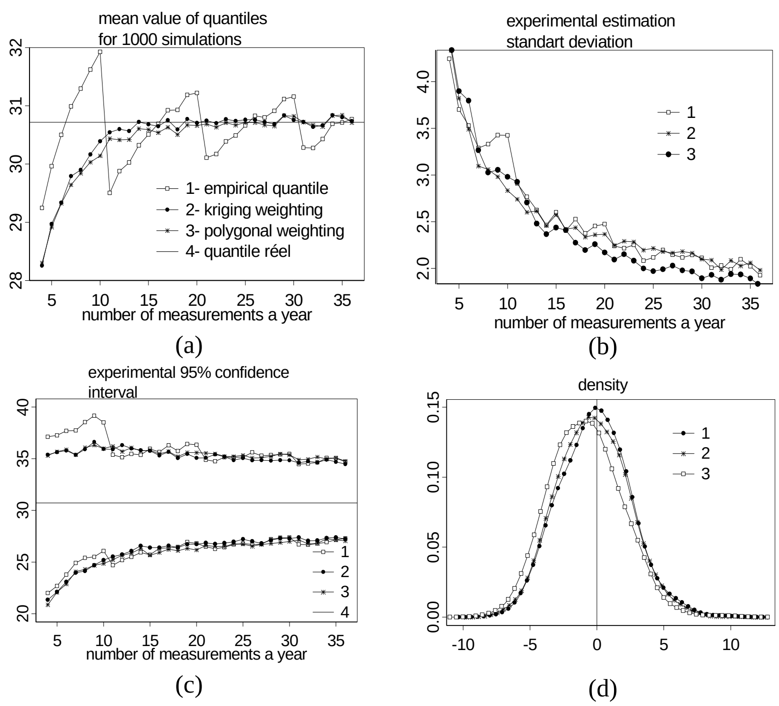


**Fig. 8.** Quantile estimation for temporal correlated variables, compared with the empirical quantity, for 1000 simulations. This empirical quantity corresponds to the mean, calculated on all the simulations, of the 90% quantiles of 365 values. All calculations are made by linear interpolation of quantiles. In abscissa for (a), (b), (c), the sample size. Upper left figure (a): average of quantiles estimation. Upper right figure (b): experimental estimation standard error. Lower left figure (c): experimental 95% confidence interval. Lower right figure (d): histogram of quantile errors for samples of size 12.

On figure 8 (a), the important bias of the empirical quantile is very well corrected by both kriging and segment of influence weighting. However, the estimation variance (figure 8 (b)) is not clearly improved for the new proposed methods and is still quite important for a sample size of less than 36 measurements a year. Actually, for 36 measurements a year, errors still represent approximately 6% of the real quantile which gives an approximate 95% confidence interval (figure 8 (c)) of $\pm 12\%$ around the real quantile because of the quasi normal distribution of errors (figure 8 (d)). For 4 measurements a year, it reaches 14% of the real quantile, and 9% for 12 measurements a year. By simulating data respecting variability and real measurements, we can determine the necessary sample size to reach a desired precision. The theoretical estimator of a confidence interval taking into account temporal correlation would be difficult to construct. Even when random variables are independent, the theoretical announced interval (Gaudoin 2002) is not satisfying because it is limited by the higher order statistic. Simulations can be a solution to evaluate errors committed on estimations. Because results are similar with kriging and segment declustering, and because the estimation variance is difficult to assess for both methods, we propose in the future to weight measurements with segment influence segments, which is easier to automate.

### *4.2.2 Preferential sampling*

Just as in paragraph 2.2.3, we compare statistical and kriging estimations on preferential sampling in Orléans over 10 years. In figure 9, on the left we present a scatter diagram between estimations on monthly sampling and preferential sampling in winter.

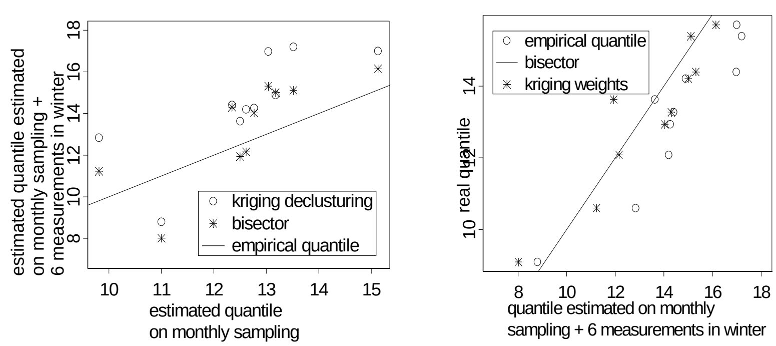


**Fig. 9.** On the left a scatter diagram presents statistical and geostatistical estimations of quantile calculated with 12 or 18 measurements a year. On the right, the scatter diagram compares 18 measurements estimations to real quantile value

Most of points are upper the bisector because of the bias created by the preferential sampling in winter. But kriging correct this bias (points are closer to the bisector and always lower than statistics estimations. In figure 9, on the right, estimations are compared to the real 90% quantile. It shows a better precision of kriging.

## Conclusion

Kriging the annual mean allows to correct the bias induced by a preferential sampling of high concentration periods. The kriging variance is lower than the predicted statistical variance of the mean of independent variables, namely because of the yearly periodic component of the variogram. Associated with a linear interpolation of the experimental quantile function, the kriging weights give an empirical estimation of quantiles practically unbiased.

The segment of influence weighting can be used to simplify the calculations.

In all cases, one or two measurements a month are not sufficient for a precise estimation of the yearly 90% quantile.

## Acknowledgements

We would like to thank the French minister of environment (MEDD) for financing the study and L.- C. Oudin and D. Maupas from the Agence de l'Eau Loire-Bretagne for their help.

## References

Bernard-Michel C, de Fouquet C (2003) Calculs statistiques et géostatistiques pour l'évaluation de la qualité de l'eau. Rapport N-13/03/G. Ecole des Mines de Paris, Centre de géostatistique.

Chilès J-P, Delfiner P (1999) Geostatistics. Modeling spatial uncertainty. Wiley series in probability and statistics.

Coles S (2001) An introduction to statistical modeling of extreme values. Springer.

Gaudoin O (2002) Statistiques non paramétriques, notes de cours deuxième année. ENSIMAG, http://www-lmc.imag.fr/lmc-sms/Olivier.Gaudoin/

Journel A G (1977) Geostatistique minière. Tome 1. Ecole des Mines de Paris. Centre de Géostatistique.

Matheron G (1970) La théorie des variables régionalisées, et des applications. Les cahiers du Centre de Morphologie Mathématique. Ecole des Mines de Paris. Centre de géostatistique.

Payne M R Farm (1993) Waste and nitrate pollution. Agriculture and the environment. John Gareth Jones. Ellis Horwood series in environmental management.

Rivoirard J (1994) Introduction to Disjunctive Kriging and Non-linear Geostatistics. Clarendon Press, Oxford.

Saporta G (1990) Probabilités, analyse de données et statistiques. Technip.