Robust regression for time series exhibiting heteroscedasticity

Internship supervised by professor You-Gan Wang.
Queensland University of Technology

March-September 2018

This master thesis constitutes a pedagogical exercice, which, in no way can engage the responsability of the host compagny or laboratory.

Master 2 Dynea. Academic year 2017-2018 Callens Aurélien

Abstract

Keywords: Robust estimation, temporal correlations, heteroscedasticy, model selection.

We propose a robust procedure for analysing times series exhibiting heteroscedasticity. We assume a parametric function for the variance with unknown parameters that are estimated by high breakdown-point estimators. Then, a weighted M-estimation for independent observations is performed to obtain the robust regression parameters. In our procedure, the tuning parameter associated with the loss function in the M-estimation is chosen to minimize the variance of the regression parameters, and the temporal correlations are accounted by adding lagged terms in the model. The efficiency of our procedure is tested on simulated data with known parameters and compared to other usual regression methods. We also illustrate the proposed method using data on chlorophyll concentration in a small tributary of the Thames River (UK). In both cases, our procedure outperforms the other methods giving estimates with lower variance in addition to homoscedastic and independent residuals.

Mot clés : Estimation robuste, corrélations temporelles, heteroscedasticité, sélection de modèles.

Nous proposons une procédure robuste pour analyser les séries temporelles présentant de l'heteroscedasticité. Pour cela, nous modélisons la variance à l'aide d'une fonction avec des paramètres inconnus, estimés de manière robuste avec des estimateurs à haut point de rupture. Une M-régression pondérée par la variance est ensuite effectuée pour obtenir des estimateurs robustes des paramètres de la moyenne. Dans notre procédure, le paramètre associé à la fonction d'objectif dans la M-régression et régulant le degré de résistance aux valeurs aberrantes, est choisi pour minimiser la variance des estimateurs. De plus, les corrélations temporelles sont prises en compte en ajoutant des termes différés dans notre modèle. L'efficacité de notre procédure est testée sur des données simulées avec des paramètres connus, et comparée a d'autres méthodes de régression. Nous illustrons également notre méthode en analysant des données de concentration de chlorophylle dans un tributaire de la Thames (UK). Dans les deux cas, notre méthode est plus fiable que les autres méthodes de régression et donne des estimateurs avec une variance plus faible en plus de résidus indépendants et homoscédastiques.

Palabras clave: Estimación robusta, correlaciones temporales, heterocedasticidad, selección de modelo.

Proponemos un procedimiento robusto para analizar series de tiempo que exhiben heterocedasticidad. Para esto, modelamos la varianza usando una función con parámetros desconocidos, robustamente estimados con estimadores de punto de ruptura alta. Luego se realiza una M-estimación ponderada por la varianza para obtener estimadores robustos de los parámetros de la media. En nuestro procedimiento, el parámetro asociado con la función objetivo en la M-estimación, que regula el grado de resistencia a valores atípicos, se elige para minimizar la varianza de los estimadores. Además, las correlaciones temporales se tienen en cuenta agregando términos diferidos en nuestro modelo. La eficacia de nuestro procedimiento se prueba con datos simulados con parámetros conocidos y se compara con

otros métodos de regresión. Finalmente, ilustramos nuestro método mediante el análisis de los datos de concentración de clorofila en un afluente del río Támesis (RU). En ambos casos, nuestro método supera a otros métodos de regresión produciendo estimadores con menor varianza además de residuales independientes y homoscedastic.

Acknowledgments

First, I wish to express my sincere thanks to Troy Farrell (Head of School of Mathematical Sciences at Queensland University of Technology), for providing me with all the necessary facilities for the research.

I would like to express my sincere gratitude to my advisor, You-Gan Wang (Professor at Queensland University of Technology) for the continuous support of my research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. This research was supported by his Australian Research Council grant.

I am extremely thankful to Benoit Liquet (Professor at University of Pau and Pays de l'Adour) for introducing me to You-Gan Wang, which allowed me to experience my first internship abroad. My sincere thanks also goes to Liya Fu (Lecturer School of teh Mathematics and Statistics school at the Xi'an Jiaotong University, China) who read and helped me improving significant part of this thesis.

I am also grateful to Mrs Amanda Kolovrat (Senior Services Coordinator in the School of Mathematical Sciences) for helping me with all administrative procedures at the beginning of this internship.

Last but not the least, I would like to thank my family and my partner, Mylene, for the unceasing encouragement, support and attention.

Table of Contents

List of Tables

List of Figures

Introduction

Heteroscedasticity is often encountered in time series regression analysis (Carroll and Ruppert, 1988) and can be differentiated in two types: impure and pure heteroscedasticity. The former type is caused by a model misspecification such as an omitted covariate. On the contrary, the latter type is inherent to the data generation process and arises in almost every fields (Carroll and Ruppert, 1988). For example, in chemical kinetics the variability depends on the time and concentration of the reactives (Box and Hill, 1974) and in fisheries research, the variability in fish production depends on the size of the spawning population (Ruppert and Carroll, 1985). Ignoring this problem and applying the least square method, would result in regression parameters with biased covariance matrix and hence would lead to erroneous prediction intervals and statistical tests.

Two methods are commonly used to cope with heteroscedasticity (Bianco et al., 2000): transform the data or perform a generalized least square analysis by assuming a parametric function for the variance. The easiest solution is to transform the data with a function such as the logarithm to make the variance constant, however one might be reluctant to apply this method due to the increased difficulty to make inference in the original scale (Giltinan et al., 1986). The second solution is to model the heterogeneity of the variance with a known parametric function (Bianco et al., 2000; Carroll and Ruppert, 1982; Davidian and Carroll, 1987; Giltinan, 1983). The common heteroscedastic model is:

$$y_i = x_i^T \beta + \sigma_i \epsilon_i,$$

where ϵ_i are the independent and identically distributed error terms with mean 0 and unknown symmetric distribution function and σ_i is the term accounting for heteroscedasticity. This term can be a power function of the mean as proposed by Box and Hill (1974): $\sigma_i = \phi |x_i^T \beta|^\gamma$, a quadratic function of the mean as introduced by Bartlett (1936): $\sigma_i = \gamma_1 (x_i^T \beta) + \gamma_2 (x_i^T \beta)^2$ or some functions of the covariates. The parameters in the variance function are not known and have to be estimated by maximum likelihood method.

The estimation of the parameters in models presenting heterogeneous variance is performed by an iterative procedure called generalized least square (GLS). A preliminary estimate of the mean parameter is obtained by least square method, the residuals of this model are then used to estimate the variance parameter. Finally, a weighted least square method is performed with the variance as weight. Modelling the heterogeneous variance allows one to obtain better estimates for the mean parameters and also to gain information on the variability of the data generation process (Davidian and Carroll, 1987; Giltinan, 1983).

Such as the least square method, the estimation method for heteroscedastic models has a low breakdown point of $1/n$, meaning that only one outlier in the response variable can have a large effect on the estimation of the mean parameters (Rousseeuw and Leroy, 2005; Zhao and Wang, 2009; Wang et al., 2018). The estimation of the variance parameters is also affected as the maximum likelihood method is very sensible to outliers (Giltinan, 1983; Stefanski et al., 1986). In practice, datasets with outliers are commonly encountered in statistical analysis.

Outliers can be found in both the response variable and the covariates but, in this work, we will assume that outliers are only present in the response variable. In the presence of outliers, robust methods must be applied. Robust methods aim to produce reliable estimates that are not seriously affected by outliers or small deviations from model assumptions (Jiang et al., 2018).

The literature on robust estimation in homoscedastic and independent case is abundant: one can cite L-estimation, M-estimation or the R-estimation (see Rousseeuw and Leroy (2005) for more details). The most documented robust estimation method is the M estimation: it consists in minimizing a loss function that is slowly varying for abnormal residuals instead of the squared residuals (Wang et al., 2007). This loss function is controlled by a tuning parameter c which "regulates the amount of robustness" (Huber, 1981). Recently, Wang et al. (2018) adapted this method to time series by accounting the temporal correlations in the estimation method. They also proposed a data-driven approach that allows one to choose the optimal "c" depending on the proportion of outliers in the data.

Contrary to the literature on robust estimation in homoscedastic case, the literature in heteroscedastic case is not abundant (Fellner, 1986; Zhao and Wang, 2009) and only few robust estimation methods are available. In 1982, Carroll and Ruppert (1982) adapted the M-estimation for the independent and heteroscedastic case. They proposed a method which estimates alternatively the mean parameter with a weighted M-estimation (with the variance as weight) and the variance parameter with a high-breakdown point estimator. Giltinan et al. (1986) also generalized homoscedastic GM-estimates to heteroscedastic regression. These GM-estimates are an extension of the M-estimates that, in addition to down-weight potential outliers in the response, also down-weights outliers in the covariate space during the estimation of the variance and mean parameters.

In this paper, we adapt the weighted M-estimation method of Carroll and Ruppert (1982) to time series by accounting for temporal correlations. Such as Wang et al. (2018), this procedure is data-dependent: the best tuning constant is chosen according to the proportion of outliers in the data. In the first part, every step of the procedure is detailed, then numerical studies are conducted to prove the efficiency of our method. Finally, we illustrate our methodology using data from a small tributary of the Thames River in Great Britain. We expect our method to be less responsive to outliers and simultaneously have the ability to cope with the problem of heteroscedasticity by robustly modelling it.

Statistical Model

The regression model

Let $y = (y_1,\ y_2,\ \dots\ , y_n)$ be the observed response measured over n equivalent time periods and $x_i\ = (x_{i1},\ \dots\ ,\ x_{ik})^T$ be the set of k associated predictors (x_{i1} is equal to 1 if the intercept is considered as a predictor). We assume the data are generated from the the following heteroscedastic linear model:

$$y_i = x_i^T \beta + \sigma_i \epsilon_i, \tag{1}$$

in which β is the vector collecting the parameters to be estimated, (ϵ_i) are the error terms following an autoregressive process of order p (AR(p)) which account for temporal correlations, and

$$\sigma_i = \phi g(x_i^T \beta, \gamma), \tag{2}$$

where $g(.)$ is a known function of the mean $(x^T\beta)$ with unknown parameter vector γ, and unknown dispersion parameter ϕ. Several choices for σ_i are possible, few examples are:

- $\sigma_i = \phi(1 + |x_i^T\beta|)^\gamma$ or $\sigma_i = \phi|x_i^T\beta|^\gamma$ (Box and Hill, 1974)
- $\sigma_i = \phi e^{\gamma x_i^T \beta}$ (Bickel et al., 1978)
- $\sigma_i = \gamma_1(x_i^T\beta) + \gamma_2(x_i^T\beta)^2$ (Bartlett, 1936)

The common approach to estimate the parameters for these heteroscedastic models is the generalized least square analysis (GLS). First, a model is fitted with the least square method. The variance parameters are then estimated thanks to the residuals with maximum likelihood method. Finally a weighted least square analysis is performed with the variance as weight $(w_i = \frac{1}{\sigma_i^2})$. This estimating approach leads to $\beta_{GLS} = (\sum x_i w_i x_i^T)^{-1}(\sum x_i w_i y_i)$.

Estimation of the parameters

In presence of outliers, the robust estimation procedure proposed by Carroll and Ruppert (1982) is desirable. This iterative method consists in robustly estimating the mean parameter by considering the variance parameter as fixed and vice versa.

For given or estimated value $\hat{\sigma}_i$, the robust M-estimate for β minimizes :

$$\sum \rho\left(\frac{y_i - x_i^T\beta}{\hat{\sigma}_i}\right), \tag{3}$$

where ρ is a loss function that is slowly varying for abnormal residuals (outliers). The most known is Huber's loss function :

$$\rho(u) = \begin{cases} \frac{1}{2}u^2 & \text{if } |u| \leq c \\ c|u| - \frac{1}{2}c^2 & \text{if } |u| > c \end{cases}. \tag{4}$$

Here, c is a tuning parameter chosen between 0 and 3 which controls the degree of robustness, the default value is $c = 1.345$ to ensure 95 % of efficiency when data are normally distributed. More example of loss functions can be found in Wang et al. (2018).
Taking derivatives of (4) leads to the following estimating equation of β:

$$U(\beta) = \sum_{i=1}^{n} \Big(\frac{x_i}{\hat{\sigma}_i}\Big)\psi\Big(\frac{y_i - x_i^T\beta}{\hat{\sigma}_i}\Big) = 0. \tag{5}$$

where $\psi(x) = min\{c, max\{x, -c\}\}$ is the derivative of Huber's loss function.

To solve this estimating function, one can rewrite $U(\beta)$ as :

$$U(\beta) = \sum_{i=1}^{n} x_i W_i r_i = 0, \tag{6}$$

where $W_i = \psi(r_i)/(r_i\hat{\sigma}_i^2)$ are weighting terms, and $r_i = (y_i - x_i^T\beta)/\hat{\sigma}_i$ are the pearson residuals. Now for a given weight W_i, the robust estimator of β can be obtained by the following formula :

$$\hat{\beta} = \{\sum_{i=1}^{n} x_i W_i x_i^T\}^{-1}\{\sum_{i=1}^{n} x_i W_i y_i\}. \tag{7}$$

An iterative approach is needed as W_i is a function of β and σ. This approach is derived of the pseudolikelihood approach and consists in fixing alternatively the parameters of the variance (γ and ϕ) and the regression parameters (β).

The variance parameters are also robustly estimated and are given by :

- A high breakdown estimator for γ :

$$\sum_{i=1}^{n} \chi\left(\frac{y_i - x_i^T\hat{\beta}}{\hat{\phi}g(x_i^T\hat{\beta},\gamma)}\right)\frac{g'(x_i^T\hat{\beta},\gamma)}{g(x_i^T\hat{\beta},\gamma)} = 0, \tag{8}$$

  where $\chi(.)$ is a bounded function. Croux (1994) and Bianco et al. (2000) suggest using $\chi(x) = min(x^2/c_1^2, 1) - 0.5$ with $c_1 = 1.041$ to obtain a 50% breakdown estimator of γ under the normality assumption.

- The MAD estimator for the dispersion parameter :

$$\hat{\phi} = \text{Median}\left\{\frac{|y_i - x_i^{\mathrm{T}}\hat{\beta}|}{g(x_i^T\hat{\beta},\hat{\gamma})}\right\}/0.6745. \tag{9}$$

Wang et al. (2007) showed that under some regularity conditions, the robust estimator $\hat{\beta}$ obtained by the iterative procedure is Fisher consistent. Moreover, when $n \to \infty$ the covariance matrix is given by (Huber et al., 1973; Huber, 1981; Carroll and Ruppert, 1982):

$$var(\hat{\beta}) = K^2\frac{[1/(n-k)]\sum\psi(r_i)^2}{[(1/n)\sum\psi'(r_i)]^2}\hat{\sigma}_i^2(x^tx)^{-1}, \tag{10}$$

where $r_i = (y_i - x_i^T\hat{\beta})/\hat{\sigma}_i$, and

$$K = 1 + \frac{p}{n}\frac{var(\psi'(r_i))}{(E\psi'(r_i))^2}. \tag{11}$$

A data-dependent tuning constant

The selection of the best value for the tuning parameter is an interesting problem: if the value is too small, too many observations will be considered as outliers and the efficiency will be lowered and if the value is too large, some outliers will be treated as normal observations and the estimators will be biased.

Such as Wang et al. (2018), we define the best tuning constant as the one which minimizes the variance of the regression parameters. Therefore we propose to do the estimating procedure with different values of c between 0 and 3 for the Huber's function (Wang et al., 2018) and choose the one which minimizes the sum of the variances of the regression parameters.

Accounting for temporal correlations

So far we only considered the independent model, we now need to incorporate the autoregressive process of order p present in the error terms. We write ϵ_i as $\sum_{j=1}^{p}(\alpha_j \epsilon_{i-j}) + \xi_i$ where ξ_i are independent errors and we rewrite the model (1) as :

$$y_i = x_i^T \beta + \sum_{j=1}^{p} \alpha_j \sigma_i \epsilon_{i-j} + \sigma_i \xi_i. \tag{12}$$

Because ϵ_i are unobserved, we propose to use the residuals from the initial model (1), say, $\hat{\varepsilon}_i$, and we now have the following linear model with roughly independent errors,

$$y_i = x_i^T \beta + \sum_{j=1}^{p} \alpha_j \hat{\sigma}_i \hat{\varepsilon}_{i-j} + \eta_i, \tag{13}$$

where $(\hat{\sigma}_i \hat{\varepsilon}_{i-1}, \hat{\sigma}_i \hat{\varepsilon}_{i-2}, ..., \hat{\sigma}_i \hat{\varepsilon}_{i-j},)$ are the augmented additional covariates including p lagged residuals, and $(\beta, \alpha_1, ..., \alpha_p)$ are the new parameters to be estimated including p lag parameters, $\hat{\sigma}_i$ is an estimate of σ_i. In the iterative procedure to be described below, this $\hat{\sigma}_i$ will be estimated from the variance function using the previous parameter estimates for (ϕ, γ, β). Here η_i represents the resulting error which should be close to $\sigma_i \xi_i$. We fit this augmented model with the optimal value of the tuning parameter to obtain the final estimate of β.
The order of the autoregressive process is determined by the ACF and PACF plots of the robustified pearson residuals of the initial model. In the application section, we will demonstrate how we choose p.

The estimation procedure

The complete estimation procedure is summarized in the following algorithm :

1. Obtain an initial robust estimate $\hat{\beta}_0$ assuming a constant variance $g(x^T\beta, \gamma) = 1$ with the default value of c (rlm function).

2. By fixing $\hat{\beta} = \hat{\beta}_0$, the robust variance parameters $(\hat{\phi}, \hat{\gamma})$ are estimated with (8) and (9) respectively.

3. By fixing the variance parameters equal to their robust estimates, we update $\hat{\beta}$ with (7).

4. Repeat steps 2 and 3 until desired convergence.

5. To find the best tuning parameter, do the steps 2-4 for several values of c between 0 and 3 and select the one which minimizes the sum of the variance of the regression parameters.

6. Once the best tuning parameter found, add the temporal correlation by following the procedure described previously.

Numerical Study

In this section, we investigate the perfomance of our procedure. We compare the mean square error of the estimates obtained by different methods such as least square (lm function in R), generalized least square method (gls function from the nlme package), Huber's method with $c = 1.345$ (rlm function from the MASS package), the weighted M-estimation with fixed tuning constant $c = 1.345$ and our data-driven method (accounting for the temporal correlations and choosing the best c).

For one simulation, we generate a multivariate normal dataset ($n = 1000$) using the model (1). In our case, $x_i^T\beta = \beta_0 + \beta_1 x_{i1}$ where $\beta_0 = \beta_1 = 10$ and x_1 comes from a uniform distribution on $(0, 1)$. For σ_i, we test two functions : the power function $\sigma_i = |x_i^T\beta|^\gamma$ with $\gamma = 0.5$ and the exponential function $\sigma_i = e^{\gamma|x_i^T\beta|}$ with $\gamma = 0.1$. For the term ϵ_i, we simulate an autoregressive process of order 1 with $\alpha = 0.8$. This autoregressive process can be written as follows : $\epsilon_i = 0.8\epsilon_{i-1} + \xi_i$ where ξ_i are independent and normally distributed errors following $N(0, 1)$. In order to add some outliers, the term ξ_i is randomly contaminated by $N(0, 8)$. Several contamination rates are considered : $\lambda = 0\%, 5\%, 10\%$.

We evaluate the relative efficiency (RE) of the different β estimators based on their mean square errors (MSE) using the least square method as a reference, that is, $RE = MSE(\hat{\beta}_{LS})/MSE(\hat{\beta})$. The larger RE value, the more efficient the estimator is (relative to the estimators obtained by the least square method). Boxplot of the simulations are presented in Appendix1.

Table 1 summarizes the results of 500 simulations. For both variance functions, the average value of the data-dependent tuning constant decreases as the contamination rate increases. This tuning constant has the expected behaviour : as the proportion of outliers becomes larger, more values should be considered as outliers therefore the tuning constant should be smaller. In both cases, the average value of α seems to digress from the true value as the contamination increases.

For the simulations with the exponential variance function, our method is less efficient comparing to the other methods in absence of contamination. However, our method outperforms all the other methods for the estimation of the mean parameters $(\hat{\beta}_0, \hat{\beta}_1)$ when

Table 1: Relative efficiency of the estimates obtained with several regression methods. The MSE of estimates obtained by the least square method are used as reference : a value larger than one indicates that the method is more efficient than LS estimation. $\bar{c}$ and $\bar{\alpha}$ are the average values estimated in the 500 simulations.

$\sigma_i = e^{\gamma\lvert x_i^T\beta\rvert}, \gamma = 0.1$, AR(1), $\alpha = 0.8$						
	$\lambda = 0\%$		$\lambda = 5\%$		$\lambda = 10\%$	
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
Huber's method with $c = 1.345$	0.9966	0.9717	1.8737	2.7203	1.5471	2.7006
Generalized least square with $\sigma_i = e^{\gamma\lvert x_i^T\beta\rvert}$	1.0042	1.2528	1.0241	1.2417	1.067	1.2975
Weighted M-estimation with $c = 1.345$	0.8094	1.0938	2.9436	3.0669	2.6296	2.708
Proposed method	0.8414	0.9157	3.4486	3.3828	2.9551	3.7975
	$\bar{c} = 2.32$	$\bar{\alpha} = 0.76$	$\bar{c} = 0.59$	$\bar{\alpha} = 0.59$	$\bar{c} = 0.38$	$\bar{\alpha} = 0.49$
$\sigma_i = \lvert x_i^T\beta\rvert^\gamma, \gamma = 0.5$, AR(1), $\alpha = 0.8$						
	$\lambda = 0\%$		$\lambda = 5\%$		$\lambda = 10\%$	
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
Huber's method with $c = 1.345$	0.9635	0.9307	2.1157	2.3603	1.692	2.2143
Generalized least square with $\sigma_i = \lvert x_i^T\beta\rvert\gamma$	0.9852	1.0174	0.9989	1.02	1.0026	1.037
Weighted M-estimation with $c = 1.345$	0.8449	0.8391	2.9402	2.6325	2.3856	2.4459
Proposed method	0.9937	1.3642	3.7137	4.3137	3.4632	5.1678
	$\bar{c} = 2.37$	$\bar{\alpha} = 0.77$	$\bar{c} = 0.6$	$\bar{\alpha} = 0.71$	$\bar{c} = 0.37$	$\bar{\alpha} = 0.67$

the data is contaminated, even with a small rate (5%). This difference in perfomance increases with the contamination rate, reaching values between 2 and 3 for the most contamined case.

With the power variance function, our method outperforms all the other methods for the estimation of the mean parameters even in the absence of contamination. This difference of efficiency also increases with the contamination rate, indicating that our method is well adapted to analyse contaminated datasets.

Case Study

Presentation of the data

In this section, we apply the proposed procedure to analyse the concentration of Chlorophyll a in the River Pang, a small tributary of the Thames River (UK).

Since the last decades, the River Thames basin is exposed to many growing pressures. In fact, the number of inhabitants in the basin increases rapidly each year, resulting in increased pollution loadings through waste water and increased water consumption. Unfortunately, these pressures and many others (intensive agriculture, global water usage) are likely to be exacerbated by the future climate change with the predicted increase of the extreme events such as droughts in the summer and floods in the winter (Bowes, 2017).

In this context, researchers decided to create the "The Thames Initiative research platform" which consists in weekly water quality monitoring for the River Thames (UK) and its major tributaries from March 2009 to February 2013. This initiative aims to detect any changes in water quality and to characterise aquatic ecology (phytoplankton communities) at the same frequency as water chemistry. The monitored parameters included nutrient fractions, anions, cations, metals, pH, alkalinity and Chlorophyll a. The data used in this thesis are available here (Centre for Ecology & Hydrology website) and a summary of the methods used to collect all the parameters can be found in the appendix 2. More details can be also found in the article of Bowes (2017).

Table 2: Summary table of all the parameters monitored weekly at the River Pang Site from October 10, 2009 to February 25, 2013 (202 observations). The NA have been excluded from all the statistical analysis (mean, standard deviation and median).

	Mean	Sd	Median	Min	Max	Number of NA
Chlorophyll a ($\mu g l^{-1}$)	2.8	4.28	2.04	0.44	50.73	2
Water Temperature ($^{\circ}C$)	10.82	3.66	11.05	0.9	18.9	6
pH	7.9	0.11	7.9	7.55	8.42	2
Suspended solids($mg.l^{-1}$)	6.22	3.75	5.03	2.21	25.67	4
Mean daily river discharge ($m^3 s{-}1$)	0.53	0.37	0.4	0.17	2.54	0
Total phosphorus ($\mu g l^{-1}$)	49.41	29.92	45	12	281	1
Ammonium ($mg.l^{-1}$)	0.04	0.03	0.04	0.001	0.17	9
Nitrite ($mg.l^{-1}$)	0.05	0.1	0.03	0	0.84	33
Nitrate ($mg.l^{-1}$)	28.05	2.67	27.93	19.39	37.05	1
Potassium ($mg.l^{-1}$)	2.87	1.62	2.5	1.7	16.5	1
Sodium ($mg.l^{-1}$)	12.13	1.51	11.9	8.8	24.2	1
Chloride ($mg.l^{-1}$)	24.58	3.35	23.61	18.6	50.06	1
Sulphate ($mg.l^{-1}$)	19.29	4.36	18.37	16	73	1
Calcium ($mg.l^{-1}$)	107.81	10.17	110.7	45.6	127.7	1
Magnesium ($mg.l^{-1}$)	3.21	0.43	3.1	2.4	4.6	1
Fluoride ($mg.l^{-1}$)	0.13	0.03	0.14	0	0.32	1

The data collected by this program has already been used to lead several studies on nutrient dynamics (Bowes et al., 2015), nutrient sources (Bowes et al., 2014) and phytoplankton dynamics (Bowes et al., 2012), etc. Our method can be particularly useful in modelling

the phytoplankton dynamics represented by the concentration of the Chlorophyll a. Some outliers may be present and the variance of the Chlorophyll concentration is not likely to be constant.

Out of the 23 monitored sites, we selected the site "River Pang at Tidmarsh" due to the weak presence of outliers in the covariate space and the presence of heteroscedasticy in the residuals after applying the least square model described in (14).

For the data on this site, all the concentrations of Phosphorus, Ammonium and Nitrites under the detection limit such as '$<$ 7', '$<$ 0.004' or '$<$ 0.1' are considered as not available (NA) as we do not know their real concentrations. After applying the linear model in (14), two outliers in the covariate space have been detected on the 19/09/2011 and 11/06/2012 (Suspended solids $>$ 100). We chose to exclude them from the analysis since our method is sensitive to outliers in the covariate space. The summary of all the variables are presented in Table 2. Hereafter, we only consider the weeks for which we had all the parameters available (162/202 weeks) for all the regression analysis and figures.

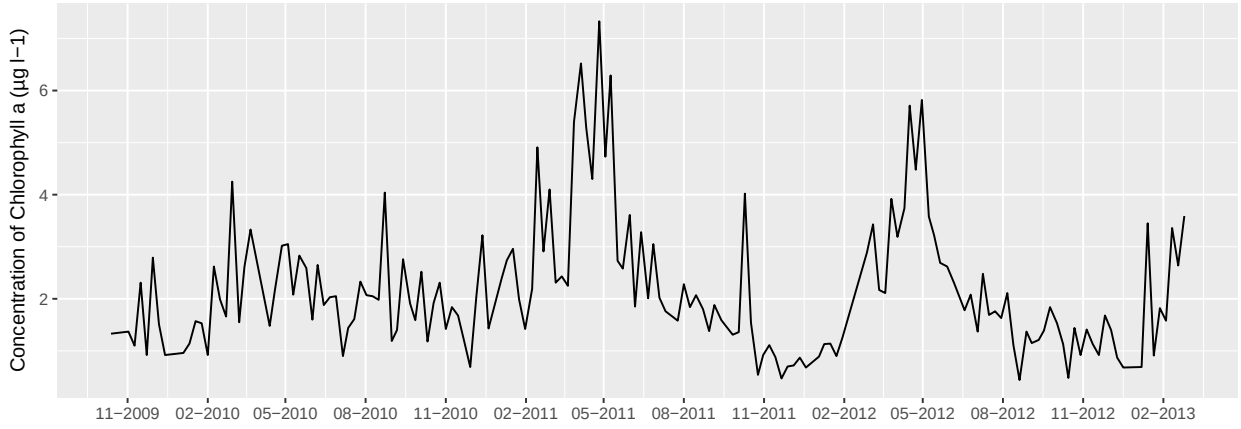


Figure 1: Time series plot of the observed concentration of Chlorophyll a in the River Pang Site from October 10, 2009 to February 25, 2013 (162 observations).

The time series of the concentration of Chlorophyll a for the River Pang is presented in the Figure 1. The concentration of the Chlorophyll seems to show some seasonality with higher values for the spring and lower values for the winter. There are two different blooms of algae during the period October 2009 – February 2013 represented by the two peaks. As expected, the blooms are situated in the Spring season, the most favourable time of the year for algae growth in this area (Bowes et al., 2012) due to the rising temperature.

Analysis

Least square method

In order to explain the chlorophyll concentration, we use all the covariates in our disposition. Each of these variables plays a key role in the development of the phytoplankton, it is therefore interesting to see which one will have a greater effect on the Chlorophyll a and

to see if their associated regression parameters will change significantly depending on the regression method used (Least square or our method). The regression model is the following:

$$\begin{aligned}\text{Chlorophyll } a = \beta_0 + \beta_1 \text{Temperature} + \beta_2 \text{pH} + \beta_3 \text{Suspended solids} + \beta_4 \text{Total phosphorus} \\ + \beta_5 \text{Ammonium} + \beta_6 \text{Fluoride} + \beta_7 \text{Nitrite} + \beta_8 \text{Nitrate} + \beta_9 \text{Sulphate} \\ + \beta_{10} \text{Potassium} + \beta_{11} \text{Calcium} + \beta_{12} \text{Mean daily river discharge} + \varepsilon,\end{aligned} \tag{14}$$

where ε is the error term. We choose not to add seasonal terms such as cosinus and sinus functions as the temperature and mean daily river flow are already highly seasonal.

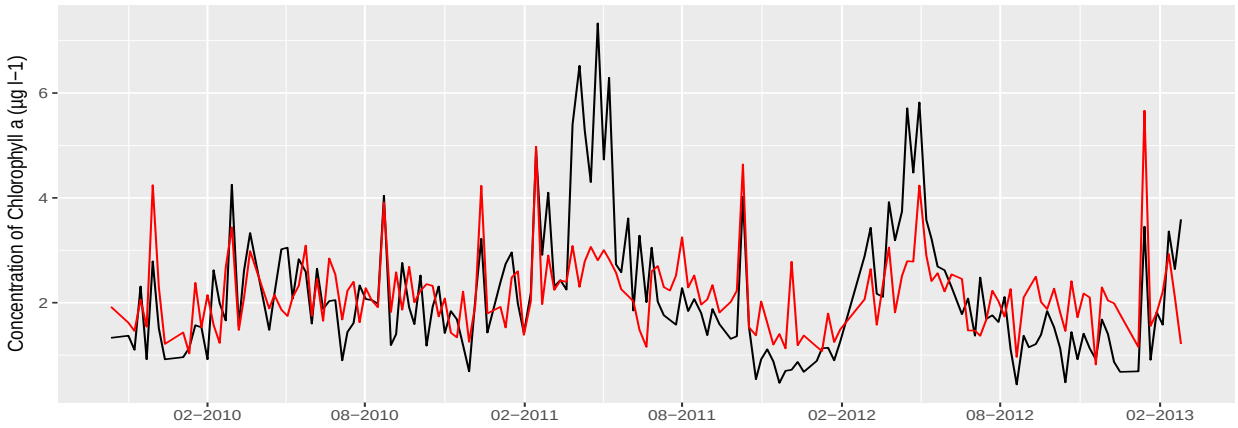
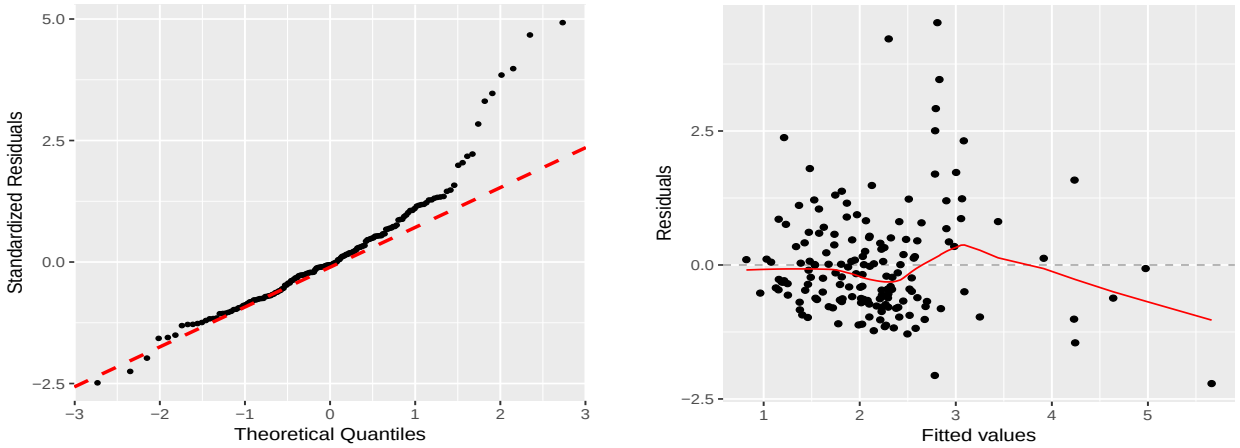


Figure 2: Time series plot of the observed concentration of Chlophyll a (in black) with predicted curves from the least square model (red line) from October 10, 2009 to February 25, 2013.

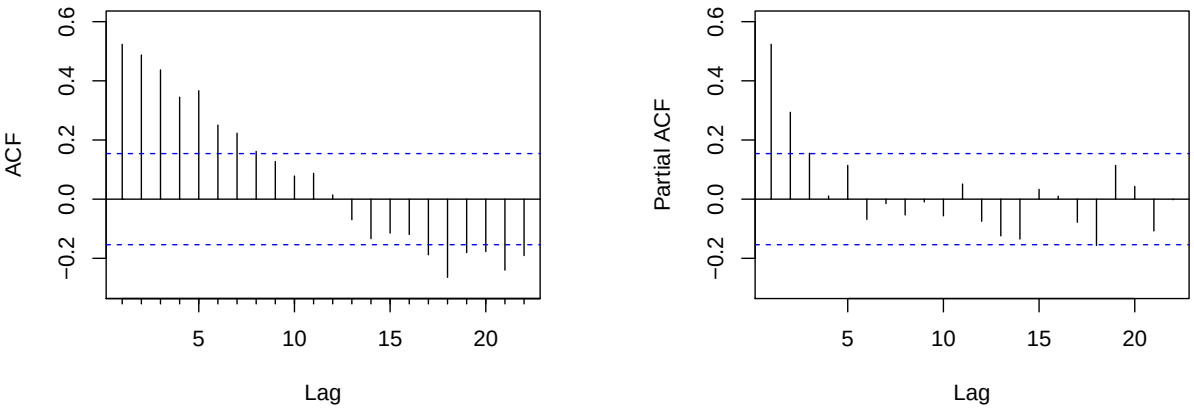
This model is fitted with the lm function from R statistical software. As it can be seen in Figure 2, the predicted values obtained from the least square analysis are not really close to the real observed concentrations. This lack of fit can be explained with the regression diagnostic plots.
Indeed, the normal probability plot in Figure 3 (a) shows that the distribution of the residuals is skewed, indicating the presence of outliers which may have influenced the estimation of the regression parameters. The residuals vs. fitted value plot (Figure 3 (b)) seems to indicate a slight heteroscedasticity with larger residuals as the fitted values increase. This heteroscedasticity does not lead to biaised estimators but to estimators with biaised covariance matrix. This could result in the underestimation of the standard errors, erroneous Z-values and therefore erroneous hypothesis tests.



(a) (b)

Figure 3: (a) Q-Q plot of the residuals of the least square model, the red line is the default qqline (R function) which passes through the first and third quartiles. (b) Residuals vs. fitted values plot for the least square model, the red line is a locally weighted scatterplot smoothing (R function: lowess(f = 2/3, iter = 3)), commonly used in the Residuals vs. fitted values plot.

In addition to the presence of outliers and heteroscedastocity, the ACF and pACF plot (Figure 4 (a) and (b)) indicate the presence of temporal correlations in the residuals of the least square model. The partial autocorrelation function plot (Figure 4 (b)) suggests a possible AR(2) model.
In this present case, our method is highly desirable as we have the presence of heteroscedasticity, temporal correlations and outliers.



(a) (b)

Figure 4: (a) Auto correlation function plot (AFC) and (b) partial Auto correlation function plot of the residuals of the least square model. The blue dashed lines represent the 95% confidence interval. If the values are beyond this line, the autocorrelations are statistically different from zero.

Our method

In the literature, we did not find works or indications on how to model the variance of the chlorophyll concentration, consequently, we use a common variance function : $\sigma_i = \phi|x_i^T\beta|^\gamma$ for the analysis.
First, our method chooses the best tuning constant. Figure 5 plots the sum of the variance of the regression parameters for a range of c values between 0 and 3. For our data, the value of the tuning parameter that minimizes the variance of the estimators is found around 1.

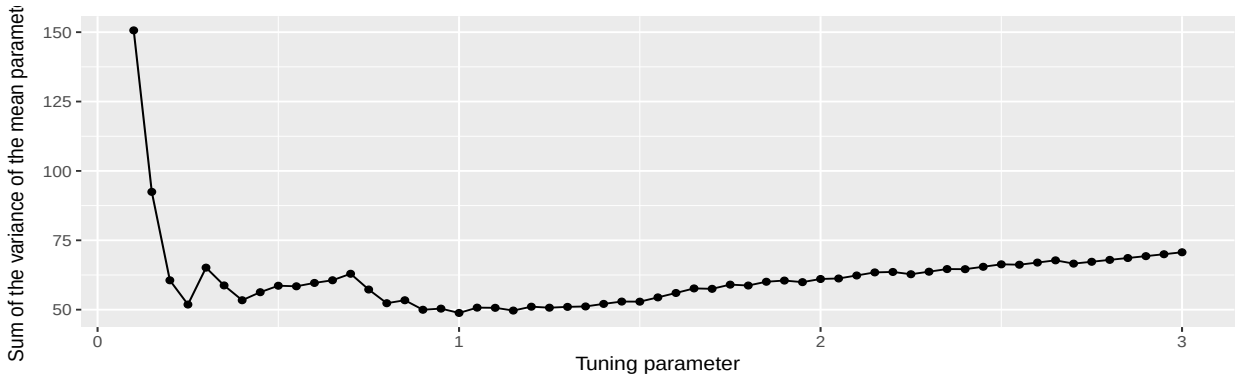
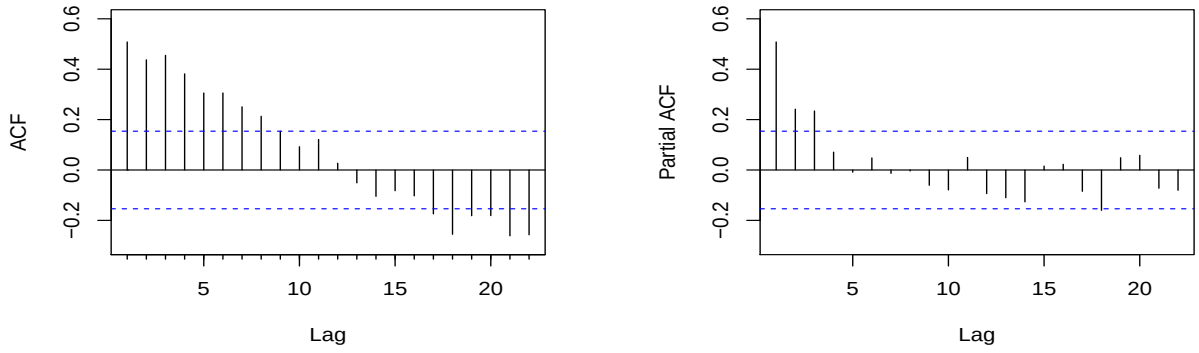


Figure 5: Sum of the variance of the estimators for different values of the tuning parameter (c between 0 and 3).

Then, we account for temporal correlations. Figure 6 shows the ACF and pACF plots of the robustified residuals after using the best c and before adding temporal correlations. The robustified residuals are defined as : $r_i = \psi(\frac{y_i - x_i^T\hat{\beta}}{\hat{\sigma}_i})$.



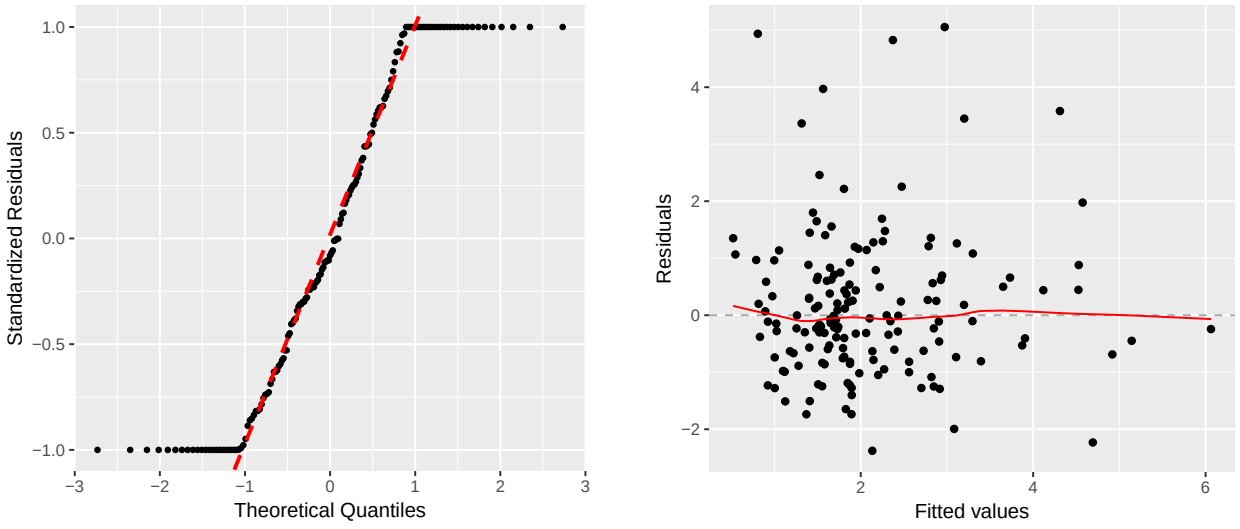
(a) (b)

Figure 6: (a) Auto correlation function plot (AFC) and (b) partial auto correlation function plot of the residuals of the proposed method with best tuning parameter ($\hat{c} = 1$) and before adding lagged terms. The blue dashed lines represent the 95% confidence interval. If the values are beyond this line, the autocorrelations are significantly different from zero.

The pACF plot suggests an AR(3), therefore, we consider that $p = 3$ and add three lagged residuals to the model (14) resulting in:

$$\begin{aligned}\text{Chlorophyll } a = \beta_0 + \beta_1 \text{Temperature} + \beta_2 \text{pH} + \beta_3 \text{Suspended solids} + \beta_4 \text{Total phosphorus} \\ + \beta_5 \text{Ammonium} + \beta_6 \text{Fluoride} + \beta_7 \text{Nitrite} + \beta_8 \text{Nitrate} + \beta_9 \text{Sulphate} \\ + \beta_{10} \text{Potassium} + \beta_{11} \text{Calcium} + \beta_{12} \text{Mean daily river discharge} + \alpha_1 \ \hat{\sigma}_i \hat{\varepsilon}_{i-1} \\ + \alpha_2 \hat{\sigma}_i \hat{\varepsilon}_i + \alpha_3 \hat{\sigma}_i \hat{\varepsilon}_{i-3} + \eta,\end{aligned} \tag{15}$$

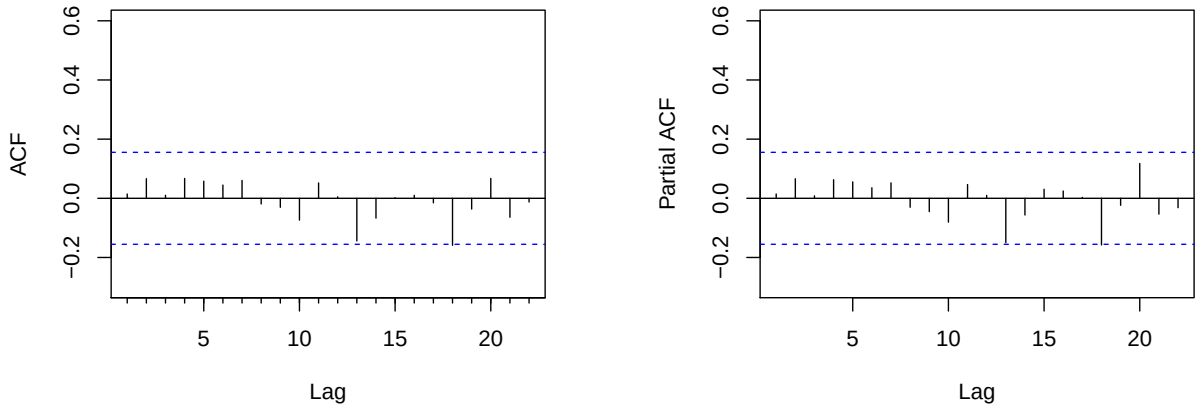
where $\hat{\sigma}_i \hat{\varepsilon}_{i-1}, \hat{\sigma}_i \hat{\varepsilon}_{i-2}, \hat{\sigma}_i \hat{\varepsilon}_{i-3}$ are lagged terms built from the initial model (14) estimated with the best tuning constant ($\hat{c} = 1$) and η are the assumed independent error terms. The term $\hat{\varepsilon}$ corresponds to the pearson residuals of the initial model and $\hat{\sigma}$ is the estimated variance fonction.



(a) (b)

Figure 7: (a) Q-Q plot of the residuals of the proposed method, the red line is the default qqline (R function) which passes through the first and third quartiles. (b) Residuals vs. fitted values plot for the method, the red line is a locally weighted scatterplot smoothing (R function: lowess(f = 2/3, iter = 3)).

The normal probability plot (Figure 3 (a)) of the residuals of the full model (15) clearly shows that our robust procedure has taken care of outliers successfully. The residuals vs. fitted values plot (Figure 3 (b)) seems to indicate that the slight heteroscedasticy has been corrected. The ACF and pACF plot (Figure 8) demonstrates that the temporal correlations in the residuals have been well accounted by our robust method.



(a) (b)

Figure 8: (a) Auto correlation function plot (AFC) and (b) partial auto correlation function plot of the residuals of the proposed method. The blue dashed lines represent the 95% confidence interval. If the values are beyond this line, the autocorrelations are significantly different from zero.

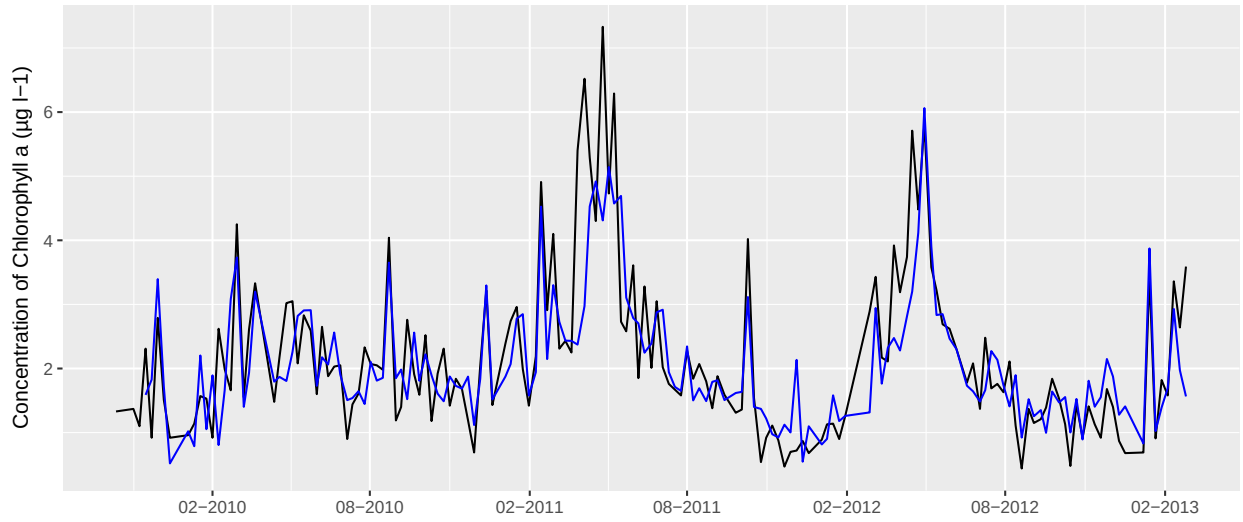


Figure 9: Time series plot of the observed concentration of Chlophyll *a* (in black) with predicted curves from the proposed method (blue line).

Finally, the predicted values obtained with our method are much closer to the observed concentrations of chlorophyll *a* (Figure 9).

Comparison with other regression methods

The results of both methods are listed in Table 3. In this table, the regression parameter of the least square method have been estimated with the lm function and

their associated covariance matrix with the vcovHAC function (sandwich package) which gives a heteroskedasticity and autocorrelation consistent estimation of the covariance. This estimation (HAC) was necessary to obtain the corrected standard errors as the residuals exhibited heteroscedasticity and temporal correlations. For our procedure, if $n \to \infty$ we know that the distribution of the regression parameter β is normal. However we do not know the distribution of the β when n is a finite sample(our case). Therefore we only present the z-values, and the significativity (z-values in bold) is just an indication and is based on the hypothesis that the β follows a normal distribution.

Table 3: Parameter estimates (β), their standard errors and z-values for the least square and the robust method. The standard errors for the least square model come from the heteroskedasticity and autocorrelation consistent estimation of the covariance matrix performed with the vcovHAC() function from sandwich package. The critical region of the significance test is $|z| > 1.96$ (significant z-values in bold). The efficiency is calculated as $\sum var\hat{\beta_{LS}} / \sum var\hat{\beta}_{met}$.

	LM			Our robust method with $\hat{c} = 1$ and $\hat{\sigma}_i = 0.41\|x^T\hat{\beta}\|^{0.5}$		
	Estimate	Std.Error	z-value	Estimate	Std.Error	z-value
Intercept	-19.8571	9.4175	**-2.1085**	-10.4085	4.5022	**-2.3119**
Temperature	0.0636	0.0339	1.8776	0.0321	0.0203	1.5825
pH	2.6412	1.1145	**2.3698**	1.5841	0.5445	**2.9091**
Suspended solids	0.1944	0.0515	**3.7780**	0.1415	0.0210	**6.7406**
Total dissolved phosphorus	-0.0093	0.0064	-1.4523	0.0045	0.0033	1.3656
Ammonium	-9.4776	4.4073	**-2.1504**	-4.1006	2.3248	-1.7639
Fluoride	-1.1960	1.7150	-0.6974	-0.8302	1.8525	-0.4481
Nitrite	1.6712	0.7040	**2.3739**	0.3028	0.6304	0.4803
Nitrate	0.0798	0.0635	1.2559	0.0406	0.0352	1.1536
Sulphate	0.0223	0.0482	0.4620	-0.0564	0.0336	-1.6768
Potassium	0.1043	0.0379	**2.7507**	0.0954	0.0414	**2.3038**
Calcium	-0.0250	0.0272	-0.9181	-0.0148	0.0127	-1.1658
Mean daily rive discharge	-0.2145	0.4012	-0.5346	-0.2291	0.2025	-1.1313
Lag 1	/	/	/	0.4190	0.0780	**5.3702**
Lag 2	/	/	/	0.1831	0.0808	**2.2662**
Lag 3	/	/	/	0.1774	0.0781	**2.2716**
				Efficiency = 3.78		

Our estimation method drastically reduced the variance of the parameters. Our robust regression is found to be 3.78 times more efficient than the usual linear regression model. Some estimators are very different between the two methods: few examples are the estimators for the intercept, Ammonium and Nitrite concentration. It is worth noting that the Ammonium and Nitrite concentration parameters are significantly different from 0 in the

least square method and not in our robust method. Finally, we can see in the Table 3 that all the lagged terms are significant. This indicates that the three previous terms contribute significantly to the output and are therefore necessary to consider.
In order to compare the efficiency of our procedure, we analysed the Chlorrophyll a concentration data with several regression methods. The Table 4 regroups three indicators for the different methods which are the sum of the variance of the estimators, the mean absolute error and the root mean square error. The first one is related to the efficiency of the method and the two last ones give some indications about the fit of the methods.

Table 4: Sum of the variance of the estimators, Mean absolute error, Root mean square error for different methods applied to the Chlorophyll a data. The method presented in this thesis is in bold.

	Efficiency	Mean absolute error	Root mean square error	Number of lags
Least square model	112.97	0.740	1.016	0
Generalized least square ($\sigma_i = \lvert x^T\beta\rvert^\gamma\rvert$)	43.73	0.753	1.067	0
Robust least square via M-estimation (c = 1.345)	99.91	0.717	1.030	0
Weighted M-estimation (c = 1.345, $\sigma_i = \phi\lvert x^T\beta\rvert^\gamma\rvert$)	93.63	0.719	1.042	0
Weighted M-estimation (c = 1.345, $\sigma_i = \phi e^{\gamma\lvert x^T\beta\rvert}\rvert$)	53.37	0.753	1.065	0
Least square model with lagged residuals	65.55	0.572	0.784	2
Generalized least square with lagged residuals ($\sigma_i = \lvert x^T\beta\rvert^\gamma\rvert$)	31.91	0.580	0.802	3
Robust least square via M-estimation with lagged residuals (c = 1.345)	49.90	0.534	0.768	3
Our proposed method ($\sigma_i = \phi\lvert x^T\beta\rvert^\gamma\rvert$ and best c)	29.86	0.528	0.772	3
Our proposed method ($\sigma_i = \phi e^{\gamma\lvert x^T\beta\rvert}\rvert$ and best c)	40.87	0.568	0.870	3

The most efficient methods are the generalized least square with lagged values and our procedure with the power variance function. They are on average twice as efficient than all the other regression methods. The methods that best fit the data, are the robust least square estimation with lagged values and our proposed procedure with the power function. Both of these methods have the lowest MAE and RMSE. For the Chlorophyll a data, the method that seems the most adapted is our proposed procedure with the power variance function as it combines low variance estimators and low MAE and RMSE.

Conclusion

We adapted the weighted M-estimation to time series. This procedure is data-dependent and allows one to choose the tuning constant that minimizes the variance of the estimators and it incorporates the temporal correlations by adding lagged terms in the covariates. The numerical study showed that this procedure outperforms the other regression methods when the data are contaminated. In fact, it gives more efficient estimates for both the mean and variance parameters. In the application with the chlorophyll a concentration dataset, we

proved that our procedure results in estimates with significantly lower variance compared to the ones obtained by least square estimation. The application of this procedure is highly desirable when time series are exhibiting heterogeneity and outliers, as it gives more reliable parameter estimates which leads to more efficient hypothesis testing.
This procedure could be improved in several ways. For example, it could be interesting to help the user, choosing the right variance funtion based on some indicators such as the AIC, BIC. Robust hypothesis testing as described in Zhao and Wang (2009) must be implemented instead of relying on the hypothesis that the mean parameter estimates for our method, follow a normal distribution. Finally, this method could be adapted to model time-series of counts by using a link function such as the generalized linear model (McCullagh and Nelder, 1989).

References

Bartlett, M. (1936). Some notes on insecticide tests in the laboratory and in the field. *Supplement to the Journal of the Royal Statistical Society*, 3(2):185–194.

Bianco, A., Boente, G., and Di Rienzo, J. (2000). Some results for robust gm-based estimators in heteroscedastic regression models. *Journal of Statistical Planning and Inference*, 89(1-2):215–242.

Bickel, P. J. et al. (1978). Using residuals robustly i: Tests for heteroscedasticity, nonlinearity. *The Annals of Statistics*, 6(2):266–291.

Bowes, M.J.;Armstrong, L. H. S. E. C. P. (2017). Weekly water quality data from the river thames and its major tributaries (2009-2013) [ceh thames initiative].

Bowes, M., Gozzard, E., Johnson, A., Scarlett, P., Roberts, C., Read, D., Armstrong, L., Harman, S., and Wickham, H. (2012). Spatial and temporal changes in chlorophyll-a concentrations in the river thames basin, uk: Are phosphorus concentrations beginning to limit phytoplankton biomass? *Science of the Total Environment*, 426:45–55.

Bowes, M., Jarvie, H., Halliday, S., Skeffington, R., Wade, A., Loewenthal, M., Gozzard, E., Newman, J., and Palmer-Felgate, E. (2015). Characterising phosphorus and nitrate inputs to a rural river using high-frequency concentration–flow relationships. *Science of the Total Environment*, 511:608–620.

Bowes, M. J., Jarvie, H. P., Naden, P. S., Old, G. H., Scarlett, P. M., Roberts, C., Armstrong, L. K., Harman, S. A., Wickham, H. D., and Collins, A. L. (2014). Identifying priorities for nutrient mitigation using river concentration–flow relationships: The thames basin, uk. *Journal of Hydrology*, 517:1–12.

Box, G. E. and Hill, W. J. (1974). Correcting inhomogeneity of variance with power transformation weighting. *Technometrics*, 16(3):385–389.

Carroll, R. and Ruppert, D. (1988). *Transformation and weighting in regression*. Chapman & Hall, Ltd.

Carroll, R. J. and Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. *The annals of statistics*, pages 429–441.

Croux, C. (1994). Efficient high-breakdown m-estimators of scale. *Statistics & Probability Letters*, 19(5):371–379.

Davidian, M. and Carroll, R. J. (1987). Variance function estimation. *Journal of the American Statistical Association*, 82(400):1079–1091.

Fellner, W. H. (1986). Robust estimation of variance components. *Technometrics*, 28(1):51–60.

Giltinan, D. M. (1983). *Bounded influence estimation in heteroscedastic linear models*. PhD thesis, Citeseer.

Giltinan, D. M., Carroll, R. J., and Ruppert, D. (1986). Some new estimation methods for weighted regression when there are possible outliers. *Technometrics*, 28(3):219–230.

Huber, P. (1981). *Robust Statistics. New York, Chichester, Brisbane*. Toronto, Singapore: John Wiley & Sons, Ltd.

Huber, P. J. et al. (1973). Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1(5):799–821.

Jiang, Y., Wang, Y.-G., Fu, L., and Wang, X. (2018). Robust estimation using modified huber's functions with new tails. *Technometrics*, (just-accepted):1–32.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC press.

Rousseeuw, P. J. and Leroy, A. M. (2005). *Robust regression and outlier detection*, volume 589. John wiley & sons.

Ruppert, D. and Carroll, R. J. (1985). *Data Transformations in Regression Analysis with Applications to Stock—Recruitment Relationships*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Stefanski, L. A., Carroll, R. J., and Ruppert, D. (1986). Optimally hounded score functions for generalized linear models with applications to logistic regression. *Biometrika*, 73(2):413–424.

Wang, N., Wang, Y.-G., Hu, S., Hu, Z.-H., Xu, J., Tang, H., and Jin, G. (2018). Robust regression with data-dependent regularization parameters and autoregressive temporal correlations. *Environmental Modeling & Assessment*, pages 1–8.

Wang, Y.-G., Lin, X., Zhu, M., and Bai, Z. (2007). Robust estimation using the huber function with a data-dependent tuning constant. *Journal of Computational and Graphical Statistics*, 16(2):468–481.

Zhao, J. and Wang, J. (2009). Robust testing procedures in heteroscedastic linear models. *Communications in Statistics—Simulation and Computation®*, 38(2):244–256.

Appendix

Appendix 1: Boxplot of the numerical study results

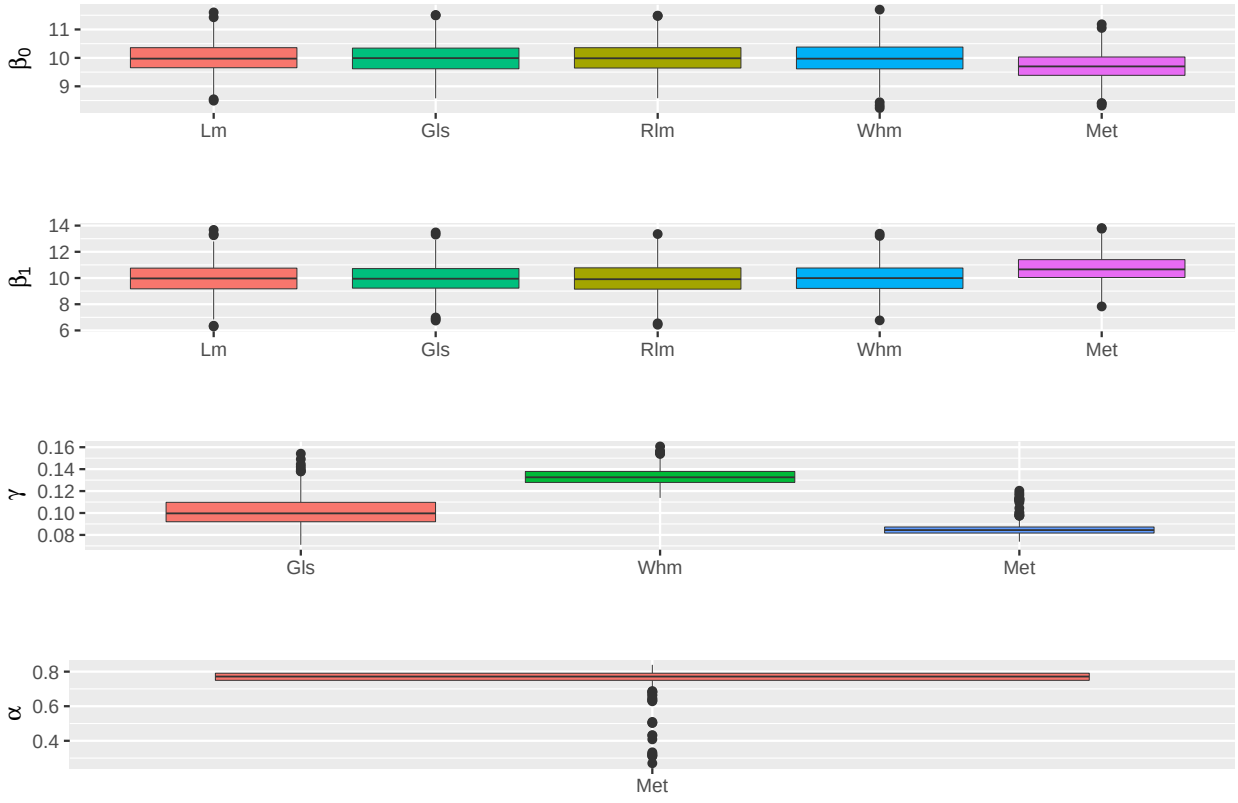


Figure A.1: Boxplot of the results of the simulation with the exponential variance function and $\lambda = 0\%$

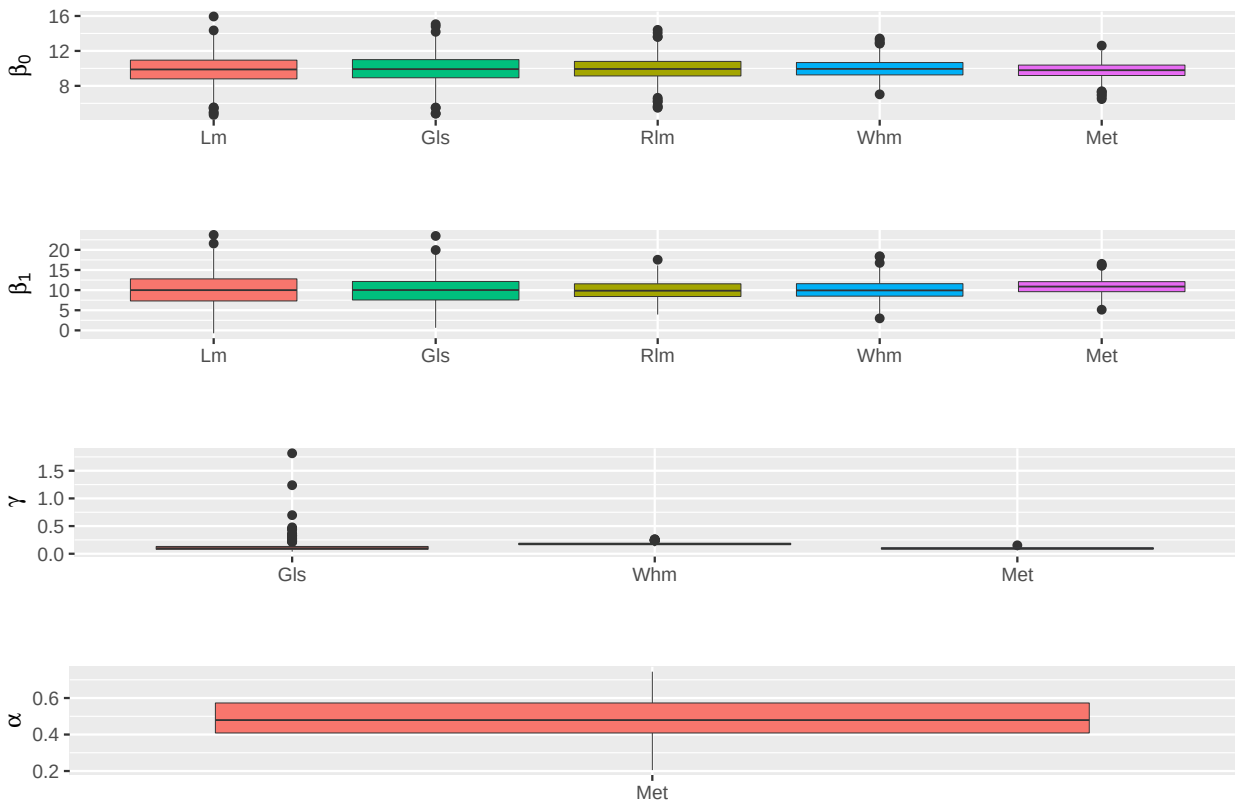


Figure A.2: Boxplot of the results of the simulation with the exponential variance function and $\lambda = 5\%$

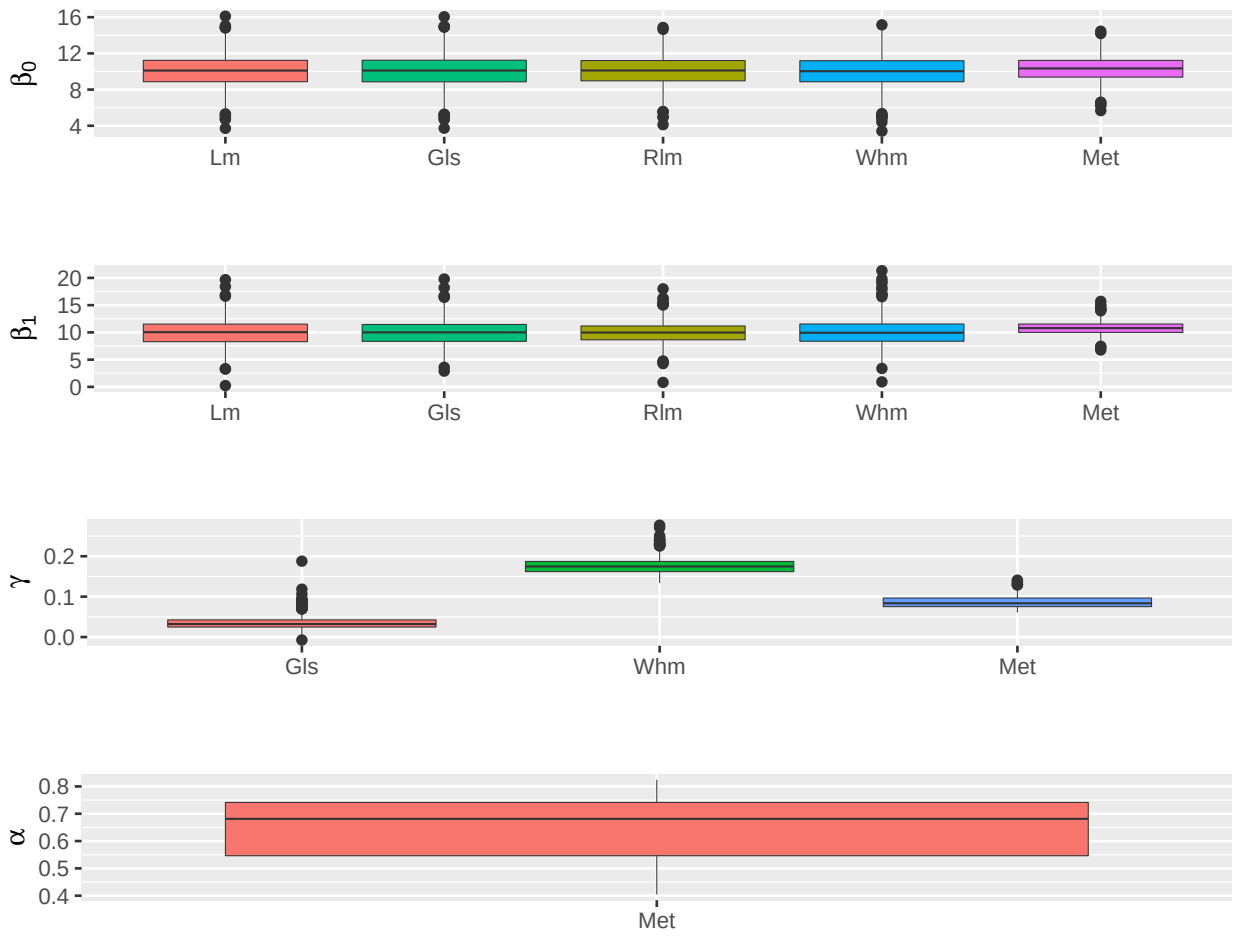


Figure A.3: Boxplot of the results of the simulation with the exponential variance function and $\lambda = 10\%$

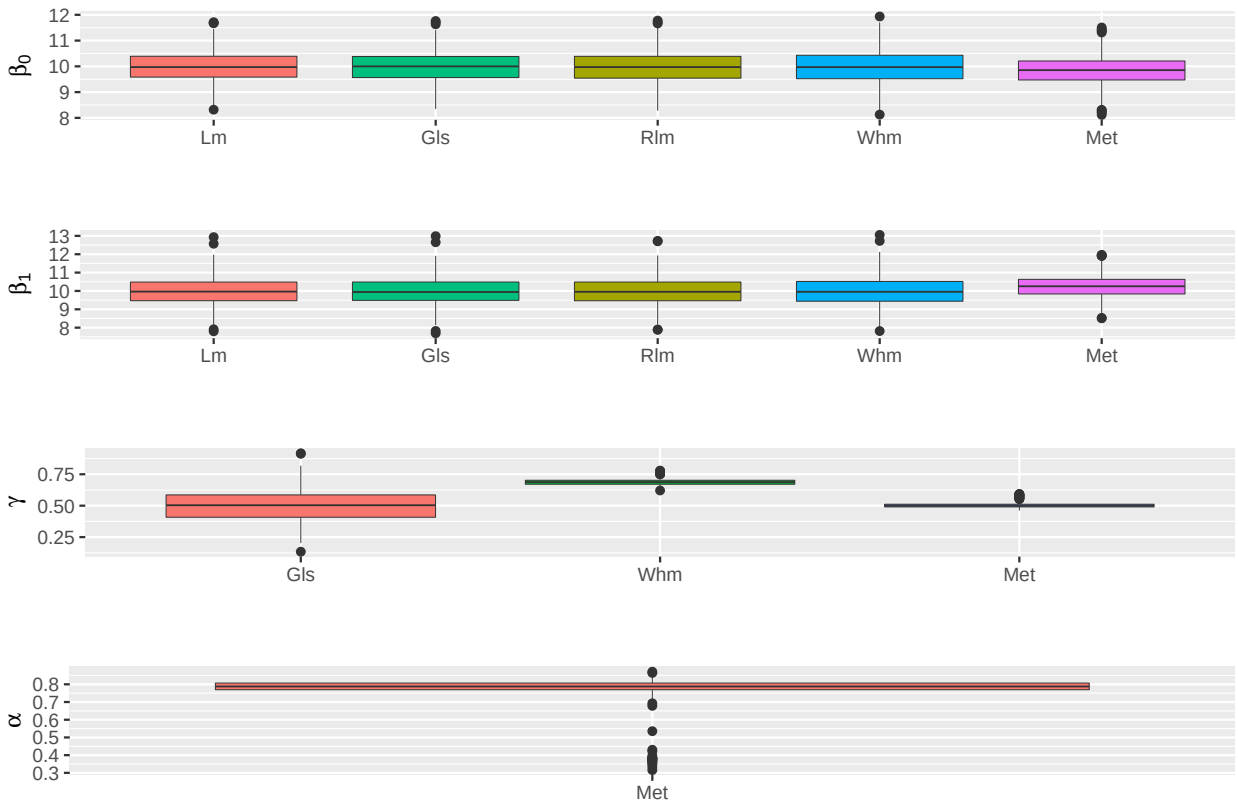


Figure A.4: Boxplot of the results of the simulation with the power variance function and $\lambda = 0\%$

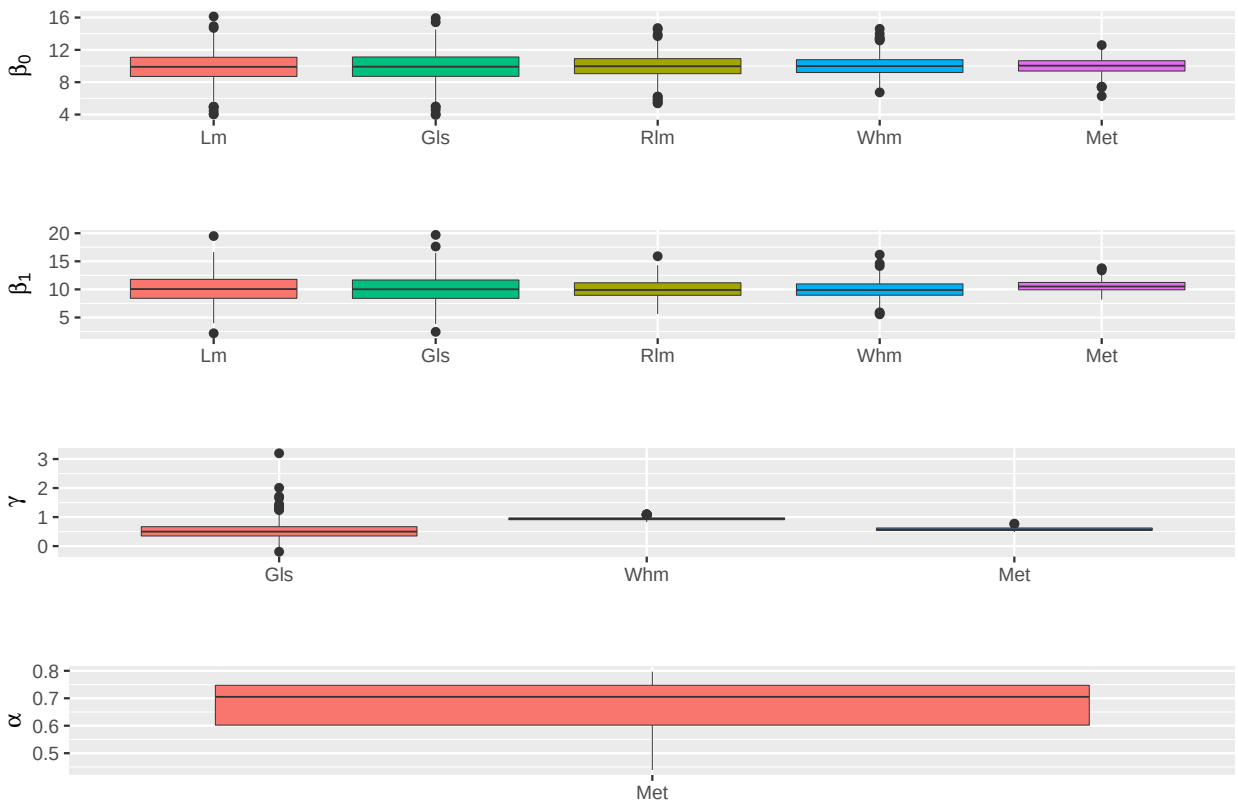


Figure A.5: Boxplot of the results of the simulation with the power variance function and $\lambda = 5\%$

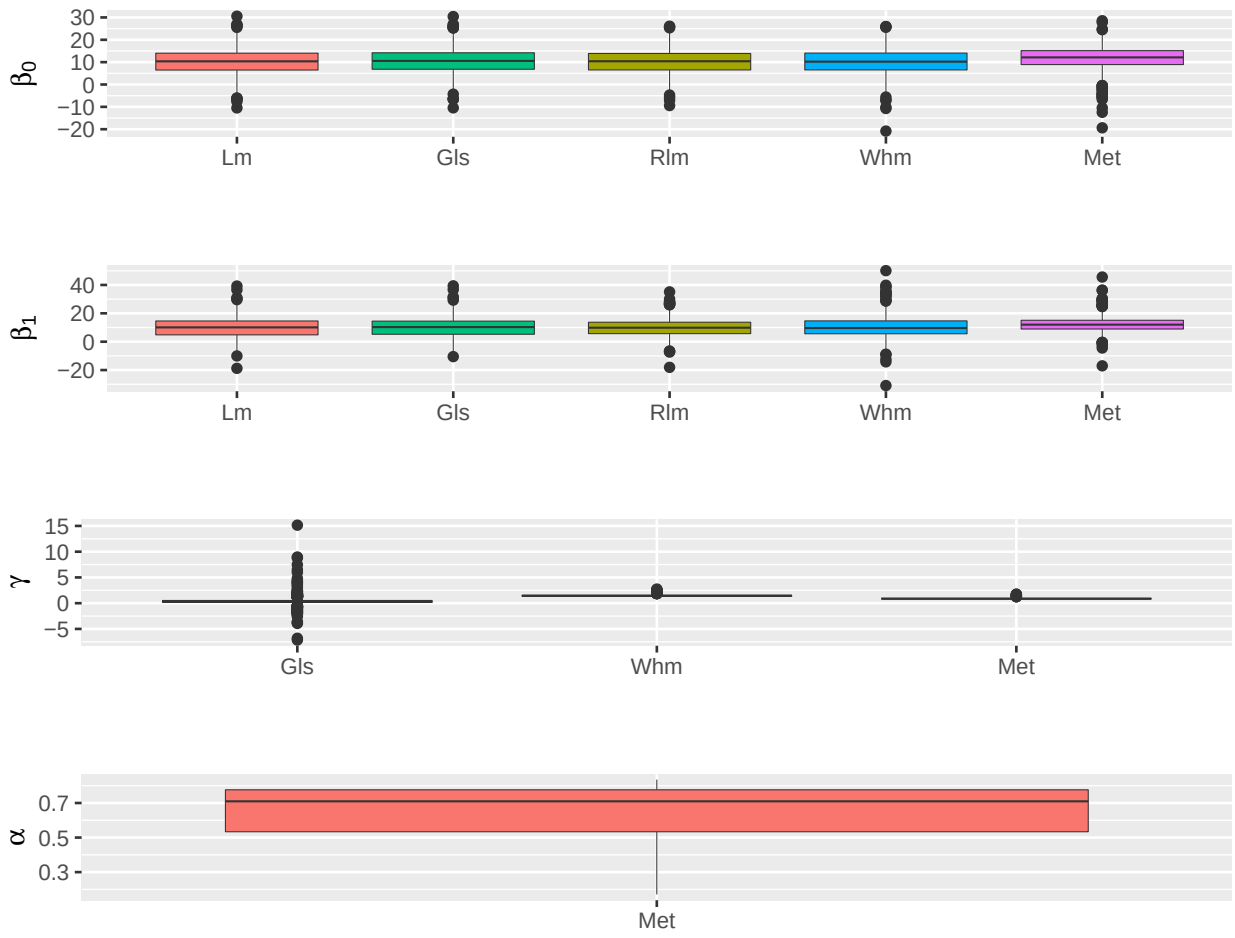


Figure A.6: Boxplot of the results of the simulation with the power variance function and $\lambda = 10\%$

Appendix 2 : Supporting information on the dataset

The following information were furnished with the dataset, all credits go to the following webpage : https://catalogue.ceh.ac.uk/documents/e4c300b1-8bc3-4df2-b23a-e72e67eef2fd and its authors (Bowes et al., 2017).

Brief description of the dataset

This data set comprises of weekly water quality monitoring data of seven sites along the River Thames, UK, and fifteen of its major tributaries from March 2009 to February 2013. Parameters measured were phosphorus and nitrogen species, dissolved reactive silicon, water temperature, pH, Gran alkalinity, suspended solids, chlorophyll and major dissolved anions (fluoride, chloride, bromide, sulphate) and cations (sodium, potassium, calcium, magnesium, boron). Dissolved and total iron, manganese, zinc, copper concentrations have also been produced from August 2010 to February 2013. The accompanying daily river flow data are also supplied. Samples were taken as part of the Centre for Ecology & Hydrology's Thames Initiative monitoring programme.

Monitoring and analytical information

Bulk samples were taken from the main flow of each river on Monday or Tuesday of each week. Subsamples were filtered immediately in the field through a 0.45 μm Whatman WCN membrane filter. On return to the laboratory, all samples were stored in the dark at 4oC, prior to analysis.

The pH was determined using a Radiometer Analytical PHM210 pH meter. The instrument was calibrated prior to use using pH 4, 7, and 10 buffer solutions traceable to NIST.

Gran alkalinity was determined by acidimetric titration to pH 4 and 3 using 0.5N H2SO4.

Suspended solids concentrations were determined by filtering a known volume (approximately 500 ml) of river water through a pre-dried Whatman GF/C filter paper. The filter paper was then re-dried (16 h at 80oC) and reweighed to determine the mass of solids in the water sample.

Chlorophyll concentrations were determined by filtering a known volume of unfiltered river water (approximately 500 ml) through a Whatman GF/C filter paper. The filter paper was then extracted in 10 ml of 90% v/v acetone/water and refrigerated overnight in the dark. Chlorophyll-a concentration was determined colorimetrically using a Beckman 750 DU spectrophotometer, using the method of Marker et al. (1980).

Total phosphorus (TP)and total dissolved phosphorus (TDP) were determined by digesting an unfiltered and 0.45 μm filtered water sample (respectively) with acidified potassium persulphate in an autoclave at 121 degree Celsius for 40 min. Acidified ammonium molybdate reagent was then added to the digested samples to produce a molybdenum–phosphorus complex. This intensely blue-coloured compound was then quantified spectrophotometrically at 880 nm (Eisenreich et al. 1975).

Soluble reactive phosphorus concentrations were determined on a filtered (0.45 μm WCN-grade cellulose nitrate membrane; Whatman, Maidstone, UK) sample , using the

phosphomolybdenum blue colorimetry method of Murphy and Riley (1962), as modified by Neal et al. (2000), using a Seal Auto Analyser 3 (Seal Analytical; Fareham, UK). SRP samples were analysed within 48 h, to minimise errors associated with sample instability.

Dissolved reactive silicon concentrations were determined by reaction with acid ammonium molybdate, to form yellow molybdosilicic acids. These were then reduced using an acidified tin (II) chloride solution to form intensely coloured silicomolybdenum blues, which were quantified spectrophotometrically using an Seal Auto Analyser 3 (Seal Analytical; Fareham, UK)(Mullin and Riley 1955).

Ammonium concentration was determined using an indophenol-blue colorimetric method (Leeks et al. 1997) using a Seal Auto Analyser 3.

Dissolved organic carbon and total dissolved nitrogen were analysed by thermal oxidation using a Thermalox analyser (Analytical Sciences Ltd.; Cambridge, UK) until December 2010 and with an Elementar Vario Cube (Elementar Ananlysensysteme GmbH; Langenselbold, Germany) from June 2011.

Major dissolved anion (fluoride, chloride, bromide, nitrite, nitrate and sulphate) concentrations were determined by ion chromatography (Dionex AS50, Thermo Fisher Scientific; Waltham, USA). Total and dissolved cation concentrations were determined on unfiltered and filtered samples respectively, by acidification, followed by analysis by inductively coupled plasma optical emission spectrometry (ICP-OES)(Perkin Elmer Optima 2100; Seer Green, UK).

All analysis (with the exception of suspended solids) was carried out alongside reference Aquacheck QC standards (LGC Standards, Teddington, UK).

The water quality data is supplied alongside mean daily flow gauging data, obtained from the National River Flow Archive http://nrfa.ceh.ac.uk/ . Most sites are situated at, or close to Environment Agency gauging stations. The exceptions are :

- the River Thames sites at Hannington Wick, Wallingford and Sonning. The flows at these sites were interpolated, based on the monitoring site's catchment area, using multiple gauging data along the length of the River Thames.
- The Jubilee River, which is an 11 km flood relief channel offshoot of the main River Thames, is supplied with flow data from the River Thames at Windsor (2km downstream of the monitoring site). The gauged flows at Windsor comprise of the amalgamated flows from the River Thames and the Jubilee River, rather than Jubilee River itself.
- The Cut and River Kennet monitoring sites are supplied with gauging station data that is some distance upstream of each monitoring site.

Format of the dataset

- The date / time is given in day/month/year Sampling time is given in hour:minute format.

- Total phosphorus concentration comprises of the dissolved and acid-extractable particulate phosphorus present in an unfiltered water sample. Soluble reactive phosphorus can also be described as filterable reactive phosphorus or soluble molybdate-reactive phosphorus. The concentration data is expressed as a concentration of phosphorus.
- Total and dissolved metal concentration data are derived from analysing unfiltered and filtered (0.45 μm) samples respectively
- Water temperature in degrees centigrade (measured at time of sampling)
- pH (no units)
- Alkalinity in micro-equivalents per litre
- Suspended solids concentration in mg dry solids per litre
- Soluble reactive phosphorus concentration in micrograms P per litre
- Total dissolved phosphorus concentration in micrograms P per litre
- Total phosphorus concentration in micrograms P per litre
- Ammonium concentration in milligrams NH4+ per litre
- Dissolved reactive silicon concentration in milligrams Si per litre
- Chlorophyll a concentration in micrograms per litre
- Dissolved fluoride concentration in milligrams F per litre
- Dissolved chloride concentration in milligrams Cl per litre
- Dissolved nitrite concentration in milligrams NO2 per litre
- Dissolved bromide concentration in milligrams Br per litre
- Dissolved nitrate concentration in mg NO3 per litre
- Dissolved sulphate concentration in mg SO4 per litre
- Total dissolved nitrogen concentration in milligrams N per litre
- Dissolved organic carbon concentration in milligrams C per litre
- Dissolved sodium concentration in milligrams Na per litre
- Dissolved potassium concentration in milligrams K per litre
- Dissolved calcium concentration in milligrams Ca per litre
- Dissolved magnesium concentration in milligrams Mg per litre

- Dissolved boron concentration in micrograms B per litre
- Dissolved iron concentration in micrograms Fe per litre
- Dissolved manganese concentration in micrograms Mn per litre
- Dissolved zinc concentration in micrograms Zn per litre
- Dissolved copper concentration in micrograms Cu per litre
- Dissolved sodium concentration in milligrams Na per litre
- Dissolved potassium concentration in milligrams K per litre
- Total calcium concentration in milligrams Ca per litre
- Total magnesium concentration in milligrams Mg per litre
- Total boron concentration in micrograms B per litre
- Total iron concentration in micrograms Fe per litre
- Total manganese concentration in micrograms Mn per litre
- Total zinc concentration in micrograms Zn per litre
- Total copper concentration in micrograms Cu per litre
- The flow data is mean daily average flow in cubic metres per second.

References

Bowes, M.J.; Armstrong, L.K.; Wickham, H.D.; Harman, S.A.; Gozzard, E.; Roberts, C.; Scarlett, P.M. (2017). Weekly water quality data from the River Thames and its major tributaries (2009-2013) [CEH Thames Initiative]. NERC Environmental Information Data Centre. https://doi.org/10.5285/e4c300b1-8bc3-4df2-b23a-e72e67eef2fd

Eisenreich, S. J., R. T. Bannerman & D. E. Armstrong, 1975. A simplified phosphorus analytical technique. Environmental Letters 9(4):45-53.

Leeks, G. J. L., C. Neal, H. P. Jarvie, H. Casey & D. V. Leach, 1997. The LOIS river monitoring network: strategy and implementation. Sci Total Environ 194-195:101-109.

Marker, A. F. H., E. A. Nusch, H. Rai & B. Riemann, 1980. The measurement of photosynthetic pigments in freshwaters and standardisation of methods: Conclusions and recommendations. Arch Hydrobiol Beih 14:91-106.

Mullin, J. B. & J. P. Riley, 1955. The colourometric determination of silicate with special reference to sea and natural waters. Anal Chim Acta 12:31-36.

Murphy, J. & J. P. Riley, 1962. A modified single solution method for the determination of phosphorus in natural waters. Analytica chemica acta 12:31-36.

Neal, C., M. Neal & H. Wickham, 2000. Phosphate measurement in natural waters: two examples of analytical problems associated with silica interference using phosphomolybdic acid methodologies. Sci Total Environ 251:511-522.