Malaysian Journal of Mathematical Sciences 8(S): 173- 184 (2014)
Special Issue: International Conference on Mathematical Sciences and Statistics 2013 (ICMSS2013)

MALAYSIAN JOURNAL OF MATHEMATICAL SCIENCES

Journal homepage: http://einspem.upm.edu.my/journal

Robust Nonlinear Regression: Case Study for Modeling the Greenhouse Gases, Methane and Carbon Dioxide Concentration in Atmosphere

[1]Hossein Riazoshams and [2*]Habshah Midi

*[1]Department of Statistics,
Stockholm University, Stockholm, Sweeden*

*[2]Department of Mathematics, Faculty of Sciences and
Institute for Mathematical Research, Universiti Putra Malaysia,
43400 UPM Serdang, Selangor, Malaysia*

E-mail: habshah@upm.edu.my

**Corresponding author*

ABSTRACT

Four nonlinear regression models are proposed for the atmospheric carbon dioxide and methane gas concentrations data, reported by United Nation 1989. Among those considered, the Exponential with Intercept is the most preferred one to model methane data due to better convergence and lower correlation between parameters. On the other hand, the scale exponential convex model is appropriate for carbon dioxide data because besides having smaller standard errors of parameter estimates and smaller residual standard errors, it is numerically stable. Due to large range of data that goes back to history to 7000 years ago, there is a big dispersion in data set, so that it made us to apply robust nonlinear regression estimation methods to have a smoother model.

Keywords: Nonlinear Regression, Robust estimates, Methane gas, Carbon Dioxide gas.

1. INTRODUCTION

Human activities are overwhelmingly dominant contribution to the current disequilibrium of the global carbon cycle. Many researchers have attempted to explore the impact of human activity on the amount of greenhouse gases in the atmosphere, such as Methane and Carbon Dioxide. They have tried to find mathematical models for the changes during time, and measured the amount of concentration of these gases trapped inside poles icebergs from thousands years ago. The United Nation Environmental Program (UNEP) (1989) reported that the atmospheric CO2 and Methane CH4 concentration data were collected from south pole whereby these gasses were trapped in icebergs from 8000 thousand years ago. Etheridge et al. (1998) presented the methane mixing ratios from 1000 A.D. to be present in Antarctic ice cores, Greenland ice cores, the Antarctic firm layer, and archived air from Tasmania, Australia.

Many authors attempted to find the vulnerabilities associated with CH4 exchange, for example see Dolman *et al.* (2008) and Etheridge *et al.* (1998) where they also discussed modeling of CO2 changes. Dolman *et al.* (2008) mentioned that the CH4 model for changes in history, is linear in pre-industrial era, exponential in industrial era, and in recent time the increase is declined (Bousquet *et al.* (2006)). Moreover, there are efforts to forecast the CO2 and CH4 concentration for future time, for example see Raupach *et al.* (2005).

Most of these researches studied the data from 1000AD to present, while data set presented by UNEP (1989) goes back to7000 BC, which have high leverage values, See Figure 1. This article attempts to fit suitable nonlinear models for Methane and Carbon Dioxide gas concentration (UNEP 1989), whereby the high leverage values are taken into consideration in the computation of robust nonlinear fitting methods. Due to linearity of the data behavior in the pre-industrial era, sharp curvature in industrialization time and high slope increase in modern era, the fitting of nonlinear model is not straight forward and some modification to the models are required.

2. ROBUST NONLINEAR REGRESSION

Consider the general nonlinear model:

$$y = f(\theta) + \varepsilon \tag{1}$$

where $y = [y_1, y_2, ..., y_n]^T$ is $n \times 1$ response vector, $f(\theta) = \left[f(\mathbf{x}_1;\theta), ..., f(\mathbf{x}_n;\theta) \right]$ is $n \times 1$ vector of function models $f(x_i;\theta)$'s, $\mathbf{x}_i = [x_{i1}, x_{i2}, ..., x_{ik}]^T$ is k dimensional predictor (design) vector, $\varepsilon = [\varepsilon_1, \varepsilon_2, ..., \varepsilon_n]^T$ is $n \times 1$ vector of errors which are usually considered to be independent identically distributed (iid) with mean zero and unknown variance σ^2, and $\theta \in \Re^p$ is p dimensional unknown parameter vector.

The Nonlinear Least Squares (NLLS) estimates of unknown parameter $\theta \in \Re^p$ are obtained by minimizing the sum of squares errors. The minimization can be computed by modified Newton method (See Bates and Watts (1988)), which uses gradient of model function. In the case of singularity the Levenberg-Marquardt method is employed (See Riazoshams (2010)).

Least Squares Estimate obviously is not robust and unduly affected by outliers. In order to reduce the effect of outliers, robust methods are put forward. Stromberg (1992, 1993) extended the robust MM-estimates for Linear Regression proposed by Yohai (1987), to nonlinear regression. This method will be used to estimate the parameters of the models considered in this article. For more theoretical detail and computation methods see Riazoshams (2010).

3. NONLINEAR MODELS

UNEP (1989) presented the Methane Gas (Figure 1) and Carbon Dioxide Gas (Figure 2) collected from the Gas trapped in icebergs in south pole from 8000 years ago. As can be seen from Figure 1 the Methane data contains high leverage points. In this respect robust methods are used to reduce the effect of high leverage points. Four nonlinear models are proposed.

Model 1. Scaled Exponential

$$y_i = p_1 + p_2 e^{\frac{x_i - p_3}{p_4}}$$

Model 2. Scaled Exponential Convex

$$y_i = p1 + e^{(p_2 - p_3 x_i)}$$

Model 3. Power Model

$$y_i = \frac{1}{p_1} - p_2 . p_3^{x_i}$$

Model 4. Exponential with Intercept

$$y_i = p_1 + e^{\frac{x_i - p2}{p3}}$$

These models are considered to be able to describe the pre-era close linearity of data, and sharp change of industrial era, and linear change in modern era with a slight decline in the rise of high values at the top. Due to such data behaviors, no any nonlinear model could be fitted easily to the data. In order to find some appropriate nonlinear models, firstly we started to fit the data with exponential model, due to the exponent behavior of the data. Then we observed that in the exponent, the location and scale parameters are needed (like p3 and p4 in Scaled Exponential). Since the data at minus infinity are asymptotically constant, a constant parameter is added to the models. As can be seen all the four models have a constant value to express the horizontal asymptote at minus infinity, and for this reason an “intercept” is added in the model.

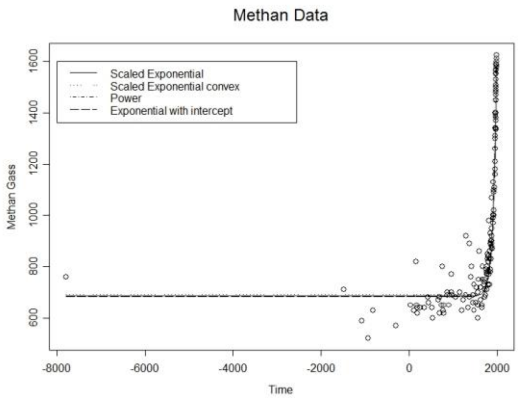

Figure 1: Four models fitted to Methane Data using robust MM-estimator

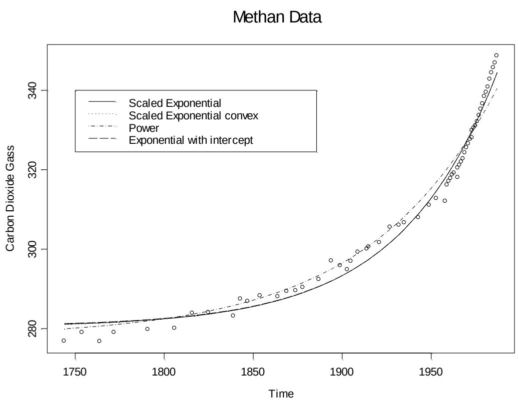

Figure 2: Four models fitted to Carbon Dioxide Data using robust MM-estimator

TABLE 1 to TABLE 4 show the results of parameter estimates, standard error, correlation of parameters, and fitted correlation for Methane data with high leverage points and without high leverage points. The data are not well behaved and convergence is hard to achieve.

TABLE 1: The robust MM and classical NLLS estimates for Scaled Exponential model, Methane Data

Data Without High Leverages							Data With High Leverages						
Parameters	NLLS	Stderror	Correlation Matrix				Parameters	NLLS	stderror	Correlation Matrix			
p1	712.04	0.1457	1	5.382 E-4	5.384 E-4	-0.585	p1	691.3089	0.1231590	1	2.37 E-5	-2.34 E-5	-0.52
p2	0.2875	4302.42		1	0.999 4	-4.260 E-4	p2	0.39243	3853.4942		1	1	-7.77 E-5
p3	1421.5	1048576			1	-4.263 E-4	p3	1400.148	741455.20			1	-7.83 E-5
p4	70.058	0.0577				1	p4	75.50851	0.0568263				1
σ	62.656						σ	65.66286					
Correlation	0.9814						Correlation	0.979283					
# Iteration	9						# Iteration	13					
Data Without High Leverages							Data With High Leverages						
Parameters	MM	stderror*	Correlation Matrix				Parameters	MM	stderror	Correlation Matrix			
p1	704.51	8.23	1	-3.7 E-05	-6.65 E-8	-0.58	p1	685.57	(*)	(*)			
p2	0.44	2509.7		1	0.999	6.1 E-05	p2	0.05					
p3	1454.57	398694			1	1.21 E-07	p3	1255.45					
p4	69.47	3.213				1	p4	74.59					
σ	56.63						σ	59.17					
correlation	0.96						Correlation	0.96					
# Iteration	3						# Iteration	7					

* stderror: abbreviation of standard error

TABLE 2: The robust MM and classical NLLS estimates for Exponential Convex model, Methane Data

Data Without High Leverages						Data With High Leverages					
Parameters	NLLS	stderror	Correlation		Matrix	Parameters	NLLS	stderror	Correlation		Matrix
p1	712.04	9.09	1	0.59	0.58	p1	691.31	8.05825	1	0.53	0.52
p2	21.54	1.45		1	0.999	p2	19.48	1.29		1	0.99
p3	0.01	0.7e-3			1	p3	0.01	0.6e-3			1
σ	62.4					σ	65.43				
correlation	0.962					correlation	0.957242				
# Iteration	12					# Iteration	13				
Data Without High Leverages						Data With High Leverages					
Parameters	MM	stderror	Correlation Matrix			Parameters	MM	stderror	Correlation		Matrix
p1	704.61	8.23	1	0.59	0.59	p1	685.55	7.267356	1	0.529	0.52
p2	21.76	1.32		1	0.999	p2	19.79	1.172492		1	0.99
p3	0.01	0.7E-3			1	p3	0.01	0.6E-3			1
σ	56.66					σ	59.17				
correlation	0.96					correlation	0.96				
# Iteration	10					# Iteration	50				

TABLE 3: The robust MM and classical NLLS estimates for Power model, Methane Data

Data Without High Leverages						Data With High Leverages					
Parameters	NLLS	stderror	Correlation Matrix			Parameters	NLLS	stderror	Correlation Matrix		
p1	1.034 E-3	5.74 E-5	1	-0.720	-0.703	p1	1.447 E-3	1.69 E-05	1	-0.53	-0.523
p2	-1.062 E-3	0.011		1	0.999	p2	-3.47 E-9	4.48 E-09		1	0.999
p3	1.006153	0.005			1	p3	1.013	6.61 E-4			1
σ	268.0398					σ	65.430				
correlation	-5.006					correlation	0.957				
# Iteration	76					# Iteration	197				
Data Without High Leverages						Data With High Leverages					
Parameters	MM	stderror	Correlation Matrix			Parameters	MM	stderror	Correlation Matrix		
p1	1.42 E-3	1.38 E-5	1	-0.59	-0.58	p1	0.001458	1.54 E-05	1	-0.529	-0.522
p2	-3.48 E-10	3.92 E-10		1	0.999	p2	-2.53 E-9	2.85 E-09		1	0.999
p3	1.014	5.62 e-4			1	p3	1.013519	0.000603			1
σ	47.156					σ	59.159891				
correlation	0.962					correlation	0.9575862				
# Iteration	841					# Iteration	195				

TABLE 4: The robust MM and classical NLLS estimates for Exponential with Intercept Model, Methane Data

Data Without High Leverages						Data With High Leverages					
Parameters	NLLS	stderror	Correlation Matrix			Parameters	NLLS	stderror	Correlation Matrix		
p1	712.0400	9.09	1	0.61	-0.58	p1	691.31	8.06	1	0.55	-0.523
p2	1508.850	24.20		1	-0.999	p2	1470.78	24.92		1	-0.998
p3	70.0600	3.60			1	p3	75.51	3.72			1
σ	62.4000					σ	65.43				
correlation	0.961754					correlation	0.957242				
# Iteration	14					# Iteration	15				
Data Without High Leverages						Data With High Leverages					
Parameters	MM	stderror	Correlation Matrix			Parameters	MM	stderror	Correlation Matrix		
p1	703.68	6.85	1	0.61	-0.58	p1	684.77	6.06	1	0.548	-0.522
p2	1512.25	17.97		1	-0.998	p2	1476.49	18.45		1	-0.998
p3	69.43	2.67			1	p3	74.55	2.75			1
σ	47.14					σ	49.37				
correlation	0.962					correlation	0.958				
# Iteration	12					# Iteration	8				

For the scaled exponential Model 1, it can be seen that the derivative with respect to p_2 is different only with a constant product of derivative with respect to p_3. This makes the columns of gradient matrix to be linearly dependent thus it is singular. This fact theoretically means that the parameters are not estimable in linear regression approximation by Taylor expansion. In this case, direct optimization using derivative free methods or Levenberg-Marquardt in singularity situation, is used.

The convergence is fast with 3 and 7 number of iteration for robust method, but the standard errors of parameters p2 and p3 without high leverage are very high, and for data with high leverages, the covariance matrix is singular, that is:

$$\hat{V}^T\hat{V} = \begin{Bmatrix} 144 & 669286.1 & -464.38 & -4322 \\ & 8470951842 & -5877486 & -56028453 \\ & & 4078 & 38875 \\ & & & 371024 \end{Bmatrix}$$

with eigenvalues,

(8.471327e+009, 4.707126e+002, 6.221316e+001, -5.416392e-007)

The negative eigenvalues show that the matrix is non singular, and the covariance matrix cannot be computed. This is probably due to the derivative of second and third parameters are linearly related and the presence of high leverage points. Since the singular gradient matrix can be solved by Levenberg Marquardt method, we suspects that high leverage points are responsible for this problem. Furthermore, for this model the correlation between p2 and p3 is almost equal to one, which is appropriate to remove it from the model, and this leads us to Model 4, Exponential with Intercept.

It can be seen from the results of Scaled Exponential Convex model from TABLE 2 that the standard errors of robust MM estimates are lower than the NLLS estimates, the correlation between p2 and p3 and the number of iterations are slightly higher for MM method than the NLLS method in both situations; in the presence and absence of outliers in a data. It is important to note that the correlation between parameters is high, but not more than the first model.

The results of Power Model 3 in the Table 3 reveal that convergence is hard to achieve, as can be seen for data without high leverages where robust method needs 841 iterations. The NLLS has a bad fit with wrong value of fitted model correlation -5.006 which is due to wrong parameter estimates, large value of residual standard error of model (268) which is possibly due to computation rounding errors (See Table 3, NLLS estimate for data without high leverages). Figure 3 clearly reveals the wrong fit of the model. The robust method for both cases works better and is more trustable, although some far points still can be seen after the high leverages have been removed from the data. It is important to note that the correlation between parameters p2 and p3 is still high.

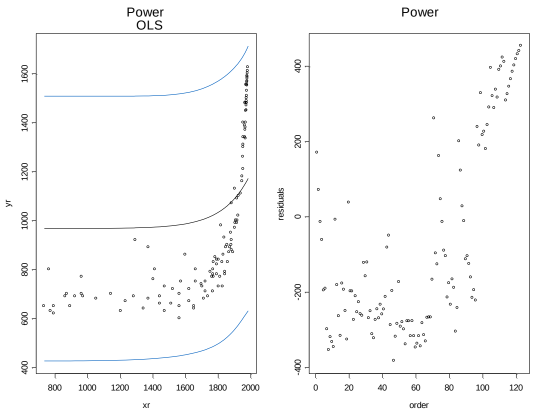

Figure 3: Classical NLLS Estimate for methane data without high leverage points

TABLE 4 shows the estimates for Model 4, the Exponential with intercept (p1 is called an intercept). As explained before, in Scaled Exponential Model, the parameters p2 and p3 have almost a linear relation, that encourage us to remove p2 which leads to Model 4. The intercept p1, is the limit value of data at (-∞), that is the amount of methane gas in ancient time. The correlation between parameters in worst case is better than other models. In the presence of high leverage, the robust estimates of the residual standard errors of Models (1-3) are very closed { 59.17 for model 1 (Table 1), 59.17 for model 2 (Table 2) and 59.16 for model 3 (Table 3) }, but their values are higher than Model 4. It is observed that the power and Scaled Exponential Convex model have bad convergence and higher correlations between parameters. Figure 1 shows the robust MM fits of the four models in the presence of high leverage, which suggests a closer fit to the data. However, Exponential with Intercept model is preferred because it has the least value of residual standard error, needs less number of iteration and better fit than other models.

The plots of Figure 1 suggests a possibility of having heteroscedastic errors since the variance of the errors is decreasing in a systematic manner with the increased in x values. We do not take into consideration this problem in the analysis since the degree of heteroscedastic errors seems to be small and it is beyond the scope of this research. To rectify this problem, future work can consider a variance model whereby a robustified model selection procedure may be used to choose a better model.

4. CARBON DIOXIDE DATA

UNEP (1989) presented the carbon Dioxide data collected from the same source as methane gas data. The data do not have outliers and in this case it's easier to fit the models.

Table 5 to Table 8 exhibit the results of parameter estimates, standard errors, correlation of parameters, and fitted correlation for Carbon Dioxide data.

Let us first focus on Scale Exponential Model of Table 5. Similar to Methane Data, the NLLS estimates cannot be computed by the modified Newton method and Levenberg Marquardt Method is then employed to compute the estimates in Table 5. The correlation of fit is high, convergence achieved after 6 iteration, but correlation between p2 and p3 is one. Similar to results of Methane Data, again this will lead us to exponential Model with intercept.

The MM estimates are fairly closed to the NLLS estimates for Scaled Exponential convex (Model 2), since the data do not have outliers. It is interesting to point out that the standard errors of the MM estimates are slightly smaller than the NLLS estimates but have higher residual standard errors. In this situation, the NLLS method is preferred.

Error! Reference source not found. shows results of power model. Similar to methane data, convergence is difficult to achieve for this model. It can be seen that the NLLS and MM methods require 100 iterations and 101 iterations, respectively. The parameter estimates seem to be very small and the model need to be rescaled.

It can be observed from Table 8 that the NLLS needs larger number of iterations than the MM method. We encountered computational problems to both NLLS and MM methods. The models were highly affected by initial values. The results of model 4 are almost similar to results of model 2, but model 2 is preferred as it has less computational problems. However, judging from the residual standard errors and the standard errors of parameter estimates, both models can be recommended for this data.

Error! Reference source not found. displays the fitting of the four models using robust MM method. The plot suggests that the power model has better fit than other models. After examining the residual plots, we observe that the errors are auto correlated but no further analysis is considered to remedy this problem, since this case is beyond the scope of this study and left for future research.

TABLE 5: The robust MM and classical NLLS estimates for Scaled Exponential model, Carbon Data

Parameters	NLLS	stderror	Correlation Matrix	Parameters	MM	Stderror	Correlation Matrix			
p1	280.3578	(*)	(*)	p1	280.3468	5.449E-4	1	-1.72E-4	-4.23E-7	-8.1678E-1
p2	0.2329			p2	0.1382	0.115375		1	1	2.0462E-4
p3	1681.586			p3	1652.462	45.494549			1	5.1831E-7
p4	54.3729			p4	54.4838	0.0014943				1
σ	2.6474361			σ	3.054331					
correlation	0.9926157			correlation	0.9850827					
# Iteration	6			# Iteration	6					

TABLE 6: The robust MM and classical NLLS estimates for Scaled Exponential Convex Model, Carbon Data

Parameters	NLLS	stderror	Correlation Matrix			Parameters	MM	stderror	Correlation Matrix		
p1	280.358	1.02204	1	0.8202	0.8162	p1	280.346	2.7900E-04	1	0.8208	0.8168
p2	32.38431	1.88758		1	0.999964479	p2	32.30756	5.1315E-04		1	0.999964388
p3	0.01839	9.480E-4			1	p3	0.01835	2.5772E-07			1
σ	2.624904					σ	3.054396				
correlation	0.985067					correlation	0.9850829				
# Iteration	10					# Iteration	9				

TABLE 7: The robust MM and classical NLLS estimates for Power Model, Carbon Data

Parameters	NLLS	stderror	Correlation Matrix			Parameters	MM	stderror	Correlation Matrix		
p1	3.632E-3	6.99E-06	1	-0.884954	-0.8809254	p1	0.0036	5.8018E-05	1	-0.8864871	-0.8825155
p2	-5.5E-11	3.36E-11		1	0.99995024	p2	-6.31e-11	0.0000E+00		1	0.999951
p3	1.0140958	3.07891E-4			1	p3	1.014	0.0027518			1
σ	3.032583					σ	3.078184				
correlation	0.990131					correlation	0.9732556				
# Iteration	100					# Iteration	101				

TABLE 8: The robust MM and classical NLLS estimates for Exponential with Intercept Model, Carbon Data

Parameters	NLLS	stderror	Correlation Matrix			Parameters	MM	stderror	Correlation Matrix		
p1	280.3581	1.022 E0	1	0.848746	-0.8162321	p1	280.52	1.9584 E-06	1	0.8478535	-0.8152082
p2	1760.812	1.1896 E+01		1	-0.9973526	p2	1761.76	2.2777 E-05		1	-0.99734132
p3	54.37217	2.802715			1	p3	54.17	5.37022 E-06			1
σ	2.624904					σ	2.22				
correlation	0.985066					correlation	0.9850167				
# Iteration	18					# Iteration	8				

5. CONCLUSION

Four models are proposed and fitted to Methane gas data presented by United Nation (1989). Three out of the four models are more feasible, however, the Exponential with Intercept is preferred due to better convergence and lower correlation between parameters. The Intercept is the limit value in antiquity, from the robust fit for exponential with intercept model, when time tends to $-\infty$ the model tends to parameter $p_1 = 648.77$ which is the value of methane consumption, 7000 years ago.

The robust MM estimator is able to balance the curve between antiquity and the recent time better than the NLLS estimator. This helps to reduce the measurement errors, because the data are collected from North and South Pole by measuring the amount of methane gas trapped inside Icebergs in Poles from 8000 years ago, and it is not guaranteed that the data are free of errors. Based on the United Nation's report UNEP (1989), only good values are presented in their graphs and no model whatsoever has been proposed to model Methane data. This research is the first attempt to model this data. New parameters can still be included in a model probably from suggestion by environmentalists.

For Carbon data, Model 2 and Model 4 fit reasonably well. The standard error of parameter estimates, and residual standard errors of the Model 2 is fairly closed to Model 4. Nonetheless, Model 4 fitting posed certain computational problems. On the other hand, the second model is numerically stable. In this respect, Model 2 is preferred to model carbon data.

REFERENCES

Bates, D. M., and Watts, D. G. (1988). Nonlinear regression analysis and its applications. New York: John Wiley & Sons.

Dolman, A. J., Freibauer, A. and Valentini, R. (2008). The continental-scale greenhouse gas balance of Europe. New York: Springer.

Etheridge, D. M., Steele, L. P., Francey, R. J. and Langenfelds, R. L. (1998). Atmospheric methane between 1000 A.D. and present: Evidence of anthropogenic emissions and climatic variability, Journal of Geophysical Research. **103**(D13): 979-993.

Raupach, M. R, Barrett, D. J, Briggs, P. R and Kirby, J. M. (2005). Simplicity, complexity and scale in terrestrial biosphere modelling. In: *Predictions in Ungauged Basins, International Perspectives on the State-of-the-Art and Pathways Forward*. (Eds). S Franks, M. Sivapalan, K Takeuchi, Y Tachikawa). IAHS Publication No. 301. IAHS Press, Wallingford, UK. p. 239-274.

Riazoshams, H. (2010). Outlier Detection and Robust estimation methods for nonlinear regression model having authocorrelated error and heteroscedastic errors, PhD thesis dissertation, School of Graduate Studies, University Putra Malaysia.

Stromberg, A. J. (1992). High Breakdown Estimators in Nonlinear Regression, in L1-Statistical analysis and related methods, ed. Y. Dodge. Amsterdam: North-Holland, 103-112.

Stromberg, A. J. (1993). Computation of High Breakdown Nonlinear Regression Parameters. *Journal of American Statistical Association*. **88**(421): 237-244.

UNEP. (1989). Environmental data report / prepared for UNEP by the GEMS Monitoring and Assessment Research Centre, London, UK, in co-operation with the World Resources Institute, Washington, D.C.

Yohai, V. J. (1987). High Breakdown point and high efficiency robust estimates for regression. *The Annals of Statistics*. **15**: 642-656.

Yves Bousquet and Serge Laplante. (2006). *Coleoptera histeridae: National Research Council Canada*. Monograph Publishing Program. Ottawa : NRC Research Press.