www.csiro.au

Rank Regression for Analyzing Environmental Data

You-Gan Wang & Liya Fu

CSIRO Mathematics, Informatics and Statistics

120 Meiers Road, Indooroopilly, QLD 4068, Australia

CSIRO

Acknowledgements

Data were kindly provided by Seqwater, Queensland, Australia.

Outline

- Background
- Descriptive Analysis
- Linear Mixed-Effects Model
- Rank Regression Model
- Results

CSIRO

Two digging tools

Which one to use?

Data Description

- **Data Collection**

Wivenhoe Dam, 1997- 2002

- **Indicators (Responses)**

Chlorophyll.a (continuous data), Total Cyanophytes (count data)

- **Covariates**

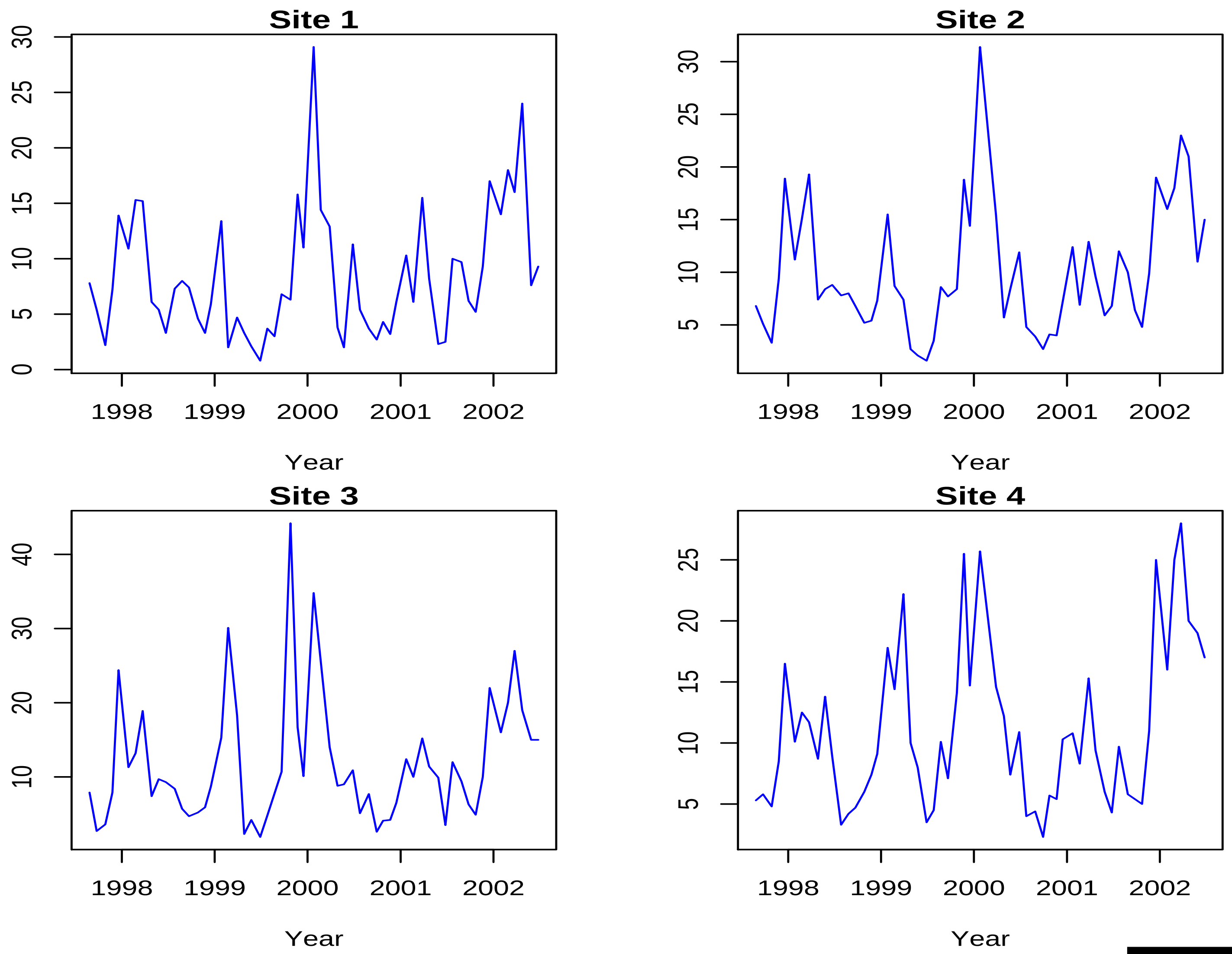
Days, Dam Level, Level Change, Rainfall

- **Purpose of this talk**

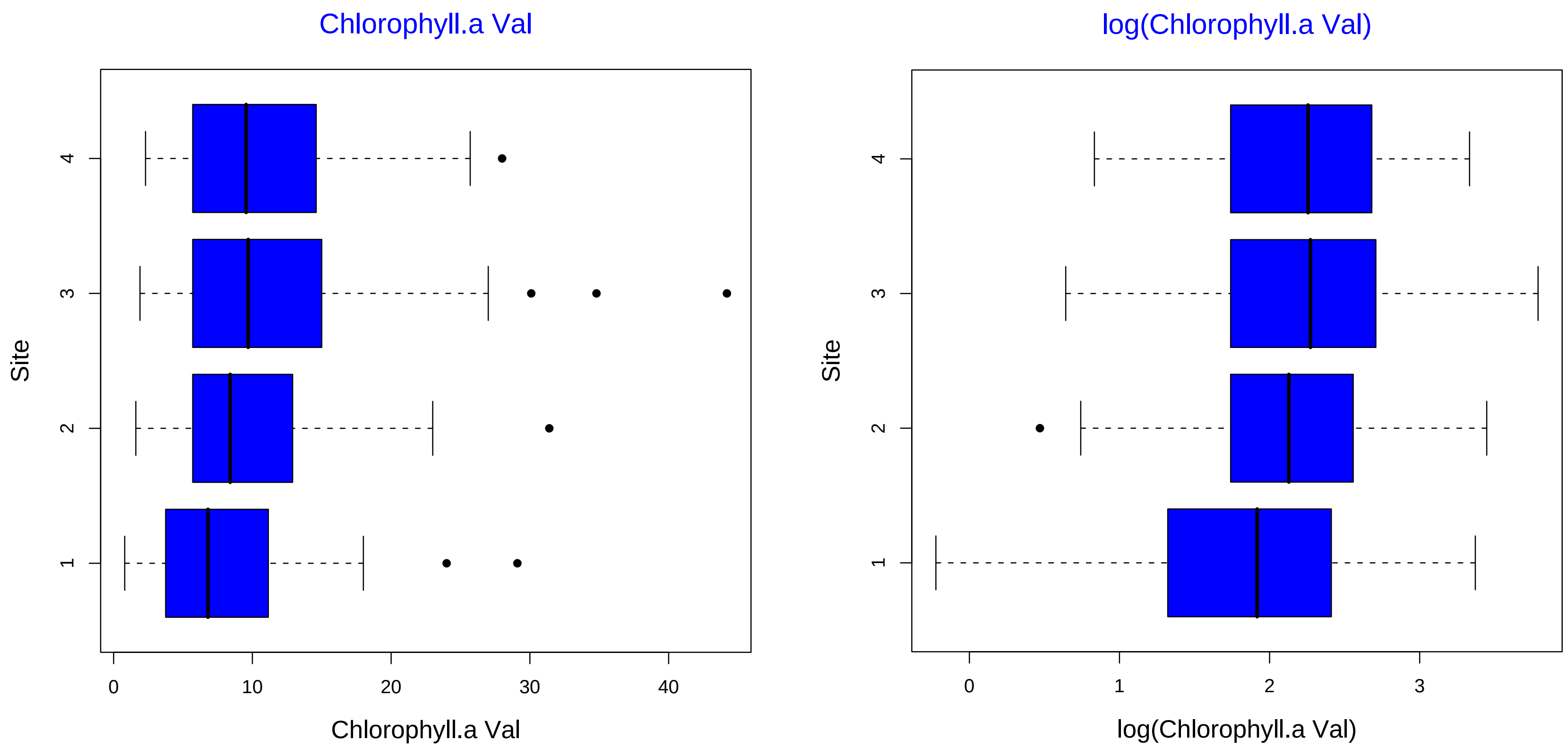
Find robust and efficient parameter estimation

CSIRO

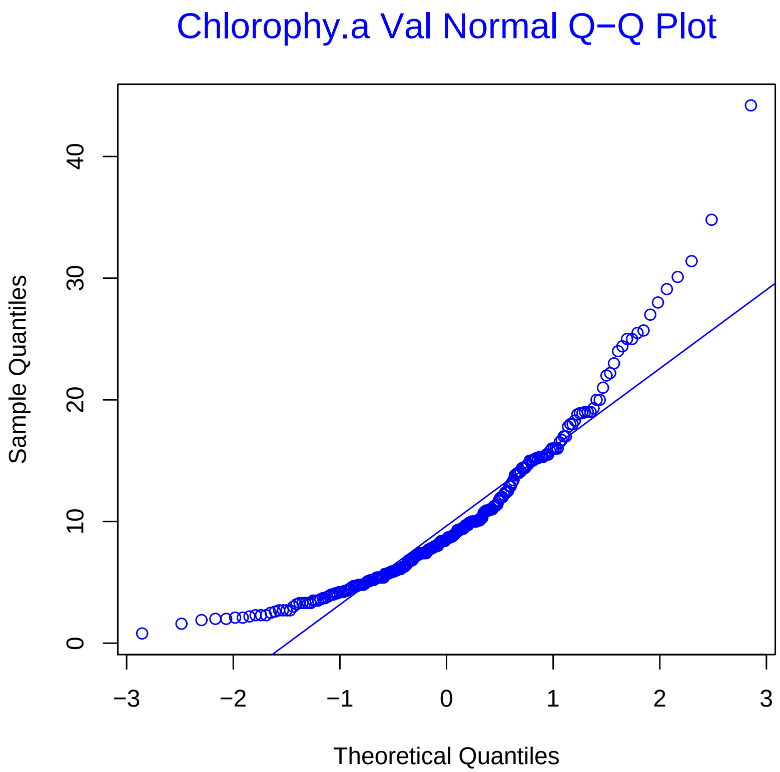
Time Series of Chlorophyll.a

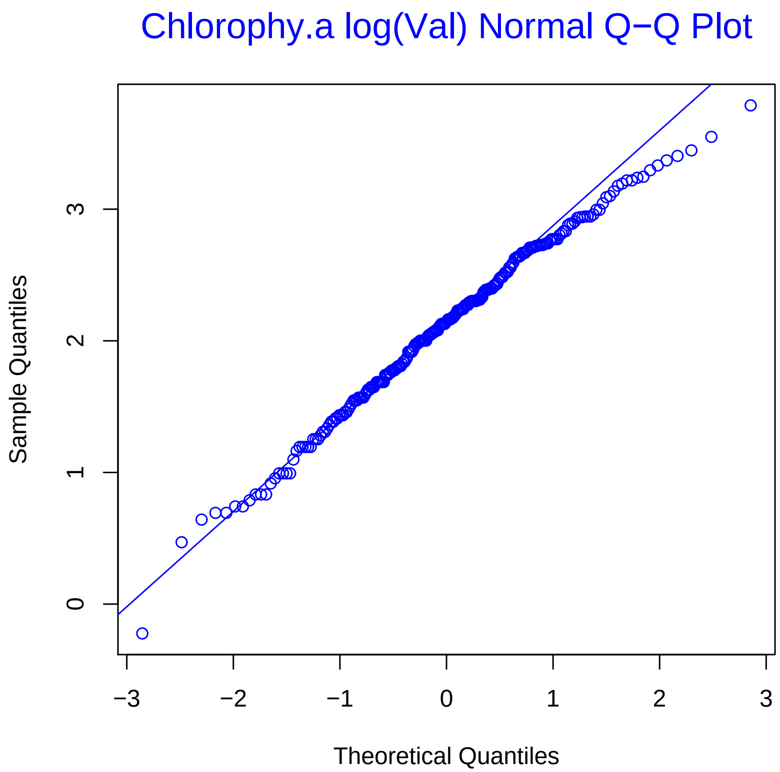


Box-Plots Chlorophyll.a

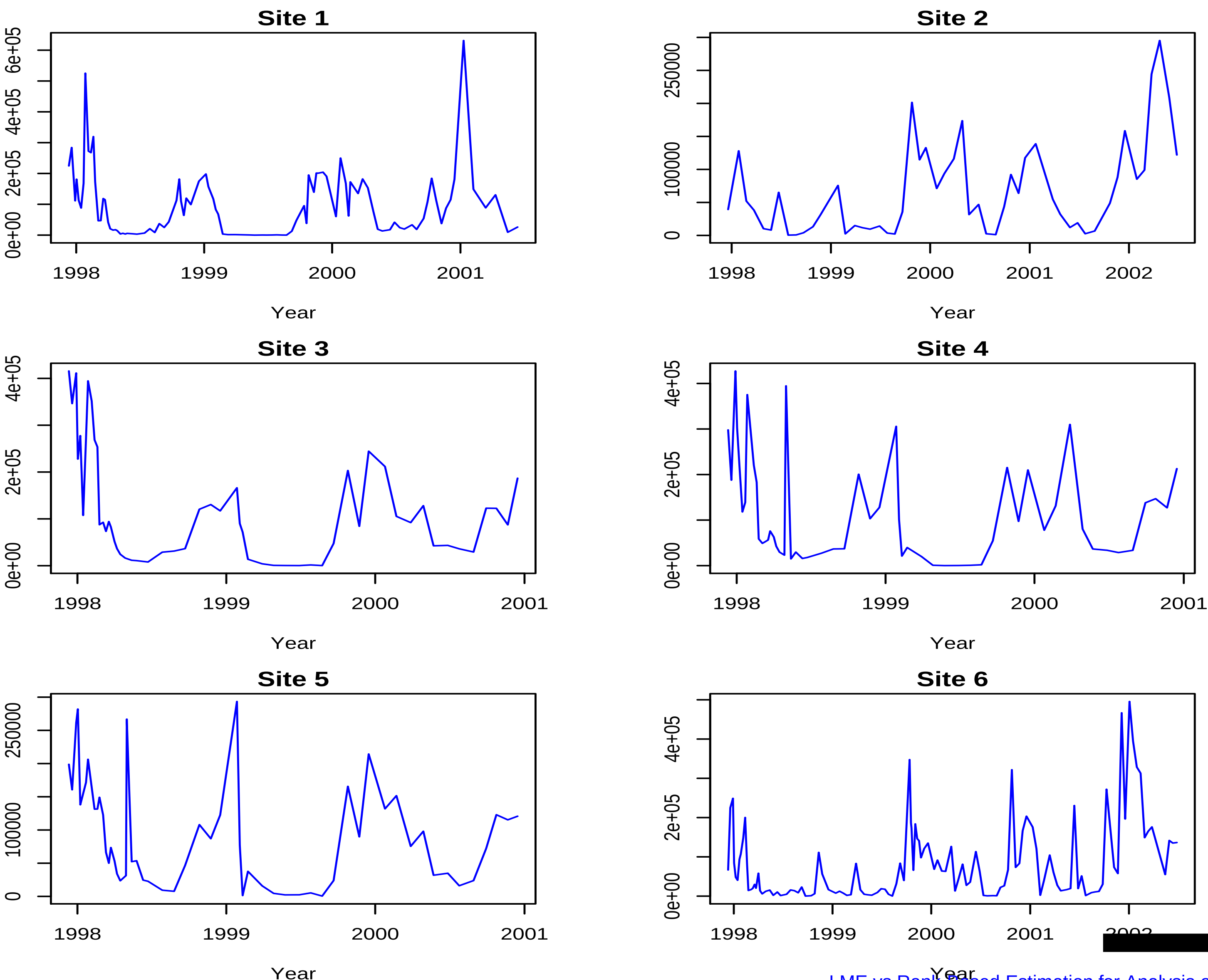


Q-Q Plots Chlorophyll.a

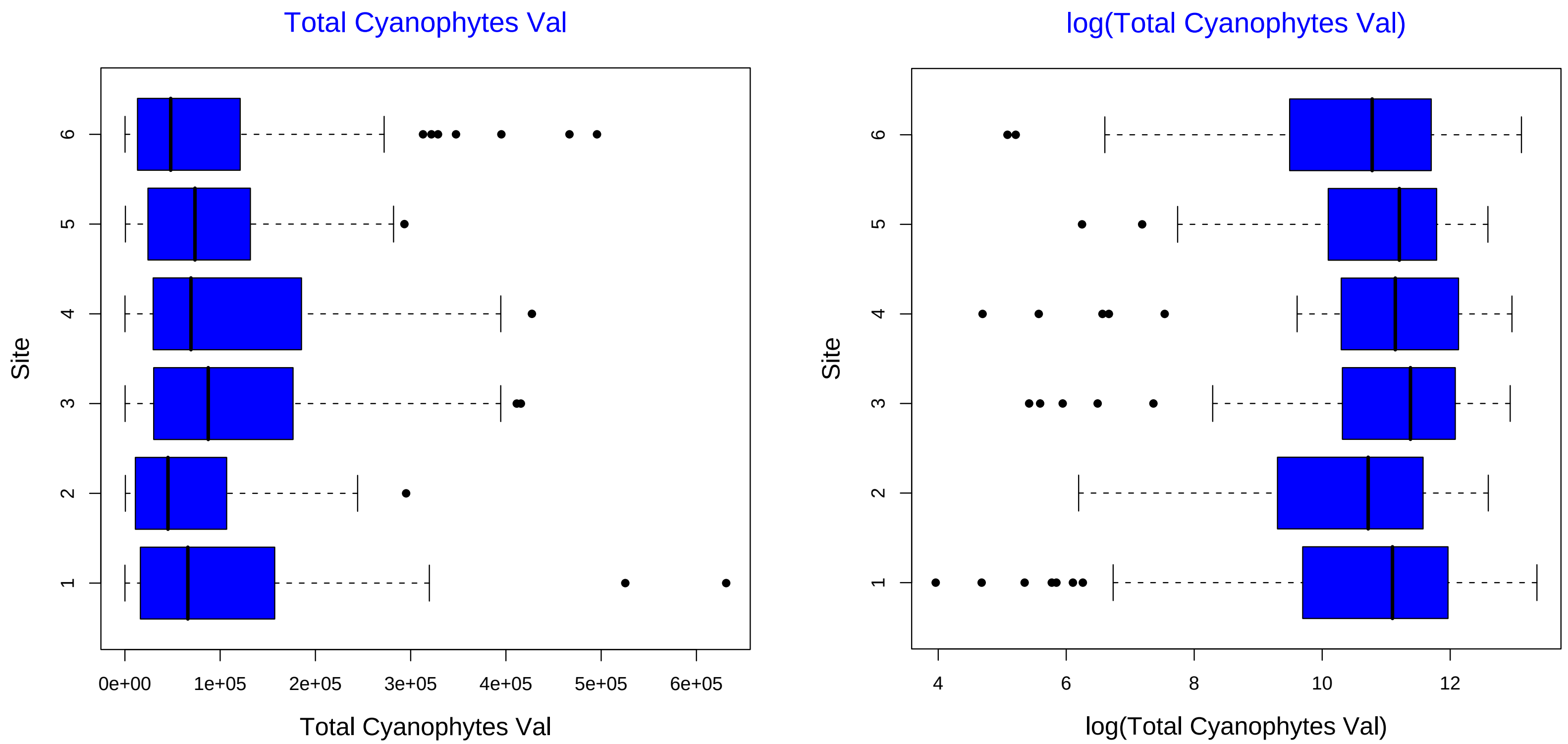




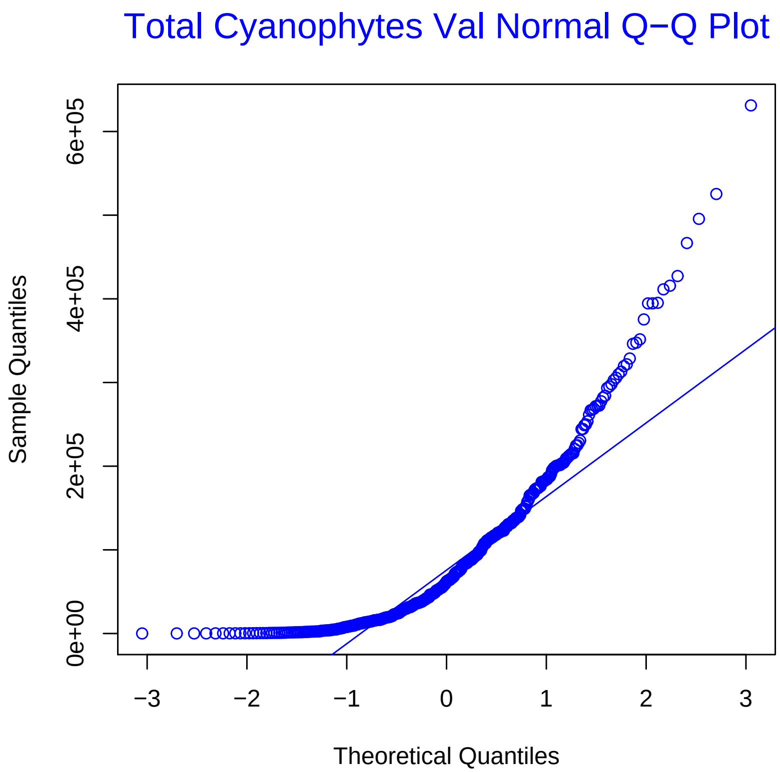
Time Series of Total Cyanophytes

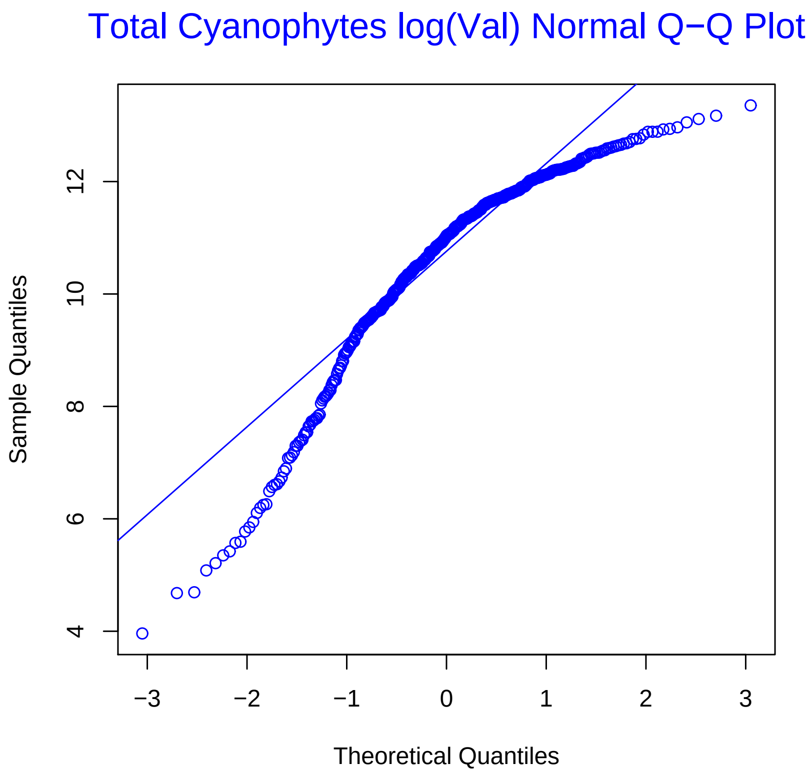


Box-Plots Total Cyanophytes



Q-Q Plots Total Cyanophytes





Linear Mixed Effects Model

Suppose observations for the i-th site are

$$Y_i = (y_{i1}, \cdots, y_{in_i}), i = 1, ..., N,$$

taken at times $T_i = (t_{i1}, t_{i2}, ..., t_{in_i})$. The linear mixed effects model is

$$\log(Y_i) = X_i\beta + Z_i\alpha_i + \epsilon_i,$$

where $X_i = (x_{i1}, ..., x_{in_i})'$ and $Z_i = (z_{i1}, ..., z_{in_i})'$ are known design matrices respectively; β are fixed effects, α_i and ϵ_i are random effects and random errors, respectively.

CSIRO

Linear Mixed Effects Model

- **Assumption:** $\alpha_i \sim N(0, \Psi)$ and $\epsilon_i \sim N(0, \Lambda_i)$
- **Estimation Method:** REML
- **Correlation Structure:** Gaussian spatial correlation

CSIRO

Rank Methods

- Robust
- Censored data (below detection limits)
- More efficient when errors have heavy-tailed distributions. To alleviate
- computational issues
- Interpretation?

Rank Regression Model

The rank regression model is $\log(Y_{ik}) = X_{ik}^{\mathrm{T}}\beta + \epsilon_{ik}$.

- **Assumption:** $\mathrm{median}(\epsilon_{ik} - \epsilon_{jl}) = 0$, for any i, j.

- **Estimation:** Residuals $e_{ik} = Y_{ik} - X_{ik}^{\mathrm{T}}\beta$, Jung and Ying (2003, Biometrika) regarded $(Y_{i1}, \cdots, Y_{in_i})$ as independent observations, and proposed minimizing the total loss function

$$\hat{\beta}_{JY} = \arg\min_{\beta} \left\{ N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} |e_{ik} - e_{jl}| \right\},$$

Incorporating Cluster Correlations

Wang and Zhu (2006, Biometrika) suggested decomposing ranks into between- and within-site ranks, and hence obtained two types of estimates.

$$\hat{\beta}_B = \arg\min_{\beta} \left\{ N^{-2} \sum_{i \neq j = 1}^{N} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} |e_{ik} - e_{jl}| \right\},$$

$$\hat{\beta}_W = \arg\min_{\beta} \left\{ N^{-1} \sum_{i=1}^{N} \sum_{k=1}^{n_i} \sum_{l=1}^{n_i} |e_{ik} - e_{il}| \right\}.$$

CSIRO

Incorporating Cluster Correlations

Combine corresponding between- and within-site estimating functions $U_B(\beta)$ and $U_W(\beta)$,

$$U_C(\beta) = (D_B, D_W)\Sigma^{-1}\begin{pmatrix} U_B(\beta) \\ U_W(\beta) \end{pmatrix},$$

where

$$\Sigma = \begin{pmatrix} U_B(\beta) \\ U_W(\beta) \end{pmatrix}.$$

CSIRO

How to Obtain $\hat{\Sigma}$

- Method 1: Perturbation method of Wang & Zhu (2006, Biometrika)

$$\begin{aligned}\tilde{U}_B(\beta) &= N^{-2}\sum_{i\neq j}\sum_{k}\sum_{l}\omega_i\omega_j(X_{ik}-X_{jl})(e_{ik}-e_{jl}),\\ \tilde{U}_W(\beta) &= N^{-1}\sum_{i}\sum_{k\neq l}\omega_i(X_{ik}-X_{il})(e_{ik}-e_{il}),\end{aligned}$$

- Method 2: $\hat{\Sigma} = ??$ (analytic expression)

CSIRO

How to Obtain $\hat{\beta}_C$

Brown and Wang (Biometrika, 2005) put forward induced smoothing method. Here we investigate this approach for rank regression.

The versions of D_B and D_W:

$$|\tilde{D}_B - D_B| \xrightarrow{a.s.} 0 \ \text{ and } \ |\tilde{D}_W - D_W| \xrightarrow{a.s.} 0$$

- **Parameter Estimation:**

$$(\tilde{D}_B, \tilde{D}_W)\hat{\Sigma}^{-1}\begin{pmatrix} U_B(\beta) \\ U_W(\beta) \end{pmatrix} = 0$$

CSIRO

Model of Water Quality Data

$$\log(\mathrm{Val}) \sim \mathrm{Intercept} + \mathrm{H}(\mathrm{Days}, \mathrm{k}=2) + (\mathrm{Level}) + (\mathrm{Cha.Level}) + (\mathrm{Rain})$$

- Days: the number of days (27/08/1997– 26/06/2002)
- Level: the dam level when the observation is collected
- Cha.Level: 30 days change on the dam level
- Rain: 14 days cumulative rainfall;
- $\mathrm{H}(\mathrm{Days}, \mathrm{k})$: is a harmonic function, and defined by following: $\mathrm{H}(\mathrm{x}, 2) = \sum_{\mathrm{k}=1}^{2}(\sin(2\mathrm{k}\pi\mathrm{x}/365.25) + \cos(2\mathrm{k}\pi\mathrm{x}/365.25))$.

CSIRO

Comparison of parameter estimation for Chlorophyll.a

	$\hat{\beta}_{lme}$	$\hat{\beta}_C$
H(Days, 2)1	-0.387	-0.439
(SE)	(0.070)	(0.017)
H(Days, 2)2	-0.352	-0.258
(SE)	(0.071)	(0.009)
H(Days, 2)3	0.207	0.207
(SE)	(0.058)	(0.036)
H(Days, 2)4	0.059	0.103
(SE)	(0.056)	(0.011)
Level	-0.0141	-0.015
(SE)	(0.053)	(0.012)
Cha.Level	-0.112	-0.156
(SE)	(0.040)	(0.022)
Rain	0.026	0.103
(SE)	(0.049)	(0.008)

Comparison of parameter estimation for Total Cyanophytes

	All Data		Outliers Removed	
	$\hat{\beta}_{lme}$	$\hat{\beta}_{C}$	$\hat{\beta}_{lme}$	$\hat{\beta}_{C}$
H(Days, 2)1	-0.100	-0.025	-0.087	-0.028
(SE)	(0.109)	(0.067)	(0.087)	(0.055)
H(Days, 2)2	-1.667	-1.472	-1.345	-1.265
(SE)	(0.110)	(0.055)	(0.090)	(0.041)
H(Days, 2)3	-0.342	-0.167	-0.298	-0.116
(SE)	(0.106)	(0.059)	(0.086)	(0.056)
H(Days, 2)4	-0.017	-0.073	0.081	0.019
(SE)	(0.107)	(0.028	(0.086)	(0.017)
Level	-0.136	0.015	0.089	0.045
(SE)	(0.083)	(0.044)	(0.067)	(0.041)
Cha.Level	-0.098	-0.319	-0.184	-0.270
(SE)	(0.081)	(0.051)	(0.072)	(0.036)
Rain	-0.308	-0.096	-0.229	-0.063
(SE)	(0.079)	(0.030)	(0.066)	(0.021)

CSIRO

Conclusions

- LME model is not always appropriate, although it is good when the data are generated from normal distributions.
- LME is much more sensitive to the outliers than rank estimation.
- Rank method is robust, and produces smaller standard errors.
- Rank methodology is computationally more intensive, but very doable in practice.

CSIRO

CSIRO Mathematics, Informatics and Statistics
120 Meiers Road, Indooroopilly, QLD 4068, Australia

You-Gan Wang

Phone: +61 7 3214 2816
Email: you-gan.wang@csiro.au
Web: www.cmis.csiro.au

www.csiro.au

Thank you

Contact Us
Phone: 1300 363 400 or +61 3 9545 2176
Email: enquiries@csiro.au Web: www.csiro.au

CSIRO