From Science of Methodology to Science of Data

Ritei Shibata

Dept. Math., Keio University

Statistical Challenges In Modern Astronomy(1992), Eds. E.D. Feogelson and G.J.Babu, Springer-Verlag

Astronomers $\rightleftarrows$ Statistician

- Galaxy Clustering
- Truncation and Censoring
- Bayesian and Image Analysis
- Time Series Analysis
- Multivariate Analysis and Other Topics
- General Discussion
  - ✓Statisticians should be involved from the beginning as real collaborators, not mere number crunchers.(Joseph Horowitz)
  - ✓The knowledge of statistics needs to come very early in the process. It is no good downstream when the paper is written. (Jasper Wall)

Astronomers $\cap$ Statistician ?

Methods in statistics

Data Stream

- Exploring Data
  - Clustering, PCA , Smoothing, Decomposition, Spectrum, …
  - Visualisation
- Cleaning Data
  - Missing, Contaminated, Disorganised,…
- Model Building
  - Distribution
  - Data Driven Model (Regression, Graphical Model, Hidden Markov, …)
  - Hybrid Model (Physical Model×Data Oriented Model,….)
- Model Fitting
  - Computational Techniques (EM, MCMC, Particle Filter, …)
  - Diagnostics
- Confirmation
  - Hypothesis Testing (p-value)
  - Model Validity
- Presentation
  - Descriptive Statistics
  - Visualisation

Statistician

Statistical theory or algorithm

Data for the test

Science of Methodology

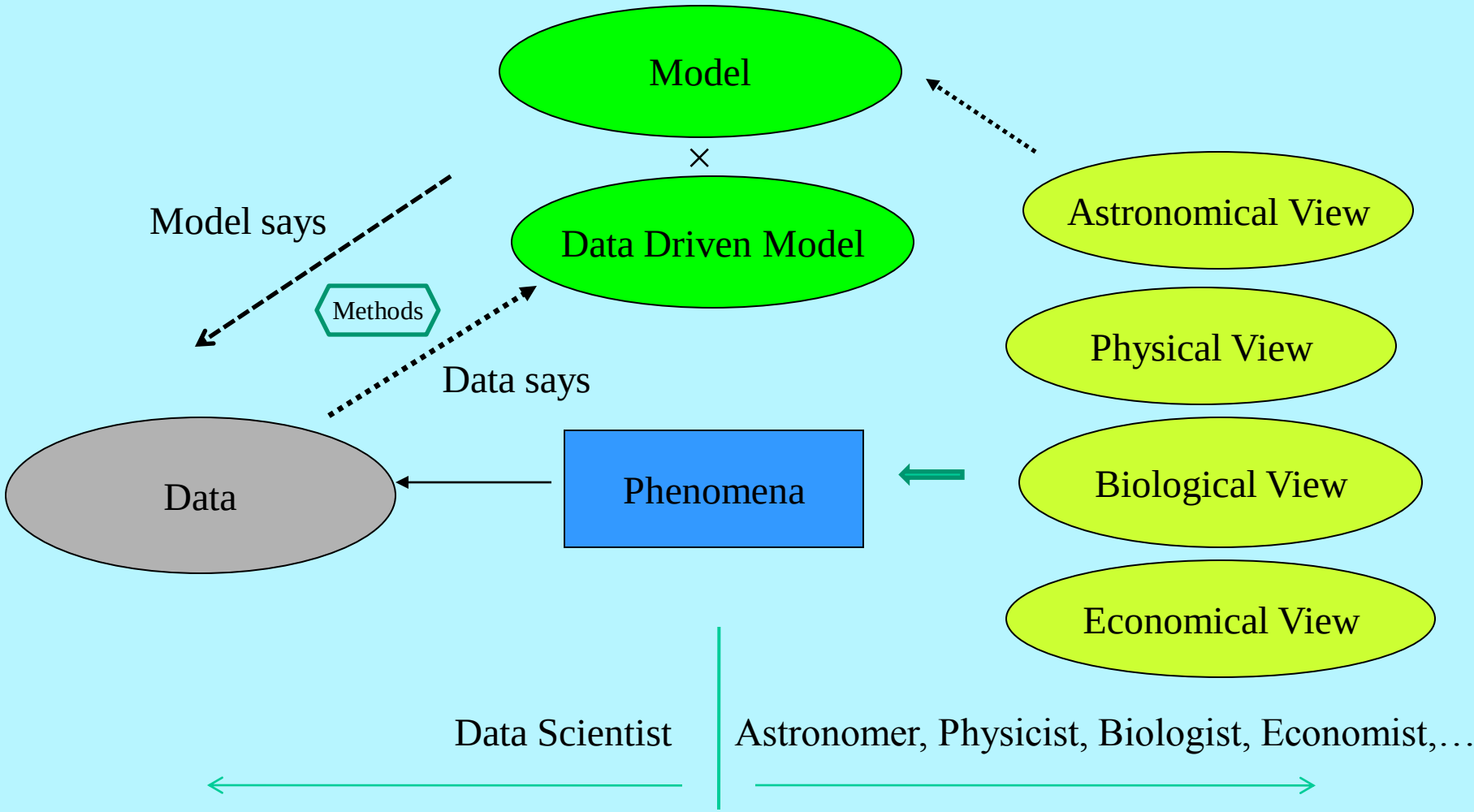
Astronomer

Fact Finding from data

Discovery from data

Develop New Theory

Data Science (Science of Data)



Exploring Data

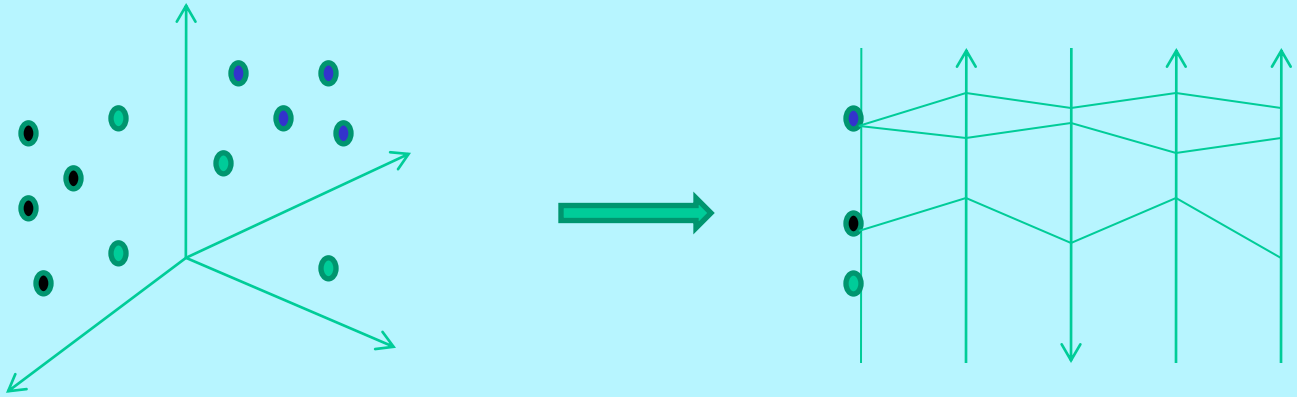
Setosa, Virginica, Versicolor

Methods: PCA, Machine Learning,…

Fisher's Iris Data (5 dimensional)

, , Setosa

	Sepal L.	Sepal W.	Petal L.	Petal W.
[1,]	5.1	3.5	1.4	0.2
[2,]	4.9	3.0	1.4	0.2
[3,]	4.7	3.2	1.3	0.2
[4,]	4.6	3.1	1.5	0.2
[5,]	5.0	3.6	1.4	0.2
[6,]	5.4	3.9	1.7	0.4
[7,]	4.6	3.4	1.4	0.3
[8,]	5.0	3.4	1.5	0.2
[9,]	4.4	2.9	1.4	0.2
[10,]	4.9	3.1	1.5	0.1
[11,]	5.4	3.7	1.5	0.2
[12,]	4.8	3.4	1.6	0.2
[13,]	4.8	3.0	1.4	0.1
[14,]	4.3	3.0	1.1	0.1
[15,]	5.8	4.0	1.2	0.2
[16,]	5.7	4.4	1.5	0.4
[17,]	5.4	3.9	1.3	0.4
[18,]	5.1	3.5	1.4	0.3
[19,]	5.7	3.8	1.7	0.3
[20,]	5.1	3.8	1.5	0.3



Orthogonal Coordinates

Parallel Coordinates

(Inselberg 1985, Wegman 1990)

PCA (Principal Component Analysis)

$$\mathbf{y} = X\mathbf{a} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_p\mathbf{x}_p$$

$$\|\mathbf{y}\|^2 \underset{\mathbf{a}}{\rightarrow} \max,$$

$\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p$: Principal Components

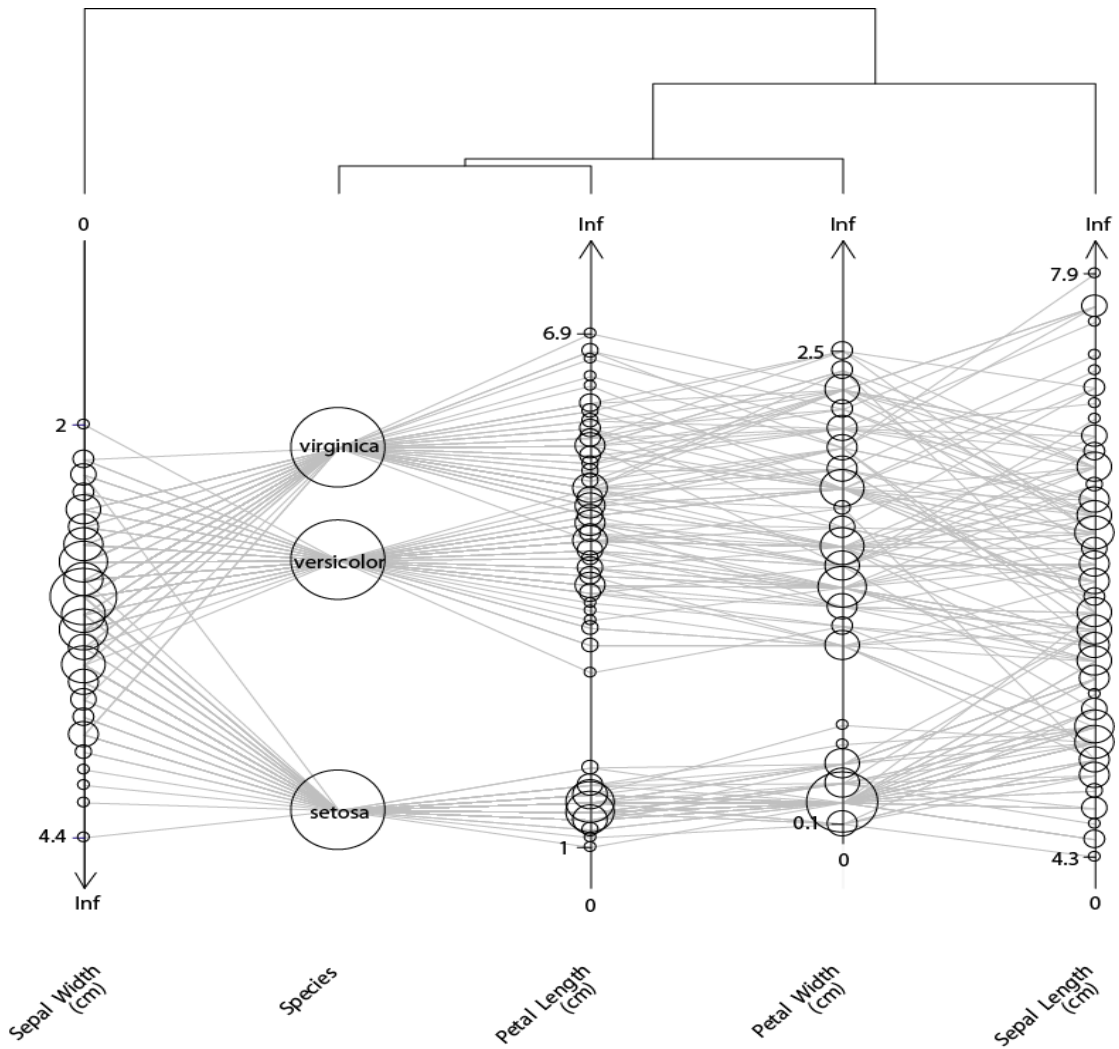
Textile Plot

$$\mathbf{y}_j = \alpha_j \mathbf{1} + \beta_j \mathbf{x}_j, \; j = 1, 2, \ldots, p$$

$$\sum_{j=1}^{p} \|\mathbf{y}_j - \boldsymbol{\xi}\|^2 \underset{\boldsymbol{\xi}, \{\alpha_j\}, \{\beta_j\}}{\rightarrow} \min$$

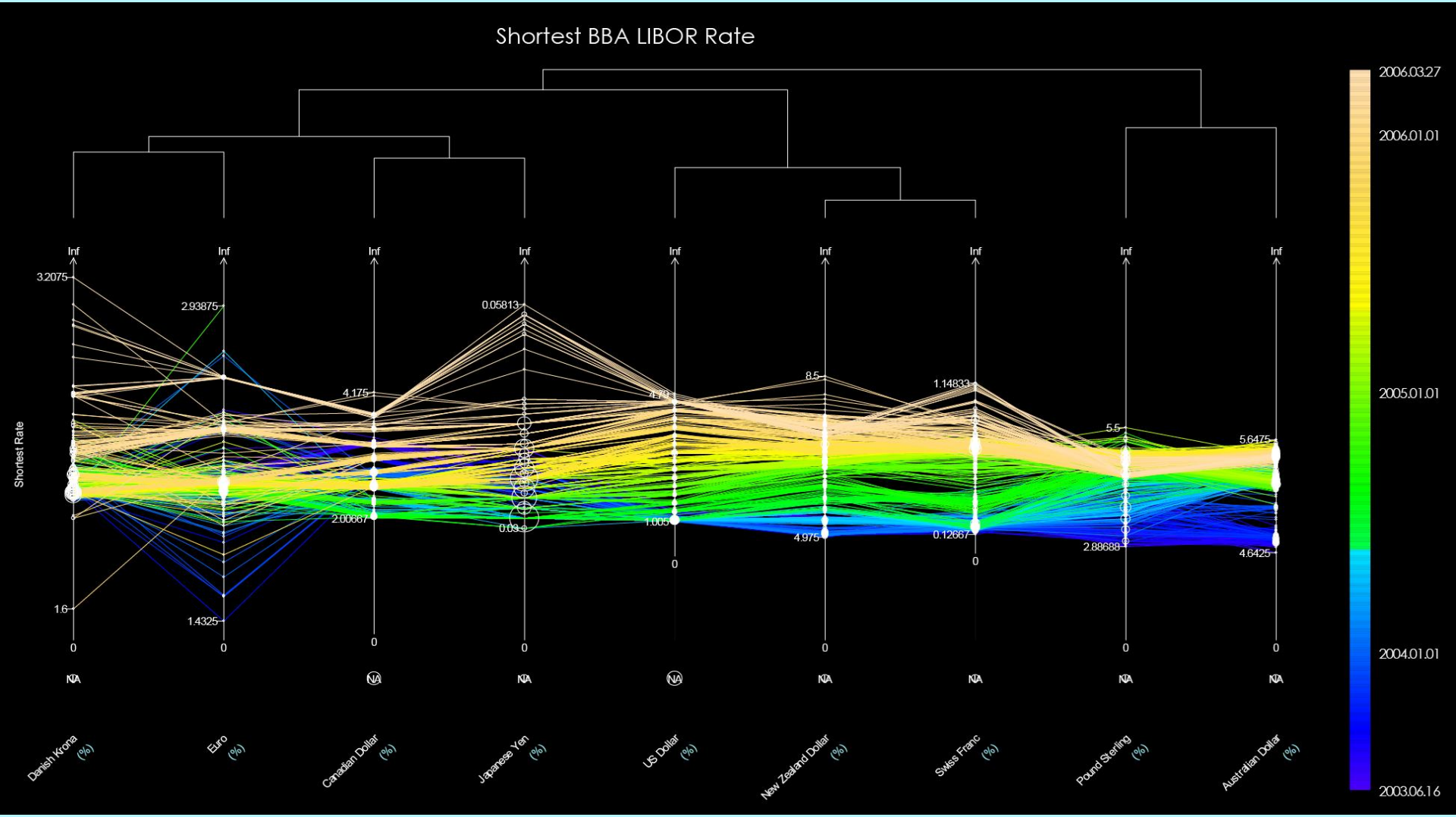
http://www.stat.math.keio.ac.jp/TextilePlot/index.html

Textile Plot



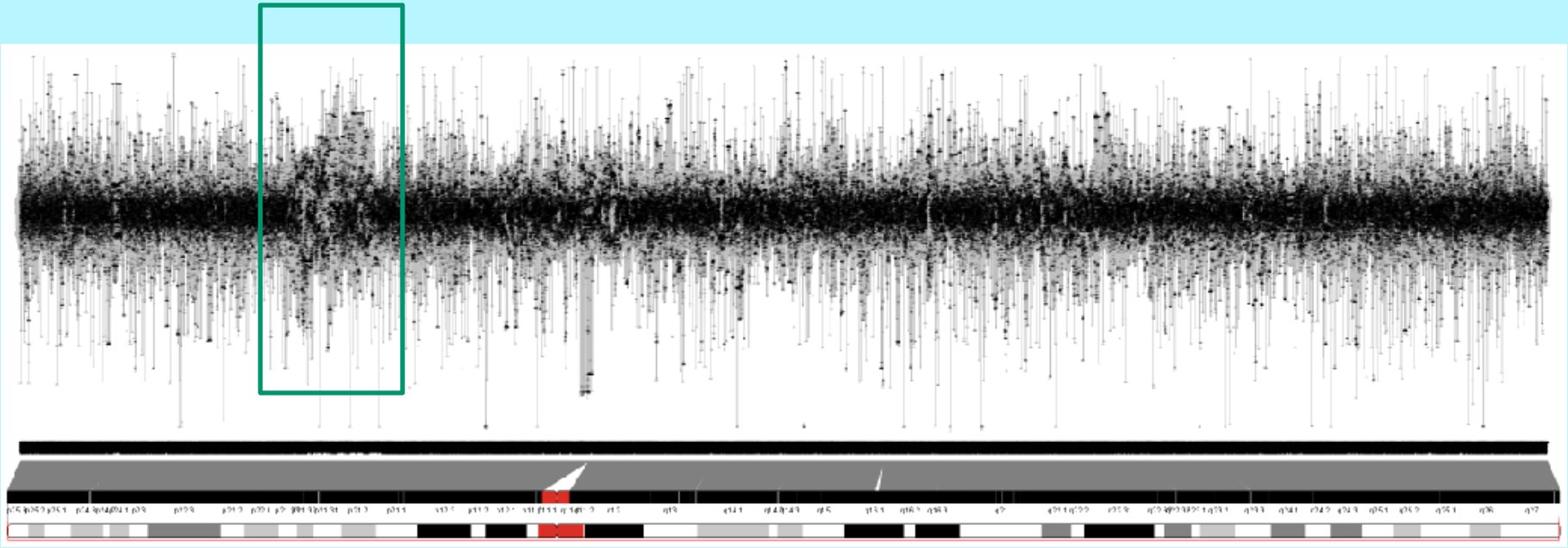
Kumasaka, N. and Shibata, R. (2008) High Dimesional Data Visualisation: the Textile Plot, *Computational Statistics & Data Analysis*, **52**, 3616-3644.

Interest Rate Data
(9 dimensional time series)



HapMap Data (100,000dimensional)

HLA (human leukocyte antigen) Region



60 caucasian’s No.6 Chromsome SNPs

Kernel PCA Algorithms

$$X : n \times p \ data\, matrix$$

$$\text{Eigen Vecotors of } X^T X \ (X^T X \ \ p \times p \ matrix)$$

$$X = UDV^T \Rightarrow X^T X = VD^2 V^T$$

$$n < p \Rightarrow U : n \times n,\ D : n \times n,\ V : p \times n$$

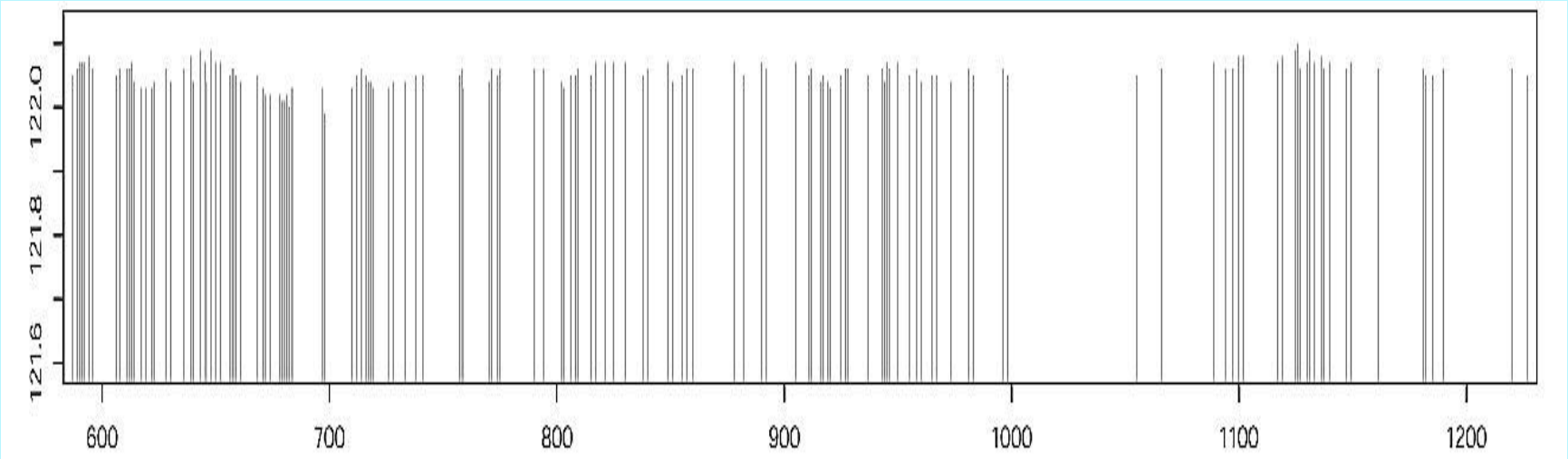
$$V = X^T U D^{-1} : Eigen\ Vectors$$

W. Wu et al.(1997), The Kernel PCA algorithms for wide data Part I , Chemometrics and Intelligent Laboratory Systems 35, 163-172

Data Driven Modelling (1)

Call Price of a dollar (New Quotes) , 219,027 Quotes for 2 weeks from 2002-06-22~

yen



seconds

$t_1, t_2, \ldots, t_n$: Clustered Point Process

$X(t_1),\ X(t_2),\ \ldots,\ X(t_n)$: Marked Point Process

Known Models

High Frequency Financial Data Analysis

$X(t_1),\ X(t_2),\ \ldots,\ X(t_n)$

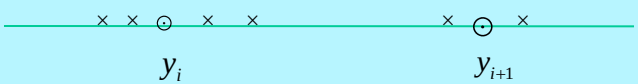
smoothing $\Rightarrow$ $X(1),\ X(2),\ \ldots,\ X(N)$: Regularly Spaced Time Series

Cox(Doubly Stochastic Poisson) Process for $t_1, t_2, \ldots, t_n$

$$P(N(A)=k) \sim Poisson(\lambda_A),\ \lambda_A : Random \sim F_A$$

Cluster Process for $t_1, t_2, \ldots, t_n$

$$N(A) = \sum_i N(A \mid y_i)$$



Modelling FX new bid prices as a clustered marked point process, *Proceedings in Computational Statistics 2006*, Ed. Alfredo Rizzi and Maurizio Vichi, 1565-1572, 2006, Physica-Verlag, Heidelberg.

Binary Clustering

$$(t_1, t_2, \cdots, t_n) = (t_1, \cdots, t_k) \cup (t_{k+1}, \cdots, t_n) : a \ \ partition$$

$$M_0 = \max_{\lambda} L(\triangle t_1, \cdots, \triangle t_n; \lambda) : \text{maximum loglikelihood of a Poisson Model}$$

vs.

$$M_1(k) = \max_{\lambda_1, \lambda_2} (L(\triangle t_1, \cdots, \triangle t_k; \lambda_1) + L(\triangle t_{k+1}, \cdots, \triangle t_n; \lambda_2)) :$$

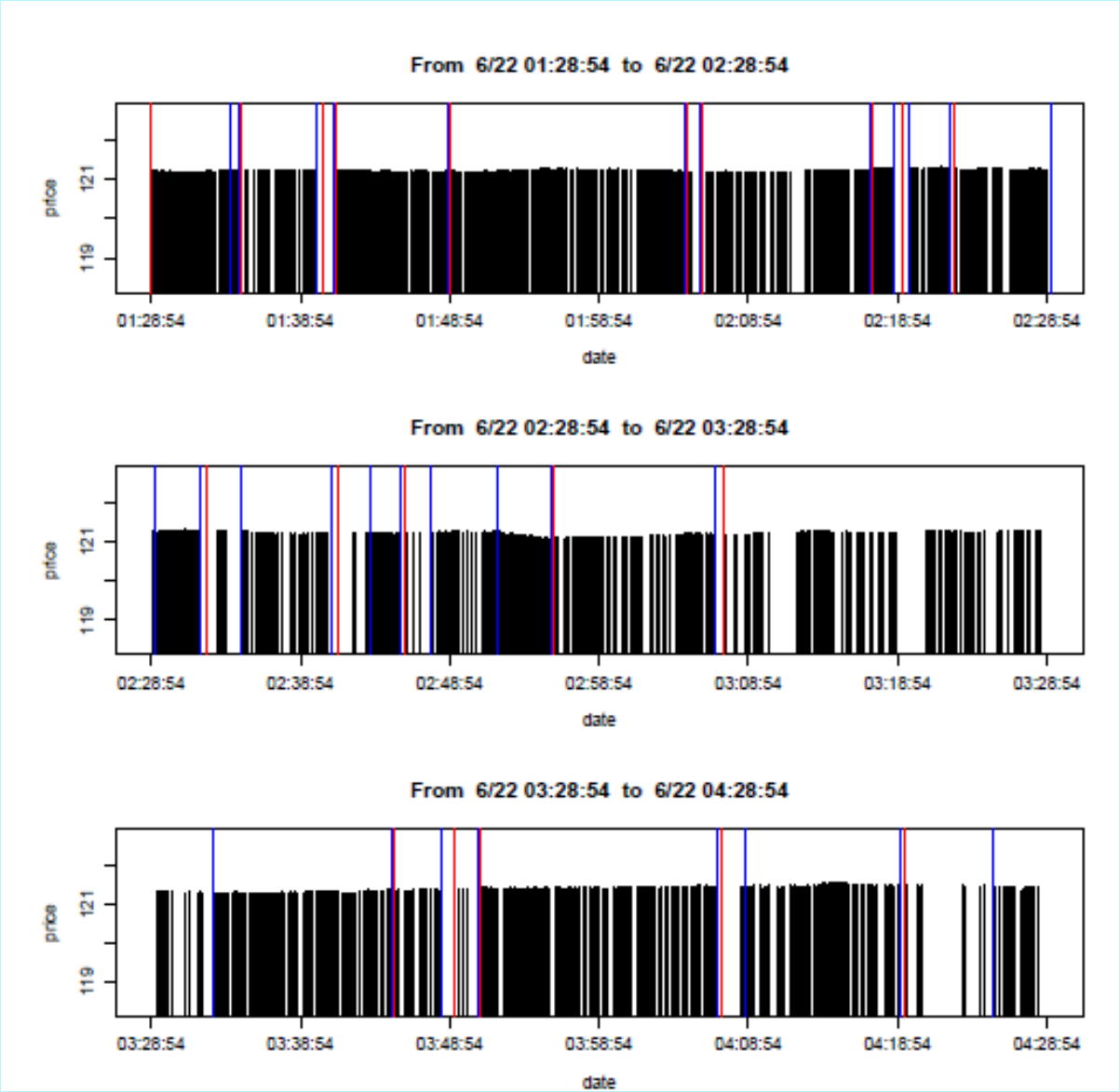
$$\text{maximum likelihood of two Poisson models}$$

$$\lambda_1 : \text{ intensity in } (t_1, t_k], \ \lambda_2 : \text{intenisty in } (t_k, t_n]$$

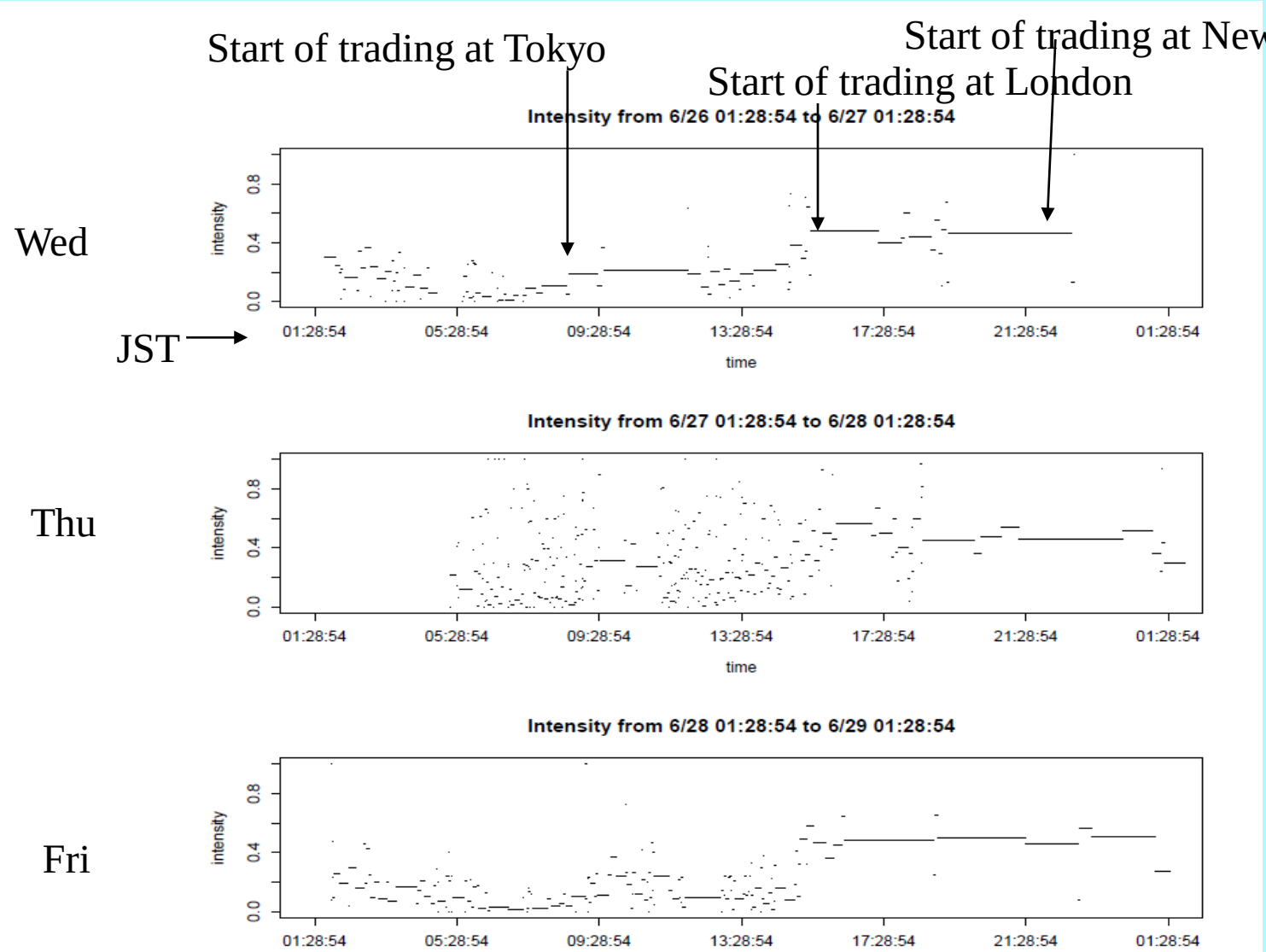
$$Find \ \ k : \frac{M_1(k)}{M_0} \to \max, \ Check \ significance$$

$$Significant \Rightarrow Divide \ \ at \ t_k$$

2245 Clusters for 219,027 quotes



Intensity of Each Cluster



Model for Mark

Well Established Continuous Time Model

$$R_i = \log X(t_i) - \log X(t_{i-1}) = \log \frac{X(t_i)}{X(t_{i-1})} : \log\ return$$

$$R_i \sim N(\mu_i \Delta t_i,\ \sigma^2 \Delta t_i) \ \text{ for } \Delta t_i = t_i - t_{i-1}\ ,\ i = 1, \ldots, n$$

$$\Longleftrightarrow \quad \log X(t) : Brownian\ Motion$$

$$\Longleftrightarrow \quad dX = \sigma X\, dB$$

Random Time Sampling

$\Delta t_i \sim \mathrm{Exp}(\lambda_i)$

$R_i \sim N(0,\ \sigma^2 \Delta t_i)$ for $\Delta t_i = t_i - t_{i-1}$, $i = 1, \ldots, n$

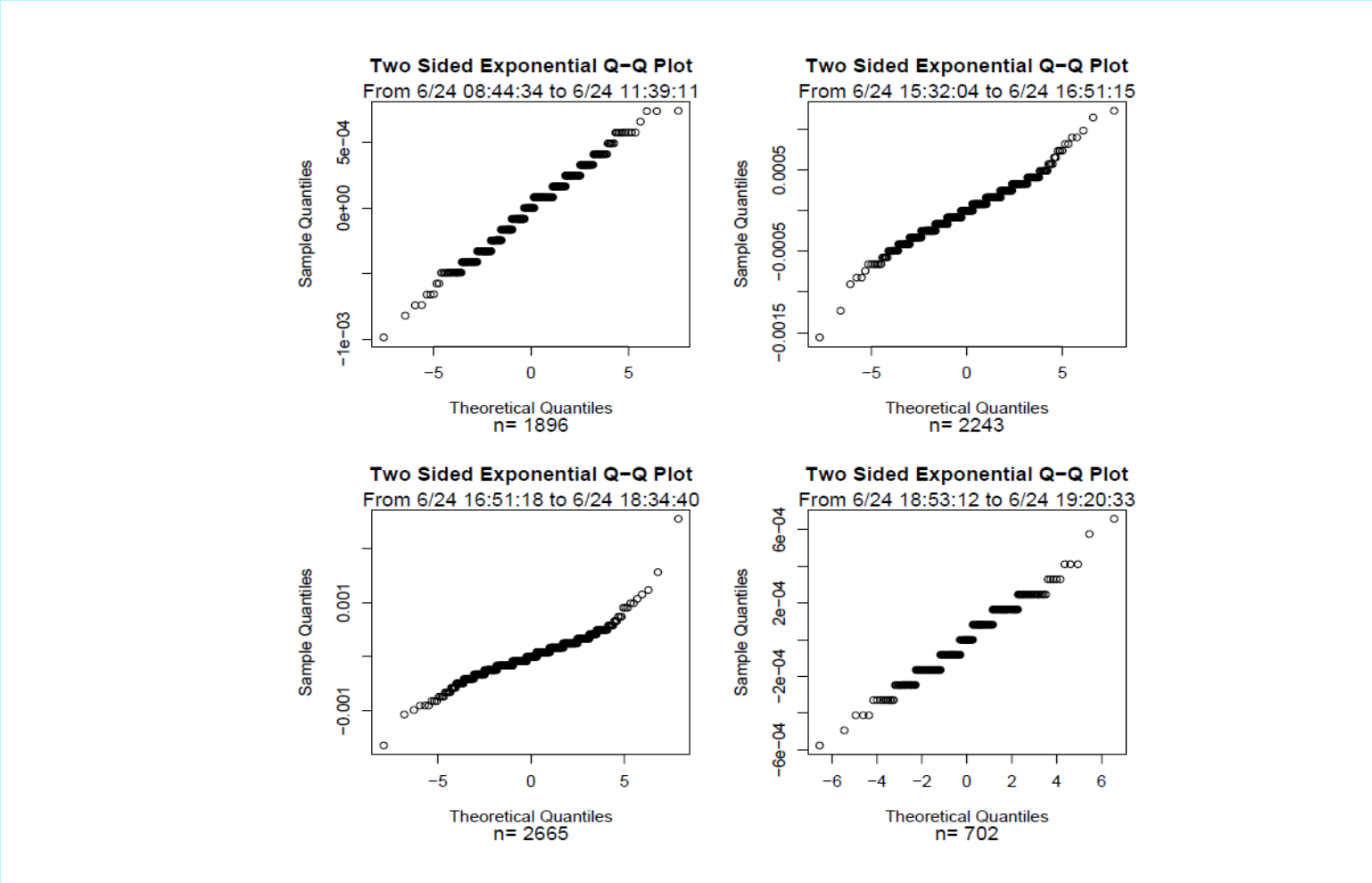
$$R_i \sim \mathrm{TExp}\left(\frac{\sqrt{2\lambda_i}}{\sigma}\right), i = 1, \ldots, n$$

$\mathrm{TExp}(\alpha)$: *Laplace Distribution* (Two Sided Exponential)

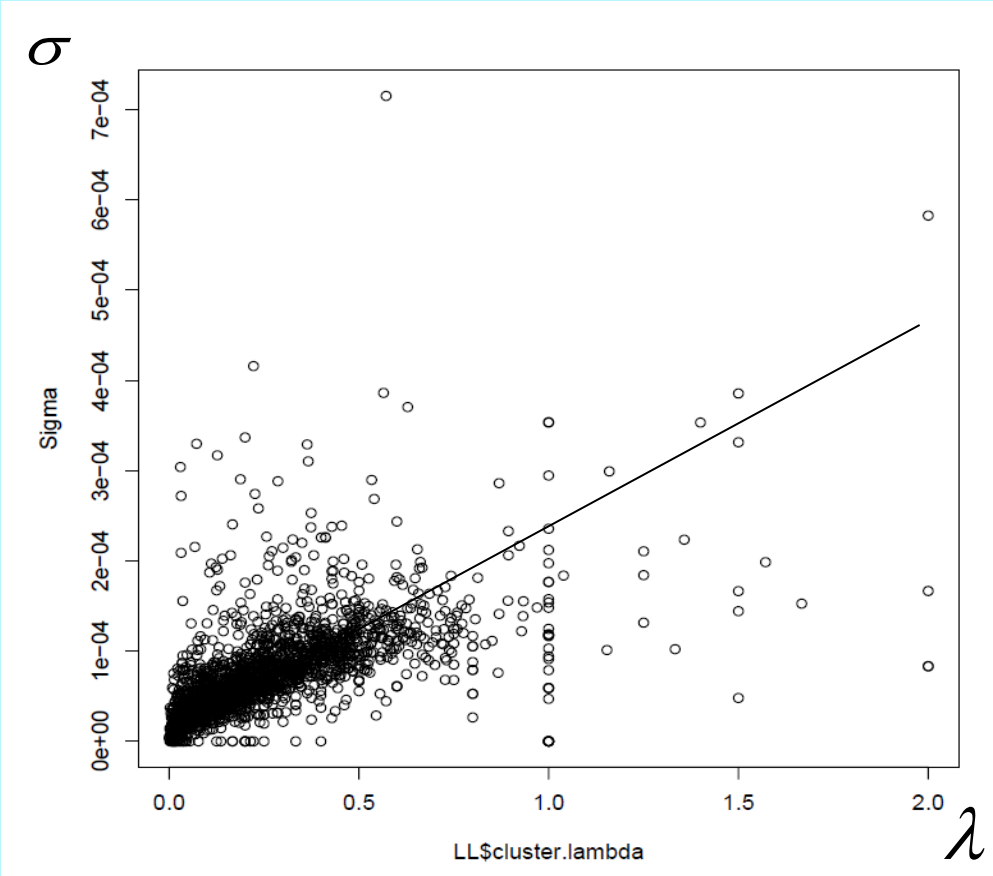
$$Density: f(x, \alpha) = \frac{\alpha}{2} e^{-\alpha |x|}$$

Special Case of Variance Gamma Model

Visual Diagnostics (QQ-Plot)



Relation between intensity and volatility



$$\sigma = 0.000411 + 0.0001245\lambda$$

Result

Binary Clustering Algorithm

$$C_1 = \{t_1, t_2, \ldots, t_{k_1}\}, C_2 = \{t_{k_1}, t_{k_1+1}, \ldots, t_{k_2}\}, \ldots : Homogeneous\,Clusters$$

$$C_i \sim \text{Poisson Point Process with inensity}\,\lambda_i \quad i = 1, 2, \ldots, 2245,$$

$$\lambda_i : \; Activity\ of\ Market$$

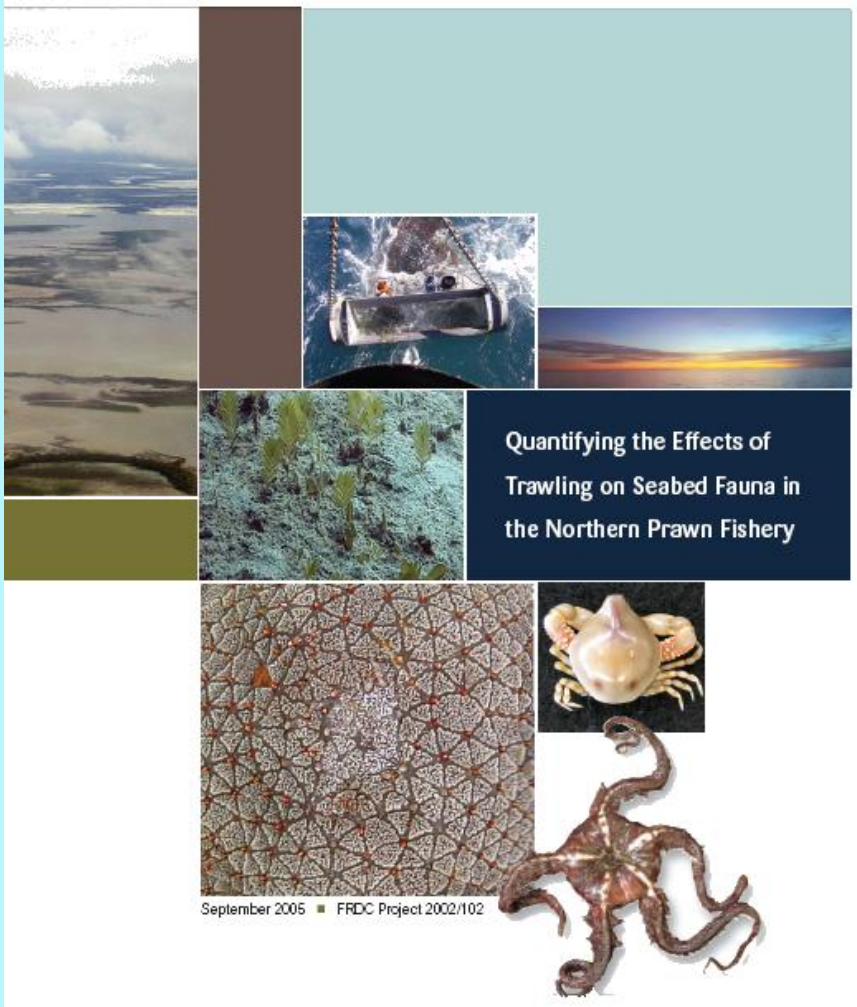
Random Sampling From Log Normal Process

$$R_i = \log X(t_i) - \log X(t_{i-1}) \sim \mathrm{TExp}(\alpha_i), \; \alpha_i = \frac{\sqrt{2\lambda_i}}{\sigma_i}$$

Relation Between Volatility and Inensity

$$\sigma_i = 0.000411 + 0.0001245\lambda_i$$

Collaboration with Marine Biologists



Quantifying the Effects of Trawling on Seabed Fauna in the Northern Prawn Fishery

Mick Haywood (Principal Investigator), Burke Hill (Co-Investigator), Anthea Donovan, Wayne Rochester, Nick Ellis, Andrzej Welna, Scott Gordon, Sue Cheers, Karl Forcey, Ian McLeod, Chris Moeseneder, Greg Smith, Fiona Manson, Ted Wassenberg and Steve Thomas

CSIRO Marine and Atmospheric Research

Petra Kuhnert, Geoff Laslett, Charis Burridge and Sarah Thomas

CSIRO Mathematical and Information Sciences.

October 2005

CSIRO

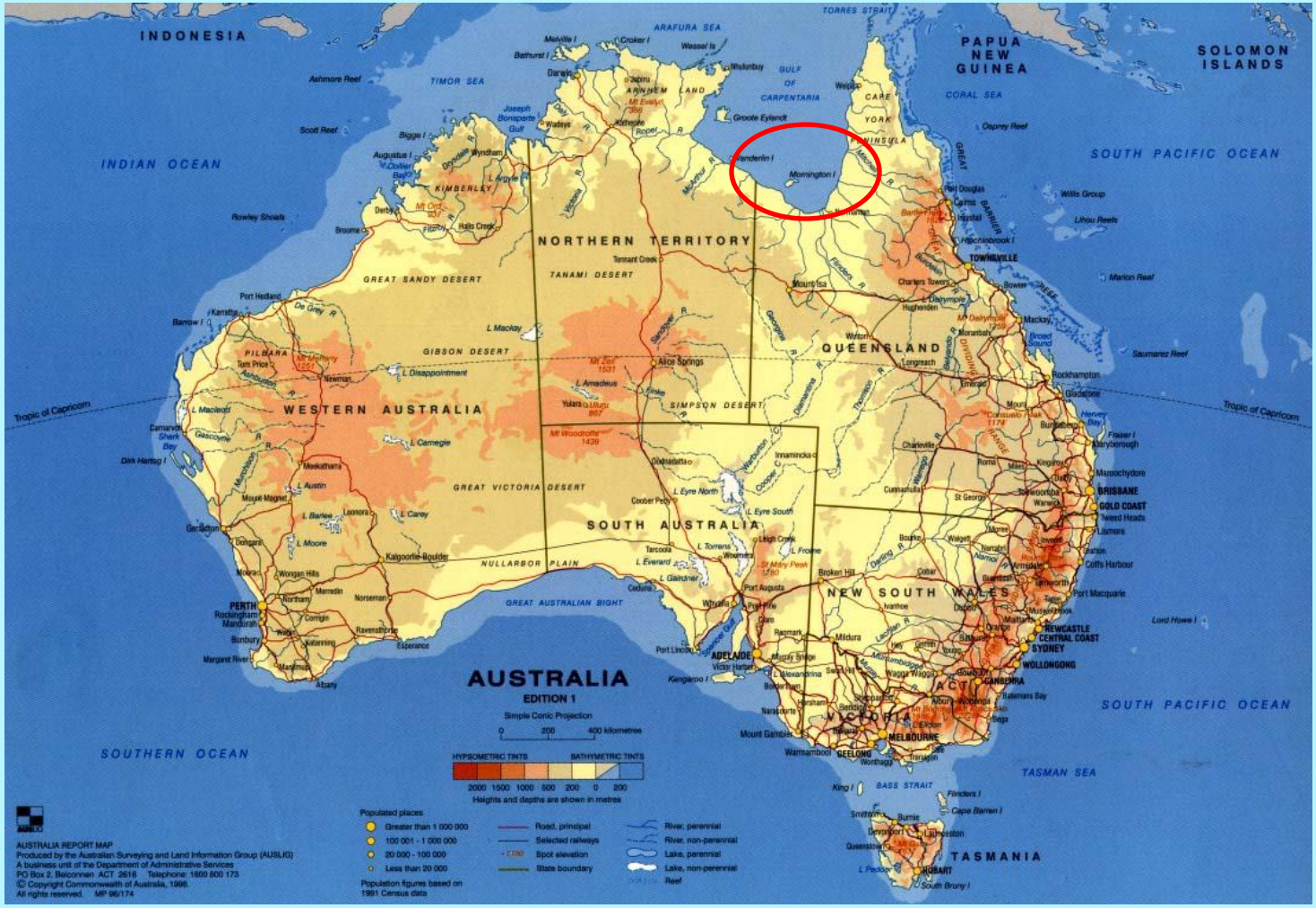
Australian Government

Fisheries Research and Development Corporation

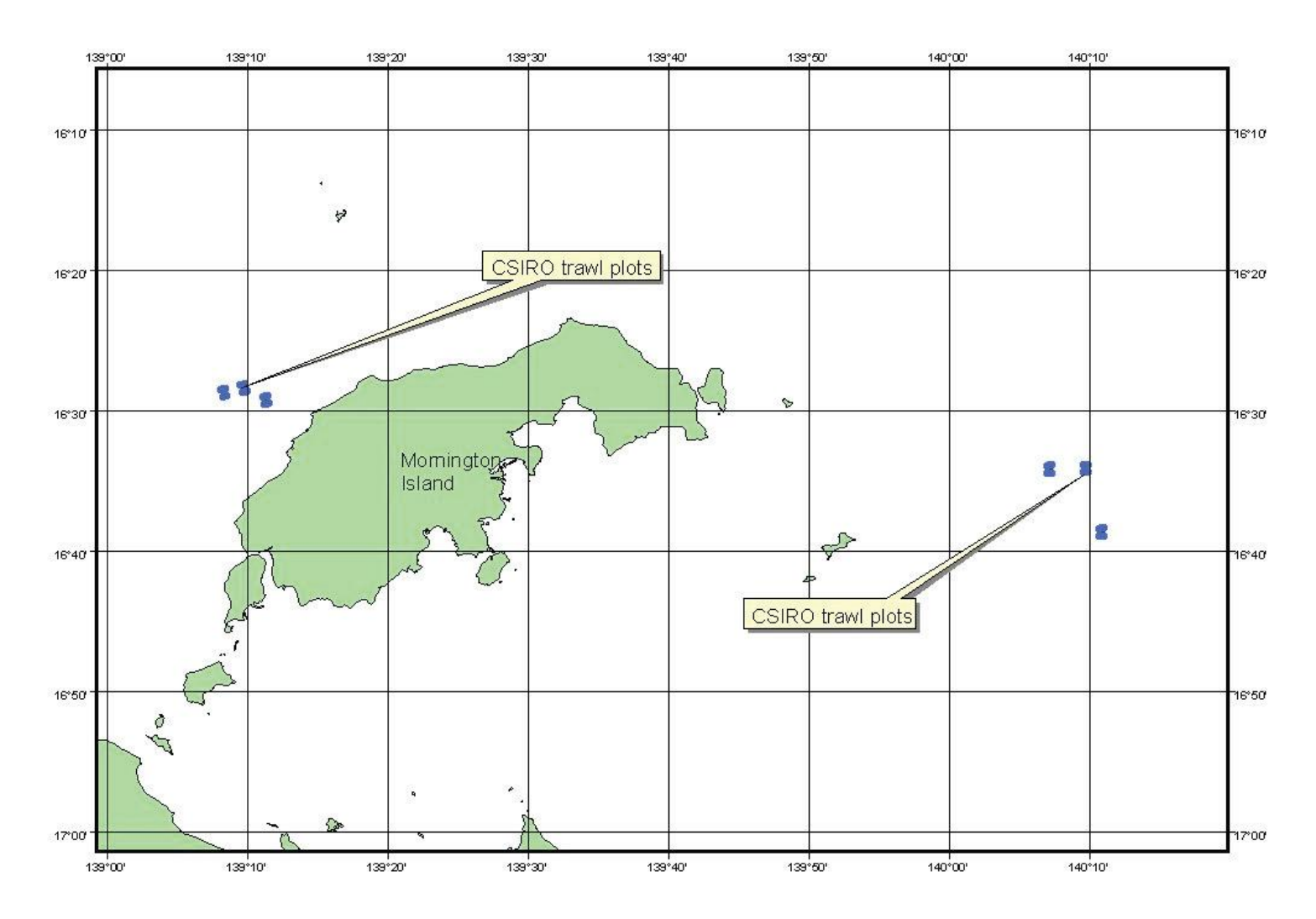
FRDC Project 2002/102

Effects of Trawling Subprogram: Quantifying the Effects of Trawling on Seabed Fauna in the Northern Prawn Fishery

Northern Prawn Fishery(NPF)

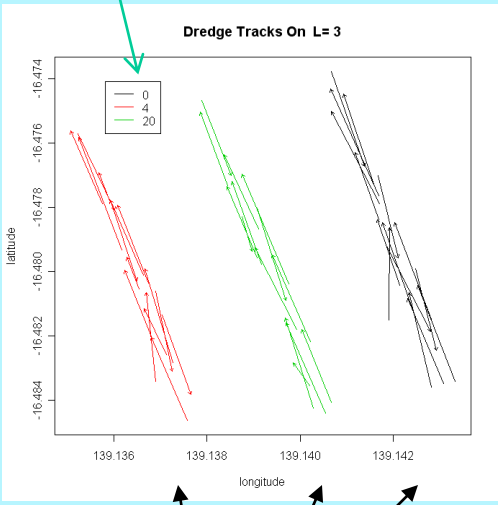


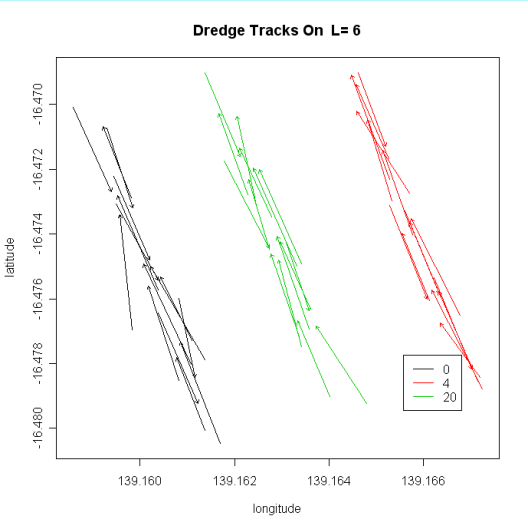
Plots for Experiment



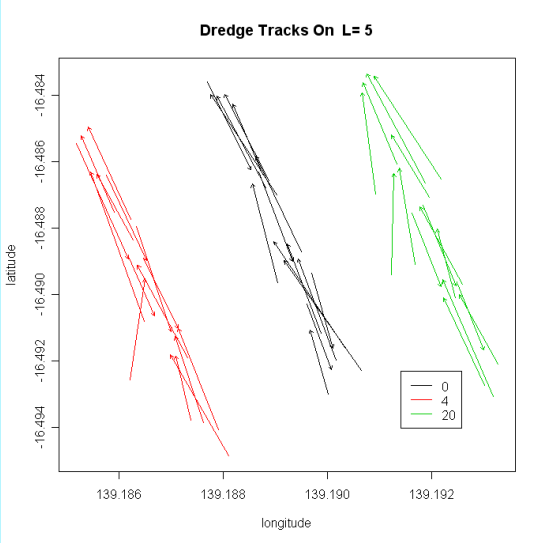
Experiment 1 (West Region)

treatments (0,4,20 trawls)





(Before After 6month 12month 18month) x 3



Dredge Data

$nperha$ ($Number\ per\ hectare$) and $gperha$ ($Gram\ per\ hectare$)
on $SPCODE \times LOCATION \times TREATMENT \times time$

$SPCODE$: $Species\ Code$ ($777\ Species\ listed$)
$LOCATION$; 3, 5, 6, 9 , 10, 12
$TREATMENT$: $Number\ of\ Trawls$; 0, 4, 20
$time$: $Dredge$; Before, Immediately After, 6month, 12month, 18month

many zeros in $nperha$ ($zero\ catches$)

Zero Inflated data

No Significant Results in the report

Problems in the precedent analysis

Based on normalised data

nperha (*Number per hectare*) and *gperha* (*Gram per hectare*)

⇓

prevent introduction of proper stochastic model

Disregarding Heterogenuity of Species

⇓

Ambiguity of the result

To be honest with data

New approach

Start from Un-normalized data

number (*Number per dredge*) , *biomass* (*Gram per dredge*)

and *swept_area* (hectare)

Analysis for each species when no trawl

Model of abundance (*number*)

Model of fertility (*biomass*)

Effect of trawls

Shift of Model

Model of Abundance

$Data$

$(N_i, \alpha_i),\ i = 1, 2, \ldots, n,\ \ (n = 54: number\ of\ dredges)$

$N_i : the\ number\ of\ a\ species\ caught$

$\alpha_i : dredge\ area\ (swept_area)$

$$N_i \sim F_{\alpha_i, \theta},\ i = 1, 2, \ldots, n$$

A standard model for zero inflated data

$$P(N=x)=\binom{k+x-1}{k-1}q^x(1-q)^k,\ x=0,1,2,\ldots,\ k>0,\ 0<q<1 : \textit{Negative Binomial}$$

Characterisation

1. $Po(\lambda),\ \ \lambda \sim G_A(\nu,\alpha)$ (*Gamma mixture of Poissons*)

2. $N = X_1 + X_2 + \cdots + X_Q,\quad P(X_i = x) = \dfrac{-1}{\log(1-\theta)}\dfrac{\theta^x}{x},\ x=1,2,\ldots,\ \ Q \sim Po(\lambda)$

   (*Poisson compound of logarithmic distribution*)

Natural ? Meaningful ?

More natural and meaningful models

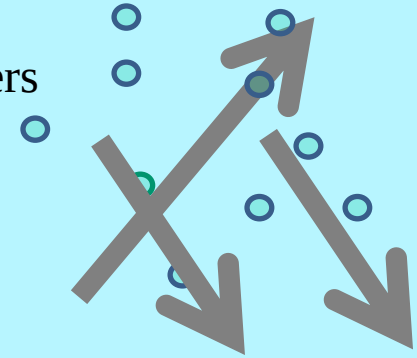
1. *Poisson*

$$N_i \sim Po(\lambda\alpha_i)$$

2. Thomas

$$N_i = X_1 + X_2 + \cdots + X_Q, \;\; Q \sim Po(\lambda\alpha_i), \;\; X_j - 1 \sim Po(\phi)$$

clusters



3. $Poisson - Stopped - Sum - Poisson\,(Neyman\,Type\;A)$

$$N_i = X_1 + X_2 + \cdots + X_Q, \;\; Q \sim Po(\lambda\alpha_i), \;\; X_j \sim Po(\phi)$$

$$\text{Poisson} \subset \text{Thomas}$$

	spcode	Thomas			Poisson		
		lambda	phi	p.value	lambda	p.value	
海綿	10009801	0.00024881	2.68942509	0.89579738	0.00091796	0.00269481	
	10100801	0.00009265	0.48615404	0.73037574	0.00013769	0.54554425	
ヒドロ虫	11000800	0.00064257	0.00000000	0.80394053	0.00064257	0.95790992	
	11000806	0.00277660	4.45498035	0.00005437	0.01514630	0.00000000	Hydroidea Hydroid
サンゴ	11185002	0.00023775	2.86109620	0.79044760	0.00091796	0.00018419	
	11185004	0.00009295	0.48139097	0.64461615	0.00013769	0.44191576	
	11185801	0.00094061	0.02470853	0.70491975	0.00096386	0.26923420	
	11185802	0.00122122	0.05233908	0.52606968	0.00128514	0.91931475	
	11287001	0.00045492	1.42142219	0.60090932	0.00110155	0.03217933	
	11305802	0.00004602	0.99461255	0.97589181	0.00009180	0.85980253	
苔虫	20322806	0.00083615	0.04294898	0.26950892	0.00087206	0.92443973	
	20322809	0.00110155	0.00000000	0.33583094	0.00110155	0.71702666	
ゴカイ	22000800	0.00164534	3.85386095	0.06638640	0.00798623	0.00000000	
２枚貝	22021800	0.00014085	0.30349176	0.39995788	0.00018359	0.12568541	
	23201003	0.00029343	0.25135926	0.32261361	0.00036718	0.07940725	
	23226804	0.00018922	0.69794920	0.37481027	0.00032129	0.00238565	
	23227800	0.00014169	0.61969787	0.57112789	0.00022949	0.33524963	
	23231004	0.00081666	0.51746063	0.87114873	0.00123924	0.04130224	
	23236801	0.00009280	0.97828532	0.78357973	0.00018359	0.28449912	
	23237005	0.00185507	2.09273499	0.18645938	0.00573723	0.00000000	
	23270003	0.00044438	0.44600848	0.50158818	0.00064257	0.07622128	
	23270004	0.00069004	1.06197241	0.23992961	0.00142283	0.00797612	
	23272803	0.00107192	1.61190984	0.95229313	0.00279977	0.00007270	
	23335802	0.00044280	0.14019940	0.73779216	0.00050488	0.47296950	
	23351001	0.00019298	0.42698874	0.41426760	0.00027539	0.04717759	
	23356800	0.00013965	0.64337096	0.36173839	0.00022949	0.19894235	
	23380800	0.00029612	0.70499181	0.95864161	0.00050488	0.02627723	
	23380803	0.00035620	0.93283964	0.61040348	0.00068847	0.14478101	
	23380806	0.00076285	0.32366481	0.89218856	0.00100975	0.61510690	
	23499800	0.00014119	0.30028990	0.73465115	0.00018359	0.27868389	

ナメクジ,海かたつむり	24057800	0.00069593	0.18713001	0.49533194	0.00082616	0.93340224	
	24086800	0.00043595	0.57925131	0.31892429	0.00068847	0.22236062	
	24125802	0.00112329	0.96128303	0.72575710	0.00220310	0.00052686	
	24125803	0.00053707	1.30742301	0.56162888	0.00123924	0.00694410	
	24145800	0.00023915	0.15153848	0.07903124	0.00027539	0.24542598	
	24200800	0.00142513	1.38325660	0.68454032	0.00339644	0.00010900	
	24200801	0.00039130	0.05567449	0.22682320	0.00041308	0.82908674	
	24200802	0.00044697	0.84835159	0.92281212	0.00082616	0.26212679	
	24200803	0.00009281	0.97807208	0.90543221	0.00018359	0.94700094	
	24202803	0.00023714	0.35481991	0.07552104	0.00032129	0.93754049	
	24208800	0.00004607	0.99270221	0.10480623	0.00009180	0.47284131	
	24215801	0.00150771	2.19642271	0.31250002	0.00481928	0.00000002	
	24215802	0.00050122	0.09886820	0.14142356	0.00055077	0.67771253	
	24231800	0.00039225	0.28712506	0.88663058	0.00050488	0.47929101	
	24395800	0.00009249	1.48116833	0.53906199	0.00022949	0.57359426	
	24420800	0.00144002	3.71720190	0.00256896	0.00679289	0.00000000	Mollusca Nudibranchs
ヒトデ	25001817	0.00009230	1.98374760	0.97285968	0.00027539	0.01047434	
	25105800	0.00139390	2.16104794	0.87501389	0.00440620	0.00000002	
	25111800	0.00118652	0.54730524	0.53818879	0.00183592	0.11395986	
	25111802	0.00187262	4.63728609	0.57346736	0.01055651	0.00000000	
	25122800	0.00057471	0.03822181	0.86835709	0.00059667	0.91018721	
	25122801	0.00047519	0.64199173	0.92840800	0.00078026	0.49912635	
	25122802	0.00184942	0.98539161	0.76876792	0.00367183	0.00124200	
	25125804	0.00024035	0.14577831	0.89268950	0.00027539	0.22099929	
クモヒトデ	25160801	0.00015455	1.37578759	0.89640535	0.00036718	0.59389678	
	25160804	0.00050459	2.00173098	0.15219585	0.00151463	0.00010346	
	25160805	0.00114563	3.64737885	0.89451441	0.00532415	0.00000000	
	25160807	0.00034497	0.06439654	0.83029418	0.00036718	0.65219449	
	25160811	0.00170737	2.68285505	0.81919310	0.00628801	0.00000000	
	25176810	0.00014142	1.27190746	0.38047834	0.00032129	0.72481122	
ウニ	25241802	0.00115220	2.46564015	0.45622487	0.00399312	0.00000000	
	25241803	0.00104027	1.60314783	0.91694440	0.00270797	0.00001799	
	25241804	0.00110238	1.16504295	0.77957991	0.00238669	0.00283116	
	25241807	0.00019666	0.63368519	0.67065842	0.00032129	0.89447202	
	25266800	0.00226823	1.73173642	0.14308971	0.00619621	0.00000047	
	25266801	0.00033800	2.66643304	0.45109436	0.00123924	0.00005751	
	25306800	0.00118435	3.53417704	0.00885776	0.00537005	0.00000000	Echinodermata Echinoidea

ホヤ	35011800	0.00268449	13.39603504	0.00000003	0.03864601	0.00000000	Ascidiacea Ascidian
	35011810	0.00061612	1.45833772	0.72768316	0.00151463	0.00010296	
	35011812	0.00004607	0.99262453	0.94853937	0.00009180	0.78819839	
	35012002	0.00045687	0.00462123	0.27651347	0.00045898	0.20502143	
魚	37118001	0.00009305	0.47985299	0.76691616	0.00013769	0.44938084	
	37118002	0.00009319	0.96999931	0.52505812	0.00018359	0.37483424	
	37296029	0.00051290	0.16331983	0.90646841	0.00059667	0.09064487	
	37313001	0.00009297	0.48105730	0.84061716	0.00013769	0.33167207	
	37351016	0.00004607	0.99270221	0.59068011	0.00009180	0.58333770	
	37427803	0.00033923	0.21769782	0.85071980	0.00041308	0.79542260	
	37463800	0.00050341	0.36761974	0.82228955	0.00068847	0.75863908	
	37465800	0.00049790	0.47492561	0.91221105	0.00073437	0.62736885	
	37465801	0.00035697	1.05722716	0.51002895	0.00073437	0.01247339	
	37465802	0.00009300	0.97410852	0.74678059	0.00018359	0.39689510	

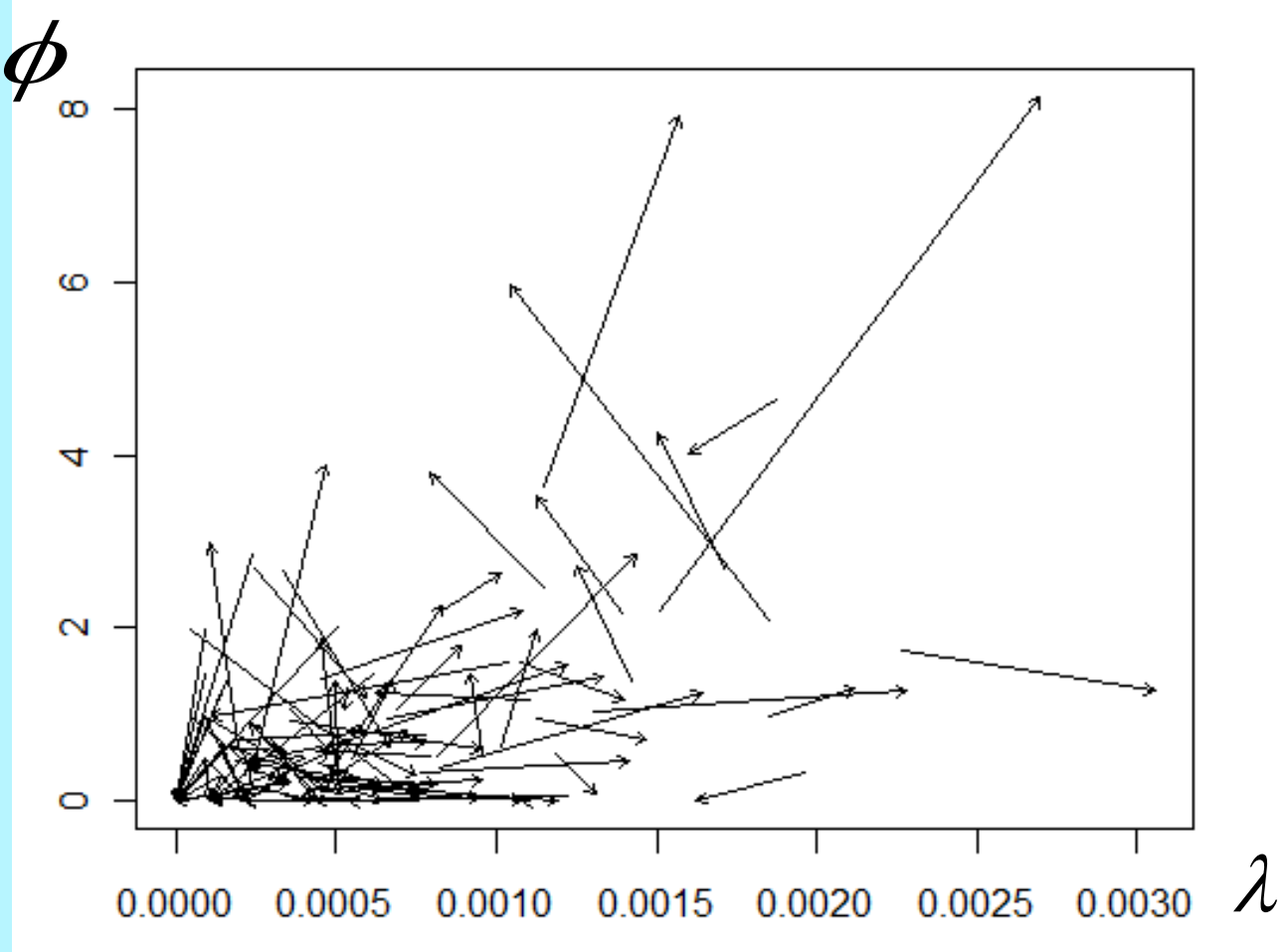
out of 123 species

Both NG: 5

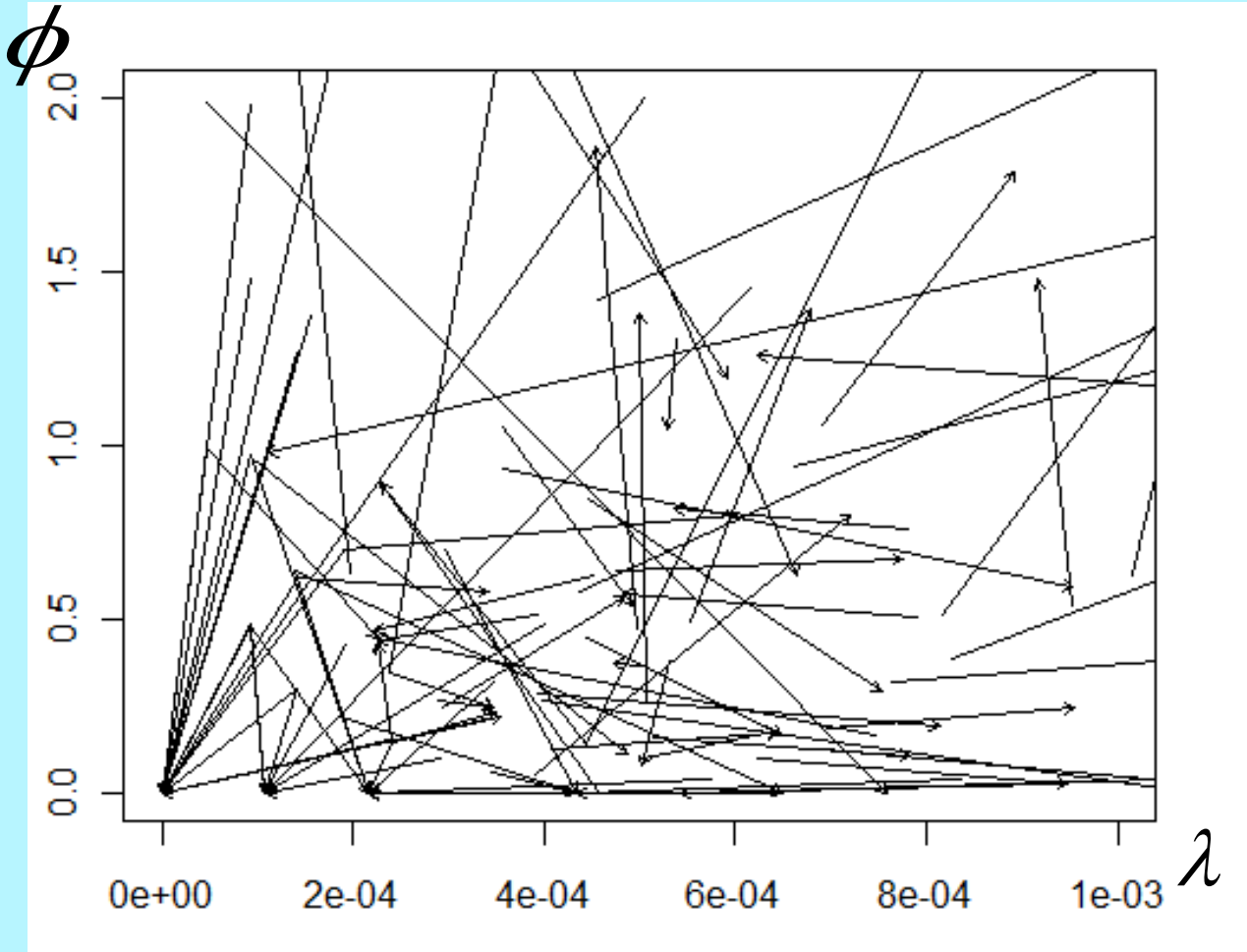
Poisson NG, Thomas OK : 40

Poisson OK, Thomas NG: 3

Effect of Trawls



Effect of Trawls (Zoom)



Model of fertility

$$Data$$

$$(m_i, n_i), i = 1, 2, \ldots, n \quad for\ each\ species$$

$$m_i : biomass$$

$$n_i : number\ caught$$

$$Distribution$$

$$m_i \sim F_{n_i, \theta}, i = 1, 2, \ldots, n$$

Growth model

Stochastic differential equation

$$dX_t = rX_t\left(1-\frac{X_t}{k}\right)dt + \sigma X_t dB, X_t > 0$$

r : rate of growth

k : limit of growth

σ : standard deviation

$B(t)$: Brownian motion

General theory

$$dX_t = a(X_t)dt + b(X_t)dB_t. \qquad P(X_t \in A) = \int_A p(t,x)dx$$

Kolmogorov's forward equation

$$\frac{\partial p(t,x)}{\partial t} = -\frac{\partial}{\partial x}\left(p(t,x)a(x)\right) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(p(t,x)b(x)^2\right), x > 0$$

$$p(t,0) = 0, \qquad \int_0^\infty p(t,x)dx = 1$$

$$\Downarrow$$

$$p(x) = \frac{C_1}{b(x)^2}\exp\left(\int \frac{2a(x)}{b(x)^2}dx\right) : \textit{stationary distribution}$$

Stationary distribution of growth model

$$dX_t = rX_t\left(1-\frac{X_t}{k}\right)dt + \sigma X_t dB,$$

$$\Downarrow$$

$$p(x) = \frac{C_1}{\sigma^2}x^{-2}\exp\left(\frac{2r}{\sigma^2}\int\left(\frac{1}{x}-\frac{1}{k}\right)dx\right)$$

$$= C_2 x^{-2+\frac{2r}{\sigma^2}}\exp\left(-\frac{2r}{k\sigma^2}x\right) = \frac{1}{\alpha^{\nu}\Gamma(\nu)}x^{\nu-1}\exp\left(-\frac{1}{\alpha}\cdot x\right) \qquad \nu = \frac{2r}{\sigma^2}-1, \alpha = \frac{k}{\nu-1}$$

Gamma Distribution

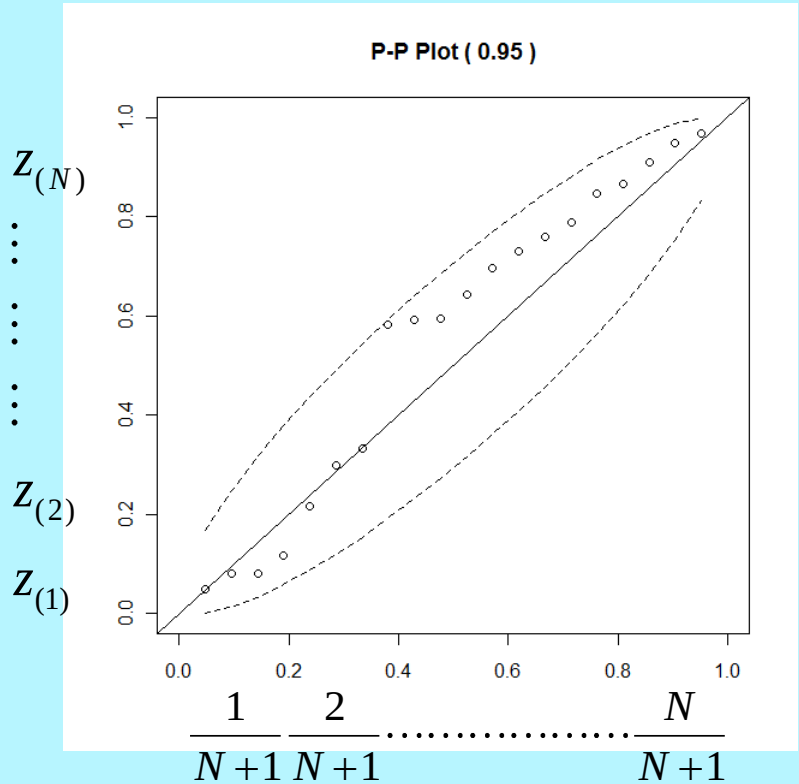
$$m_i \sim F_{n_i,\theta} = G_A(n_i\nu, \alpha),\ i = 1, 2, \ldots, n$$

Diagnostics

Visual Diagnostics (Probability-Probability Plot)

$$z_i = F_{n_i \nu, \alpha}(m_i),\ i = 1, \cdots, n$$

$$z_{(i)} \sim Beta(i, N - i + 1)$$

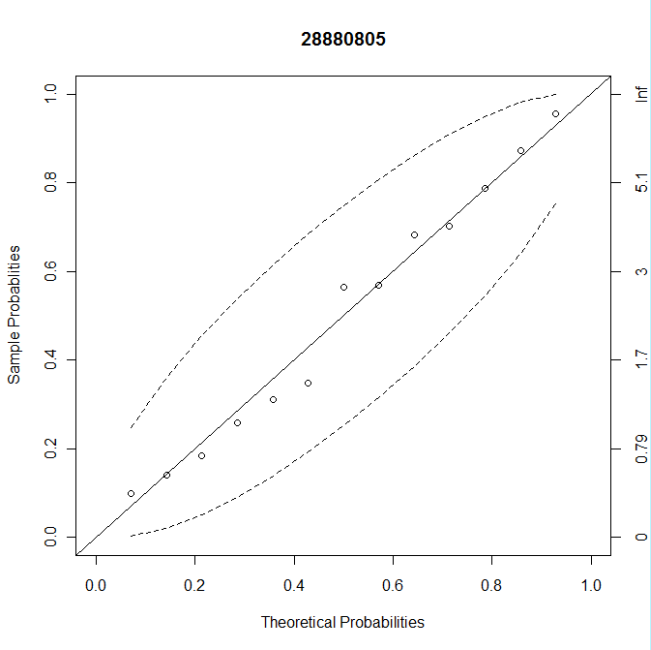


Numerical Diagnostics ($p - value\ of$ Kolmogorov-Smirnov Test Statistic)

$$D = \sup_x \left| F_n(x) - x \right|$$

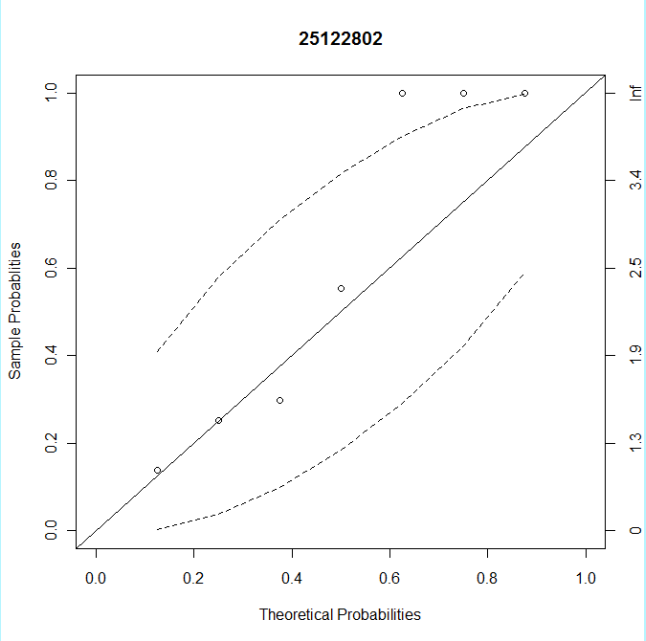
$$F_n(x) = \#\{z_i \le x\} / n \ (i = 1, 2, \cdots, n)$$

Examples



SPCODE: 28880805
(Cruscacea)
Location: east
Time: Before

ν : 1.096
α : 2.921
p-value: 0.989



SPCODE: 25122802
(Asteroidea)
Location: west
Time : Before

ν : 3.543
α : 0.679
p-value: 0.153

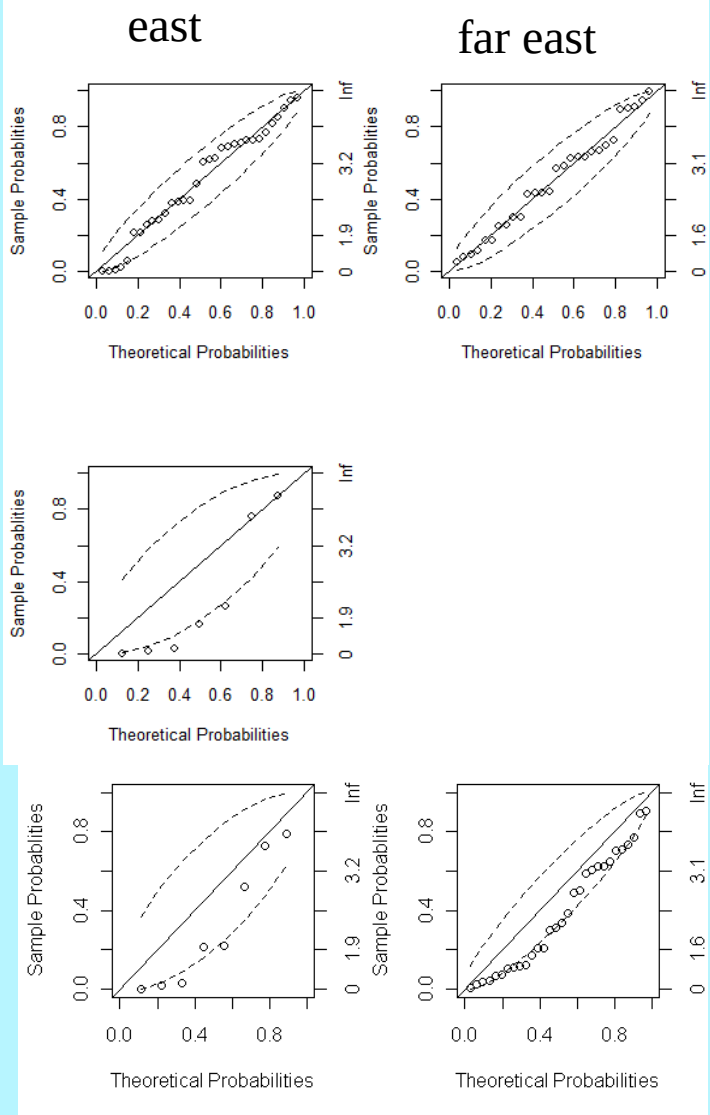
Trawling effect

Example
species : 24200800

No trawl

4 trawls

20 trawls



lighter
in stochastic order

Suggestions

- Without prejudice but with real data
- Honest with data
- Take time for exploring data
- Deep insight into the phenomena behind data
- Good models
  - Simple, Clear cut, Generality,…
- Fact Finding

- No manuals
- Collaboration with experienced data scientists