Combining cheap massive commercial data and unbiased scientific survey: a zero inflated model under preferential sampling

Marie-Pierre Etienne[1], Eric Parent[1], Jean-Baptiste Lecomte[1], Robyn Forrest[2]

([1]) AgroParisTech / INRA
([2]) Pacific Biological Station - Nanaimo, BC- Canada

AgroParisTech — Institut des sciences et industries du vivant et de l'environnement — Paris Institute of Technology for Life, Food and Environmental Sciences

INRA — Science & Impact

AppliBugs Juin 2013

Fisheries Management

Reliable stock assesment require reliable relative abundance indices :

- Unbiased or with constant multiplicative bias,
- As precise as possible (low variability)
- Acceptable for stakeholders

Major data sources:

- Scientific survey data,
- Commercial fisheries data.

Scientific Survey data

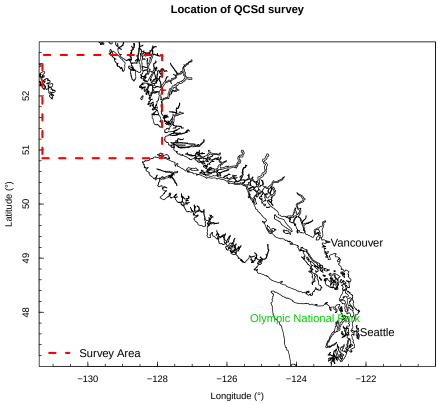


Figure: Queen Charlotte area - DFO (GroundFish division) - Focus on Dover Sole

Scientific Survey data

- Scientific campaigns are organized regularly to monitor the species of interest.
- Mostly random sampling or stratified random sampling design.
- Produce unbiased but highly variable and expensive abundance indices series.
- Stakeholders have difficulty to accept random sampling : "why sample some zone where there is no fish"?

Scientific Survey data

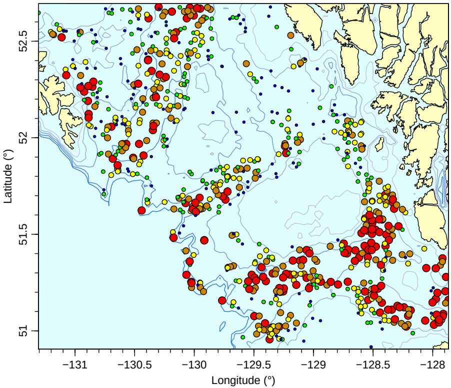


Figure: QCSd - The records are the weights of Dover Sole caught.

Commercial Fisheries data

- Cheap and massive data.
- Roughly used, produce biased abundance indices.
- Stakeholders are part of the collection process.

Commercial Fisheries data

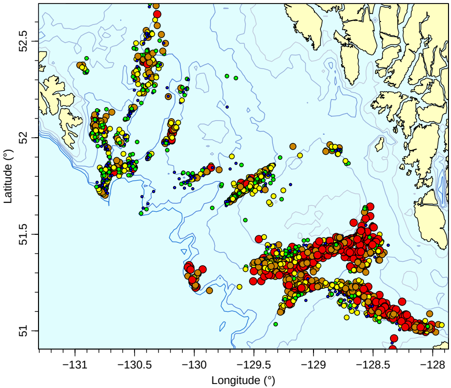
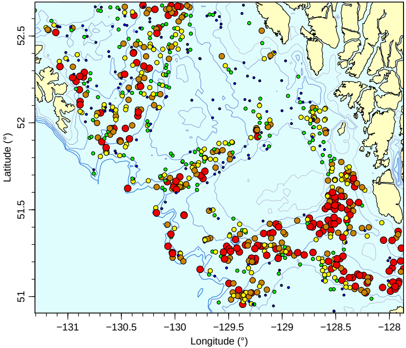
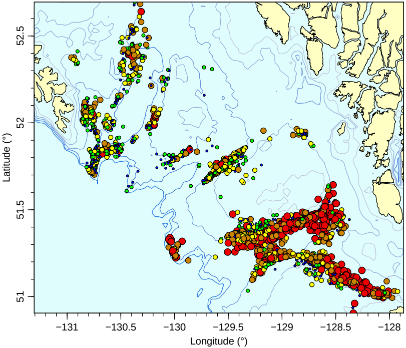


Figure: Com. Fish.: The records are the weight of Dover Sole Catch

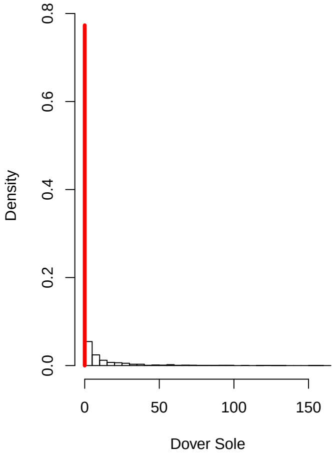
Two data sources





FISHING.ID	YEAR	LAT	LONG	SWEPT.AREA	DOVER.SOLE
1700	1999	51.43333	-129.2117	0.770	0.0
6550	2005	51.06167	-128.2867	0.610	210.1
...					

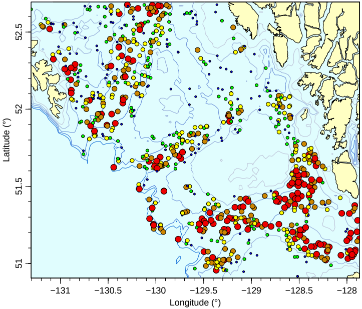
Zero-Inflated data



Continuous Zero Inflated data

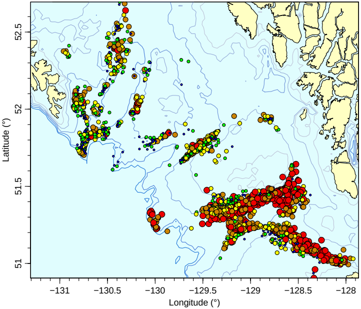
- Classically, high proportion of zeros,
- Appart from 0, continuous biomass data

Data are spatially correlated



- The biomass repartition is somehow continuous,
- Data are spatially correlated,
- This correlation should be accounted for.

Location of Commercial Fisheries catch



- Fishermen target specific species,
- Location of the catch are highly related to the amount of local biomass,
- This information must be accounted for.

Modelling challenges

The resulting model should represent,

- Zero inflated,
- Spatially correlated,
- and preferentially sampled,

data,
for building a relative abundance index.

Modelling challenges

The resulting model should represent,

- Zero inflated,
- Spatially correlated,
- and preferentially sampled,

data,
for building a relative abundance index.

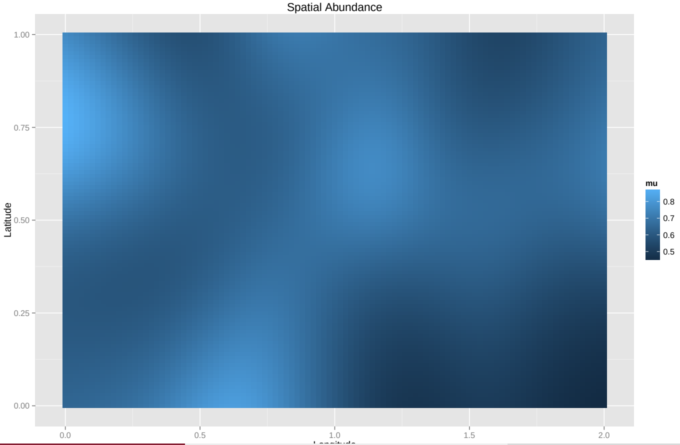
Taking benefits of hierarchical modelling

Biomass Model: Log Gaussian Cox Process

Let $\mu(s)$ be the local abundance and define the intensity of an inhomogenous Poisson Process which may be thought as the fish repartition.

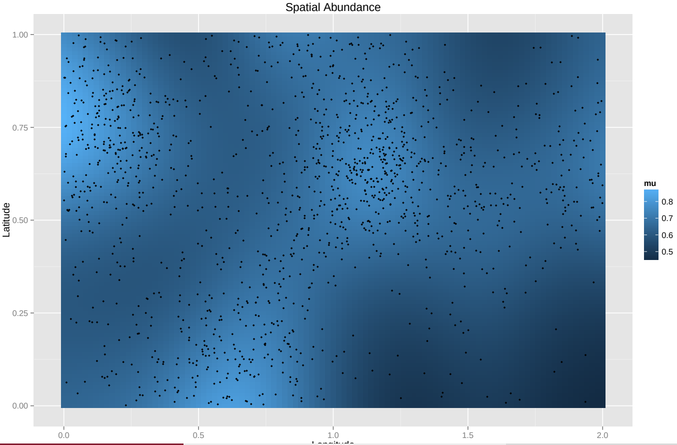
Biomass Model: Log Gaussian Cox Process

Let $\mu(s)$ be the local abundance and define the intensity of an inhomogenous Poisson Process which may be thought as the fish repartition.



Biomass Model: Log Gaussian Cox Process

Let $\mu(s)$ be the local abundance and define the intensity of an inhomogenous Poisson Process which may be thought as the fish repartition.



Biomass Model: Log Gaussian Cox Process

Let $\mu(s)$ be the local abundance and define the intensity of an inhomogenous Poisson Process which may be thought as the fish repartition.
To account for heterogeneity,

$$log(\mu(s)) = \alpha_0 + Z(s),$$

where $Z(s)$ is a gaussian random field (GRF) with covariance function $c(s,t) = \exp - \frac{d(s,t)^2}{2\phi^2}$

Observation Process : Compound Poisson Process

One fishing event, for a given swept area A :

$$N(A) \sim \mathcal{P}\left(\int_A \mu(s)ds\right),$$

is the number of fish caught.

Observation Process : Compound Poisson Process

One fishing event, for a given swept area A :

$$N(A) \sim \mathcal{P}\left(\int_A \mu(s)ds\right),$$

is the number of fish caught.
Approximation:

$$\int_A \mu(s)ds \approx |A|\mu_{s_A}.$$

$$Y(A) = \sum_{i=1}^{N(A)} \xi_i,$$

where ξ_i are iid random variable (weight).

Observation Process : Compound Poisson Process

Commercial

$$Y_s^C = \sum_{i=1}^{N_s^C} \xi_{s,i},$$
$$N_s^C \sim \text{Poisson}(||A_s|\mu_s),$$
$$\xi_{s,i}^C \sim \text{Exp}(\rho),$$

$$\mathbb{E}(Y_s^C) = \frac{\mu_s}{\rho}$$

Scientific

$$Y_s^S = \sum_{i=1}^{N_s^S} \xi_{s,i},$$
$$N_s^S \sim \text{Poisson}(|A_s|\mu_s),$$
$$\xi_{s,i}^S \sim \text{Exp}(\rho^S), \quad \rho^S = q\rho$$

$$\mathbb{E}(Y_s^S) = \frac{q\mu_s}{\rho}$$

$$\mathbb{P}(Y_s^i = 0) = \exp\left(-|A_s|\mu(s)\right)$$

called LOL model in Ancelet & al, 2010; Lecomte & al 2013

Full model specification

Process model :

$$log(\mu(s)) = \alpha_0 + Z(s),$$

where $Z(s)$ GRF with covariance function $c(s,t) = \exp -\frac{d(s,t)^2}{2\phi^2}$.

Data model :

$$Y_s^k \sim LOL(\mu_s, \rho^k)$$

But dimension issues when the number of observations increase.
Reduction dimension using random basis function.

A 2 D discrete convolution of a gridded (latent) structure

The points of the grid are denoted $g = 1..G$.

$$X(g) \underset{iid}{\sim} N(0, \sigma_x^2)$$

Convolution kernel K_θ between any data point s and grid location g

$$K_\theta(s, g) = \exp - \frac{d^2(s, g)}{\phi^2}$$

Discrete convolution for site $s, s = 1..S$:

$$Z(s) = \sum_{g=1}^{G} K_\theta(s, g) X(g) + m(s)$$

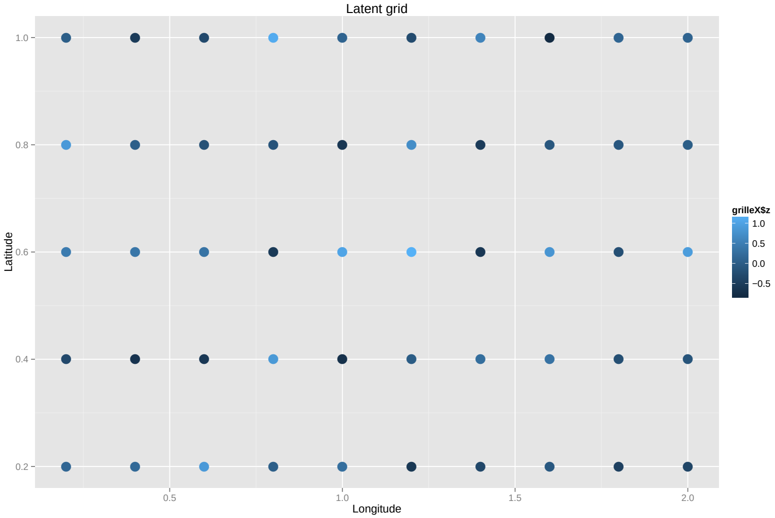
$$m(s) = \alpha_0 + \alpha_1 \times Depth(s) + \dots$$

Preferential sampling

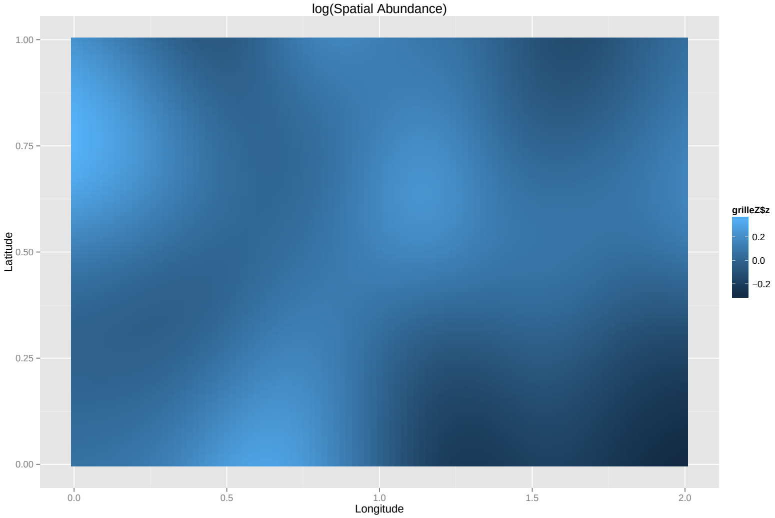
Commercial fisheries focus on area with high abundance. The position of the commercial catch are modeled as an inhomogenous Poisson point process conditionned to have *NCom* points.

$$(S_1^C, \ldots, S_{NCom}^C) \sim IPP(\mu(s))$$

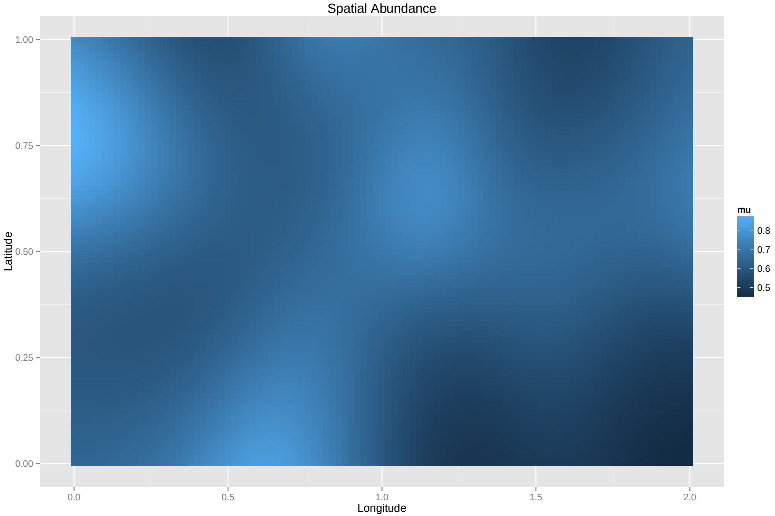
Full model specification with graphics



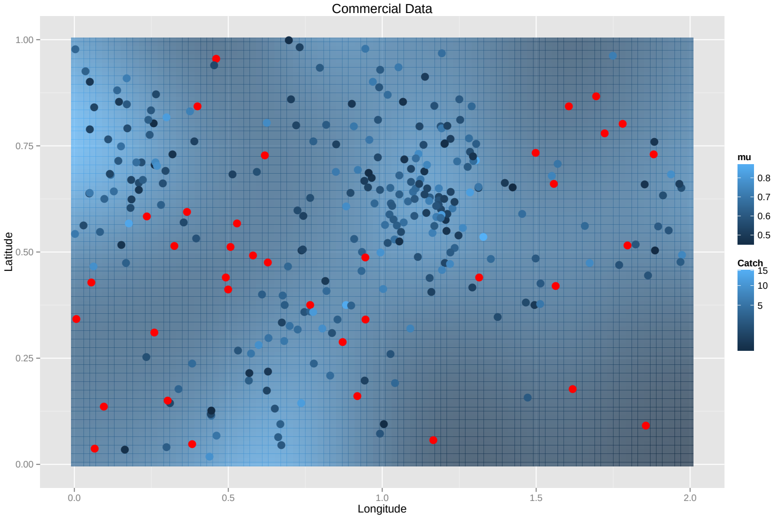
Full model specification with graphics



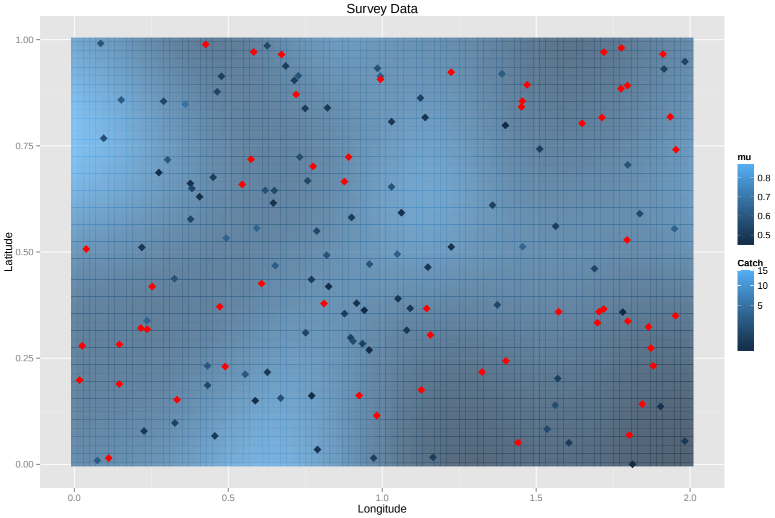
Full model specification with graphics



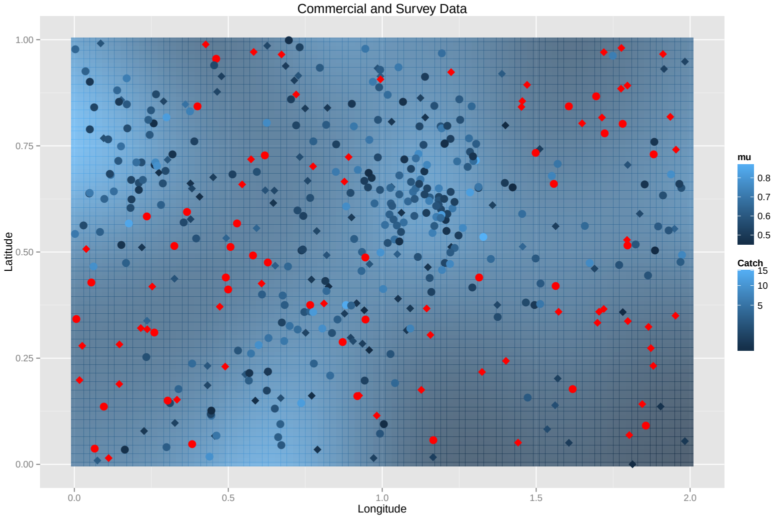
Full model specification with graphics



Full model specification with graphics



Full model specification with graphics



Model Summary

Data:
$\mathbf{S}^C = (S_1^C, \ldots, S_{nCom}^C)$, Commercial locations of catch : Poisson Process
$\mathbf{Y}^C = (Y_1^C, \ldots, Y_{nCom}^C)$, Actual commercial catch : LOL model
$\mathbf{Y}^S = (Y_1^S, \ldots, Y_{nScien}^S)$, Actual survey catch: LOL model

Latent layer:
$\mathbf{Z} = K_\phi \mathbf{X}$, with $\mathbf{X} = (X_1, \ldots, X_G)$ Independant centered gaussian variables, with variance σ^2.

Parameters:
$\theta = (\sigma^2, \phi, \boldsymbol{\alpha}, \rho^C, \rho^S)$

Indice:
$I = \int_s \mu(s) ds$

Problems - Likelihood

Complete likelihood

$$[\mathbf{Y}^C, \mathbf{Y}^S, \mathbf{S}^C, \mathbf{X}|\theta] = [\mathbf{Y}^C|\mathbf{S}^C, \mathbf{X}, \theta]\ [\mathbf{Y}^S|\mathbf{S}^S, \mathbf{X}, \theta]\ [\mathbf{S}^C|\mathbf{X}, \theta][\mathbf{X}|\theta]$$

Likelihood

$$[\mathbf{Y}^C, \mathbf{Y}^S, \mathbf{S}^C|\theta] = \int_{\mathbf{X}} [\mathbf{Y}^C, \mathbf{Y}^S, \mathbf{S}^C, \mathbf{X}|\theta] d\mathbf{X}$$

Computing the likelihood

Computing the likelihood

Monte Carlo approximation

$$MC : [\mathbf{Y}^C, \mathbf{Y}^S, \mathbf{S}^C|\theta] \approx \frac{1}{M}\sum_{m=1}^{M}[\mathbf{Y}^C, \mathbf{Y}^S, \mathbf{S}^C|\mathbf{X^m}, \theta], \quad \mathbf{X^m} \sim [X|\theta]$$

Computing the likelihood

Monte Carlo approximation

$$MC : [\mathbf{Y}^C, \mathbf{Y}^S, \mathbf{S}^C|\theta] \approx \frac{1}{M}\sum_{m=1}^{M}[\mathbf{Y}^C, \mathbf{Y}^S, \mathbf{S}^C|\mathbf{X^m}, \theta], \quad \mathbf{X^m} \sim [X|\theta]$$

Importance sampling approximation

$$IS : [\mathbf{Y}^C, \mathbf{Y}^S, \mathbf{S}^C|\theta] \approx \frac{1}{M}\sum_{m=1}^{M}\frac{[\mathbf{Y}^C, \mathbf{Y}^S, \mathbf{S}^C|\mathbf{X^m}, \theta][\mathbf{X^m}|\theta]}{q_\theta(\mathbf{X^m})}, \quad \mathbf{X^m} \sim q_\theta(.)$$

Inference on parameters

Inference on parameters

- Numerical optimisation of the likelihood.

Inference on parameters

- Numerical optimisation of the likelihood.
- Full Metropolis Hasting algorithm

Inference on parameters

- Numerical optimisation of the likelihood.
- Full Metropolis Hasting algorithm

Inference on parameters

- Numerical optimisation of the likelihood.
- Full Metropolis Hasting algorithm
- Pseudo Marginalized MCMC Algorithm - Andrieu & Roberts (2009)

Inference on parameters

- Numerical optimisation of the likelihood.
- Full Metropolis Hasting algorithm
- Pseudo Marginalized MCMC Algorithm - Andrieu & Roberts (2009)
    - propose $\theta^* \sim q(\cdot|\theta)$ and $X^* \underset{iid}{\sim} \prod_{m=1}^{M} q^S(Z^{(m)}|\theta^*)$

Inference on parameters

- Numerical optimisation of the likelihood.
- Full Metropolis Hasting algorithm
- Pseudo Marginalized MCMC Algorithm - Andrieu & Roberts (2009)
  - propose $\theta^* \sim q(\cdot|\theta)$ and $X^* \underset{iid}{\sim} \prod_{m=1}^{M} q^S(Z^{(m)}|\theta^*)$
  - accept them with probability $\rho((\mathbf{X},\theta),(\mathbf{X}^*,\theta^*)) = \frac{\tilde{\pi}(\theta^*,\mathbf{Z}^*)}{\tilde{\pi}(\theta,\mathbf{Z})} \times \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}$

with $\tilde{\pi}(\theta,\mathbf{Z})$ given by the importance sampling quantity

$$\tilde{\pi}(\theta,\mathbf{Z}) = \frac{1}{M}\sum_{m=1}^{M} \frac{[\mathbf{Y}^C,\mathbf{Y}^S,\mathbf{S}^C,\mathbf{Z}^{(m)},\theta]}{q^S(\mathbf{Z}^{(m)}|\theta)}$$

Finding good importance function q_θ

Moment based method with kernel smoothing :

Finding good importance function q_θ

Moment based method with kernel smoothing :

1. Using $\mathbb{E}(Y_s) = \frac{\mu(s)A_s}{\rho}$,

Finding good importance function q_θ

Moment based method with kernel smoothing :

1. Using $\mathbb{E}(Y_s) = \frac{\mu(s)A_s}{\rho}$,

$$\hat{Z}(s) = log\left(\frac{\rho\hat{Y}_s}{A_s}\right), \quad \hat{Y}(s) = K_{smooth}\mathbf{Y}$$

2. And $P(Y_s = 0) = \exp\{-|A_s|\mu(s)\}$,

Finding good importance function q_θ

Moment based method with kernel smoothing :

1. Using $\mathbb{E}(Y_s) = \frac{\mu(s)A_s}{\rho}$,

$$\hat{Z}(s) = log\left(\frac{\rho \hat{Y}_s}{A_s}\right), \quad \hat{Y}(s) = K_{smooth}\mathbf{Y}$$

2. And $P(Y_s = 0) = \exp\{-|A_s|\mu(s)\}$,

$$\tilde{Z}_s = \log\left(-\log\left(\tilde{p}_s\right)/|A_s|\right), \quad \tilde{p}_s = \frac{\#\text{Neighbours with } 0}{\#\text{Neighbours}}$$

Finding good importance function q_θ

Moment based method with kernel smoothing :

1. Using $\mathbb{E}(Y_s) = \frac{\mu(s)A_s}{\rho}$,
$$\hat{Z}(s) = log\left(\frac{\rho\hat{Y}_s}{A_s}\right), \quad \hat{Y}(s) = K_{smooth}\mathbf{Y}$$
2. And $P(Y_s = 0) = \exp\{-|A_s|\mu(s)\}$,
$$\tilde{Z}_s = \log(-\log(\tilde{p}_s)/|A_s|), \quad \tilde{p}_s = \frac{\#\text{Neighbours with } 0}{\#\text{Neighbours}}$$
3. Finally, $Z = KX$,
$$\hat{X} = (K'K)^{-1}K'(p\tilde{Z} + (1-p)\hat{Z})$$

Finding good importance function q_θ

Moment based method with kernel smoothing :

1. Using $\mathbb{E}(Y_s) = \frac{\mu(s)A_s}{\rho}$,

$$\hat{Z}(s) = log\left(\frac{\rho \hat{Y}_s}{A_s}\right), \quad \hat{Y}(s) = K_{smooth}\mathbf{Y}$$

2. And $P(Y_s = 0) = \exp\{-|A_s|\mu(s)\}$,

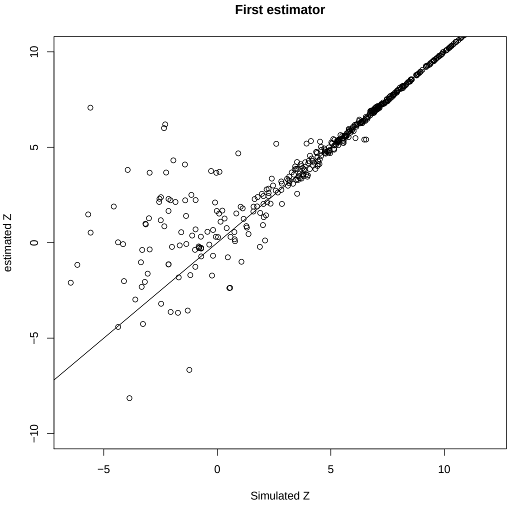
$$\tilde{Z}_s = \log\left(-\log\left(\tilde{p}_s\right)/|A_s|\right), \quad \tilde{p}_s = \frac{\#\text{Neighbours with } 0}{\#\text{Neighbours}}$$

3. Finally, $Z = KX$,

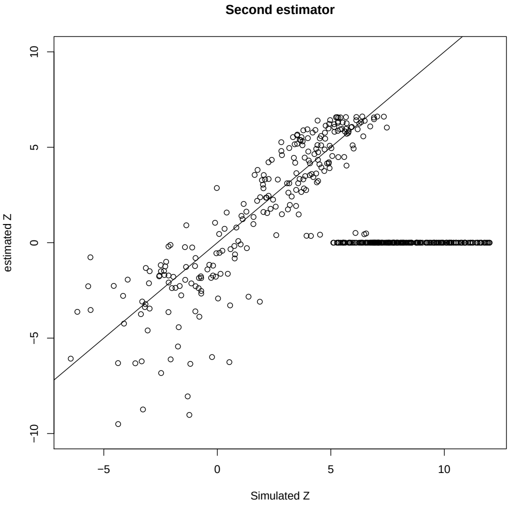
$$\hat{X} = (K'K)^{-1}K'(p\tilde{Z} + (1-p)\hat{Z})$$

$$X^* \sim \mathcal{N}(\hat{X}, \Sigma_X)$$

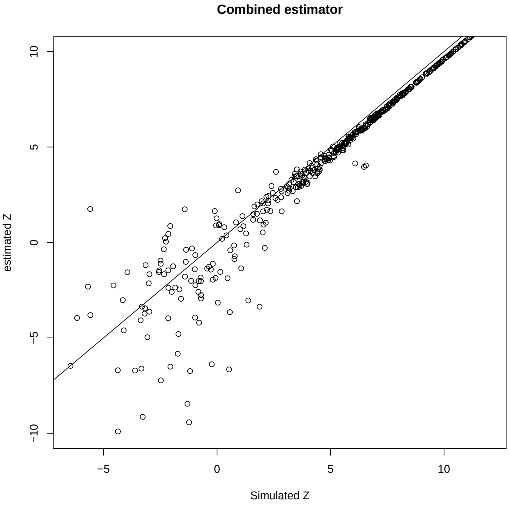
Finding good importance function q_θ



Finding good importance function q_θ



Finding good importance function q_θ



And now

Mixing all the ingredients and baking the cake

[1] C. Andrieu , A. Doucet, R. Holenstein (2010) Particle Markov chain Monte Carlo methods, *J. Roy. Stat. Ass.*, vol. 73, iss. 3, pp. 269-342.

[2] C. Andrieu, G.O. Roberts (2009) The pseudo marginal approach for efficient Monte Carlo simulations, *The Annals of Statistics*, Vol. 37, No. 2, 697–725.

[3] D. Higdon. (1998) A process-convolution approach to modelling temperatures in the North Atlantic Ocean, *Environmental and Ecological Statistics*, Vol. 5, No. 2, 173–190.

[4] Ancelet, Etienne, Benoit, Parent 2010 Modelling zero inflated data with an exponentially compound Poisson Process EES, vol17, iss 3 pp. 347.

[5] R. Menezes, T. Su, P.J. Diggle (2010). Geostatistical inference under preferential sampling, *The Annals of Statistics*, Vol. 59, No. 2, 191–232.