Modelling and Analyzing

Unevenly Spaced Measurements

David R. Brillinger

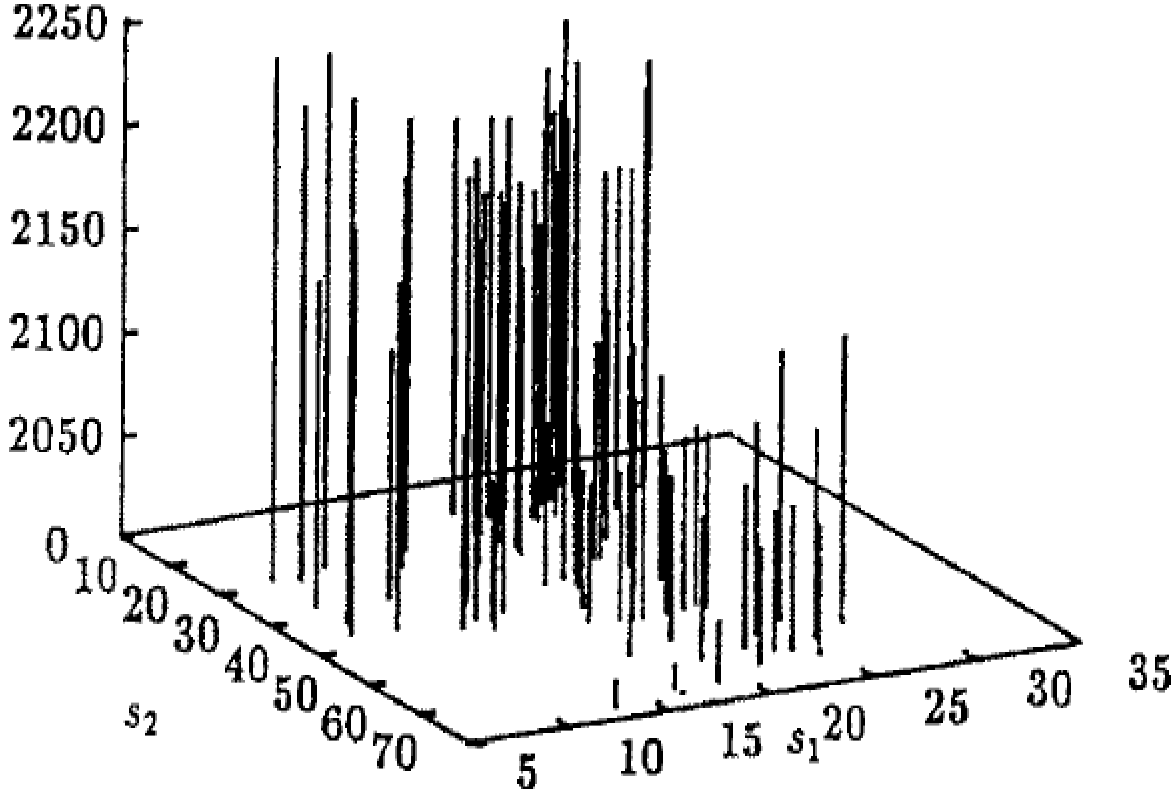
Statistics Department

University of California, Berkeley

$$2\pi \neq 1$$

Shumway Lecture, UC Davis, 15 April 2010

Shumway analysis.



Water heights, Saratoga Valley, Wyoming

Yucel and Shumway (1996) *Stochastic Hydrology & Hydrolics*

Underlying surface, $y(\mathbf{s})$, $\mathbf{s}$ in R^2

Spatial marked point process data, $\{\mathbf{s}_j, M_j\}$, $M_j = y(\mathbf{s}_j)$

$y(\mathbf{s}) = \mathbf{v}(\mathbf{s})'\beta + x(\mathbf{s}) + e(\mathbf{s})$, $\mathbf{s}$**:** location

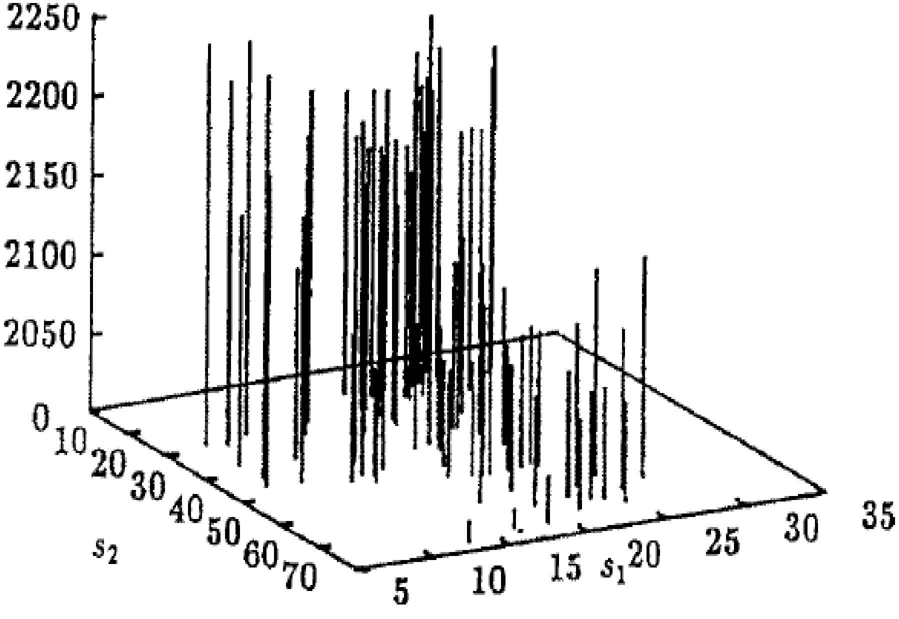
$\mathbf{v}$: known regressor function

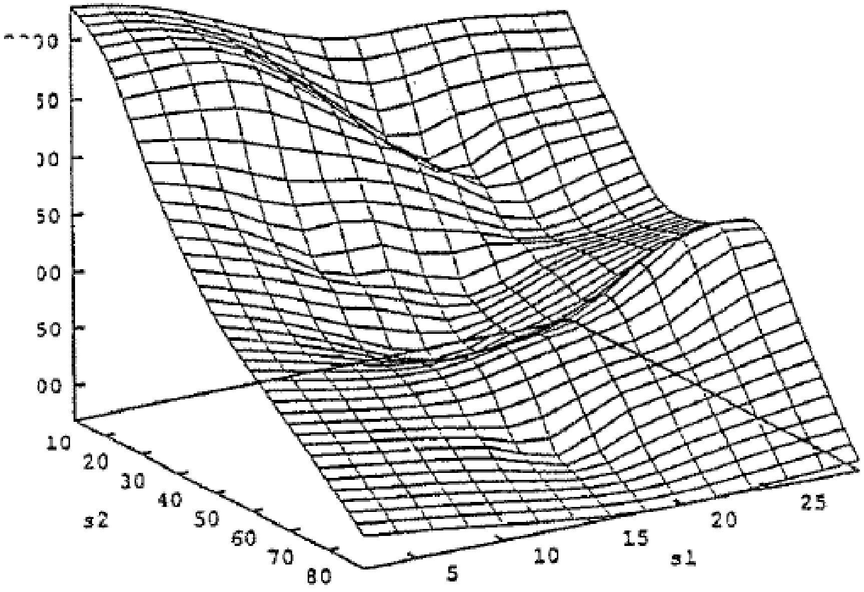
x: signal, stationary Gaussian, autcov $C(t|\phi)$, spectrum $f(\Lambda|\psi)$

e: stationary Gaussian white noise

mle obtained via ECM after approximating C

Estimated $\mathbf{v}(\mathbf{s})'\beta + x(\mathbf{s})$





Method and program checked by simulation

Other Shumway work.

"Some applications of the EM algorithm to analyzing incomplete time series data." Pp 290-324 in *Time Series Analysis of Irregularly Observed Data*, ed. E. Parzen (1984)

"Dynamic mixed models for irregularly observed time series." *Resenhas* 4, 433-456 (2000).

missing values in equispaced time series

state space models

Many important random process methods and programs expect equi-spaced data values.

Variants needed and effects to be understood.

Causes: missing values, irregular sampling, explanatories differ, measurement mechanism details, triggered observation, enforced gaps, ...

Reasons may include speed of computation.

Uneven/ irregular measurements described via point process of times, locations, ...

Process $\{Y(t)\}$ defined for all t, but data $(\tau_k, Y(\tau_k))$ available. (Hybrid process)

Suppose stationary, mean $E(Y(t)) = c_Y$

Given the sequence of times $\tau_1, \ldots, \tau_N$ for the interval $0 < t \leq T$, properties of average

$$[Y(\tau_1) + \ldots + Y(\tau_N)]/N \qquad (*)$$

as estimate of c_Y?

Take τ's as fixed or random?

Formal setup.

$N(t)$: number of τ's in the interval $(0,t]$

$dN(t)$: number in interval $(t,t+dt]$ 0 or 1

Sampled process, $\{Y(\tau_k)\}$,

$$dX(t) = Y(t)dN(t)$$

estimate (*) of c_N

$$\int_0^T Y(t)\, dN(t)/N(T) \qquad (*)$$

Dirac comb

$$dN(t) = \sum_k \delta(t - \tau_k)dt$$

$$dX(t) = \sum_k Y(\tau_k)\, \delta(t - \tau_k)dt$$

$\delta(x)$: density of r. v. $X = 0$ with prob 1

Stochastic stationary case.

Suppose there exist C's of bounded variation with

$$\text{cum}\{dN(t+u_1),...,dN(t+u_k),dN(t)\} = dC_{N...N}(u_1,...,u_k)dt$$

"cum": moment measure of joint dependence

C's define parameters and provides mixing conditions

One sees that

$$E\{N(T)\} = TC_N$$

$$E\{T^{-1} \int_0^T [N(t+u)-N(t)]dN(t)\} = uC_N^2 + C_{NN}(u)$$

Back to estimating c_Y . Suppose given $Y(\tau_k)$ for (0,T].

If assume {Y(t)} and {N(t)} independent, then average (*) asymptotically normal with mean c_Y and large sample variance

$$T^{-1}[\int c_{YY}(u)du + \int c_{YY}(u)\, dC_{NN}(u)/C_N{}^2]$$

In homogenious Poisson case variance is

$$T^{-1}\int c_{YY}(u)du$$

One can look for {N(t)} to obtain small variance.

Irregular sampling can be an advantage with good choice.

View τ_k as fixed. cp. GHA

Assumption:

$\tau_1, \tau_2, \ldots$ increasing postive numbers

$|N(s)-N(t)| \leq A + B|s-t|$, A, B finite

$\lim N(T)/T = C_N$

$\lim T^{-1} \int_0^T [N(t+u)-N(t)]dN(t) = M_{NN}(u)$ as $T \to \infty$

Large sample variance now

$T^{-1} \int c_{YY}(u)\, dM_{NN}(u)/C_N^2$

In practice $Y(\tau_k)$ often interpolated to obtain equispaced values

Irregular sampling can be an advantage

Can reduce aliasing problem

Estimate at a one frequency also estimate at second

Occurs if signal not sampled often

Suppose the series {Y(t)} and the point process {N(t)} are independent and stationary with pertinent moments existing.

Define the process with stationary increments

$dX(t) = Y(t)dN(t)$

When Y has mean 0

$$f_{XX}(\lambda) = c_N^2 f_{YY}(\lambda) + \int f_{NN}(\lambda - \alpha) f_{YY}(\alpha) d\alpha \quad \text{[CHECK]}$$

Can solve for ...

No aliasing when ...

Works Poisson

No interpolation is needed.

One can use a periodogram to assess if a process is bandlimited white noise when only irregularly sampled values available and for residual analysis. (Missing seasonal?)

Having fit a model to sampled data one can examine the periodogram of the residuals for whiteness

Amazonas river flow.

Solimoes flows from Peru

Divides into, Parana do Careiro and Amazon near Manaus

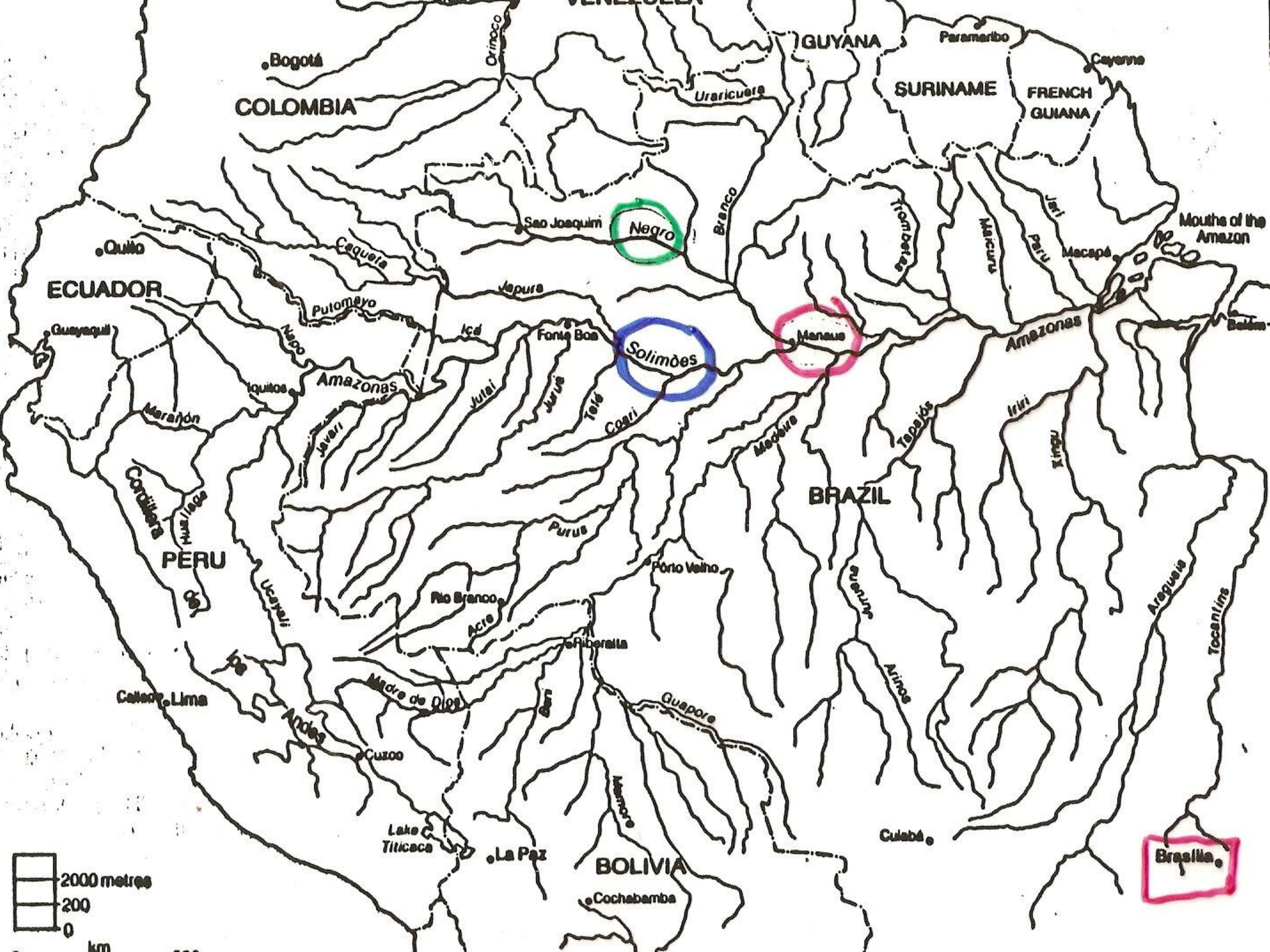
Change of flow in Parana could damage habitat

Does proportion of discharge entering Parana depend on Solomoes discharge level?

Solimoes flow increasing?

Careiro and Solimoes relationship changing?

Difficulty: flows measured at different irregular times



GUYANA
Paramaribo
Cayenne
SURINAME
FRENCH GUIANA
Bogotá
COLOMBIA
Orinoco
Uraricuera
Branco
Trombetas
Maicuru
Jari
Paru
Macapá
Mouths of the Amazon
Sao Joaquim
Negro
Quito
Caqueta
ECUADOR
Japura
Putomayo
Guayaquil
Napo
Iça
Fonte Boa
Solimões
Manaus
Amazonas
Belém
Iquitos
Marañon
Jutai
Juruá
Tefé
Coari
Javari
Madeira
Tapajós
Iriri
Xingu
Cordillera
Huallaga
BRAZIL
Purus
PERU
Porto Velho
Ucayali
Juruena
Araguaia
Tocantins
Rio Branco
Acre
Riberalta
Arinos
Callao
Lima
Madre de Dios
Guapore
Beni
Andes
Cuzco
Mamoré
Lake Titicaca
La Paz
BOLIVIA
Cuiabá
Brasília
Cochabamba
2000 metres
200
0
km

Rio Negro - Solimoes confluence

Single ship - measurements at different times and locations

Flow estimates via

discharge = area of section * velocity

Measurement times $\{\sigma_j\}$ and $\{\tau_k\}$

X(t): Solimoes $\{X(\sigma_j)\}$ explanatory

Y(t): Careiro $\{Y(\tau_k)\}$ response

Models/questions.

$Y(t) = g(X(t)) + \text{noise}$ g(.) linear?

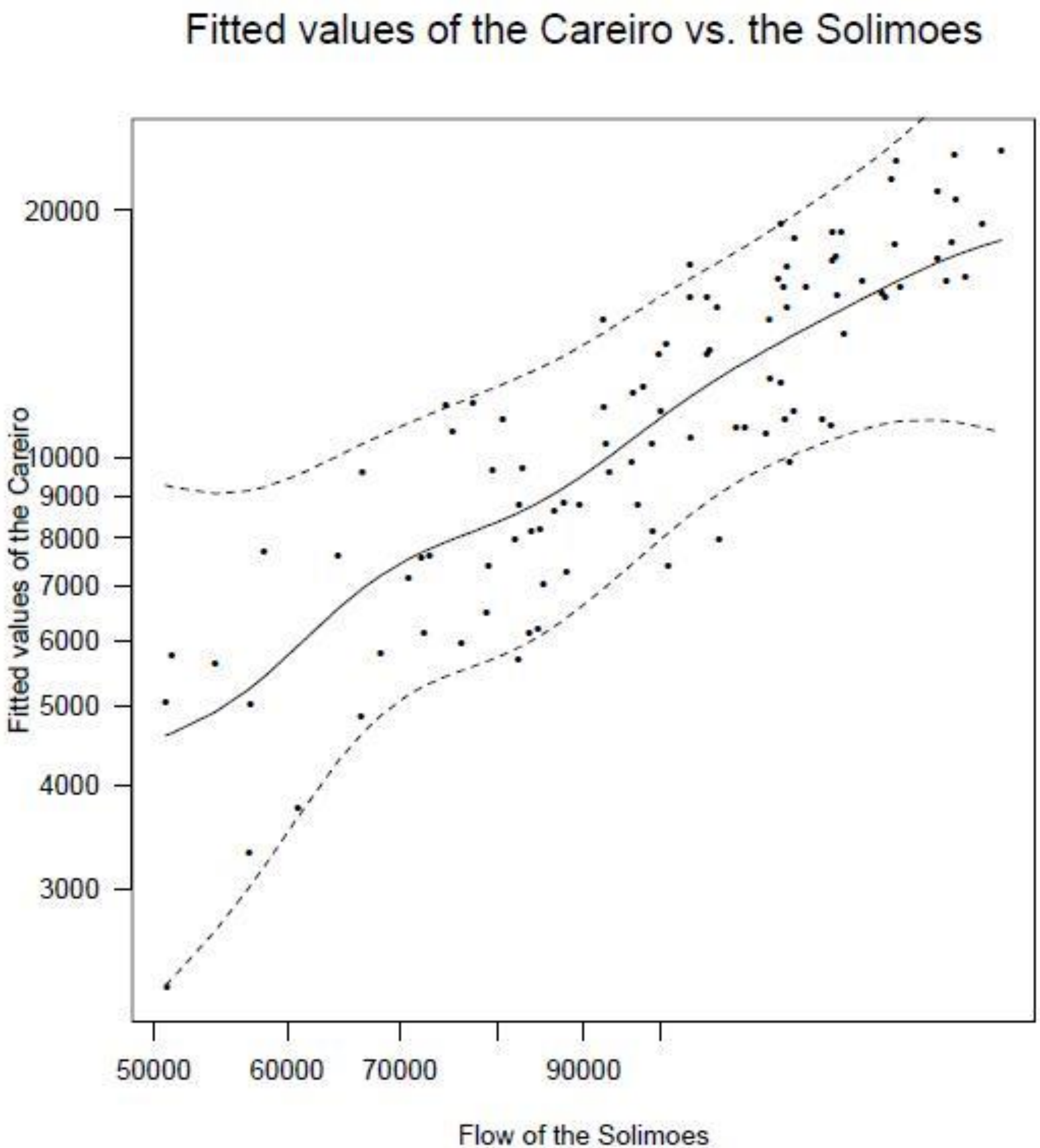
$Y(t)/X(t) = h(X(t)) + \text{noise}$ h(x) constant? h(x,t)?

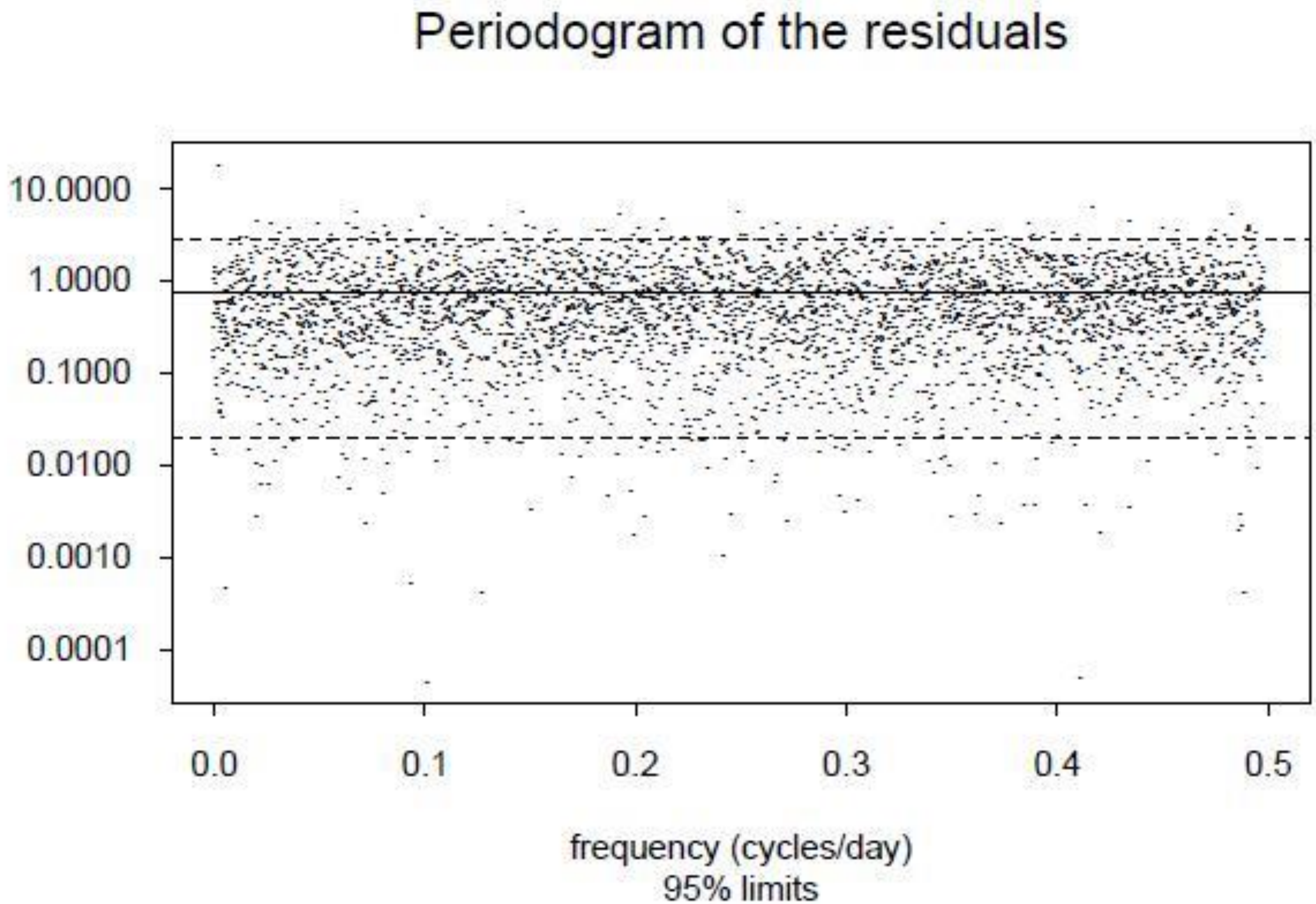
Time lag, δ, to handle travel time Manacapuru to Careiro

$$Y(t) = g(X(t-\delta)) + \text{noise}$$

Interpolation. Smoothing to estimate $X(\tau_k - \delta)$

Smoothing to estimate $g(X)$


Fitted values of the Careiro vs. the Solimoes
Fitted values of the Careiro
Flow of the Solimoes
20000
10000
9000
8000
7000
6000
5000
4000
3000
50000
60000
70000
90000


Periodogram of the residuals
10.0000
1.0000
0.1000
0.0100
0.0010
0.0001
0.0
0.1
0.2
0.3
0.4
0.5
frequency (cycles/day)
95% limits

Error of interpolation.

$n=M(T)$ $\quad F_n(\sigma) = n^{-1}\#\{\sigma_j \le \sigma\}$ $\quad W_n(\sigma)=b_n^{-1}W(b_n^{-1}\sigma)$

$$d_n = \sup |F_n(\sigma) - F(\sigma)|$$

Interpolant.

$$X(\sigma) \sim \sum W_n(\sigma-\sigma_j)\, X(\sigma_j) \,/ \sum W_n(\sigma-\sigma_j)$$

$$= \int W_n(\sigma-s)\, X(s)\, dF_n(s) \,/ \int W_n(\sigma-s)\, dF_n(s)$$

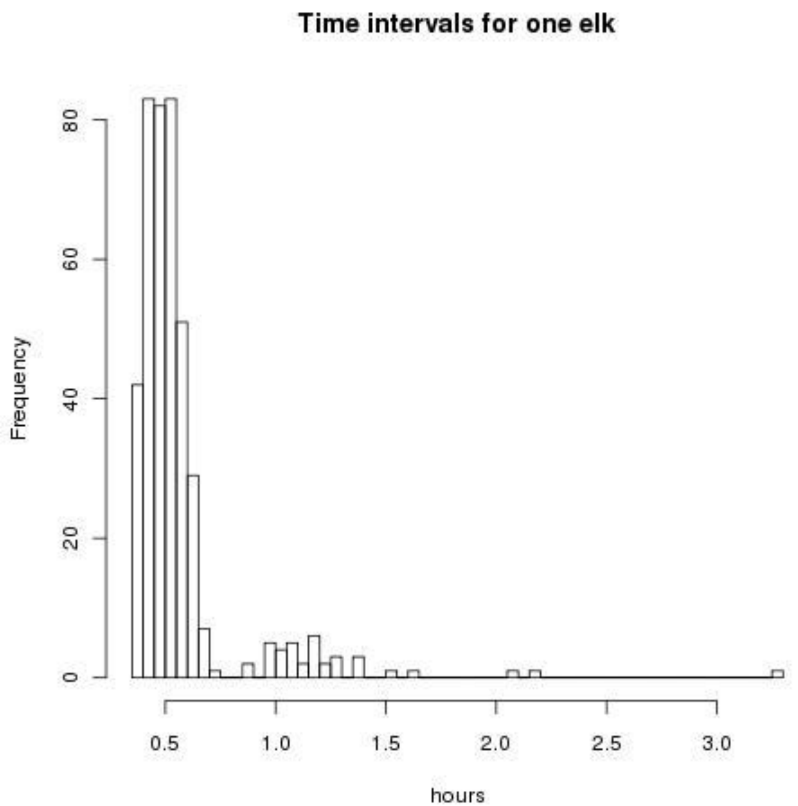
$$\sim X(\sigma)f(\sigma)/f(\sigma) + O(b_n^2) + O(b_n^{-1}d_n) \quad \text{uniformly}$$

Work with σ_j/T as necessary

Animal tracks. 94 elk 1991 - 1993 LORAN

Starkey Experimental Forest and Range, Eastern Oregon

time differences: irregular, and varied with animal



Vector field model. Location $\mathbf{r} = (x,y)$

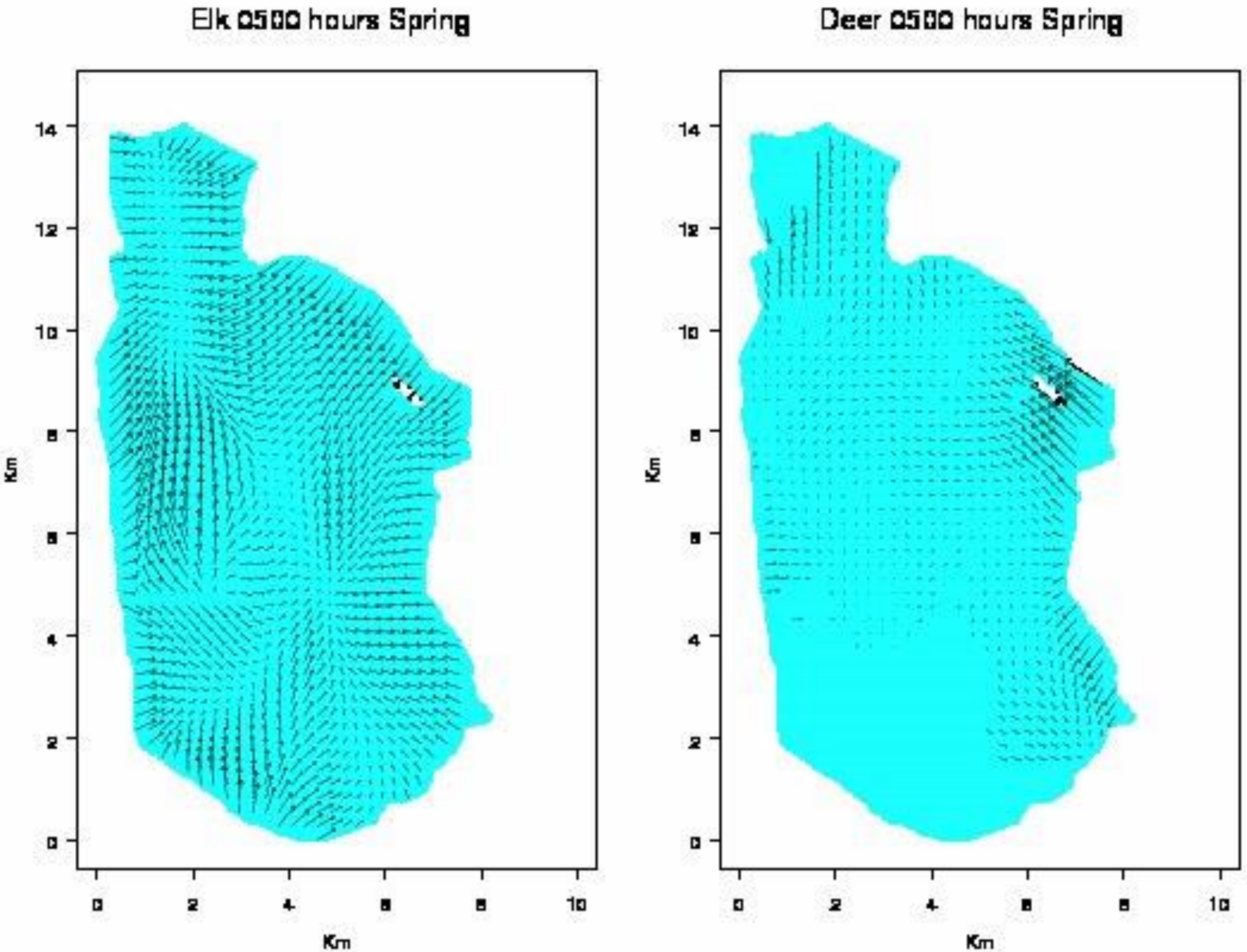
$$(\mathbf{r}(t_{i+1})\mathbf{-r}(t_i))/(t_{i+1}\mathbf{-}t_i) = \boldsymbol{\mu}(\mathbf{r}(t_i),\langle t_i\rangle) + \boldsymbol{\sigma}\,\mathbf{Z}_{i+1}/\sqrt{(t_{i+1}\mathbf{-}t_i)}$$

LHS: velocity, $\langle t\rangle$: time of day

Motivated by SDE

$$d\mathbf{r}(t) = \boldsymbol{\mu}(\mathbf{r}(t),\langle t\rangle)\,dt + \boldsymbol{\sigma}\,d\mathbf{B}(t)$$

Robust fit via gam()



Elk 0500 hours Spring
Deer 0500 hours Spring
Km
Km
Km
Km

Residuals provided evidence for 24 hour period

included in model

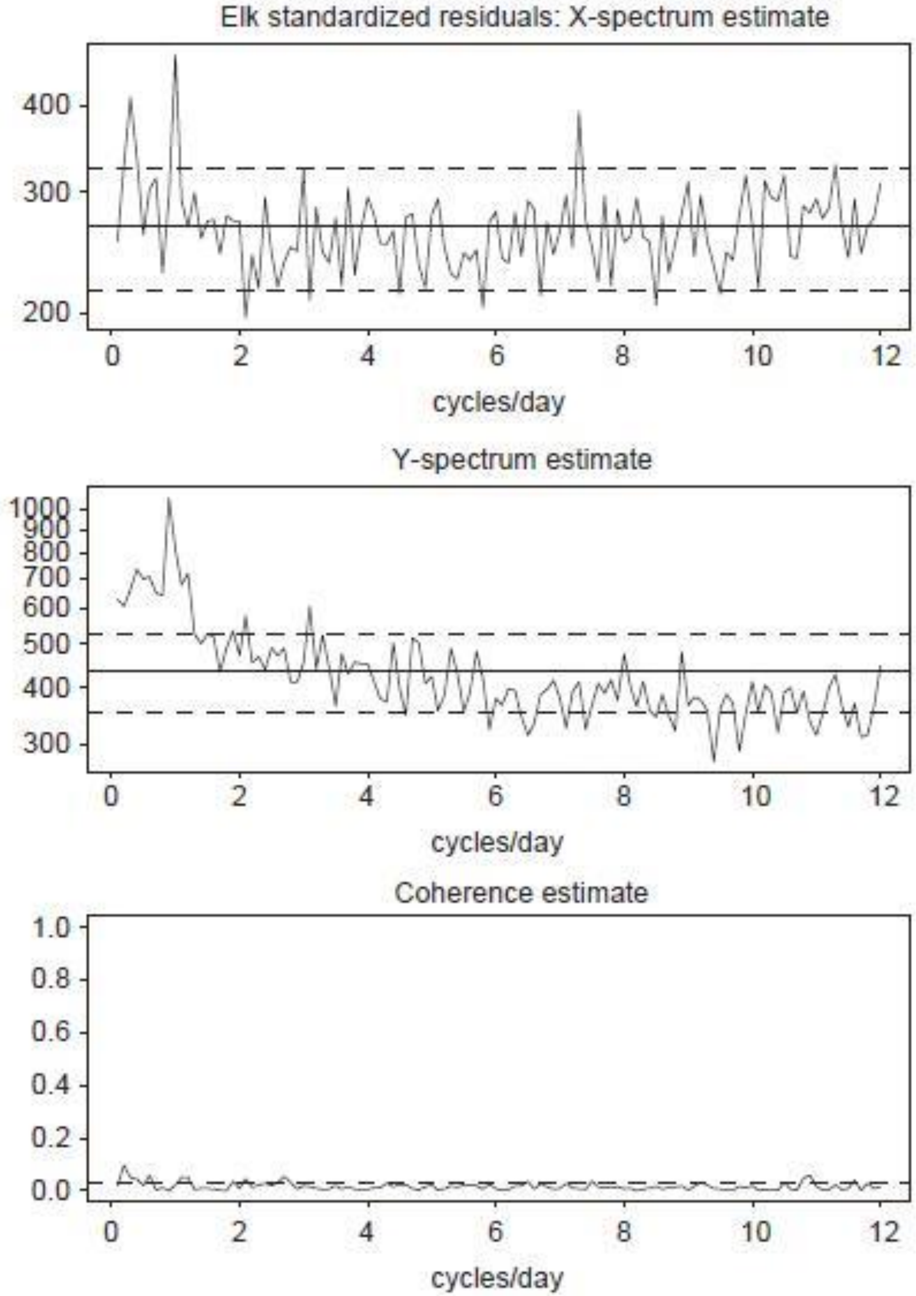
spectra of hybrid estimated by averaging periodograms

$$(2\pi T)^{-1}|\sum \exp\{-i\lambda\tau_k\} r_{residual}(\tau_k)|^2$$

Fundamental relation

$$c_N^2 f_{noise}(\lambda) + \int f_{NN}(\lambda - \alpha) f_{noise}(\alpha) d\alpha \qquad (*)$$

flat if noise spectrum bandlimited white [CHECK]



Elk standardized residuals: X-spectrum estimate
400
300
200
0
2
4
6
8
10
12
cycles/day
Y-spectrum estimate
1000
900
800
700
600
500
400
300
0
2
4
6
8
10
12
cycles/day
Coherence estimate
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
cycles/day

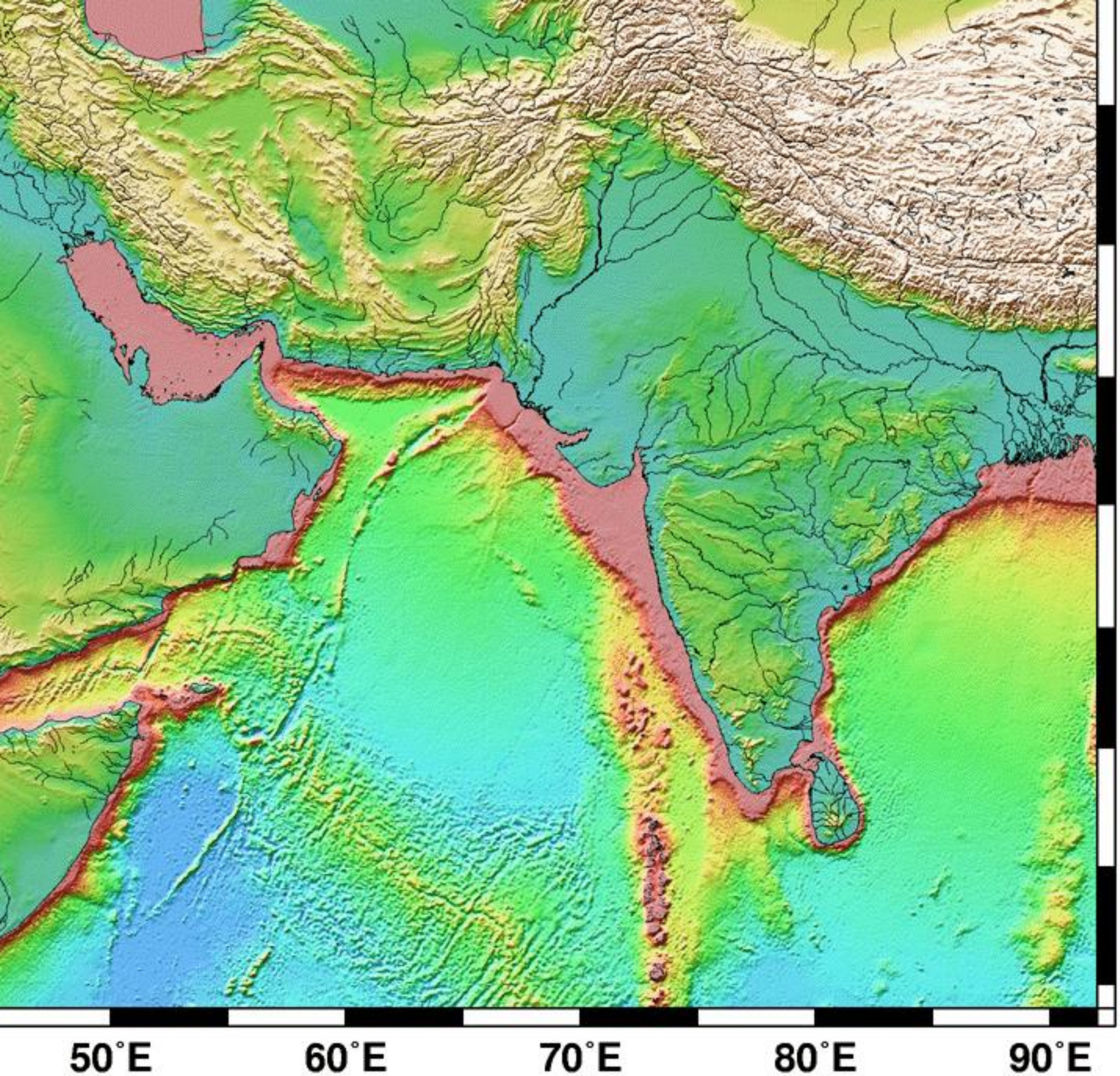
Evidence for low frequency departure for N-S component

Whaleshark tagging study. Off Kenya

To study ecology, e.g. where they travelled and foraged, and when?

Goal species protection.

Brent


50˚E
60˚E
70˚E
80˚E
90˚E

29 June - 19 July, 2008 Indian Ocean

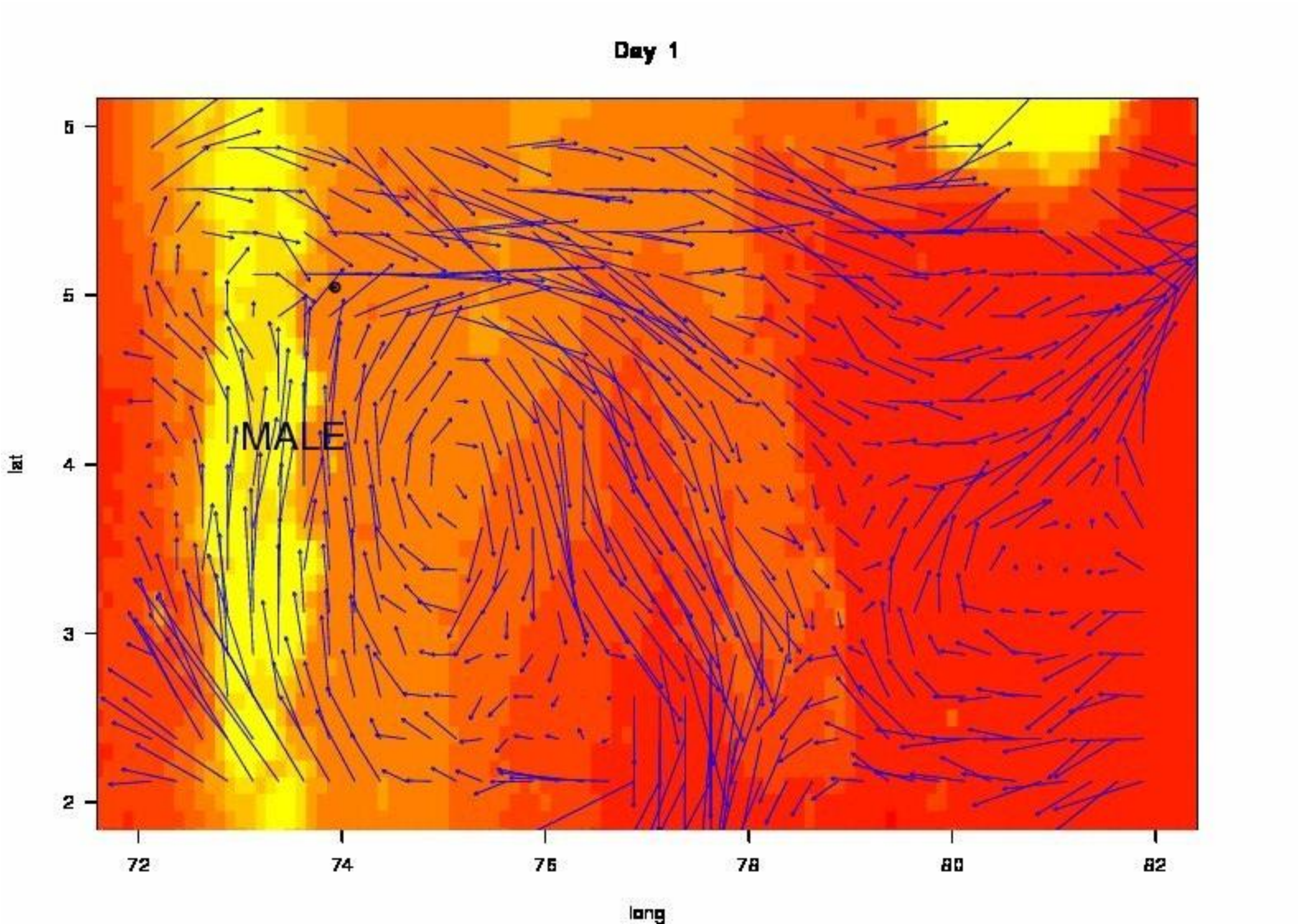
One shark lost its tag, drifted til batteries expired

Brent realized could check NOAA's sea surface current model results

Unequally spaced location times, about 250 time points

Geosrophic currents for June 29, 2008 (ten day composite)

some days missing (clouds)



display of sequence of locations

Functional stochastic differential equation (FSDE)

$$d\mathbf{r(t)} = \boldsymbol{\mu}(H(t),t)dt + \boldsymbol{\sigma}(H(t),t)d\mathbf{B}(t)$$

H(t): a history based on the past, $\{\mathbf{r}(s), s \leq t\}$

Process Markov when $H(t) = \{\mathbf{r}(t)\}$

Interpretation

$$\mathbf{r}(t) - \mathbf{r}(0) = \int_0^t \boldsymbol{\mu}(H(s),s)ds + \int_0^t \boldsymbol{\sigma}(H(s),s)d\mathbf{B}(s)$$

Approximation

$$\mathbf{r}(t_{i+1})-\mathbf{r}(t_i)=\boldsymbol{\mu}(H(t_i),t_i)(t_{i+1}-t_i)+\boldsymbol{\sigma}(H(t_i),t_i)\sqrt{\{t_{i+1}-t_i\}}\,\mathbf{Z}_{i+1}$$

Analysis.

How was the irregular sampling handled?

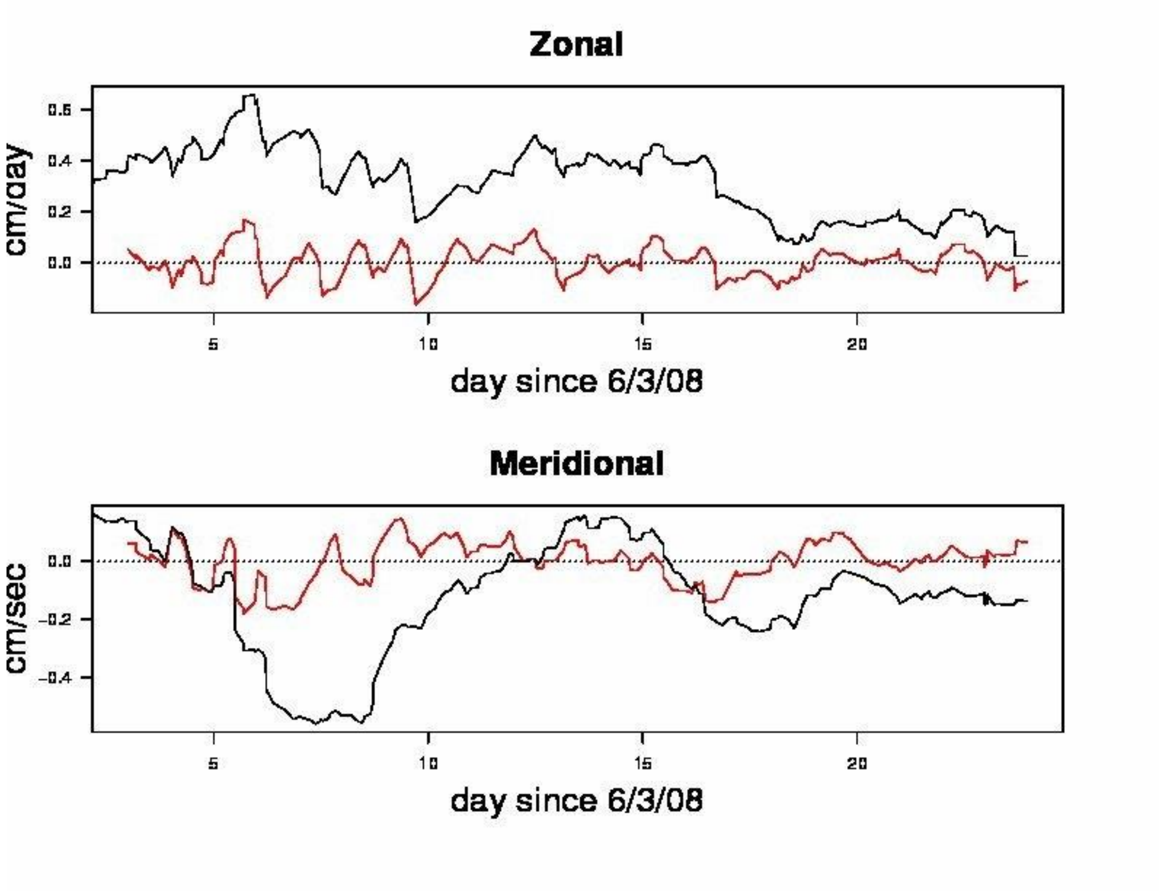
Interpolated

Reduced tag data to 46 contiguous 12 hour periods

- median values

Estimated local zonal and meridional speeds from tag locations and ... [Get straight]

Tag currents and residuals from fits



Current work.

Moving particles.

Are they interacting?

Pairwise interation model

$$d\mathbf{r}_i(t) = \Psi(\mathbf{r}_i(t))dt + \Sigma_{j\neq i}\ \Phi(\mathbf{r}_i(t) - \mathbf{r}_j(t))dt + \sigma d\mathbf{B}_i(t)$$

$i = 1,\ldots,p$ for some functions Ψ, Φ

Different time spacings for different models

Other cases.

P.p. case: deletions/thinning

t in 0, ± 1, ± 2,...

missings, gaps, ...

Discussion.

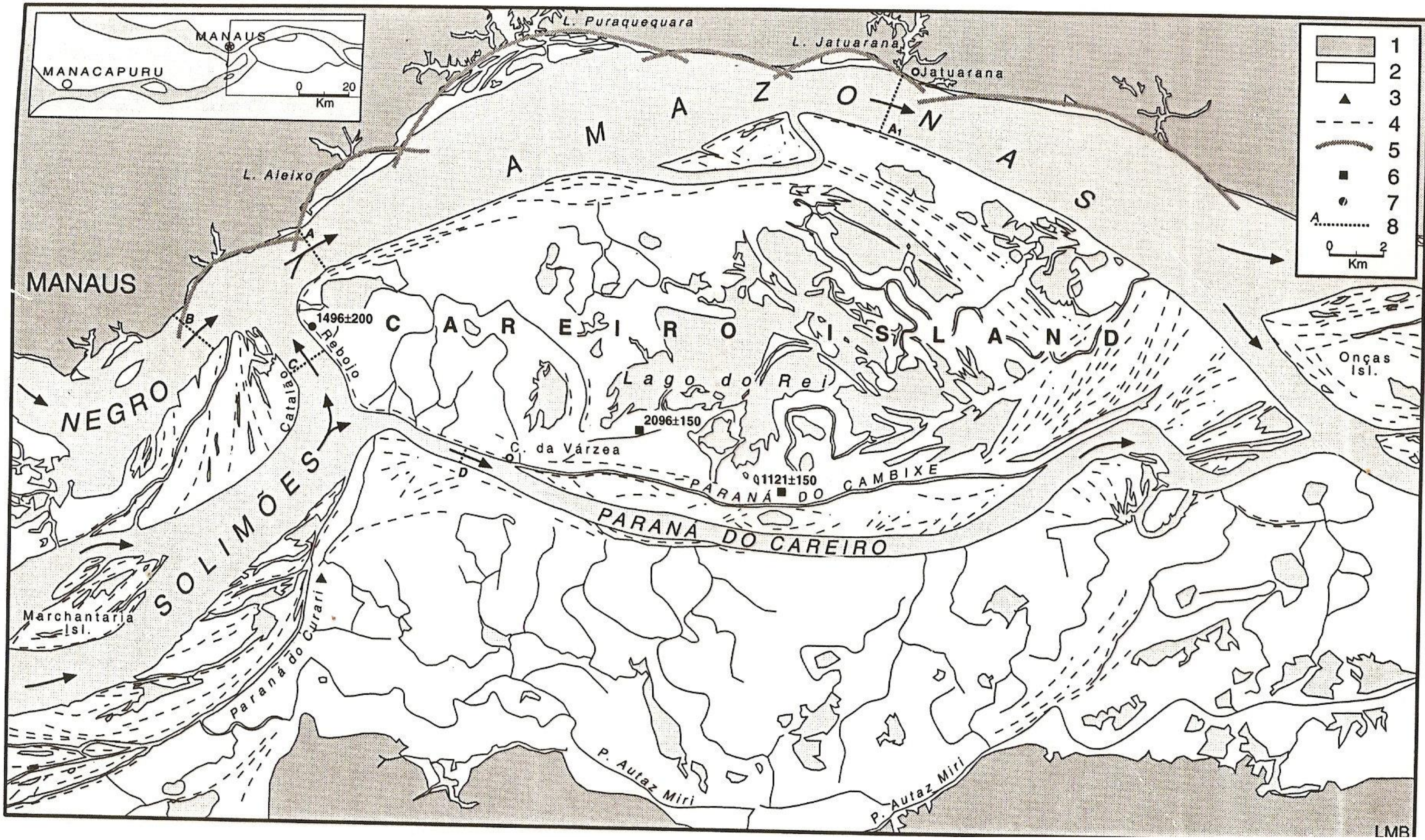
Sometimes can model the missingness

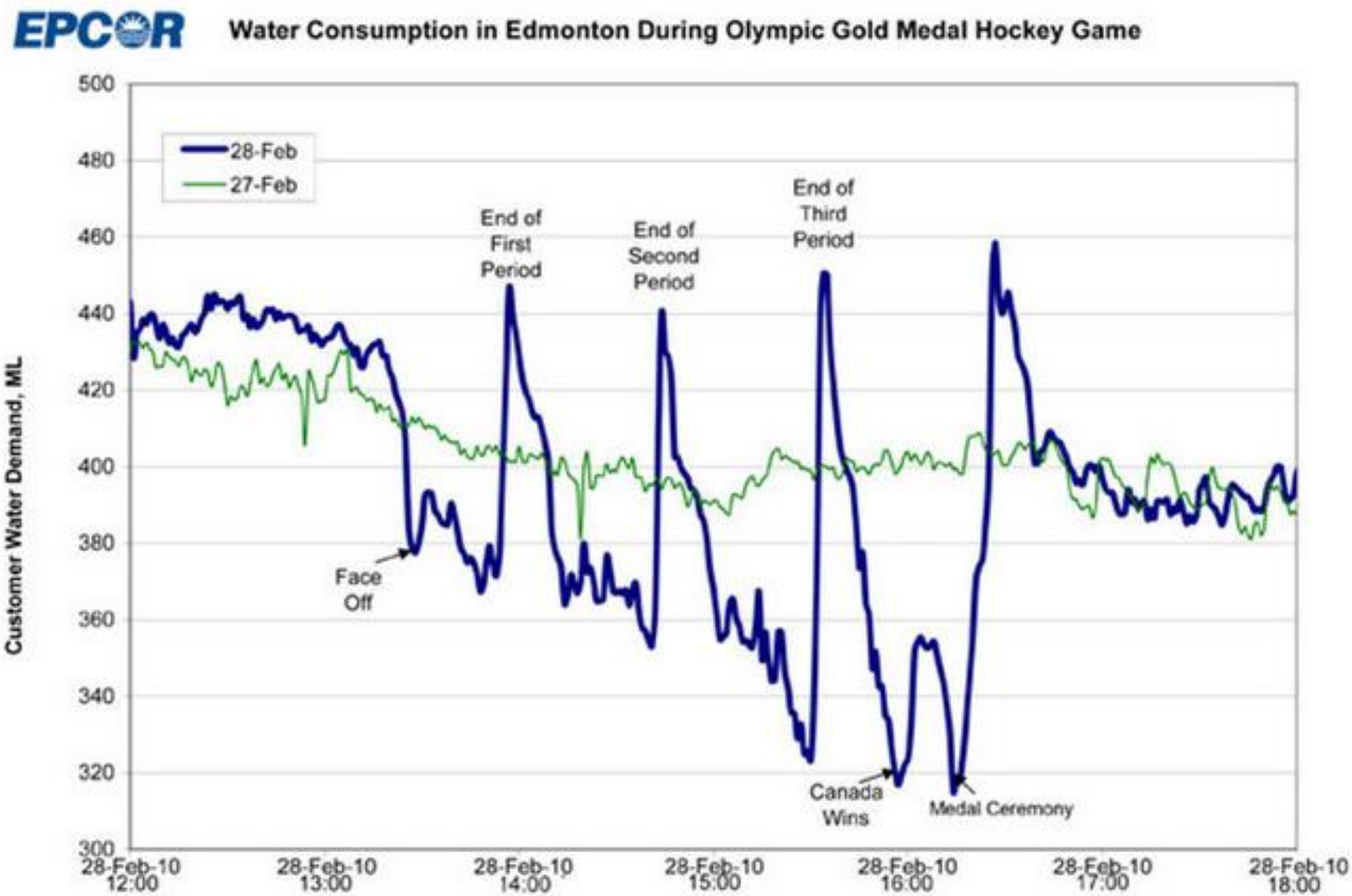
There are FFTs for the irregular case

Common features of examples

Acknowledgements.

B. Stewart, ...


L. Puraquequara
L. Jatuarana
Jatuarana
MANAUS
MANACAPURU
0 20
Km
1
2
3
4
5
6
7
8
0 2
Km
AMAZONAS
L. Aleixo
MANAUS
NEGRO
SOLIMÕES
Catalão
Rebojo
1496±200
CAREIRO ISLAND
Lago do Rei
2096±150
C. da Várzea
1121±150
PARANÁ DO CAMBIXE
PARANA DO CAREIRO
Onças Isl.
Marchantaria Isl.
Paraná do Curari
P. Autaz Miri
P. Autaz Miri
LMB


EPCOR
Water Consumption in Edmonton During Olympic Gold Medal Hockey Game
28-Feb
27-Feb
Customer Water Demand, ML
500
480
460
440
420
400
380
360
340
320
300
End of First Period
End of Second Period
End of Third Period
Face Off
Canada Wins
Medal Ceremony
28-Feb-10 12:00
28-Feb-10 13:00
28-Feb-10 14:00
28-Feb-10 15:00
28-Feb-10 16:00
28-Feb-10 17:00
28-Feb-10 18:00