Understanding rapidly varying processes

Sofia C. Olhede
University College London

ERC Research Fellow
Scientific Director, UCL Big Data Institute

Cantab Institute
Cambridge, 30 November 2016

With co-authors, such as A. P. Guillaumin, A. M. Sykulski, J. J. Early, J. M. Lilly

Announcement — Theory of Big Data 3

Theory of Big Data 3

26–28 June 2017, London, UK

1. Challenges in Spatial & Temporal Analysis
2. High-Dimensional Estimation and Learning
3. Privacy-preserving inference
4. Tensors and statistical modelling

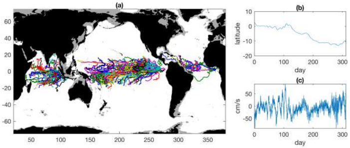
- Organizers: Sofia Olhede & Patrick Wolfe (UCL)
- Confirmed speakers: Jianqing Fan, Ming Yuan, Arthur Gretton, Arnak Dalalyan, Guy Nason, Heather Battey....
- Mailing list and additional information: http://www.ucl.ac.uk/bigdata-theory

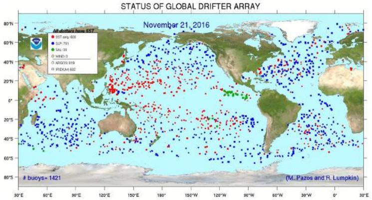
Fundamental question

How can we model rapidly-varying processes in time?

Ingredients

1. Mechanisms that generate data
2. Structure that facilitates analysis
3. Tools that can be understood





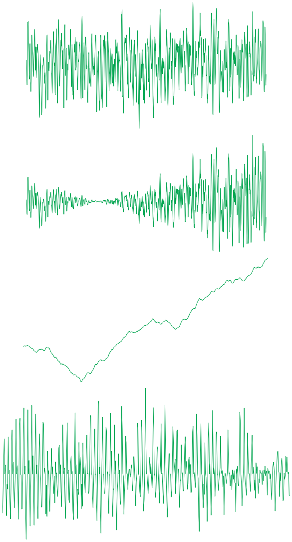
- We start from the analysis of a single time series $\{X_t\}$.
- To understand this object, we wish to model its **mean**

$$\mathbb{E}(X_t) = \mu(t), \tag{1}$$

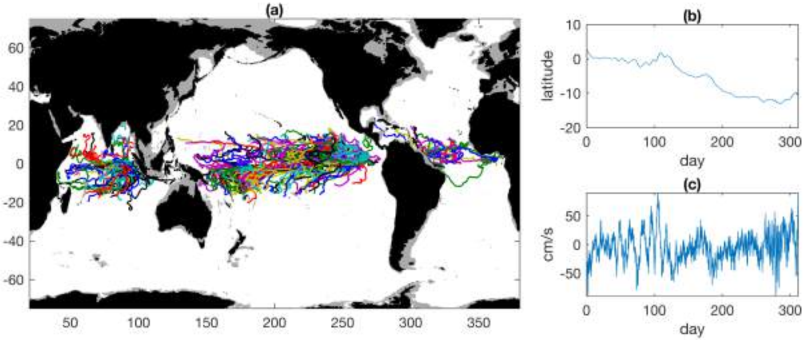
and its **covariance**

$$\mathbb{C}\text{ov}(X_t, X_{t-\tau}) = c(\tau, t). \tag{2}$$

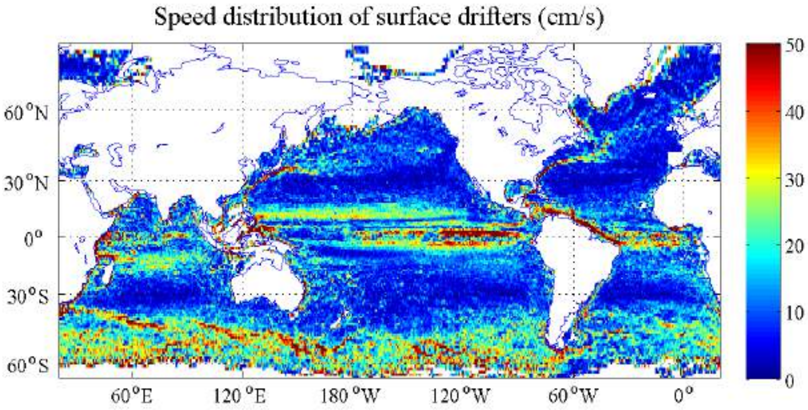
- $c(\tau, t)$ describes **evolving dependence**.
- If this is evolving rapidly, then we need to analyze smaller portions of data together.



Rapid variation



Global Drifter Program

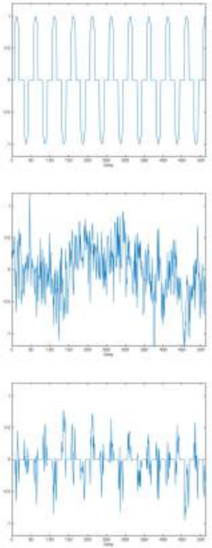


How to model time series variation

- Simplest model is **modulation** (Parzen (1963) & Priestley (1965)).
- Modulation can make Y_t quite nonstationary.
- Let $\{g_t\}$ be a (known?) deterministic sequence, and X_t a stationary latent processes. Take

$$Y_t = g_t X_t, \quad t \in \mathbb{Z}. \tag{3}$$

- Why not divide by g_t?
- g_t may be zero and/or we observe Y_t superimposed with yet another process.



More modelling

- We can always calculate

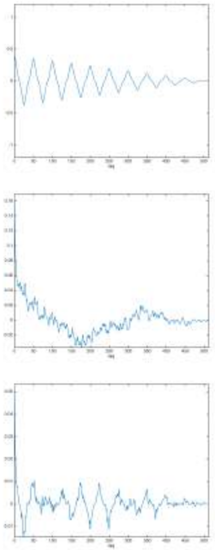
$$\hat{c}_Y^{(N)}(\tau) = \frac{1}{N} \sum_{t=0}^{N-\tau-1} Y_t Y_{t+\tau}. \qquad (4)$$

- This has expectation

$$\bar{c}_Y^{(N)}(\tau) = c_g^{(N)}(\tau) \cdot c_X^{(N)}(\tau). \qquad (5)$$

- Leads to the natural notion of an asymptotically stationary process. $\{Y_t\}$ is an **asymptotically stationary process** (Parzen) if there exists a fixed function $\gamma(\tau)$ such that

$$\lim_{N\to\infty} \bar{c}_Y^{(N)}(\tau) = \gamma(\tau).$$



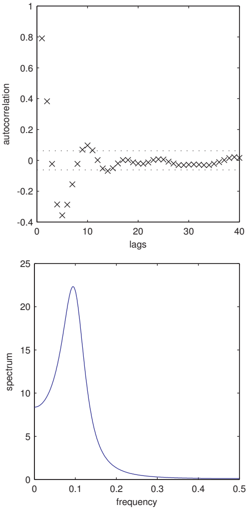
Fourier analysis

- As X_t is stationary it admits the representation

$$X_t = \mu + \int_{-\frac{1}{2}}^{\frac{1}{2}} dZ_x(f) e^{2i\pi ft}, \qquad (6)$$

where the basic quantity of interest is the spectrum $S(f)$.

- Here $S(f)\, df = \mathbb{E}\{|dZ_x(f)|^2\}$, and $dZ_x(f)$ is uncorrelated across f.
- Traditional theory claims X_t is "nearly stationary", where $S(f)$ is evolving very slowly, and $S(f) \mapsto S_t(f)$.
- This has the consequence of $dZ_x(f)$ nearly uncorrelated with $dZ_x(f')$ if $|f - f'| >> \epsilon$.



Whittle Pseudo-Likelihood

- If X_t was Gaussian, then we could infer its parameters using its likelihood function

$$\ell_T(\theta) = -\frac{1}{2}\log|\Sigma(\theta)| - \frac{1}{2}X^T\Sigma(\theta)^{-1}X, \quad \Sigma(\theta) = \mathbb{E}\,XX^T. \tag{7}$$

- Would like to form

$$\widehat{\boldsymbol{\theta}}^{(t)} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell_t(\boldsymbol{\theta}).$$

- Instead commonly the Whittle likelihood is used:

$$\ell_W(\theta) = -\sum_{\omega\in\Omega_N}\left\{\log S_X(\omega;\theta) + \frac{\hat{S}_X^{(N)}(\omega)}{S_X(\omega;\theta)}\right\}, \tag{8}$$

  where Ω_N is the set of Fourier frequencies $2\pi l/N$ where $l = 0, \dots, N-1$.
- Computationally efficient; convenient; but far from exact (see Sykulski et al (2016), Anitescu *et al.* (2012), Stein *et al.* (2013), Dutta and Mondal (2014)). Speed versus computation.

And what of modulation?

- If we calculate the DFT $J_Y(\omega)$, then its empirical variance has expectation

$$\overline{S}_Y^{(N)}(\omega;\boldsymbol{\theta}) = \mathbb{E}\Big\{\hat{S}_Y^{(N)}(\omega) \mid g_0,\cdots,g_{N-1};\boldsymbol{\theta}\Big\}.$$

- This has form

$$\overline{S}_Y^{(N)}(\omega;\boldsymbol{\theta}) = \int_{-\pi}^{\pi} S_X(\omega-\lambda;\boldsymbol{\theta})\Big|G^{(N)}(\lambda)\Big|^2 d\lambda,\ \forall\omega\in[-\pi,\pi), \tag{9}$$

and

$$G^{(N)}(\omega) = \frac{1}{\sqrt{N}}\sum_{t=0}^{N-1} g_t e^{-i\omega t}.$$

- We can compute $\overline{S}_Y^{(N)}(\omega;\boldsymbol{\theta})$ in $N\log(N)$ once G has been precomputed.

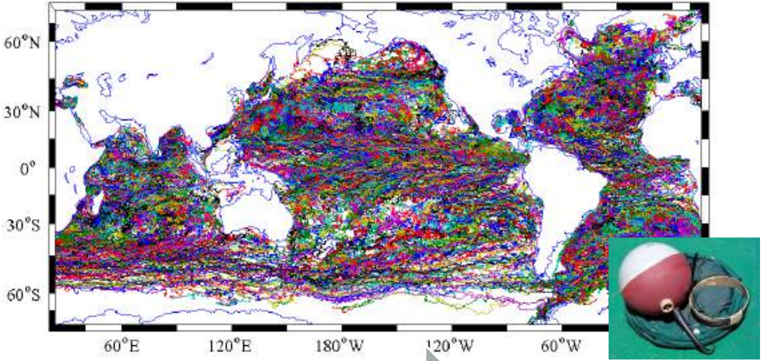
Debiased

- We calculate

$$\ell_M(\theta) = -\sum_{\omega \in \Omega_N} \left\{ \log \overline{S}_Y^{(N)}(\omega;\theta) + \frac{\hat{S}_Y^{(N)}(\omega)}{\overline{S}_Y^{(N)}(\omega;\theta)} \right\}, \qquad (10)$$

  where Ω_N is the set of Fourier frequencies $2\pi l/N$ where $l = 0, \ldots, N-1$.
- This set of frequencies can be restricted when suitable using local frequencies (Robinson (1995)), and time-frequencies (van Bellegem and Dahlhaus (2006)).

Global Drifter Program

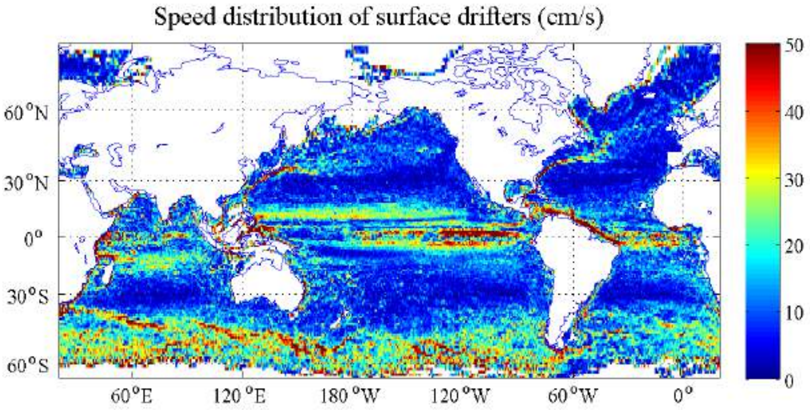


- 10,000+ drifters
- Data going back to 1979
- Over 60 million data points

Objective

Useful Summary Statistics

Global Drifter Program



Modelling oceanographic data

- Data from the Global Drifter Program (GDP, www.aoml.noaa.gov/phod/dac).
- The measurements include position, and often sea surface temperature, salinity and atmospheric pressure. In total, over 11,000 drifters have been deployed, with approximately 100 million position recordings obtained.
- The analysis of this data is crucial to our understanding of ocean circulation (Lumpkin 2007), which is known to play a primary role in determining the global climate system (Andrews, 2012).
- The Lagrangian velocity time series is modelled as a complex-valued time series, with the following 6-parameter power spectral density:

$$S(\omega) = \frac{A^2}{(\omega - f)^2 + \lambda^2} + \frac{B^2}{(\omega^2 + h^2)^\alpha}, \tag{11}$$

where ω is given in cycles per day.

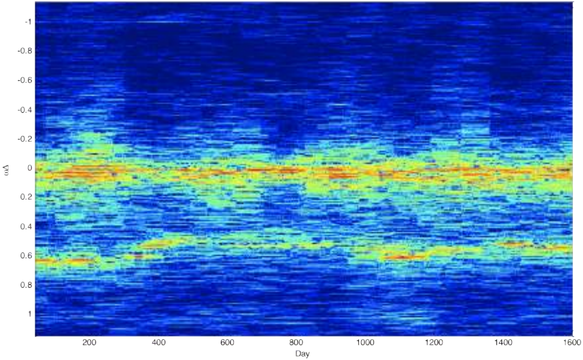
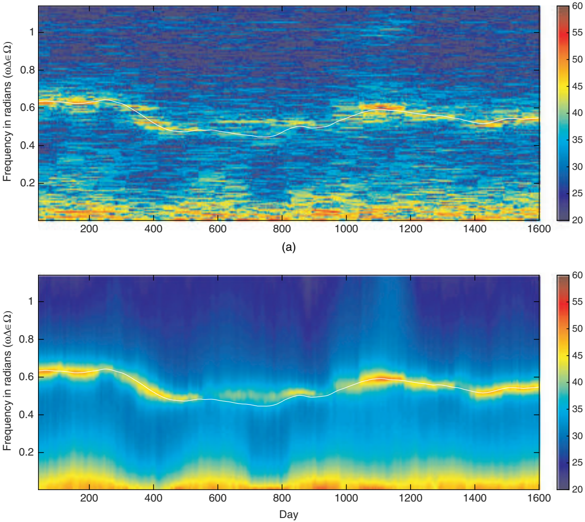


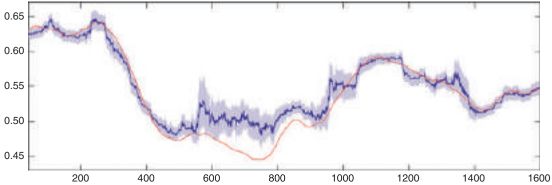
Figure: Time-frequency of bivariate data.

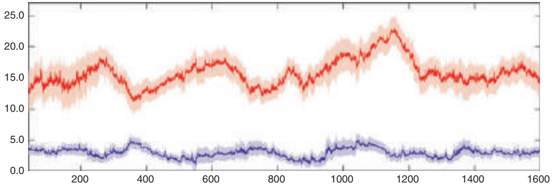


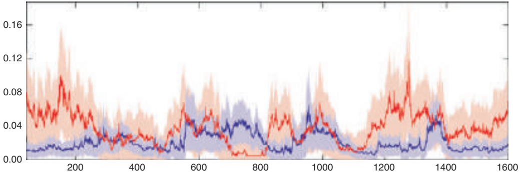
Frequency in radians (ω∆∈Ω)
1
0.8
0.6
0.4
0.2
200
400
600
800
1000
1200
1400
1600
60
55
50
45
40
35
30
25
20
Day

(a)

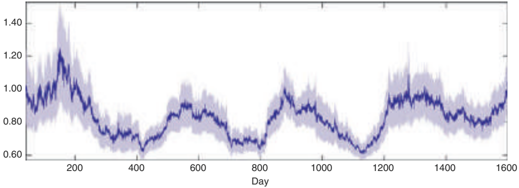
(b)


0.65
0.60
0.55
0.50
0.45
200
400
600
800
1000
1200
1400
1600
(a)


25.0
20.0
15.0
10.0
5.0
0.0
200
400
600
800
1000
1200
1400
1600
(b)

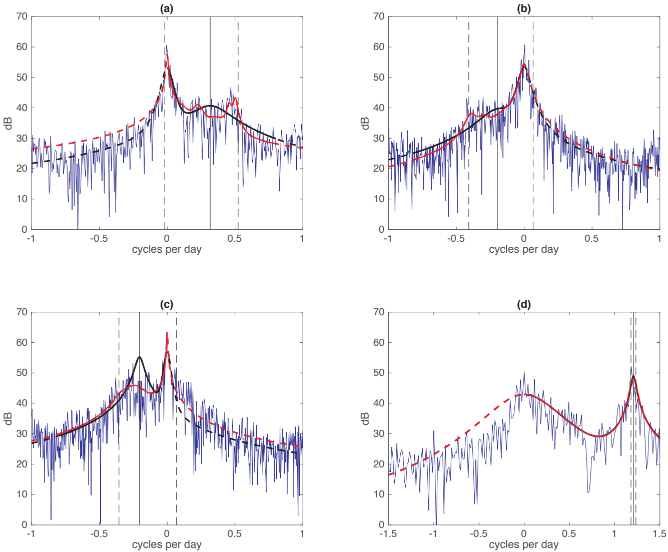

0.16
0.12
0.08
0.04
0.00
200
400
600
800
1000
1200
1400
1600

(c)

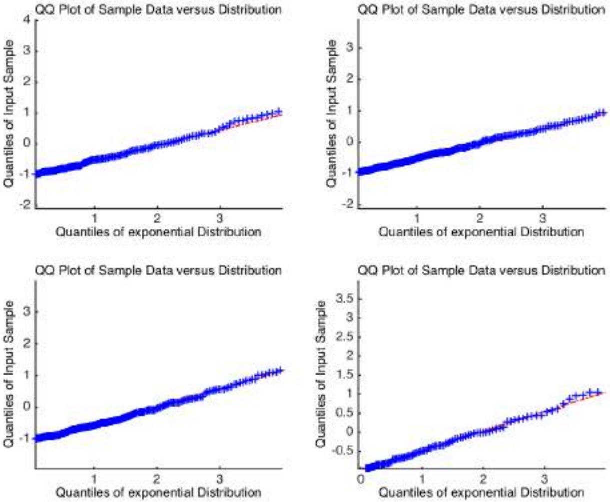

1.40
1.20
1.00
0.80
0.60
200
400
600
800
1000
1200
1400
1600
Day

(d)

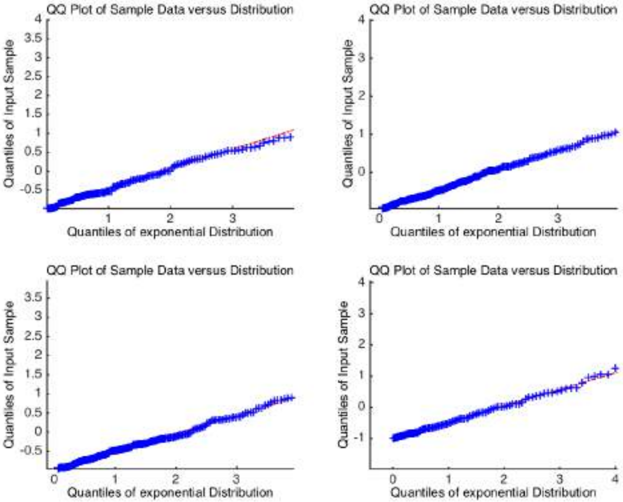
Real data



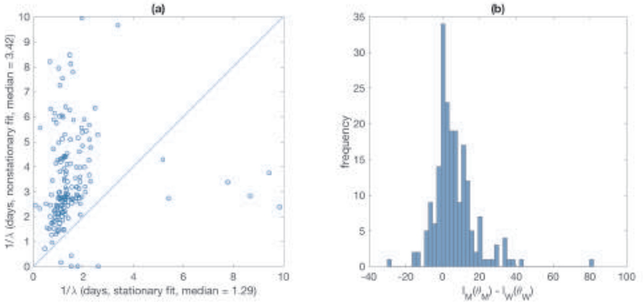
Real data


QQ Plot of Sample Data versus Distribution
Quantiles of Input Sample
Quantiles of exponential Distribution
QQ Plot of Sample Data versus Distribution
Quantiles of Input Sample
Quantiles of exponential Distribution
QQ Plot of Sample Data versus Distribution
Quantiles of Input Sample
Quantiles of exponential Distribution
QQ Plot of Sample Data versus Distribution
Quantiles of Input Sample
Quantiles of exponential Distribution

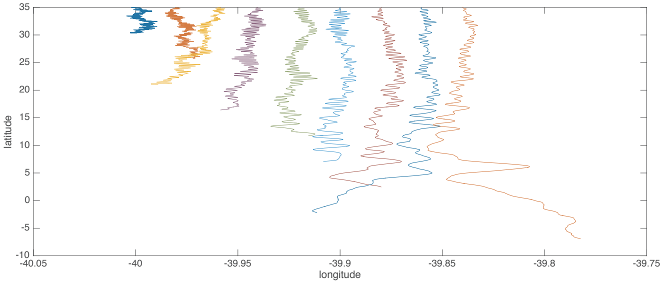
Real data


QQ Plot of Sample Data versus Distribution
Quantiles of Input Sample
4
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
1
2
3
Quantiles of exponential Distribution
QQ Plot of Sample Data versus Distribution
Quantiles of Input Sample
4
3
2
1
0
0
1
2
3
Quantiles of exponential Distribution
QQ Plot of Sample Data versus Distribution
Quantiles of Input Sample
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
0
1
2
3
Quantiles of exponential Distribution
QQ Plot of Sample Data versus Distribution
Quantiles of Input Sample
4
3
2
1
0
-1
0
1
2
3
4
Quantiles of exponential Distribution

Real data


(a)
1/λ (days, nonstationary fit, median = 3.42)
1/λ (days, stationary fit, median = 1.29)
(b)
frequency
$l_M(\theta_M) - l_W(\theta_W)$

Real data



35
30
25
20
15
10
5
0
-5
-10
latitude
-40.05
-40
-39.95
-39.9
-39.85
-39.8
-39.75
longitude

Simulated data

Sample size (N)	128	256	512	1024	2048	4096
Stationary frequency domain likelihood						
Bias (r)	-2.3481e-02	-3.2400e-02	-4.8112e-02	-6.9807e-02	-9.3332e-02	-1.1161e-01
Variance (r)	1.8163e-03	1.0760e-03	1.1422e-03	1.5550e-03	1.4045e-03	8.2890e-04
MSE (r)	2.3677e-03	2.1258e-03	3.4570e-03	6.4280e-03	1.0115e-02	1.3286e-02
Bias (σ)	2.5577e-02	5.4988e-02	8.9480e-02	1.3241e-01	1.7432e-01	2.0651e-01
Variance (σ)	3.3898e-03	2.8178e-03	3.3471e-03	4.4660e-03	3.9885e-03	2.1609e-03
MSE (σ)	4.0440e-03	5.8415e-03	1.1354e-02	2.1999e-02	3.4376e-02	4.4809e-02
CPU time (sec)	1.3083e-02	1.7776e-02	2.5743e-02	4.3666e-02	5.0948e-02	8.6940e-02
Nonstationary frequency domain likelihood						
Bias (r)	-4.6158e-03	-2.0129e-03	-1.4184e-03	-2.9047e-04	-2.6959e-04	8.8302e-05
Variance (r)	1.6508e-03	7.5379e-04	3.9819e-04	2.0710e-04	1.0674e-04	5.3236e-05
MSE (r)	1.6721e-03	7.5784e-04	4.0020e-04	2.0719e-04	1.0681e-04	5.3244e-05
Bias (σ)	-1.4999e-02	-8.8581e-03	-4.4302e-03	-2.5292e-03	-1.4125e-03	-9.1703e-04
Variance (σ)	2.2543e-03	1.1989e-03	6.4245e-04	3.4775e-04	2.0113e-04	1.0759e-04
MSE (σ)	2.4793e-03	1.2774e-03	6.6208e-04	3.5415e-04	2.0312e-04	1.0843e-04
CPU time (sec)	1.6814e-02	2.0272e-02	3.1397e-02	5.5925e-02	8.9997e-02	2.4147e-01

$$Z_t = rZ_{t-1} + \epsilon_t. \tag{12}$$

here g_t is a phase-shift.

Definition (Modulated process with highly significant correlation contribution)

Assume that Y_t is a modulated process. We say that Y_t is a modulated process with a highly significant correlation contribution if for any τ there exists two constants $N_\tau \geq \tau$ and $\alpha_t > 0$ such that for $N \geq N_\tau$,

$$\left| \frac{1}{N} \sum_{t=0}^{N-1-\tau} g_t g_{t+\tau} \right| \geq \alpha_\tau. \tag{13}$$

With this definition, the performance of the Whittle likelihood can be understood.

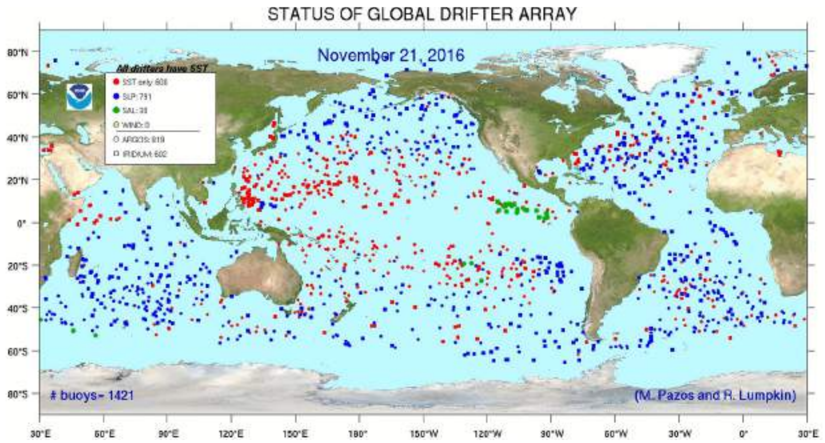
Summary

- Traditional local stationary facilitates straight averaging of summary statistics, thereby facilitating inference.
- This opens up new and interesting questions in asymptotic statistics, which feed back into other areas
- Well-motivated theory drives new algorithms, interpretations for approaches that already see wide use in data science

References:

- Analysis of nonstationary modulated time series with applications to oceanographic flow measurements, (arXiv:1605.09107, with A. P. Guillaumin, A. M. Sykulski, J. J. Early, J. M. Lilly)
- The De-Biased Whittle Likelihood for Second-Order Stationary Stochastic Processes (arXiv:1605.06718, with A. M. Sykulski, & J. M. Lilly)
- Lagrangian Time Series Models for Ocean Surface Drifter Trajectories (with A. M. Sykulski, J. M. Lilly, E. Danioux, J. Royal Statistical Society Series C 65(1) (2016) 29-50).

Questions?



STATUS OF GLOBAL DRIFTER ARRAY
November 21, 2016
All drifters have SST
SLP: 791
WIND: 0
ARGOS: 819
IRIDIUM: 602
buoys= 1421
(M. Pazos and R. Lumpkin)
80°N
60°N
40°N
20°N
0°
20°S
40°S
60°S
80°S
30°E
60°E
90°E
120°E
150°E
180°
150°W
120°W
90°W
60°W
30°W
0°
30°E