Wavelet Tools for Scaling Data

Patrice Abry, Patrick Flandrin and Darryl Veitch

CNRS — ENS Lyon, France and SERC, Melbourne, Australia

thanks to

P. Chainais (CNRS — ENS Lyon), L. Huang (SERC, Melbourne),

J. Cleary and J. Micheel (WAND, Univ. of Waikato, New Zealand)

Scaling Data

- "$1/f^{\alpha}$ spectrum"
- self-similarity
- long-range dependence
- (multi-)fractal behaviour
- cascades

Common property of scale invariance

- no characteristic scale (in a given range)
- invariant relations between scales

Multiresolution analysis — 1.

Basic idea

“signal = (low-pass) approximation + (high-pass) detail”

+

iteration

- successive approximations (at coarser and coarser resolutions)
  $\sim$ aggregated data
- details (difference in information between different resolutions)
  $\sim$ increments

Multiresolution is a natural language for scaling processes.

Multiresolution analysis — 2.

Wavelet-based formalization

A MultiResolution Analysis (MRA) of $L^2(\mathcal{R})$ is given by

1. a series of nested approximation spaces $\ldots V_1 \subset V_0 \subset V_{-1} \ldots$ such that their intersection is zero and their closure is dense in $L^2(\mathcal{R})$;
2. a dyadic scaling relation between approximation spaces :

$$X(t) \in V_j \Leftrightarrow X(2t) \in V_{j-1} \,;$$

3. a scaling function $\varphi(t)$ such that all of its integer translates $\{\varphi(t-n), n \in \mathcal{Z}\}$ form a basis of V_0.

Wavelet decomposition — 1.

Given a resolution depth J, a signal $X(t) \in V_0$ admits therefore the decomposition :

$$\underbrace{X(t)}_{\text{signal}} = \underbrace{\sum_k a_X(J,k)\,\varphi_{J,k}(t)}_{\text{approximation}} + \underbrace{\sum_{j=1}^{J}}_{J \text{ scales}} \underbrace{\sum_k \overbrace{d_X(j,k)}^{\text{wavelet coeffs.}} \psi_{j,k}(t)}_{\text{details}},$$

with $\{\xi_{j,k}(t) := 2^{-j/2}\,\zeta(2^{-j}t - k), j$ and $k \in \mathcal{Z}\}$, for $\xi = \varphi$ and ψ.

The wavelet $\psi(.)$ is constructed in such a way that its integer translates form a basis of W_0, the complement of V_0 in V_{-1}.

Wavelet decomposition — 2.

- The wavelet coefficients $d_X(j,k)$ are obtained as

$$d_X(j,k) := \langle X, \psi_{j,k} \rangle \,.$$

- From a practical point of view, they can be computed recursively with efficient pyramidal algorithms (faster than FFT).

- An important property of a wavelet is its number of vanishing moments, i.e., the number $N \geq 1$ such that

$$\int t^k \, \psi(t) \, dt \equiv 0, \quad \text{for } k = 0, 1, \ldots N-1 \,.$$

Wavelets and self-similarity — 1.

If a process $X = \{X(t), t \in \mathcal{R}\}$ is self-similar, i.e., if

$$\{X(t), t \in \mathcal{R}\} \stackrel{d}{=} \{c^{-H} X(ct), t \in \mathcal{R}\}$$

for any $c > 0$, its wavelet coefficients exactly reproduce the self-similarity through :

$$\{d_X(j,k), k \in \mathcal{Z}\} \stackrel{d}{=} \left\{2^{j(H+1/2)} d_X(0,k), k \in \mathcal{Z}\right\}.$$

Wavelets and self-similarity — 2.

- For processes whose wavelet coefficients have finite second-order statistics (e.g., fractional Brownian motion), one has :

$$\log_2 \mathbb{E} d_X^2(j,k) = j(2H+1) + \log_2 \mathbb{E} d_X^2(0,k).$$

- For processes whose wavelet coefficients may have infinite second-order statistics, but for which $\mathbb{E}\log_2 |d_X(j,k)|$ exists (e.g., linear fractional stable processes), one has :

$$\mathbb{E}\log_2 |d_X(j,k)| = j(H+1/2) + \mathbb{E}\log_2 |d_X(0,k)|.$$

Estimation of H in a Logscale Diagram

Key features for estimation

- Admissibility (mean value zero) $\Rightarrow$ stationarization of nonstationary processes with stationary increments (e.g., fractional Brownian motion)

- Number of vanishing moments high enough $\Rightarrow$ almost decorrelation in the wavelet domain, scale by scale :

$$\mathbb{E} d_X(j,n) d_X(j,m) \propto \int \frac{|\Psi(2^j f)|^2}{|f|^\alpha} e^{i2\pi(n-m)f} \, df.$$

LRD in $X(.)$ can be turned into SRD in $d_X(j,.)$.

- Corollary : detrending

Beyond 2nd order scaling

$$T_X(a) := 2^{-j/2}\, d_X(j,n)\Big|_{j=\log_2 a}$$

$$\mathbb{E}|T_X(a)|^q \sim a^{Hq} = \quad \exp\{Hq \ln a\} \qquad \text{(monoscaling)}$$

$$\downarrow$$

$$\exp\{H(q) \ln a\} \qquad \text{(multiscaling)}$$

$$\downarrow$$

$$\exp\{H(q) n(a)\} \qquad \text{(cascade)}$$

Beyond 2nd Order Scaling

- Self-Similarity: $\mathbb{E}|d_X(j,k)|^q = C_q (2^j)^{qH} = C_q \exp\left(qH \ln(2^j)\right)$
  - A single scaling parameter H
  - Power-laws
- Multi-Scaling: $\mathbb{E}|d_X(j,k)|^q = C_q (2^j)^{H(q)} = C_q \exp\left(H(q) \ln(2^j)\right)$
  - A collection of parameters: H(q)
  - Power-laws
- Infinitely Divisible Cascade: $\mathbb{E}|d_X(j,k)|^q = C_q \exp\left(H(q) n(2^j)\right)$
  - No Power-Law !
  - order q / scale 2^j separability

Note : Scale : $a = 2^j$

Normalization : $d_X(j,k) = \langle x, 2^{-j/2} \psi_{j,k} \rangle$

Cascade

Castaing 90, 96, Arnéodo et al., 97

- Self-Similarity :$(a < a')$, $P_a(d) = \frac{1}{\alpha_0} P_{a'}(\frac{d}{\alpha_0})$, $\alpha_0 = \left(\frac{a'}{a}\right)^H$

- Cascade : $(a < a')$, $P_a(d) = \int G_{a,a'}(\ln \alpha) \frac{1}{\alpha} P_{a'}(\frac{d}{\alpha})\, d \ln \alpha$

  - $G_{a,a'}$ kernel or propagator of the cascade
  - $G_{a,a'}(\ln \alpha) = \delta(\ln \alpha - \ln \alpha_0)$ $\rightarrow$ Self-Similarity (Kolmogorov, 41)

- $u = \ln |d|$: $\underline{P}_a(\ln |d|) = \int G_{a,a'}(\ln \alpha) \underline{P}_{a'}(\ln |d| - \ln \alpha)\, d\ln \alpha$

$$\Rightarrow \underline{P}_a = G_{a,a'} * \underline{P}_{a'} \qquad \text{Convolution}$$

Infinitely Divisible Cascade

- No Characteristic scale :

If $a = a_0 < a_1 < \ldots < a_n = a'$ $\underline{P}_{a_{k-1}} = G_{a_{k-1},a_k} * \underline{P}_{a_k}$

Then $\underline{P}_a = G_{a,a'} * \underline{P}_{a'}$ with $G_{a,a'} = G_{a_0,a_1} * \ldots * G_{a_{n-1},a_n}$

- Infinite divisibility (or Continuous Self Similarity) :

$$G_{a,a'}(\ln \alpha) = [G_0(\ln \alpha)]^{*\{n(a)-n(a')\}}$$

$$\tilde{G}_{a,a'}(q) = \left[\tilde{G}_0(q)\right]^{n(a)-n(a')}$$

$$\mathbb{E}|d_X(j,k)|^q = C_q \exp\left[H(q)n(2^j)\right], \quad H(q) = \ln \tilde{G}_0(q)$$

$\rightarrow$ Separability order q / Scale 2^j

- Arbitrariness : $\ln \mathbb{E}|d_X(j,k)|^q = H(q)n(2^j) + K_q$

$$H(q)n(2^j) + K_q = [H(q)/\beta]\left[\beta\, n(2^j) + \gamma\right] + [K_q - \beta H(q)/\gamma]$$

$$= H'(q)n'(2^j) + K'_q$$

Scale Invariant Infinitely Divisible Cascade

- Scale Invariance : Set $n(a) = \ln a$, Then,

$$\begin{aligned} \tilde{G}_{a,a'}(q) &= \exp\left[\left(\ln \tilde{G}_0(q)\right)(n(a) - n(a'))\right] \\ &= \left(\frac{a}{a'}\right)^{\left(\ln \tilde{G}_0(q)\right)} \\ \mathbb{E}|d_X(j,k)|^q &= C_q(2^j)^{\ln \tilde{G}_0(q)} \\ &\rightarrow \text{Multi-Scaling} \end{aligned}$$

- Multifractal Analysis : $\mathbb{E}|d_X(j,k)|^q = C_q(2^j)^{\zeta_q}$, $2^j \rightarrow 0$
  - $\tilde{G}_0(q) = \exp(\zeta_q)$,
  - $n(a) = \ln(a)$

e.g., Multinomial stochastic cascades, Mandelbrot

Infinitely Divisible Cascade : Model testing

$$H(q) = \ln \tilde{G}_0(q),$$

- Power-laws are back !

$$\mathbb{E}|d_X(j,k)|^q = C_{q,p} \left(\mathbb{E}|d_X(j,k)|^p\right)^{(H(q)/H(p))}$$

$$\ln \mathbb{E}|d_X(j,k)|^q = \frac{H(q)}{H(p)} \ln \mathbb{E}|d_X(j,k)|^p + \kappa_{q,p}$$

- Extended Self-Similarity

- Key-Quantities : $S_q(j) = \frac{1}{n_j} \sum_{k=1}^{n_j} |d_X(j,k)|^q$
  - Estimators for $\mathbb{E}|d_X(j,k)|^q$,
  - $d_X(j,k)$ stationary, weak statistical dependence,
  - Statistics of $\ln S_q(j)$: e.g., able to estimate $\mathrm{Var}\ \ln S_q(j)$
- Model testing :

Check straight lines in $\ln S_q(j)$ versus $\ln S_p(j)$ plots.

Consider the variances of the $\ln S_q(j)$!

Infinitely Divisible Cascade : Estimation

p is an arbitrary reference

- $H(.)$: $\ln \mathbb{E}|d_X(j,k)|^q = H(q)/H(p) \ln \mathbb{E}|d_X(j,k)|^p + \kappa_{q,p}$

  Weighted Linear Regression in $\ln S_q(j)$ versus $\ln S_p(j)$ plots

  $$\hat{H}(q)/H(p) = \text{slope}_{q,p}$$

  Consider the variances of the $\ln S_q(j)$!

- $n(.)$: $\ln \mathbb{E}|d_X(j,k)|^q = H(q)n(2^j) + C_q$

$$H(p)\hat{n}(2^j) = \langle \frac{H(p)}{\hat{H}(q)} \left(\ln S_q(j) - \langle \ln S_q(j) - \frac{\hat{H}(q)}{H(p)} \ln S_p(j) \rangle_j \right) \rangle_q$$

Consider the variances of the $\ln S_q(j)$!

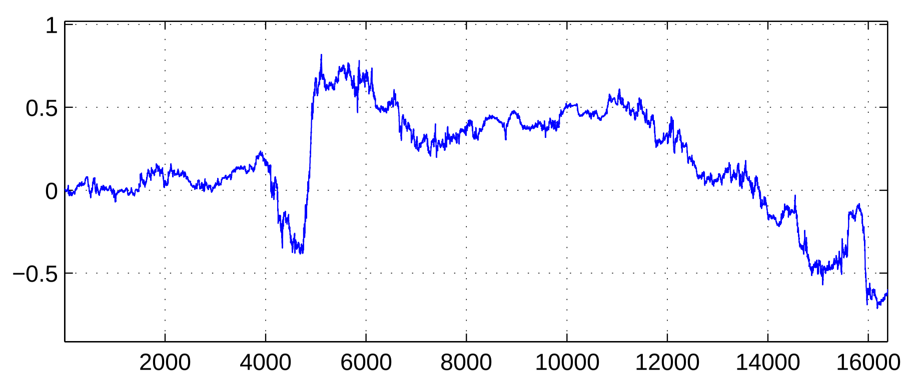
Note : Arbitrary Convention : $H(q)n(2^j) \equiv (H(q)/H(p)) \left(H(p)n(2^j) \right)$

Infinitely Divisible Cascade : An Example

fractional Brownian motion in Multifractal time

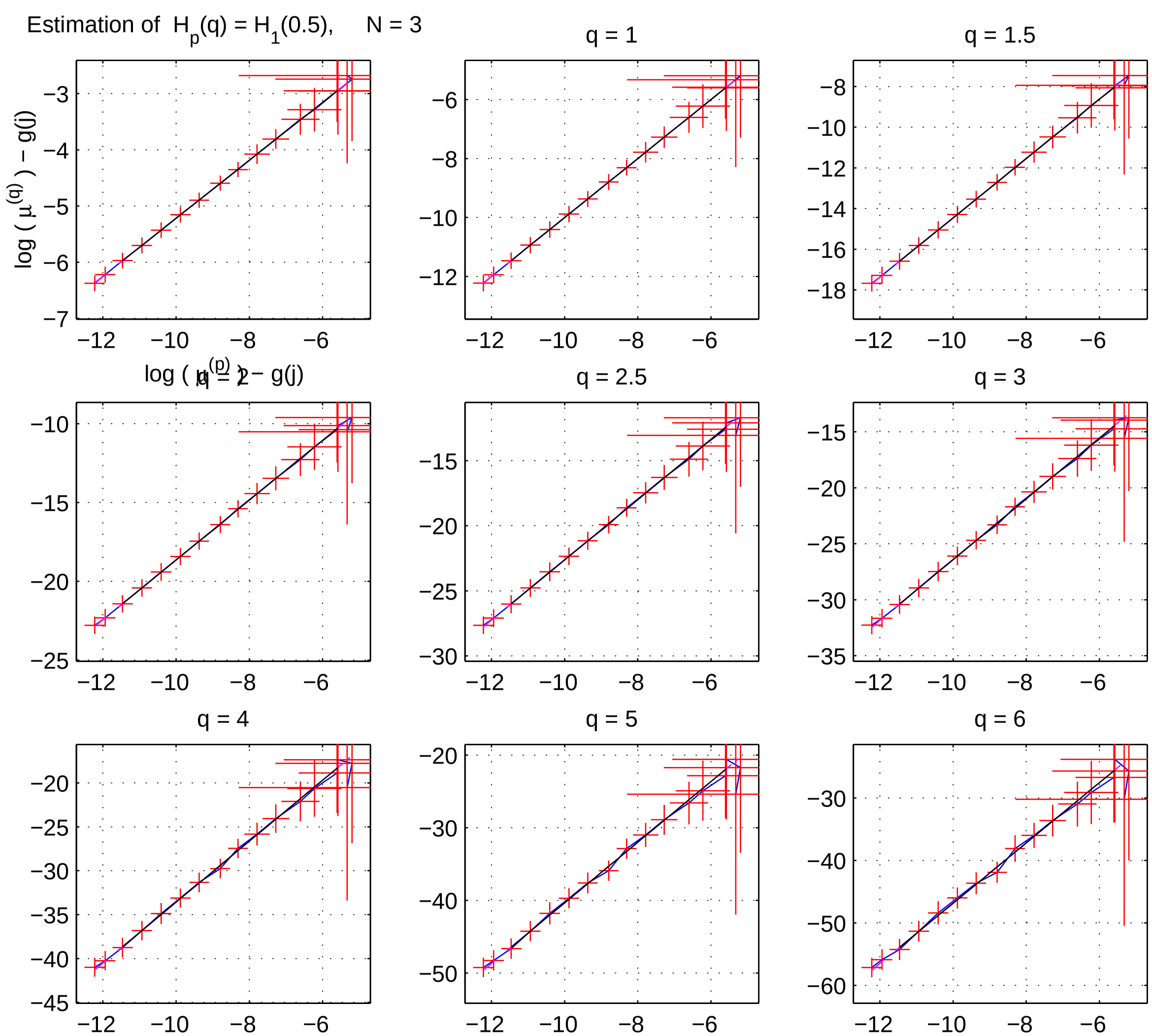
Mandelbrot 97, Riédi, 99

- Let $\mu(t)$ be the measure of a binomial multiplicative cascade,
- Let $\mathcal{M}(t) = \int^t d\mu(s)$ be its distribution function,
- Let $B_H(t)$ be a fBm with self-similarity parameter H,
- Define the fBm in Multifractal time as : $\mathcal{B}(t) = B_H(\mathcal{M}(t))$,
- Then, $\mathcal{B}(t)$ is a process with rich scaling.

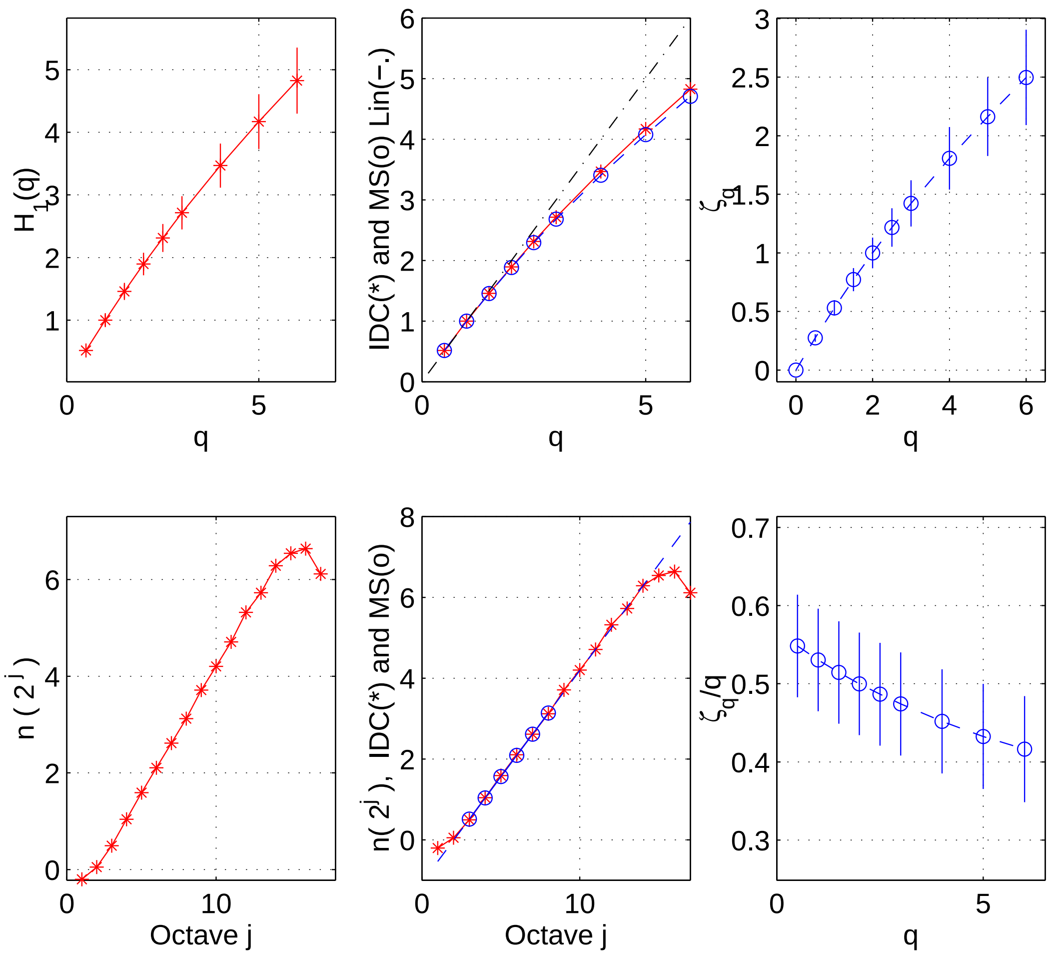


Thanks to P. Gonçalvès for Matlab synthesis codes for the MfBm

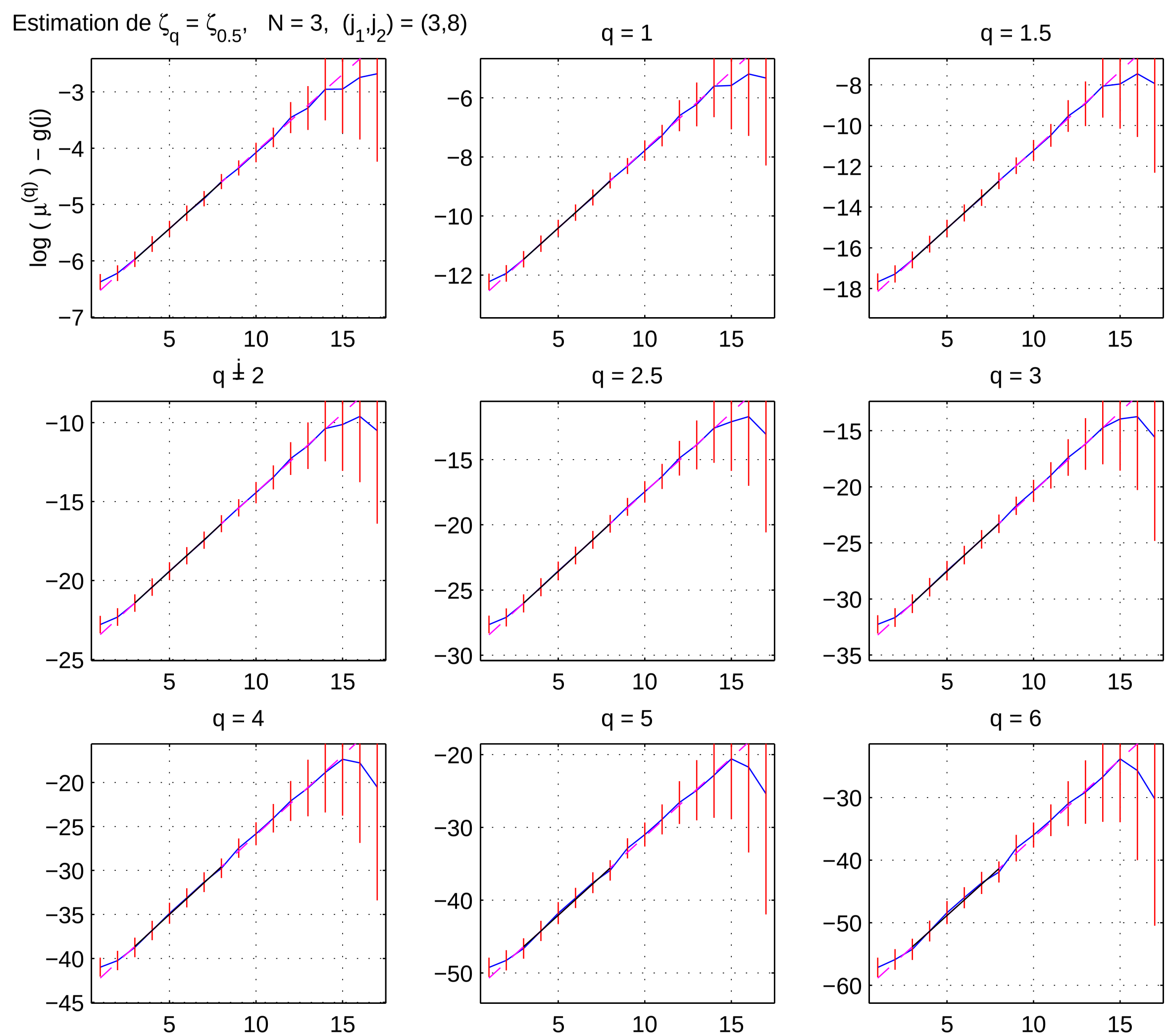
Cascade Analysis : $\log_2 S_q(j)$ versus $\log_2 S_p(j)$



Scaling Analysis : Summary Plot



Multi-Scaling (or Multifractal) Analysis : $\ln S_q(j)$ versus j plots



The Data: TCP/IP over ATM

Data courtesy of Prof. Cleary and WAND, University of Waikato NZ.

(Special thanks to Jörg Micheel at WAND and Li Dong Huang of SERC for time series extraction)

The Measurement Equipment:

- Measurement of OC3 ATM link (155 Mbits/s).
- Cell capture (64 byte records) and timestamping on high performance "DAG2.1" adaptor cards designed and built at WAND.
- $\sim 0.1\mu s$ timestamping and no losses.
- GPS based drift correction of clocks.

The Raw Data:

- Important link, external and internal traffic, at Auckland University.
- Busiest two hour period: 6pm - 8pm, Thursday July 8th, 1999.
- One VC, IP traffic filtered, 137 Mbytes of raw data.
- Only first cell of each IP packet captured: header + 40 bytes.
- TCP connections can be reconstructed, data payloads erased.

Two Extracted Time Series

From a set of raw data many different time series can be extracted. Here we consider two:

Arrivals: The number of new TCP connections in 10 ms intervals.

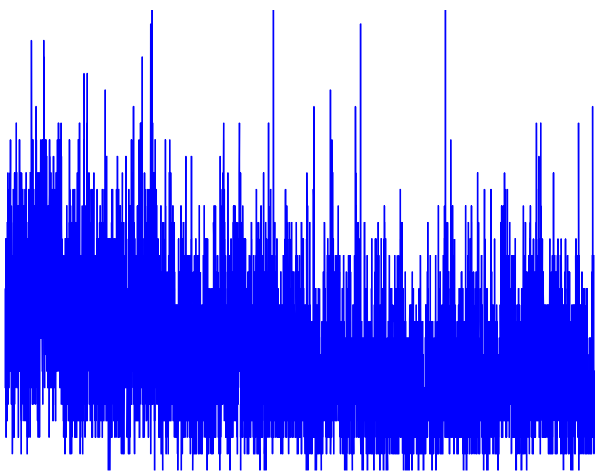
- series is time indexed and non-negative integer valued.
- series is $n = 720,000$ long.
- low data density, 90.2% zeros (average traffic rate 1.13 Mbits/s).

Durations: Successive durations of TCP connections.

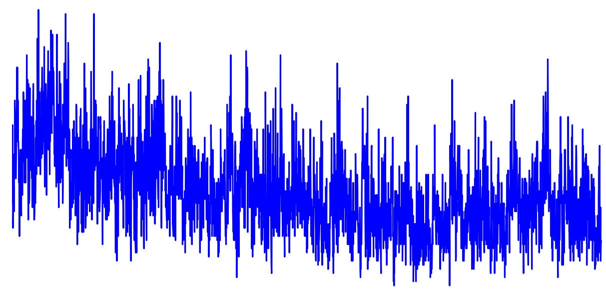
- series is intrinsically discrete and positive real-valued.
- series is $n = 66,370$ long.
- mean duration is ~ 3 minutes.

Connection arrivals: Aggregated series

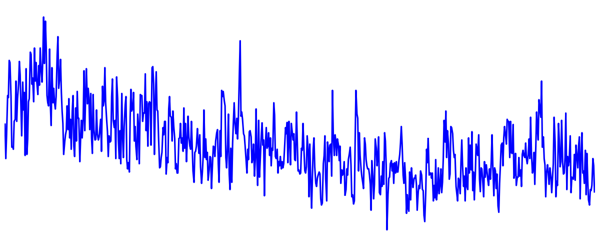
Agg level 64, Len = 11250



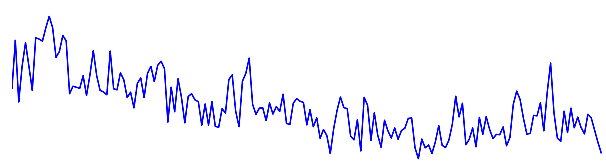
Agg level 256, Len = 2812



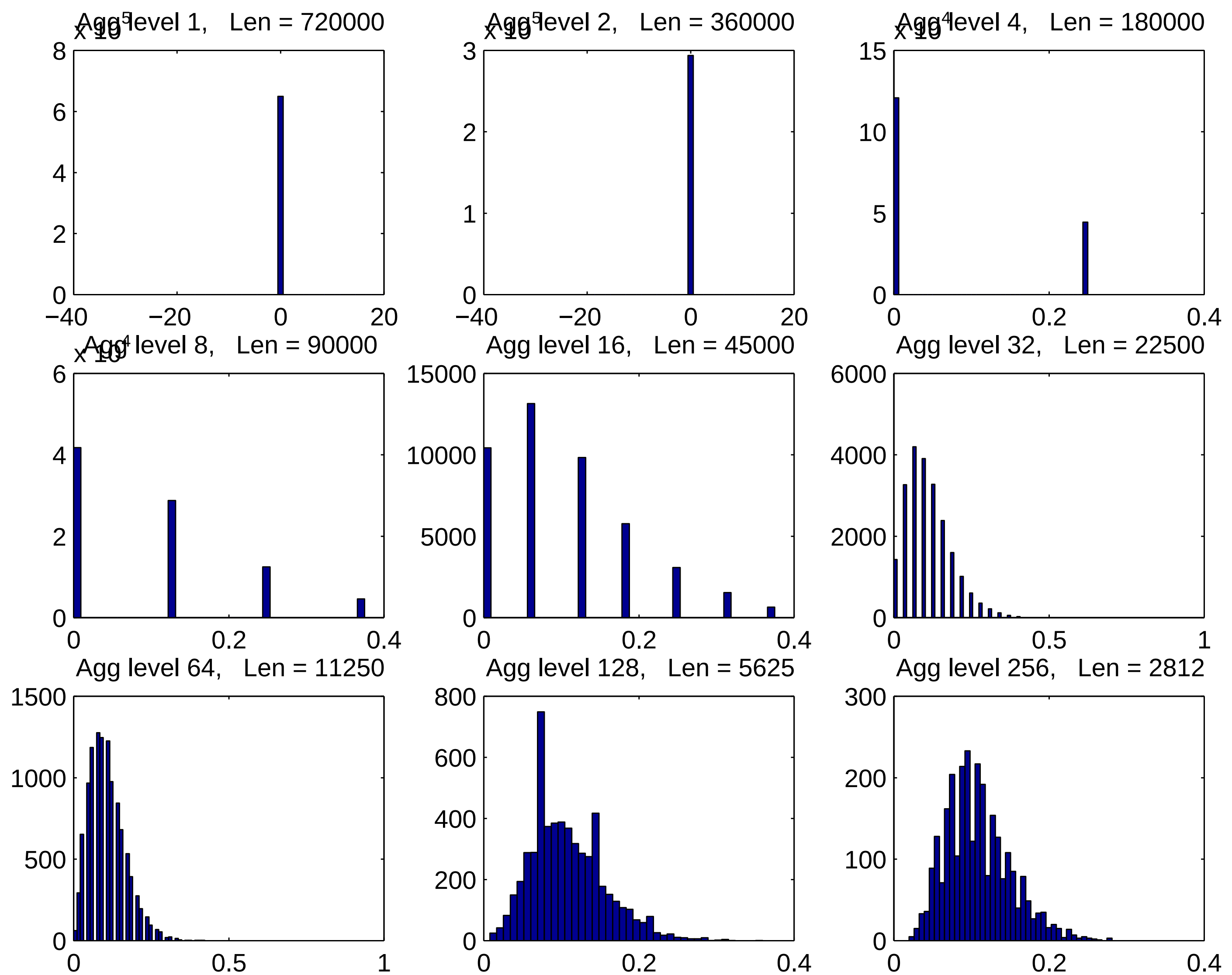
Agg level 1024, Len = 703



Agg level 4096, Len = 175

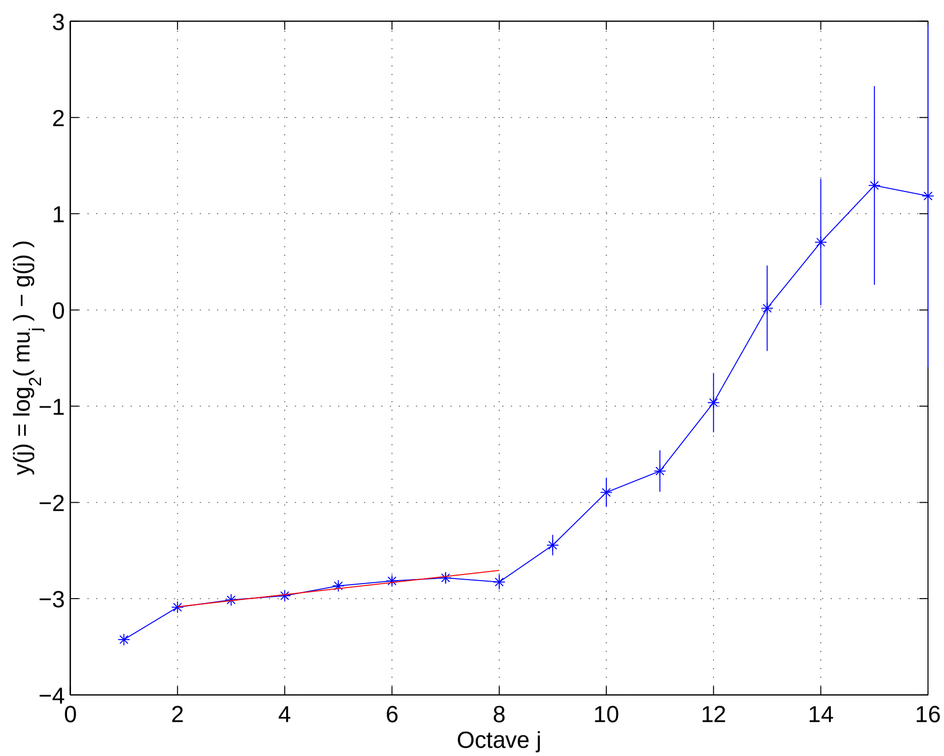


Connection arrivals: 1D Marginal



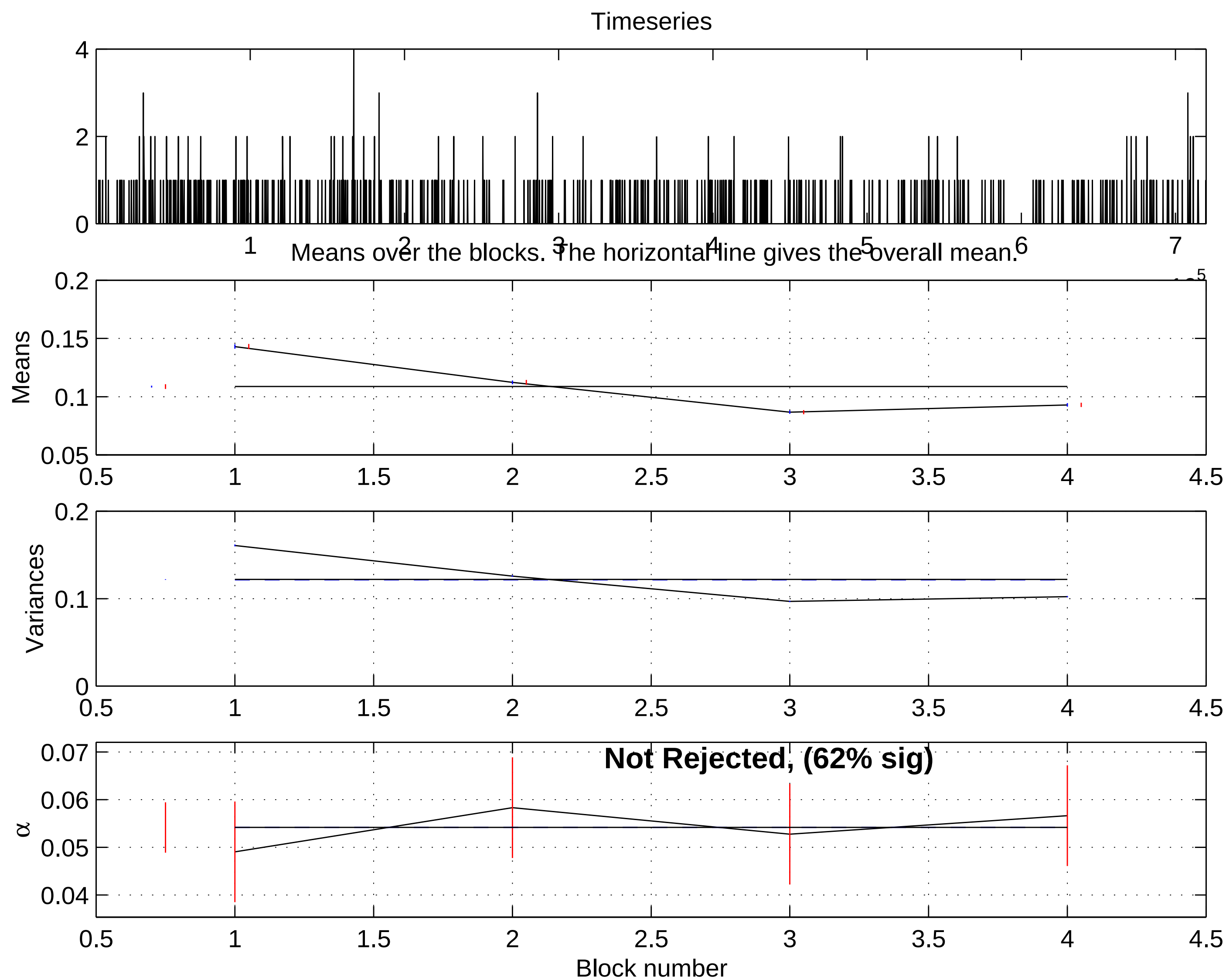
Connection arrivals: The Logscale Diagram

Spectral Estimate: $\log_2(S_2(j)) = \log_2\left(\frac{1}{n_j}\sum_{k=1}^{n_j} |d_X(j,k)|^2\right)$ vs j



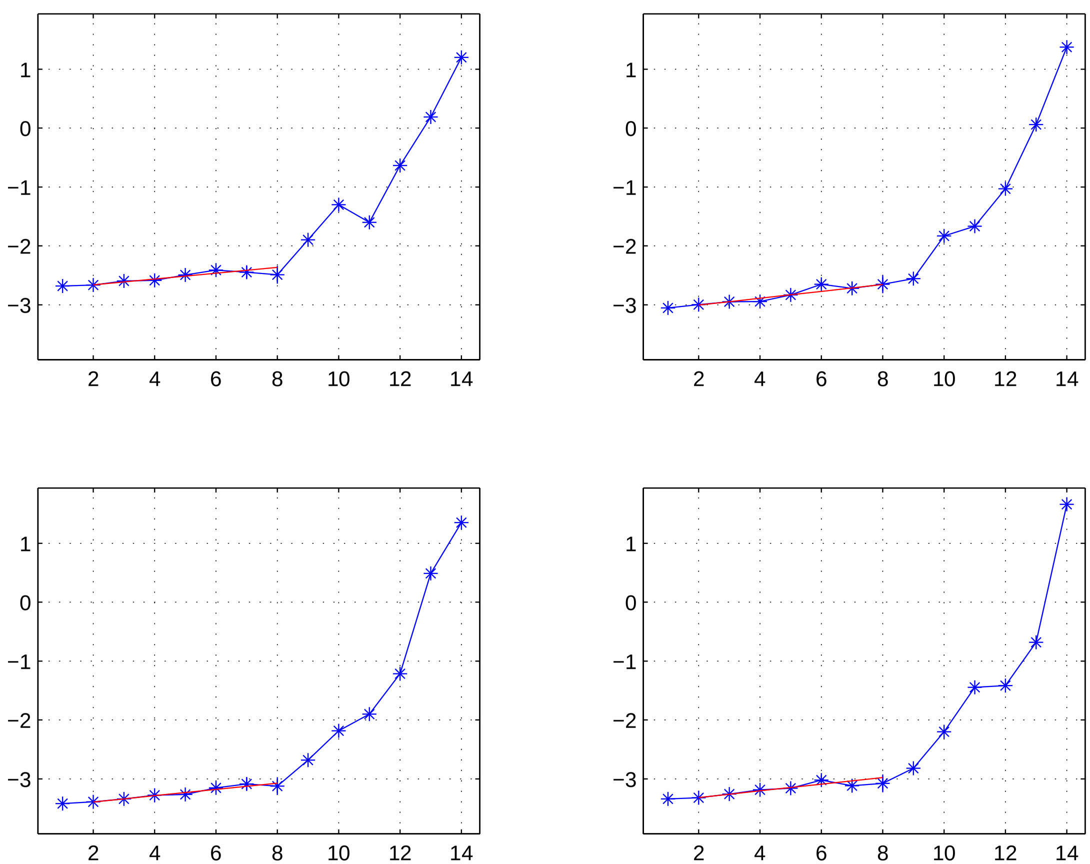
- Small scales: slope $\sim 0.063 \pm 0.005$ [discontinuous sample path]
- Large scales: slope $\sim 0.49 \pm 0.04$ [long range dependence]

Stationarity check: First and second order



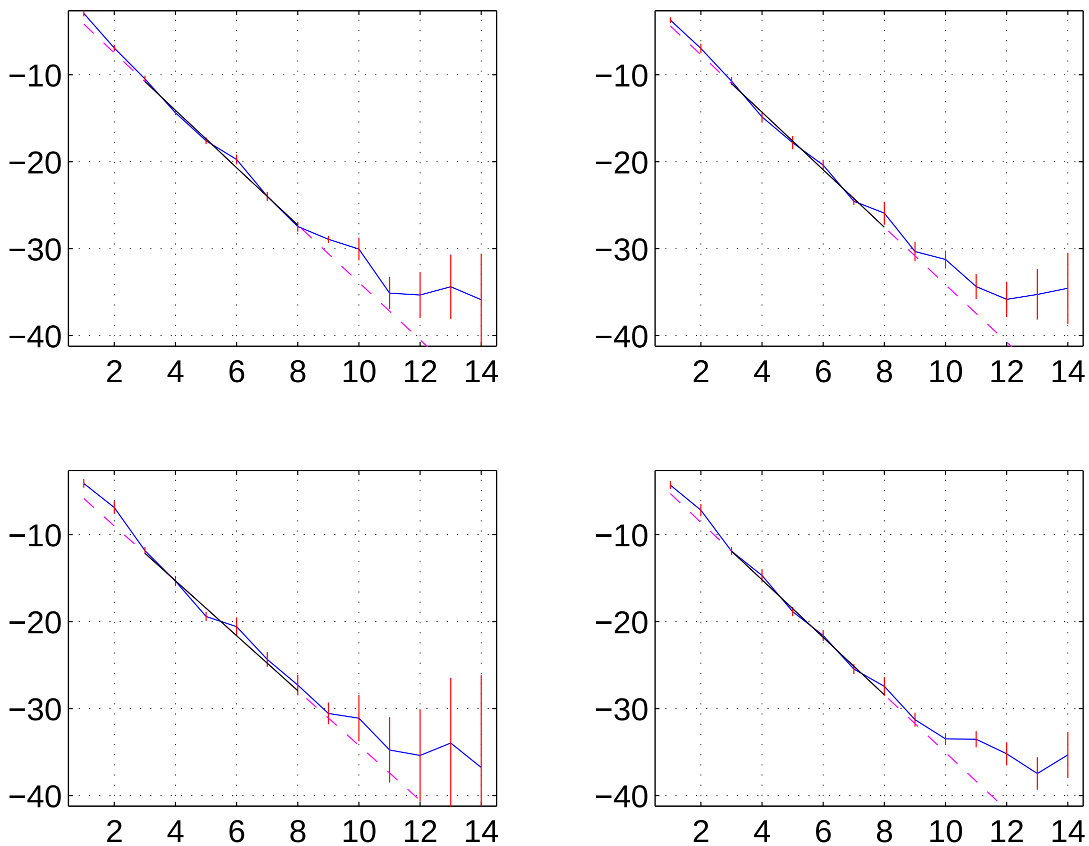
Stationarity check: Second order scaling

$$\log_2(S_2(j)) = \log_2\left(\frac{1}{n_j}\sum_{k=1}^{n_j}|d_X(j,k)|^2\right) \quad vs \quad j$$



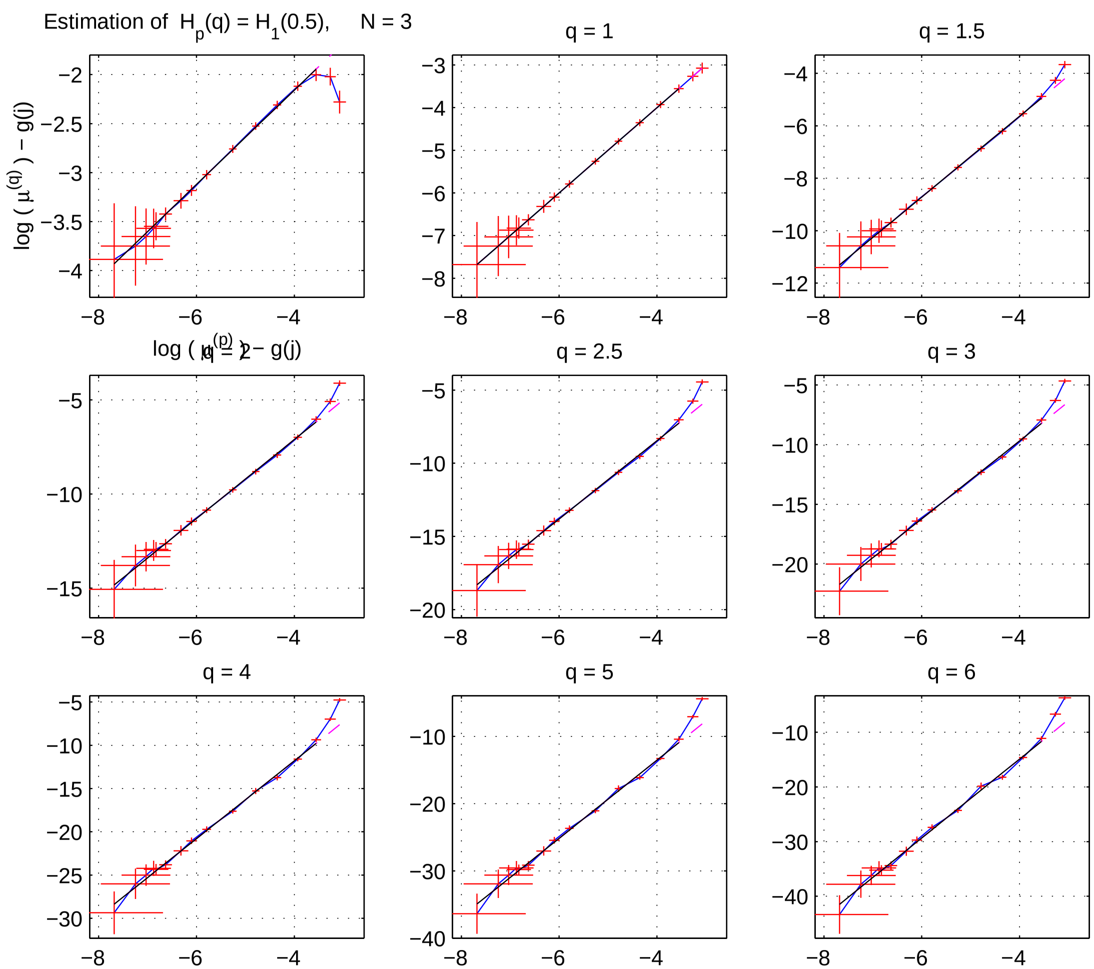
Stationarity check: Higher order scaling $(q = 6)$

$$\log_2(S_q(j)) = \log_2\left(\frac{1}{n_j}\sum_{k=1}^{n_j}|d_X(j,k)|^q\right) \quad vs \quad j$$



Cascade Analysis: Divisibility and Estimation

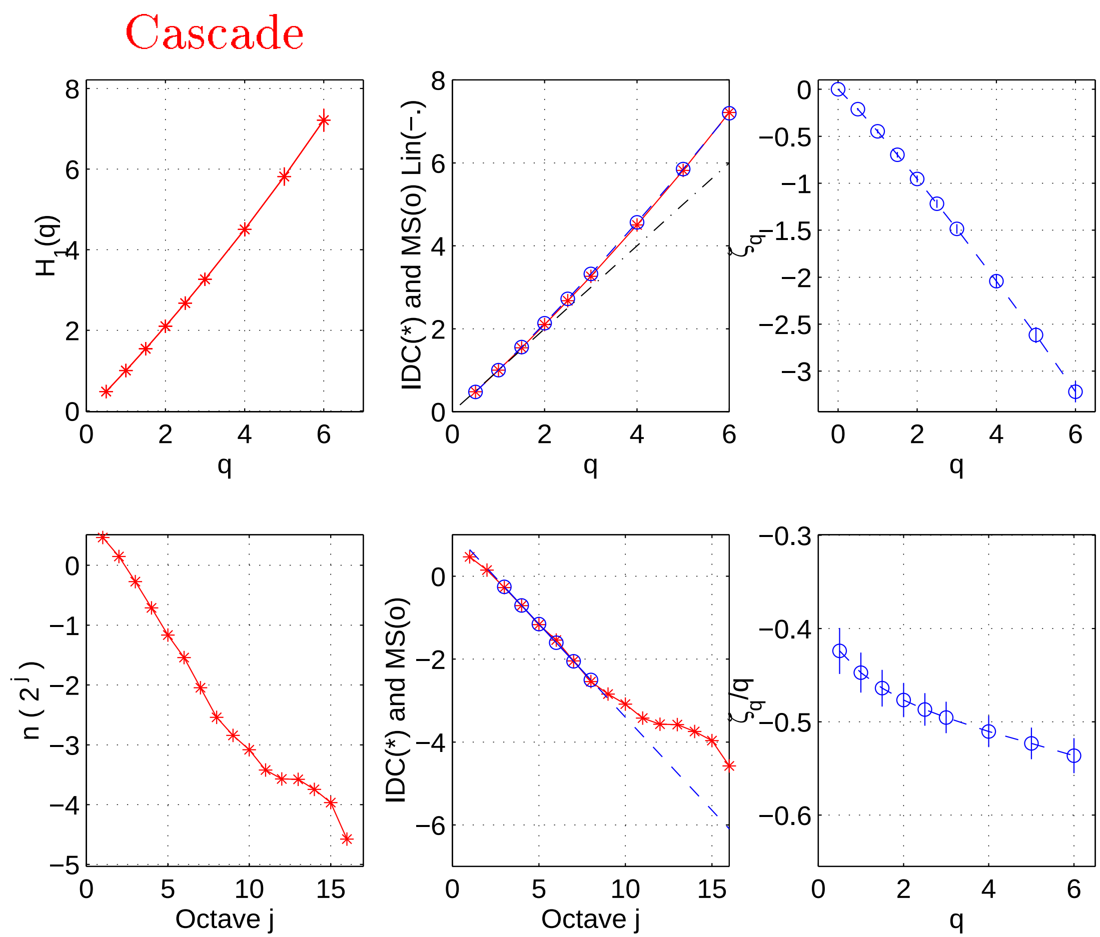
Over all scales: $\log_2(S_q(j))$ *vs* $\log_2(S_1(j))$, $q = 0.5, \cdots, 6.$



Cascade Analysis: Observations

$H(q)$ $\longrightarrow$

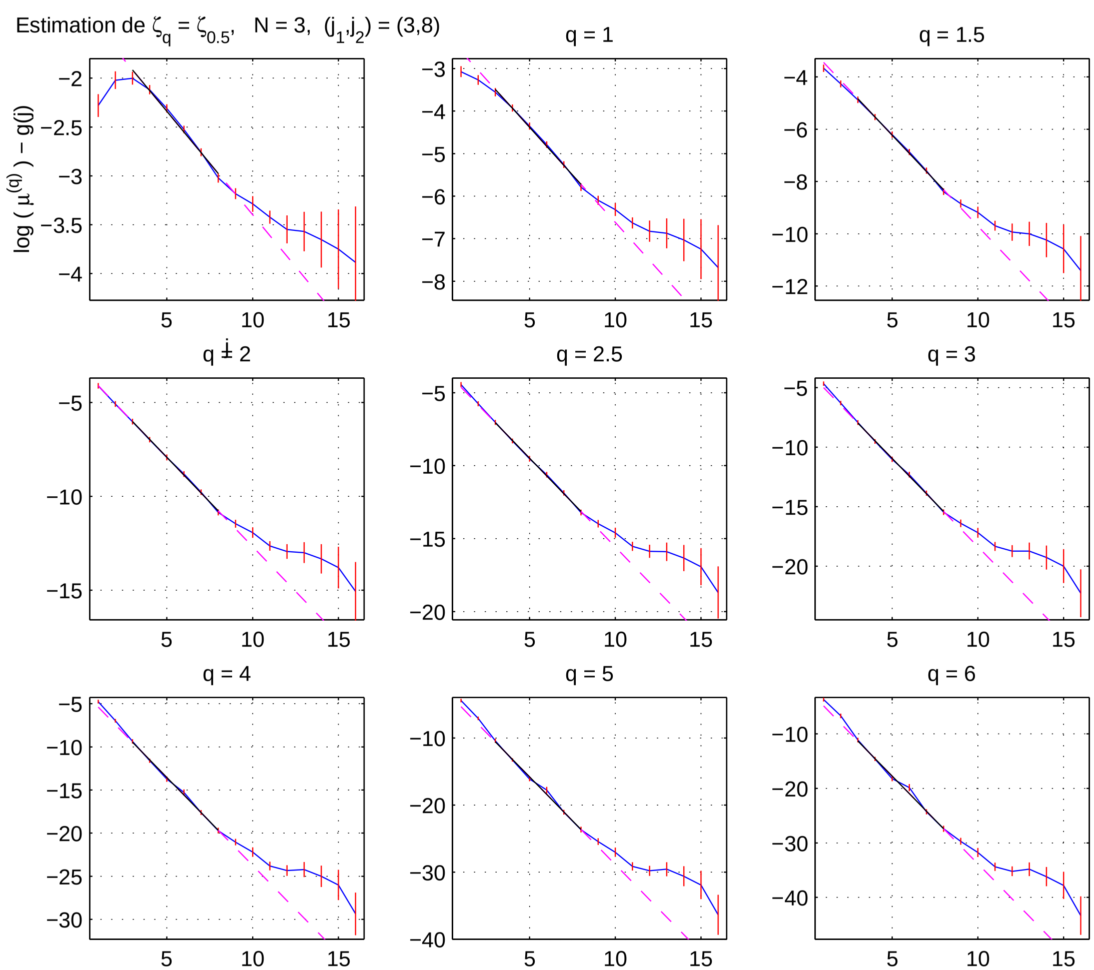
$n(2^j)$ $\longrightarrow$



Conclusion: IDC model fits well over all scales, $n(a)$ is not $\log(a)$.

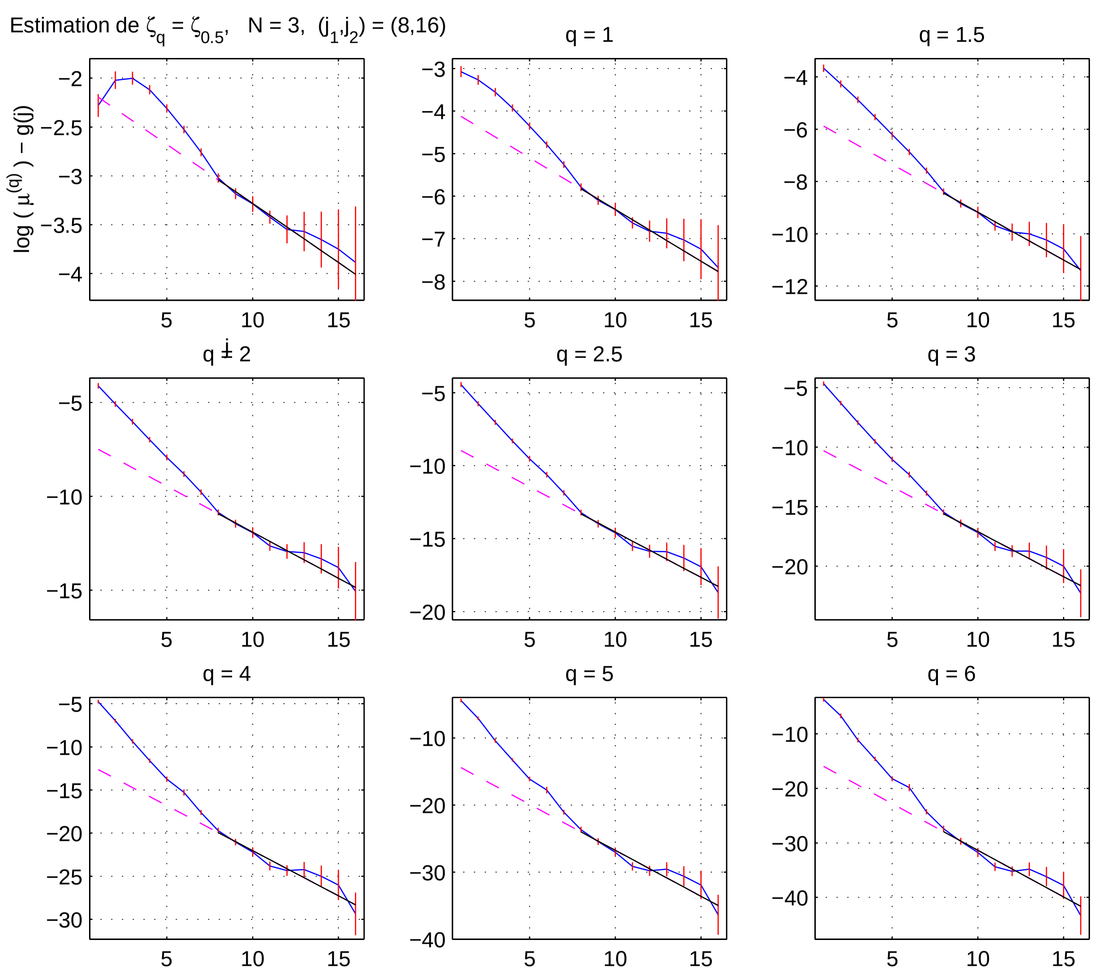
Multiscale Analysis: Small scales

$$\log_2(S_q(j)) = \log_2\left(\frac{1}{n_j} \sum_{k=1}^{n_j} |d_X(j,k)|^q \right) \ vs \ \ j, \quad q = 0.5, \cdots, 6.$$



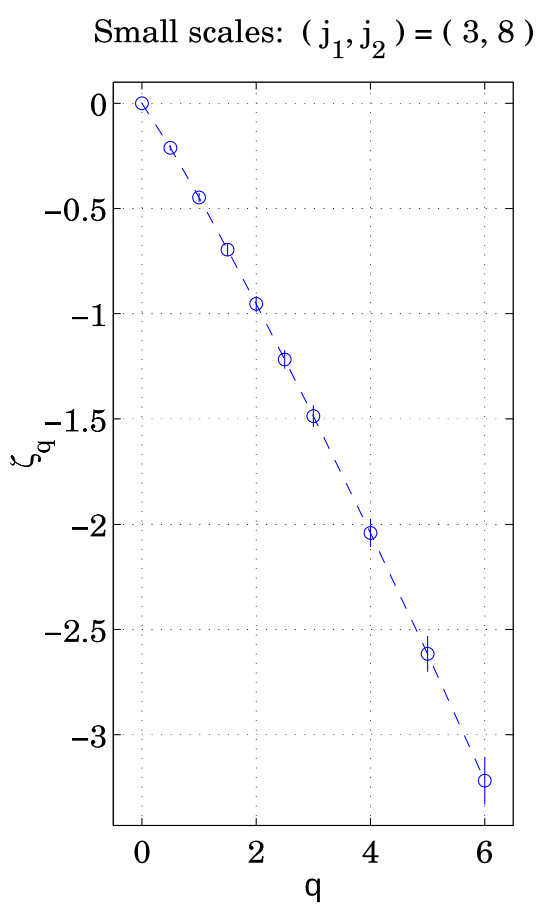
Multiscale Analysis: Large scales

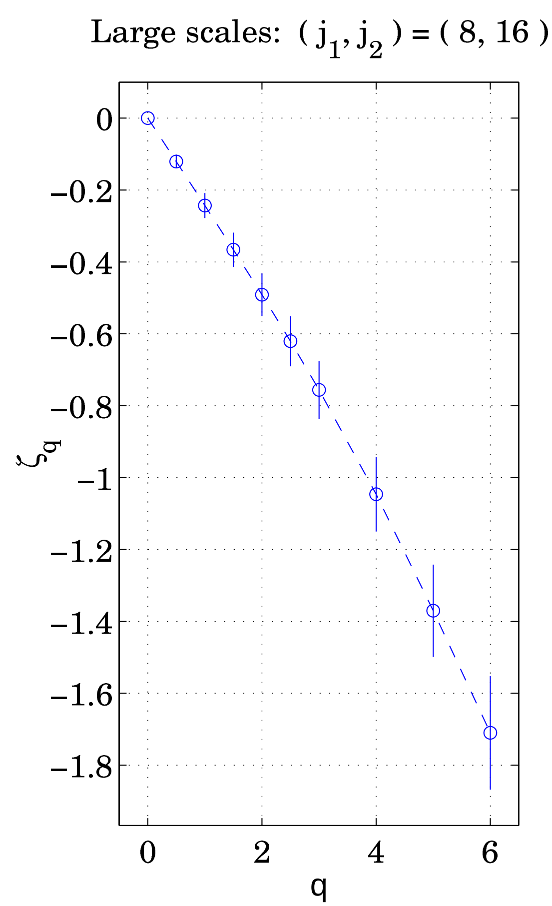
$$\log_2(S_q(j)) = \log_2\left(\frac{1}{n_j}\sum_{k=1}^{n_j}|d_X(j,k)|^q\right) \; vs \;\; j, \quad q = 0.5, \cdots, 6.$$



Multiscale Diagrams

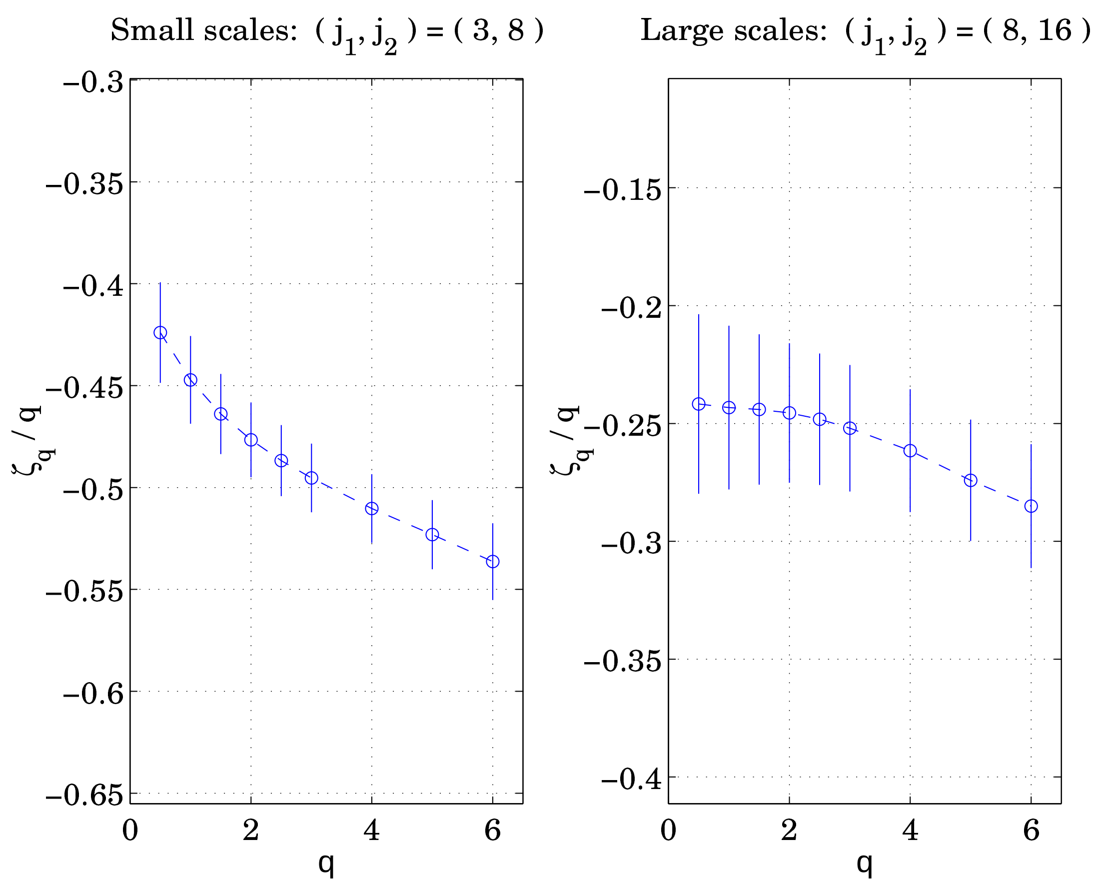
ζ_q against q





Linear Multiscale Diagrams

ζ_q/q against q



Conclusion: Small scales: Non-trivial multiscaling, eg Multifractal.

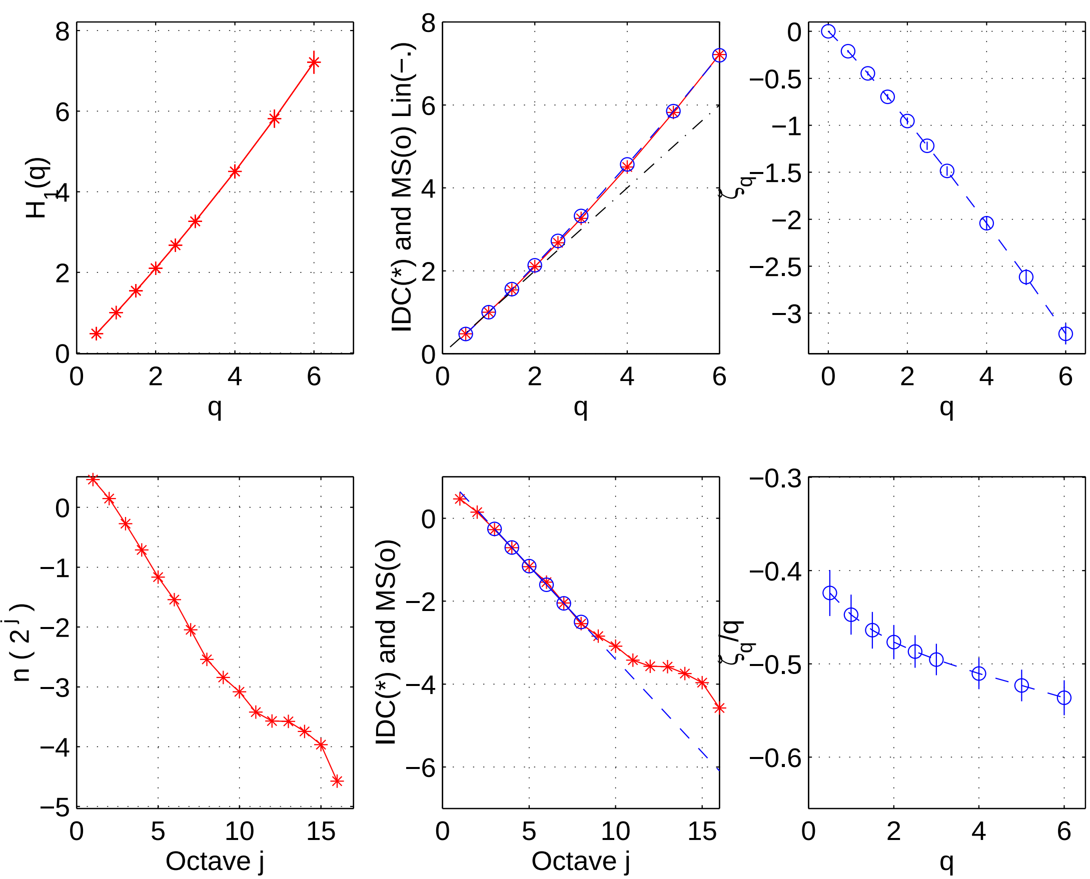
Large scales: Trivial multiscaling, eg H-ss model.

Comparison: Small scales (Connection Arrivals)

Cascade | Comparison | Multiscale

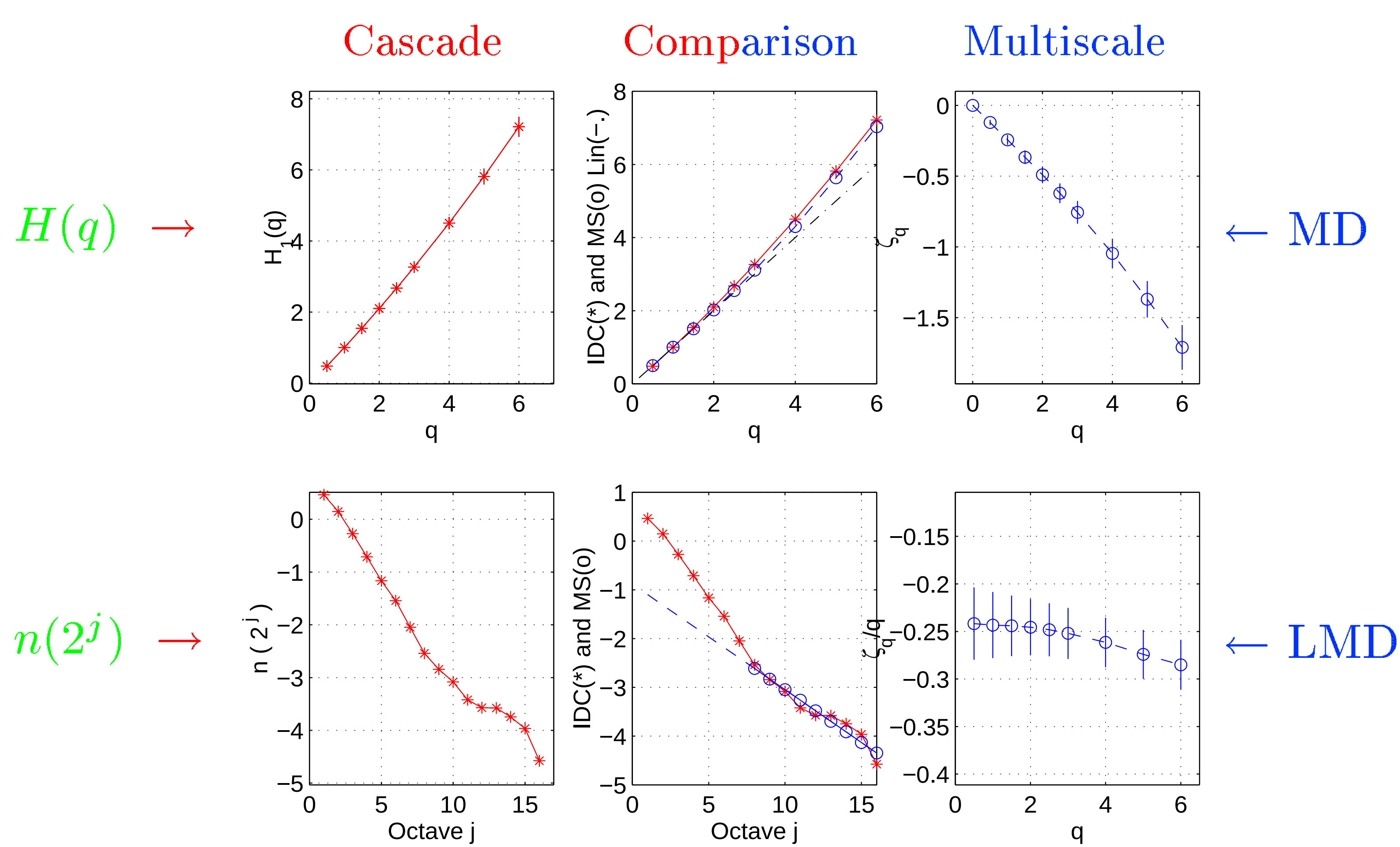
$H(q)$ → ... ← MD

$n(2^j)$ → ... ← LMD



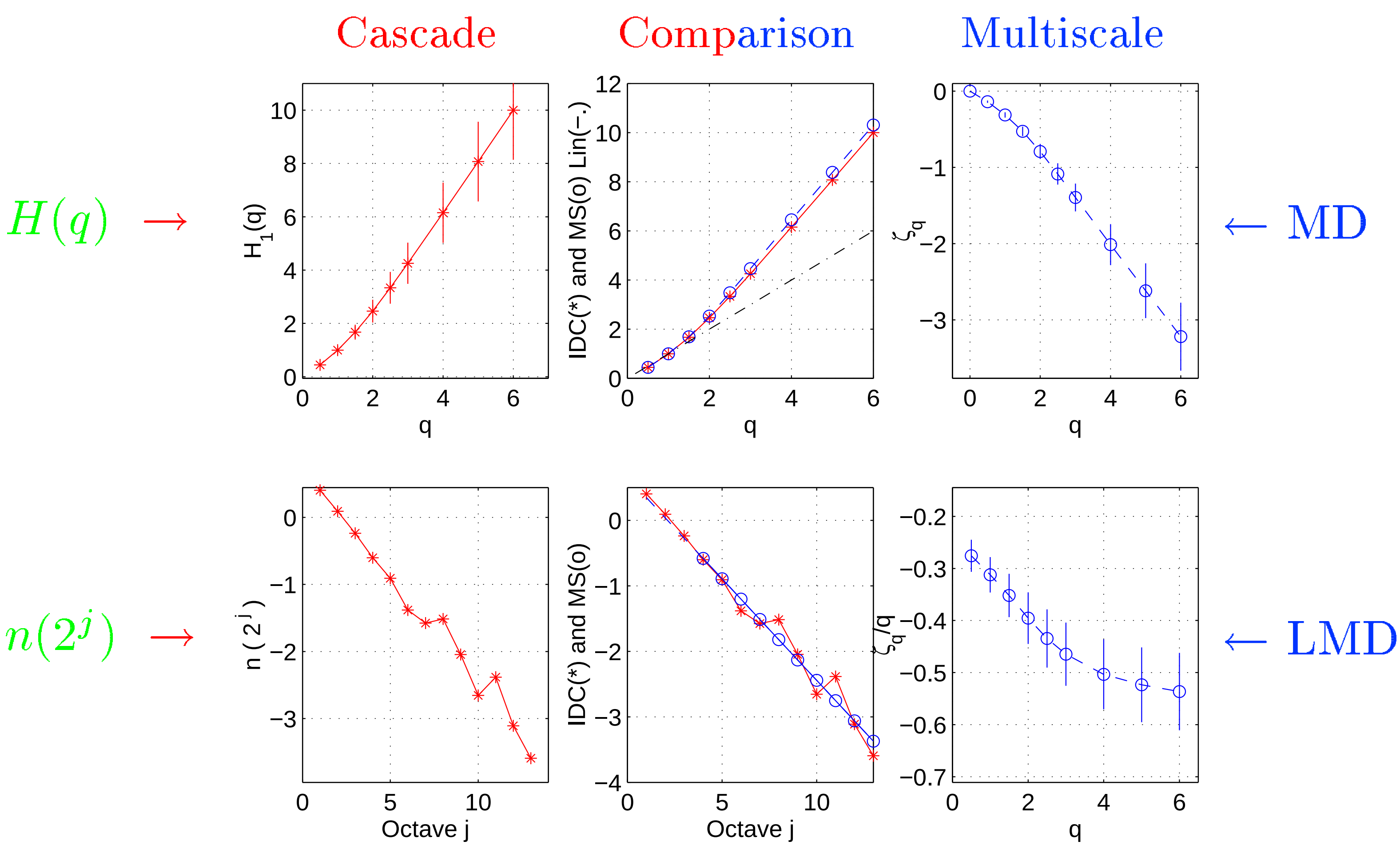
Conclusion: Cascade reduces to Multiscale over small scales.

Comparison: Large scales (Connection Arrivals)



Conclusion: Cascade reduces to Multiscale over large scales!

Comparison: Small scales (Connection **Durations**)



Conclusion: For Durations we see a single multiscaling range.

Conclusions

- Single IDC observed for TCP connection arrivals and durations. Independent multiscaling models often observed in two scale ranges. Cascade model reveals they are equivalent: If $n(a)$ is "piecewise log" then the two $\zeta(q)$ are simple multiples.

- Infinitely divisible cascades generalise multiscale analysis

$$E \log_2 S_q(j) \sim H(q) n(2^j) + C_j .$$

  When $n(a) = c \log(a) + d$, IDC reduces to the multiscale ζ_q analysis.

- Wavelets provide a statistically effective and flexible basis for scaling analysis of diverse types.

Matlab code for second order scaling analysis, and documentation, available at:
http://www.serc.rmit.edu.au/∼darryl