Phân tích số liệu

+ Chuẩn hóa số liệu – Data standardization

+ Nội suy dữ liệu bị mất (interpolation of data gaps)

+ Loại bỏ xu thế (tuyến tính) – Trend removal

+ Bộ lọc số - digital filtering

+ Hồi qui tuyến tính đơn giản

+ Phân tích Fourier cơ bản

+ *Trình diễn* số liệu

Một số phép tính thống kê đơn giản

Đối với tính hiệu ngẫu nhiên rời rạc $\{x_i\}$ i = 1, 2…n

- **Tính trung bình (mean, average):** $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Độ lệch chuẩn:** $$SD = \sigma = \sqrt{\frac{1}{n-1}\sum_{1}^{n}(x-\bar{x})^2}$$

- **Phương sai (variance):** $$\sigma^2 = \frac{1}{n-1}\sum_{1}^{n}(x-\bar{x})^2$$

Khử xu thế trong số liệu (trend removal)

Đôi khi trong số liệu đo đạc tồn tại các nhiễu động:

- Xu thế không xác định

- hoặc sóng thành phần có tần số thấp với

chiều dài sóng > khoảng thời gian đo được $T_r = N\Delta t$

→ Các nhiễu này phải được loại bỏ trước khi xử lý số liệu

→ Thường kỹ thuật này áp dụng cho xử lý số liệu sóng được đo trong vùng ảnh hưởng bởi thủy triều

Khử xu thế trong số liệu (trend removal)

Kỹ thuật loại bỏ xu thế tuyến tính này thường là điều chỉnh (số liệu gốc) bằng đa thức bậc thấp sử dụng pp bình phương tối thiểu (least square)

Giả sử chuỗi số liệu là u_n được điều chỉnh với một đa thức bậc K như sau

$$\tilde{u} = \sum_{k=0}^{K} b_k (n\Delta t)^k \qquad n = 1, 2, \ldots N$$

Khử xu thế trong số liệu (trend removal)

Điều chỉnh bình phương tối thiểu giữa số liệu và đa thức như sau

$$Q = \sum_{n=1}^{N} (u_n - \tilde{u}_n)^2 = \sum_{n=1}^{N} \left[u_n - \sum_{k=0}^{K} b_k (n\Delta t)^k \right]^2$$

Khử xu thế trong số liệu (trend removal)

Lấy đạo hàm toàn phần của Q theo b_k và cho bằng 0 thu được

K + 1 phương trình dạng

$$\sum_{k=1}^{K} b_k \sum_{n=1}^{N} (n\Delta t)^{k+m} = \sum_{n=1}^{N} u_n (n\Delta t)^m \quad m = 0,1,...K$$

Với các hệ số hồi quy $\{b_k\}$ phải được giải.

VD: K = 0, phương trình trên thành

$$b_0 \sum_{n=1}^{N} u_n (n\Delta t)^0 = \sum_{n=1}^{N} u_n (n\Delta t)^0 \Rightarrow b_0 = \frac{1}{N} \sum_{n=1}^{N} u_n = \bar{u}$$

Khử xu thế trong số liệu (trend removal)

VD: K = 1, phương trình trên thành

$$b_0\sum_{n=1}^{N}(n\Delta t)^m + b_1\sum_{n=1}^{N}(n\Delta t)^{1+m} = \sum_{n=1}^{N}u_n(n\Delta t)^m \quad m = 0,1$$

Chú ý: $\sum_{n=1}^{N} n = \frac{N(N+1)}{2}$ *và* $\sum_{n=1}^{N} n^2 = \frac{N(N+1)(2N+1)}{6}$

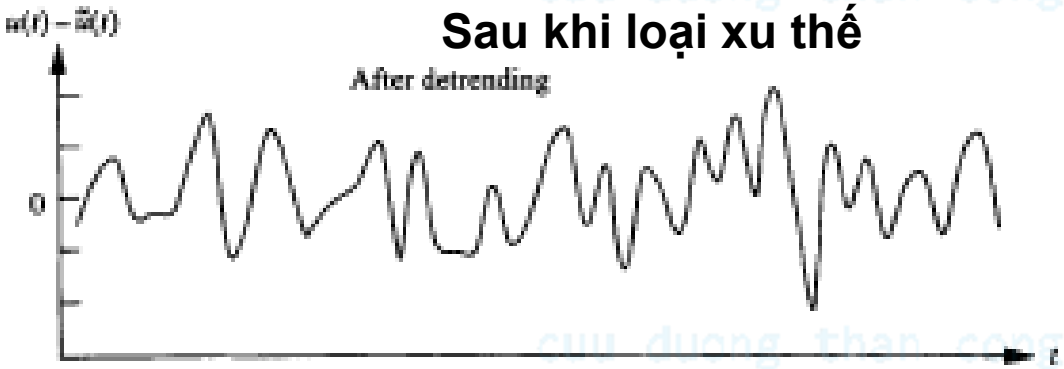
$$b_0 = \frac{2(2N+1)\sum_{n=1}^{N}u^n - 6\sum_{n=1}^{N}nu^n}{N(N-1)}$$

$$b_1 = \frac{12\sum_{n=1}^{N}nu^n - 6(N+1)\sum_{n=1}^{N}u^n}{\Delta tN(N-1)(N+1)}$$

Khử xu thế trong số liệu (trend removal)

Số liệu gốc

Sau khi loại xu thế



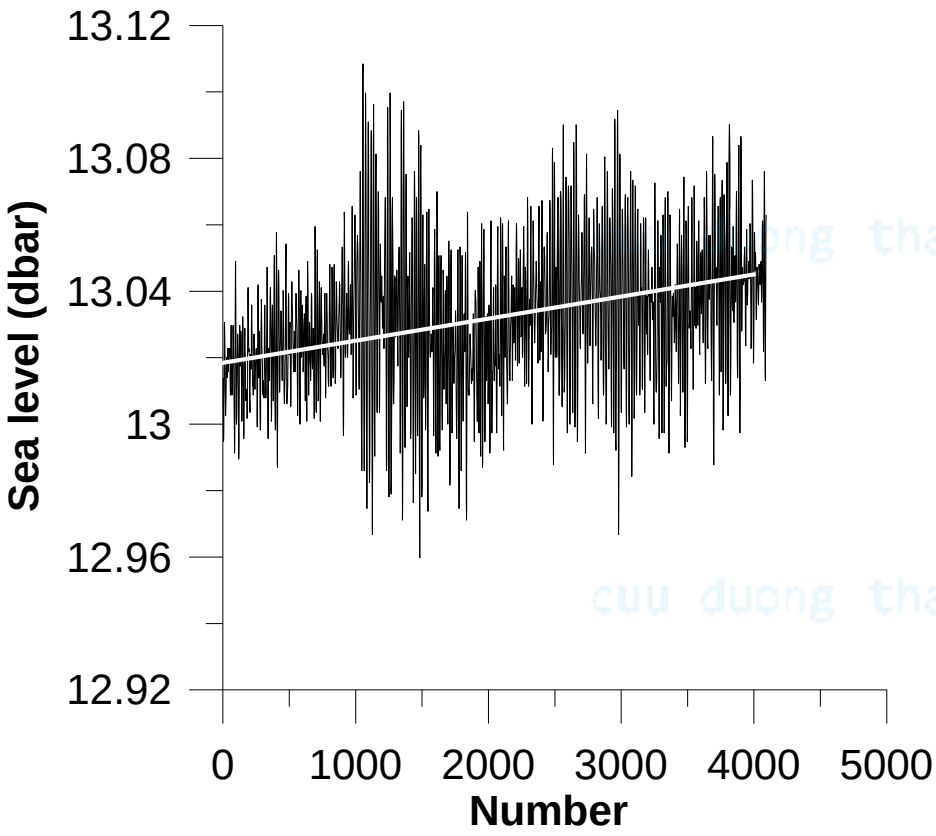
Xu thế tuyến tính

$$\tilde{u} = b_0 + b_1 t$$

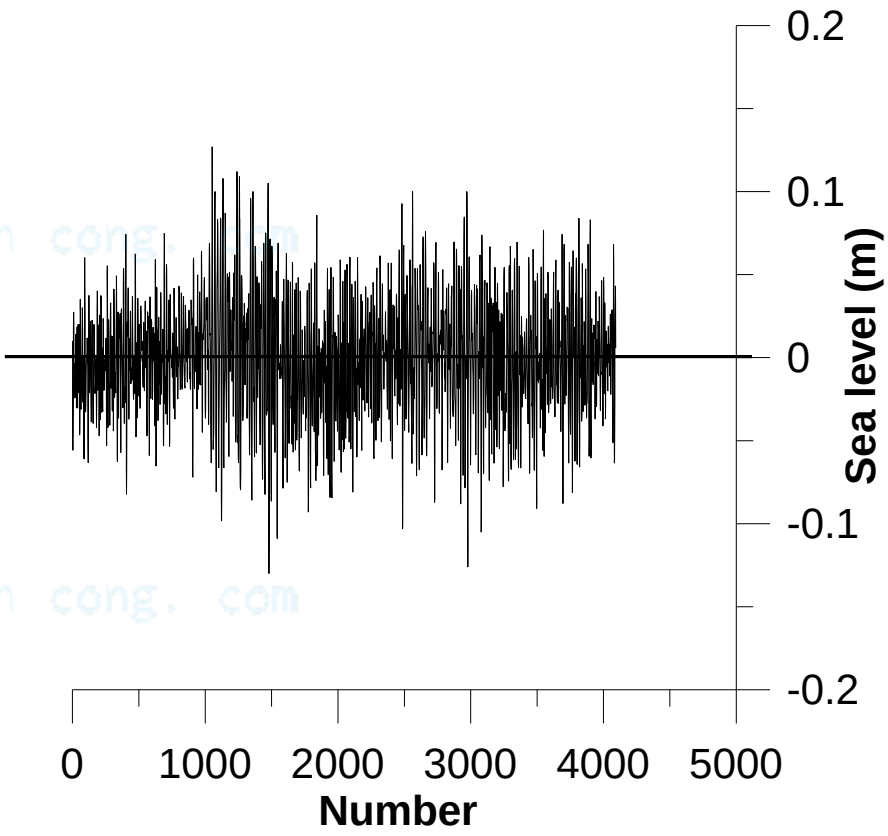
Bendat and Pierson (2010)

Khử xu thế trong số liệu (trend removal)

Ký đồ sóng trước khi khử triều



Ký đồ sóng sau khi khử triều



Nội suy số liệu (interpolation of data gaps)

Dữ liệu thực đo thường có những ‘gaps’ do

- Máy đo bị lỗi hoặc đo sai
- Cách đo đạc không đầy đủ
- Số liệu có chỗ không tin cậy

Các gaps này thường được nội suy trước khi đưa vào phân tích số liệu

Nội suy số liệu (interpolation of data gaps)

Trước khi chọn kỹ thuật nội suy các câu hỏi bên dưới nên được xác định:

- các số liệu nào được sử dụng để nội suy?
- các hàm nội suy nào nên dùng (hàm tuyến tính, đa thức bậc cao, spline bậc 3…)

Nội suy tuyến tính (linear interpolation)

- **Nội suy tuyến tính là dạng nội suy đường thẳng và được sử dụng rộng rãi để nội suy số liệu.**
- **Thủ tục này là 'fitting straight line' giữa hai số liệu.**
- **Cho chuỗi số liệu y(x), thủ tục nội suy như sau**

$$y(x) = y(a) + \frac{x-a}{b-a}\left[y(b) - y(a)\right]$$

$$= \frac{(b-x)y(a) + (x-a)y(b)}{b-a}$$

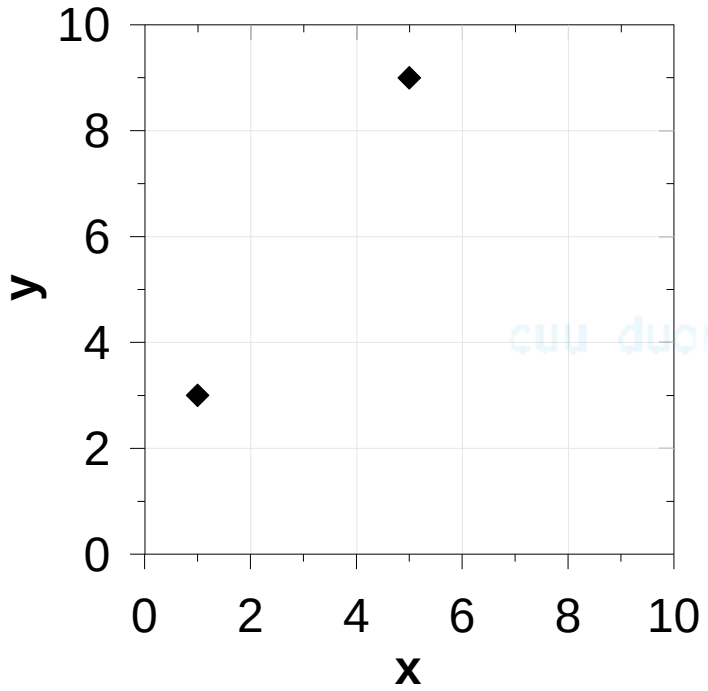
$x_{start} = a$

$x_{end} = b$

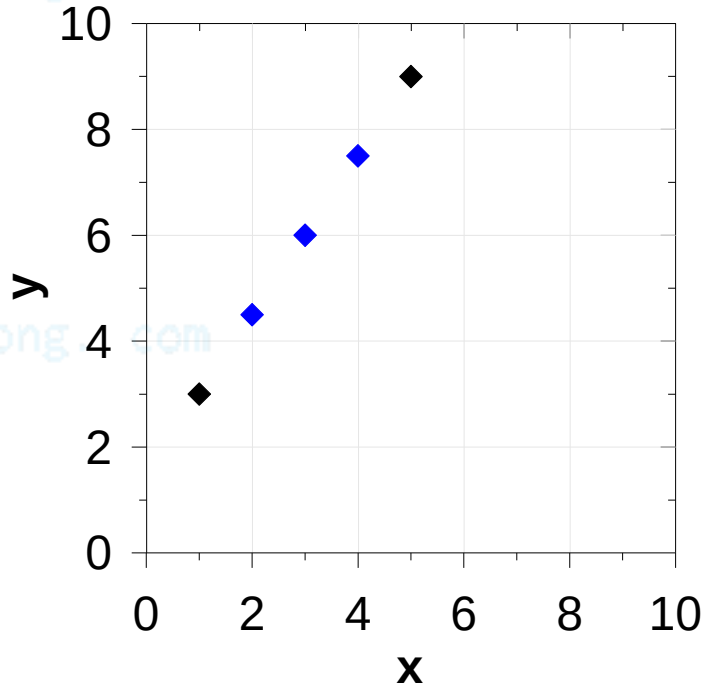
Nội suy tuyến tính: Ví dụ

Vd: có hai điểm $x_a = 1, y_a = 3$; $x_b = 5, y_b = 9$

Nội suy số liệu y tại x = 2, 3, 4



Nội suy tuyến tính thu được
$y_2 = 4.5, y_3 = 6, y_4 = 7.5$

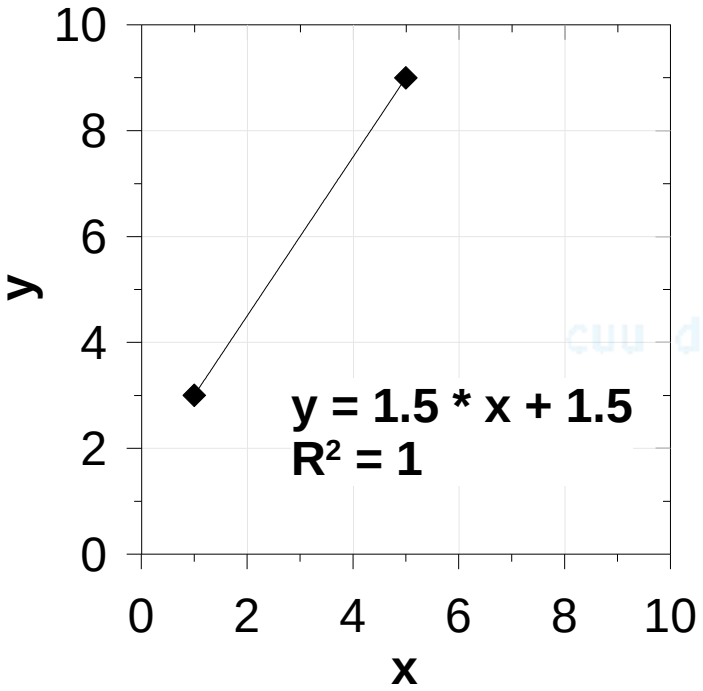


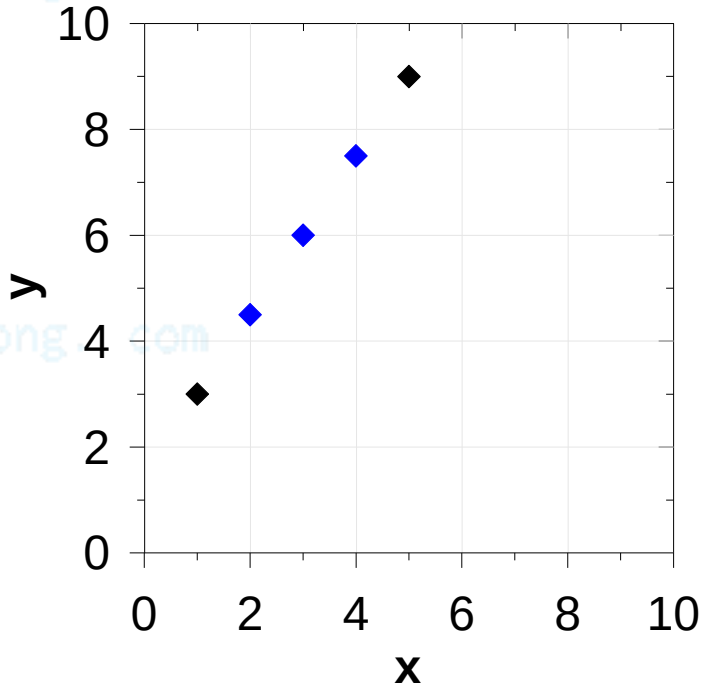
Nội suy tuyến tính: Ví dụ

Vd: có hai điểm $x_a = 1, y_a = 3$; $x_b = 5, y_b = 9$

Nội suy số liệu y tại x = 2, 3, 4

Dùng phương trình fit 2 điểm trên:
$y_2 = 4.5, y_3 = 6, y_4 = 7.5$





Nội suy đa thức (Polynomial interpolation)

Khi chúng ta muốn nội suy số liệu hơn 2 điểm, ta cần dùng đa thức nội suy bậc cao ví dụ đa thức Lagrange

Đa thức nội suy Lagrange dùng để

- Tìm một đa thức nội suy y(x) bậc N, sao cho

- Đi qua tất cả các điểm đã có (x_i, y_i), i = 1, 2, N+1

- Đa thức này có dạng tổng quát

$$y(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_N x^N = \sum_{k=0}^{N} a_k x^k$$

Nội suy đa thức (Polynomial interpolation)

Dưới dạng tổng quát

$$y(x) = \sum_{k=1}^{N+1} \left[y_i \left(\prod_{\substack{k=1 \\ k \neq i}}^{N+1} \frac{x - x_k}{x_i - x_k} \right) \right]$$

Mục tiêu của nội suy bằng đa thức Lagrange

- **Tìm một đa thức bậc N sao cho**

- **Đa thức này bắt buộc phải đi qua N + 1 số liệu đã có**

- **Nội suy bất kỳ điểm x nào nằm giữa các điểm đã có**

Nội suy đa thức (Polynomial interpolation)

Đa thức trên biến đổi thành

$$y(x) = \sum_{i=1}^{N+1} y_i \left[Q_i(x) / Q_i(x_i) \right]$$

$$Q_i(x) = (x - x_1)(x - x_2)...(x - x_{i-1})(x - x_{i+1})...(x - x_{N+1})$$

Đối với bất kỳ x, phương trình trên có thể biểu diễn

$$y(x) = y_1 \frac{(x - x_2)(x - x_3)...(x - x_{N+1})}{(x_1 - x_2)(x_1 - x_3)...(x_1 - x_{N+1})} +$$

$$y_2 \frac{(x - x_1)(x - x_3)...(x - x_{N+1})}{(x_2 - x_1)(x_2 - x_3)...(x_2 - x_{N+1})} +$$

$$+ ... \quad y_N \frac{(x - x_1)(x - x_2)...(x - x_N)}{(x_{N+1} - x_1)(x_{N+1} - x_3)...(x_{N+1} - x_N)}$$

Nội suy đa thức (Polynomial interpolation)

Vd: cho tập hợp 4 điểm (x_i, y_i), i = 1,…, 4

Có các số liệu như sau: (0, 2), (1, 2), (2, 0), (3, 0)

Nội suy bằng PP Lagrange với đa thức bậc N = 3

Phương trình cuối biểu diễn như sau

$$y(x) = 2\frac{(x-1)(x-2)(x-3)}{(0-1)(0-2)(0-3)} + 2\frac{(x-0)(x-2)(x-3)}{(1-0)(1-2)(1-3)} + 0 + 0$$

$$y(x) = \frac{2}{3}x^3 - 3x^2 + \frac{7}{3}x + 2$$

Bộ lọc số: digital filters

Đôi khi chúng ta cũng áp dụng các bộ lọc số nhằm:

Loại bỏ các sóng / dao động có chu kỳ cao hơn

Bộ lọc thông thấp (Lowpass filter)

Loại bỏ các sóng / dao động có chu kỳ thấp hơn

Bộ lọc thông cao (Highpass filter)

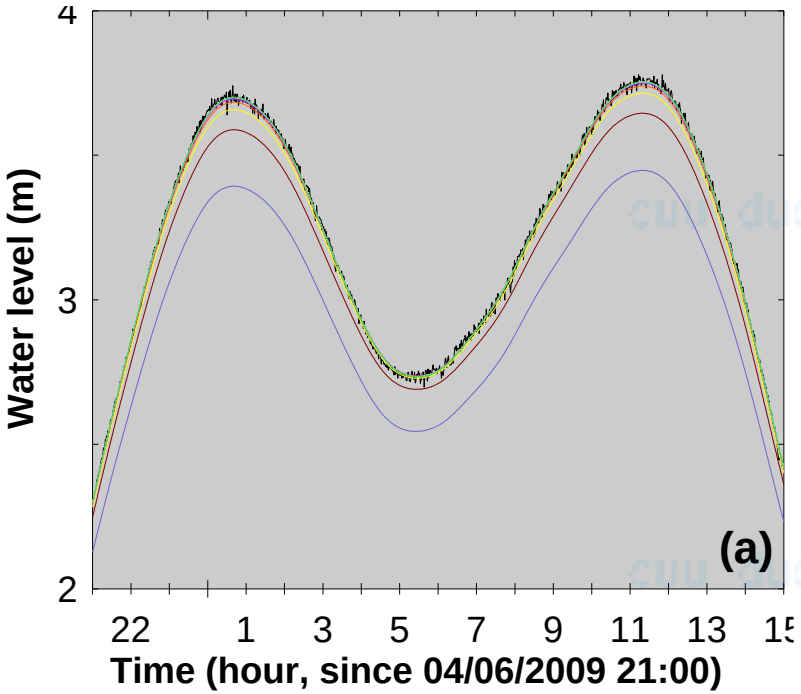
Loại bỏ các sóng / dao động trong một dải chu kỳ

Bộ lọc băng thông [thông dải] (Bandpass filter)

Bộ lọc số: digital filters

Số liệu thủy triều thực đo trong thời gian 1 tháng

Với các số liệu được lọc bằng bộ lọc Buterworth thông thấp



Gồm các dao động tần số cao như sóng gió, sóng tàu…

- Có thể dùng bộ lọc thông thấp để loại bỏ các nhiễu này

- Cũng có thể dùng bộ lọc băng thông để loại thêm các dao động tầng số dài hơn thời gian đo

Hồi qui tuyến tính đơn giản

Tài liệu Đọc Thêm

- Nguyễn Văn Tuấn, 2015. Phân tích dữ liệu với R. NXB Tổng hợp TP. Hồ Chí Minh.

Hồi qui tuyến tính đơn giản

Ví dụ:

Ta có chuỗi số liệu để hiệu chỉnh OBS

Tìm mối tương quan tuyến tính giữa số liệu mẫu nước và số liệu thô của máy đo độ đục

No	Water sample (mg/l)	OBSdata (FTU)	No	Water sample (mg/l)	OBSdata (FTU)
1	70	166.2	21	82.5	130.9
2	113.5	171.3	22	83.7	132.2
3	156.5	250.3	23	143	217.4
4	50.7	132.3	24	148.5	251.8
5	87	168.3	25	51.3	83
6	306	446	26	41.8	79.6
7	176.5	248.5	27	22.2	61.9
8	349	555.2	28	73	140.2
9	179	286	29	92.3	184.4
10	244	309.1	30	38.7	100.2
11	49.3	123.1	31	55.8	116.7
12	29.8	63.6	32	141	222
13	126	187.5	33	35	70.1
14	44	102.8	34	117	216.1
15	60.7	124.2	35	293	480.3
16	215	329.2	36	65.7	122.8
17	27.3	75	37	75	120.2
18	159	201.6	38	105.5	178.2
19	357.5	470.3	39	43	99.7
20	148	252.3	40	25	49.7

Kỹ thuật chuẩn máy đo độ đục bằng các số liệu độ đục thu thập tại hiện nơi khảo sát (các mẫu nước)

Hồi qui tuyến tính đơn giản

Mô hình hồi qui tuyến tính:

$$y_i = \alpha + \beta x_i + \varepsilon_i ; i = 1,2,...n$$

Phương trình này mô tả:

- **Số liệu máy đo độ đục có liên quan với số liệu mẫu nước**
- **Liên quan này thông qua hằng số α và hệ số β và sai số ε**

- **α: chặn (intercept)**
- **β: độ dốc**

- **α, β: là 2 hệ số hồi qui**
- **ε ~ luật phân phối chuẩn, với mean = 0**

Hồi qui tuyến tính đơn giản

Ước lượng của phương trình tuyến tính trên

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

các ước số α β được tính

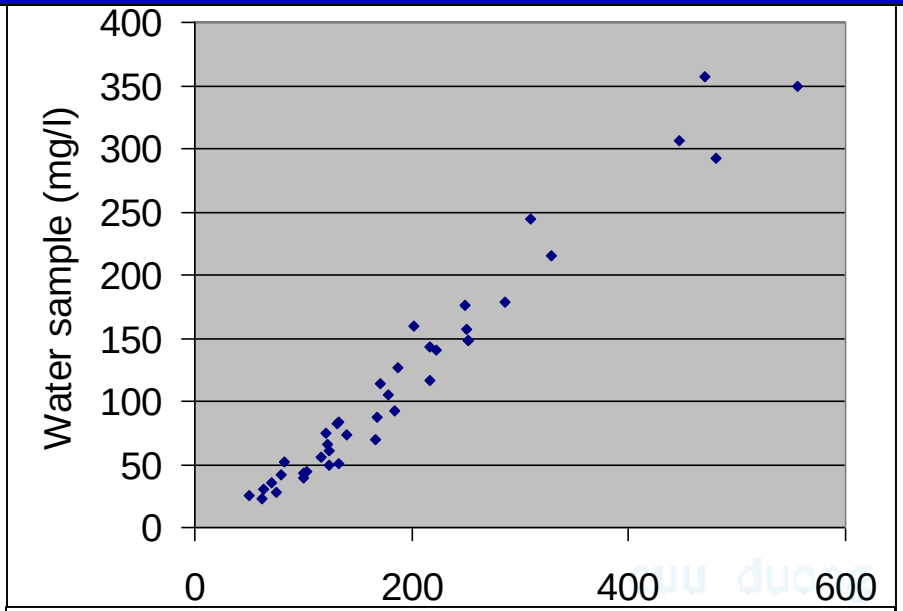
và

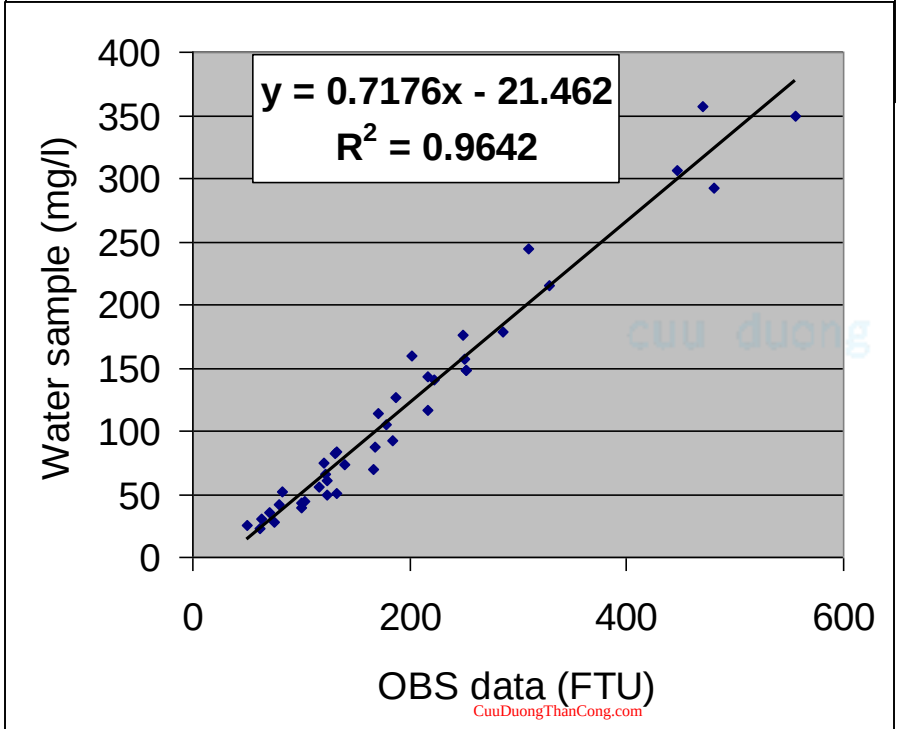
$$\left.\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\,\bar{x} \end{aligned}\right\} \Rightarrow \begin{cases} R^2 = \dfrac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}}_i)^2}{\sum_{i}^{n}(y_i - \bar{y})^2} \\ \varepsilon_i = y_i - \hat{y}_i \end{cases}$$

Tổng bình phương phần dư (error sum of squared or residual sum of squared)

$$SSE = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = \sum_{i=1}^{n}\varepsilon_i^2$$

Hồi qui tuyến tính đơn giản





No	Water sample (mg/l)	OBSdata (FTU)	No	Water sample (mg/l)	OBSdata (FTU)
1	70	166.2	21	82.5	130.9
2	113.5	171.3	22	83.7	132.2
3	156.5	250.3	23	143	217.4
4	50.7	132.3	24	148.5	251.8
5	87	168.3	25	51.3	83
6	306	446	26	41.8	79.6
7	176.5	248.5	27	22.2	61.9
8	349	555.2	28	73	140.2
9	179	286	29	92.3	184.4
10	244	309.1	30	38.7	100.2
11	49.3	123.1	31	55.8	116.7
12	29.8	63.6	32	141	222
13	126	187.5	33	35	70.1
14	44	102.8	34	117	216.1
15	60.7	124.2	35	293	480.3
16	215	329.2	36	65.7	122.8
17	27.3	75	37	75	120.2
18	159	201.6	38	105.5	178.2
19	357.5	470.3	39	43	99.7
20	148	252.3	40	25	49.7

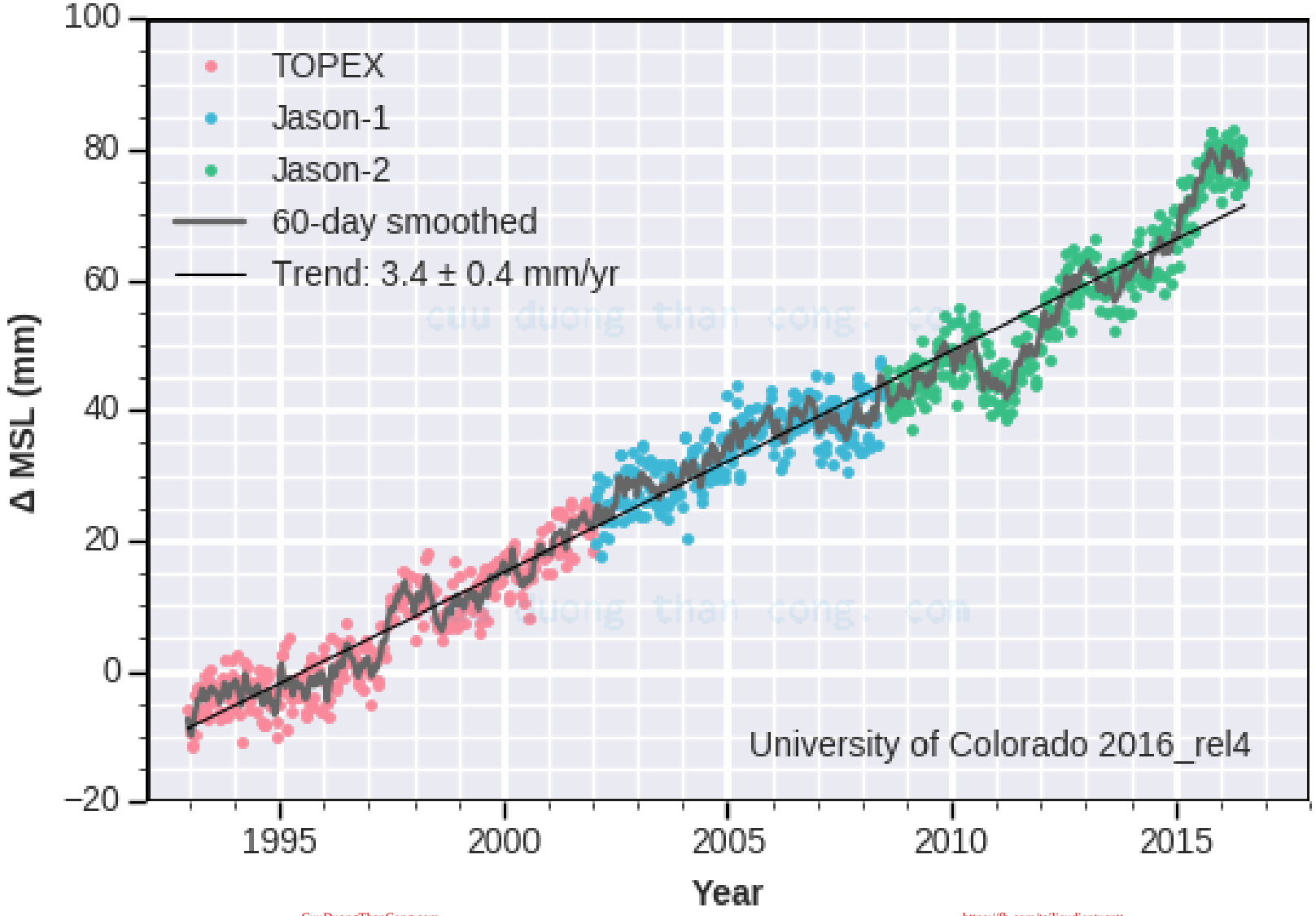
$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

$$= 0.72 * X - 21.5$$

Hồi qui tuyến tính đơn giản

Ví dụ: tốc độ mực nước biển dâng toàn cầu từ ảnh vệ tinh

(nguồn University of Colorado)



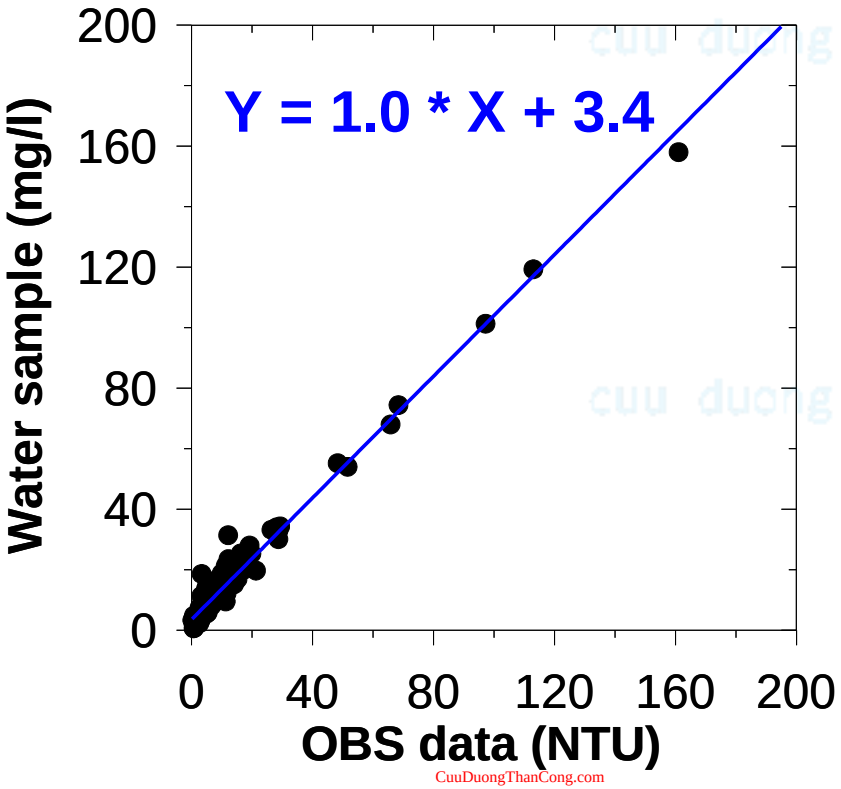
Hồi qui tuyến tính đơn giản: Vấn đề chuẩn hóa thiết bị đo

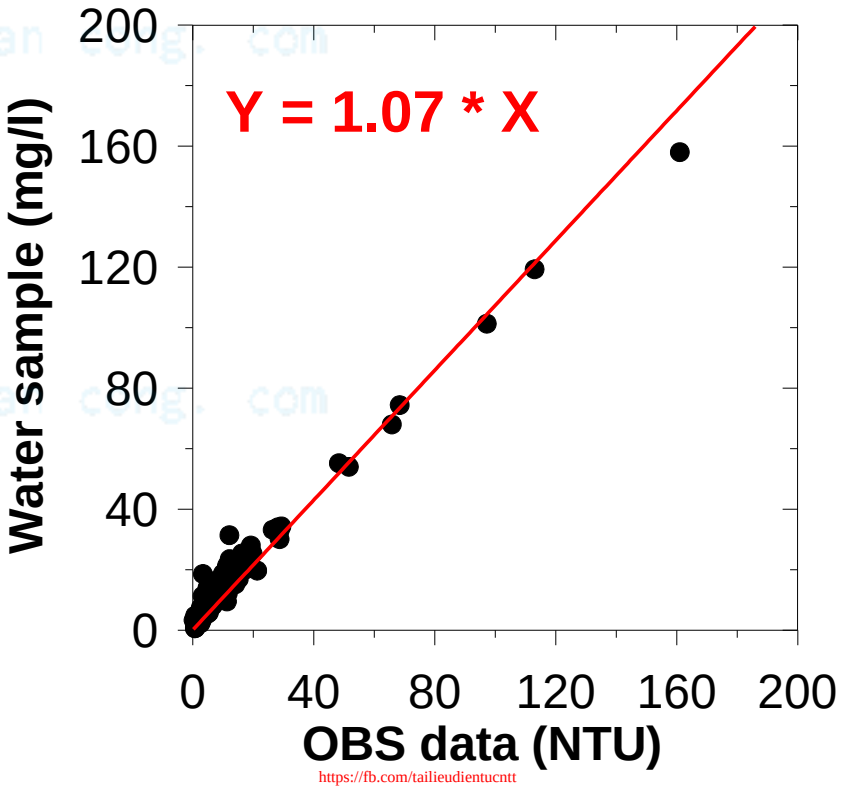
Trong phương tương quan

$$y = \alpha + \beta x$$

Hệ số α nên được chọn ra sao?

$$\begin{cases} \alpha = 0 \\ \alpha \neq 0 \end{cases}$$





Hồi qui tuyến tính đơn giản: Vấn đề chuẩn hóa thiết bị đo

Trong phương tương quan $y = \alpha + \beta x$

Hệ số α nên được chọn ra sao?

$$\begin{cases} \alpha = 0 \\ \alpha \neq 0 \end{cases}$$