4. 相関・回帰

- 4.0 相関関係とは？ Correlation？
- 4.1 相関係数 Correlation coefficient
- 4.2 自己相関 Auto-correlation
- 4.3 相互相関 Cross-correlation
- 4.4 相関解析の実例 examples
- 4.5 相関の有意性
  - 相関係数の検定 test of correlation coef.
  - 自由度の見積もり effective number of DOF
- 4.6 回帰 regressions

4.1 相関係数 Correlation coefficient

二変数を $x_i, y_i (i = 1, \cdots, n)$ とすると、その関係の度合いを定量的に表したもの、相関係数 (correlation coefficient) r は次のように定義される。

$$r = \frac{\sum_{i=1}^{n} x'_i y'_i}{\sqrt{\sum x'^2 \sum y'^2}}$$

$$x'_i = x_i - \overline{x}$$

共分散

分散

例えば

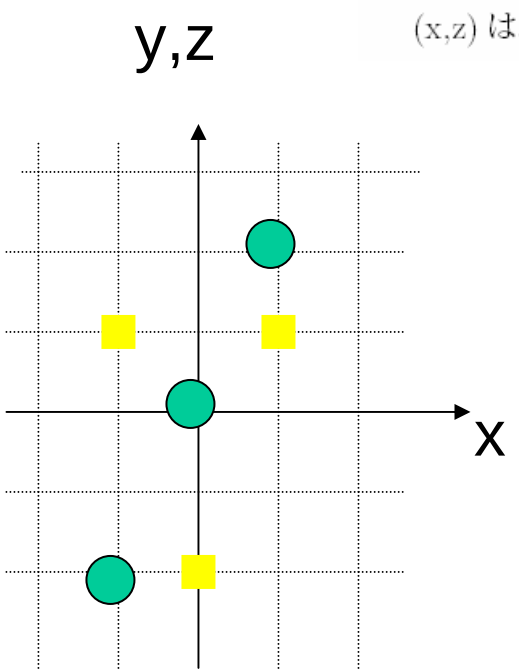
x= [-1, 0, 1]
y= [-2, 0, 2]
z= [1, -2, 1]

$$r_{xy} = \frac{(-1\cdot -2)+0\cdot 0+1\cdot 2}{\sqrt{(-1\cdot -1)+0\cdot 0+(1\cdot 1)}\sqrt{(-2\cdot -2)+0\cdot +(2\cdot 2)}} = \frac{4}{\sqrt{2}\sqrt{8}} = 1$$

x と y はどんな値・単位でも、傾向が似てさえいれば相関が高い。

$$r_{xz} = \frac{(-1\cdot 1)+0\cdot (-2)+1\cdot 1}{\sqrt{(-1\cdot -1)+0\cdot 0+(1\cdot 1)}\sqrt{(1\cdot 1)+(-2\cdot 2)+(1\cdot 1)}} = \frac{0}{\sqrt{2}\sqrt{6}} = 0$$

これを散布図上で見ると、(x,y) は右肩上がりの直線上にのっているが、(x,z) はバラバラに分布している。



相関係数と散布図

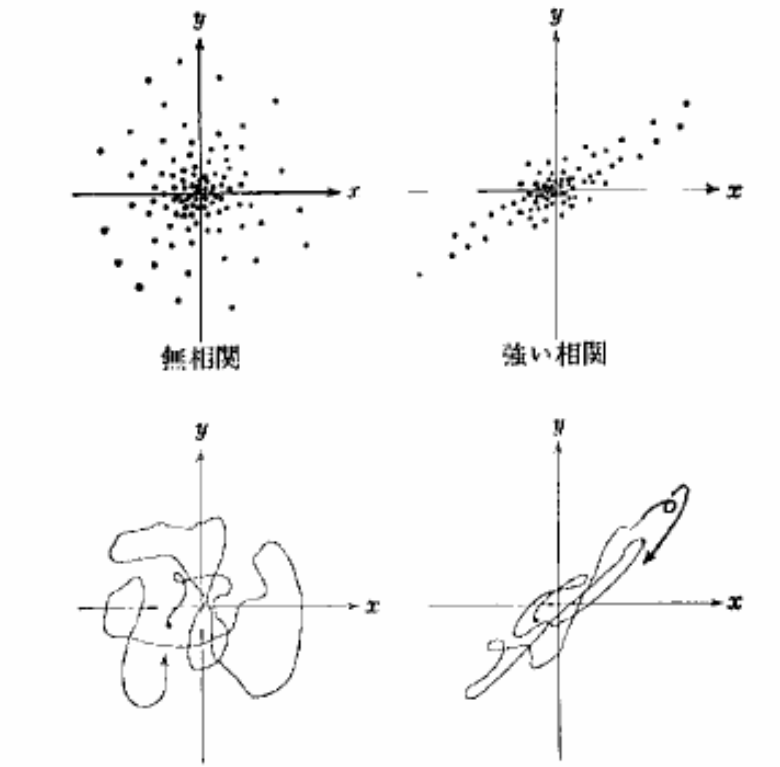


図 2.1　不規則変量 x, y の相関

- $r > 0$ 正の相関
- $r \simeq 0$ 無相関
- $r < 0$ 負の相関

4.2 自己相関関数 （auto-correlation function）

$$C(t,\tau) = E[x(t)x(t+\tau)]$$

アンサンブル平均

定常確率過程では時間平均で置き換えることができて

時間平均

$$C(\tau) = \overline{x(t)x(t+\tau)}$$

Covariance function

$$= lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} x(t)x(t+\tau)dt$$

$C(\tau)$ をラグ 0 の値でわって正規化したものを自己相関係数という。

τ ：lag

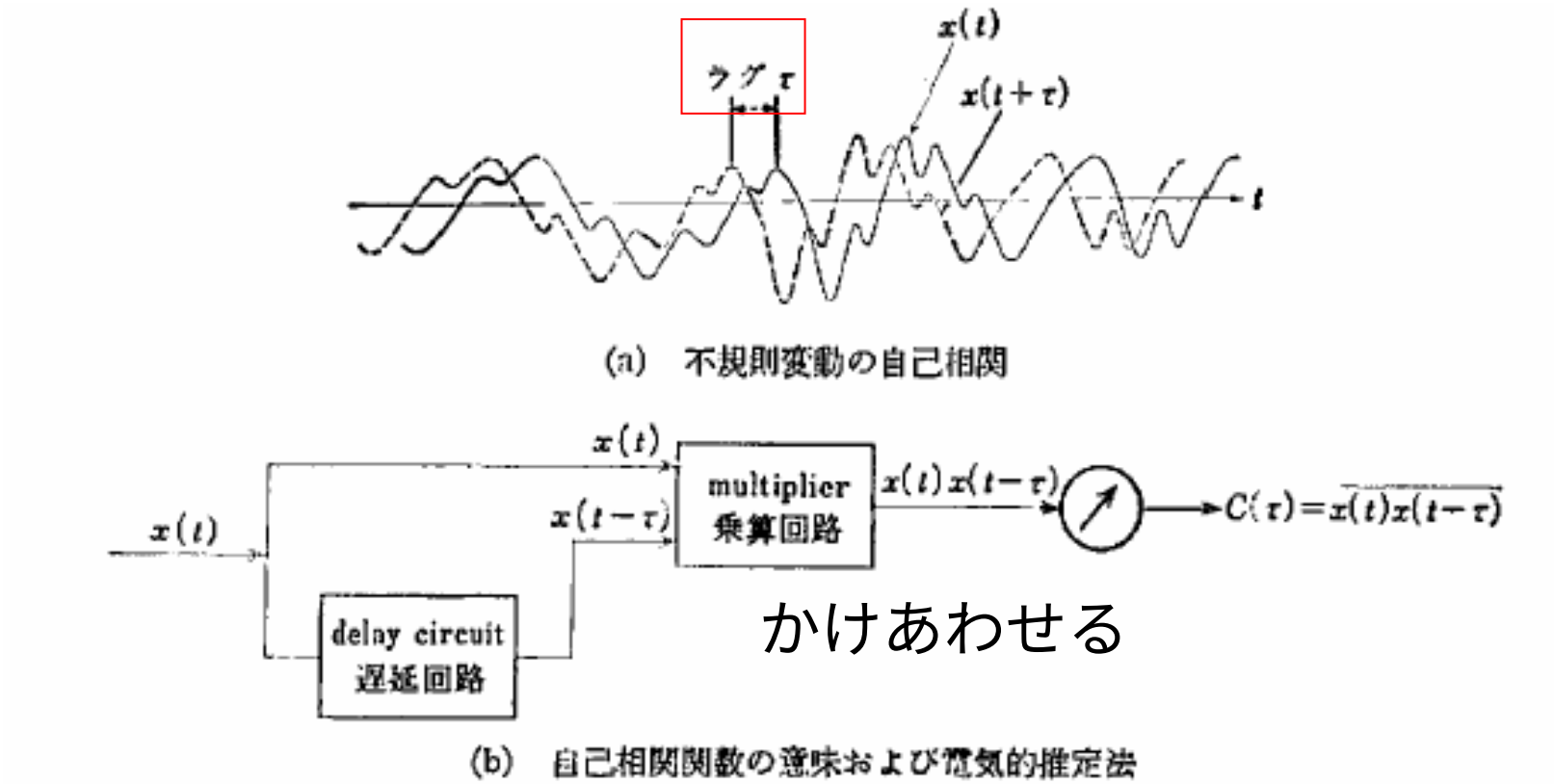
$$R(\tau) = \frac{C(\tau)}{C(0)}$$

Autocorrelation function

$$R(0) = 1$$

τ ずらしたものをyとする。
τ をどんどん変える。

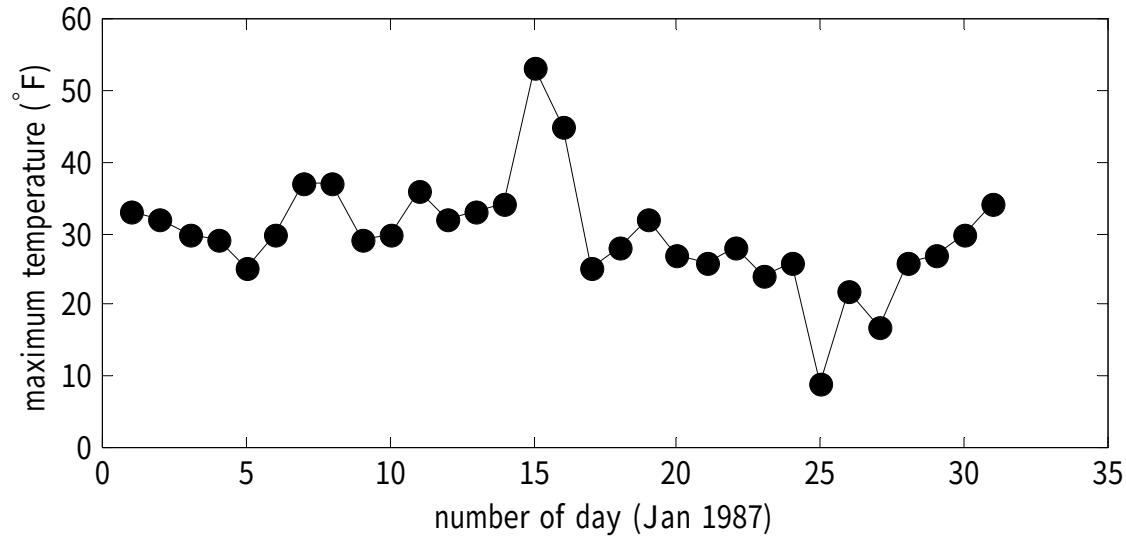
ずらす

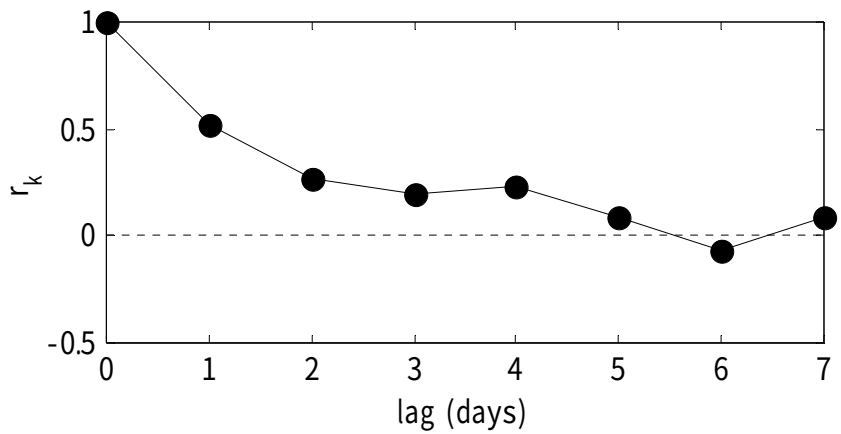


かけあわせる

ずらして

例題 イサカの1987年1月の日平均気温の自己相関係数





代表的な時系列関数と自己相関関数の形

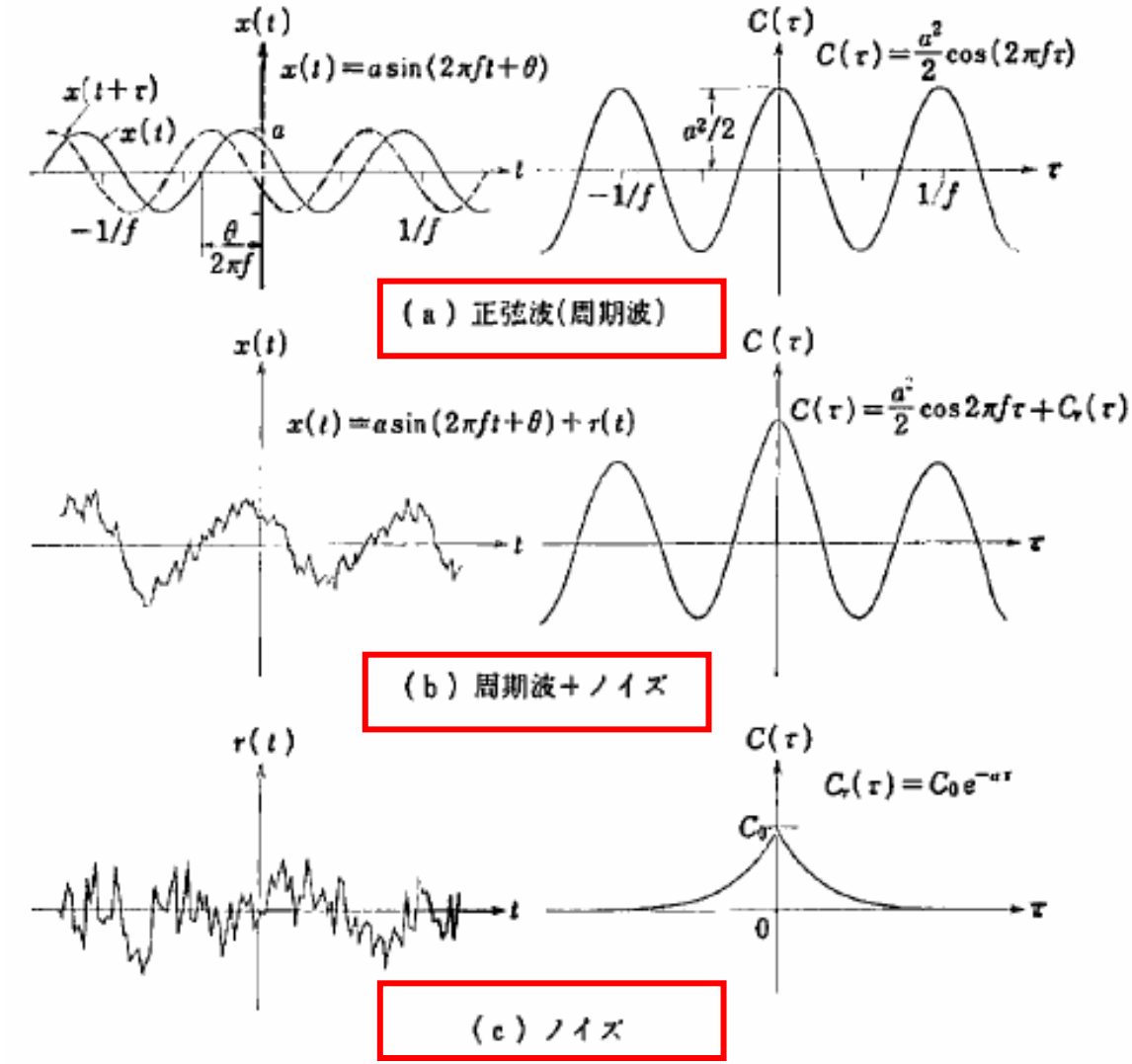


図 2.3 信号と自己相関関数

1) 正弦波の自己相関関数

$$x(t) = asin2\pi ft$$

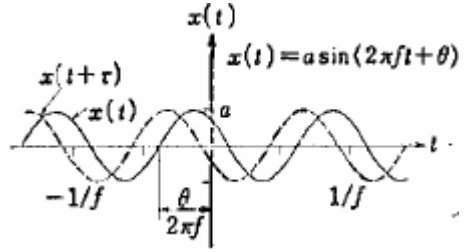
$$x(t+\tau) = asin2\pi f(t+\tau)$$

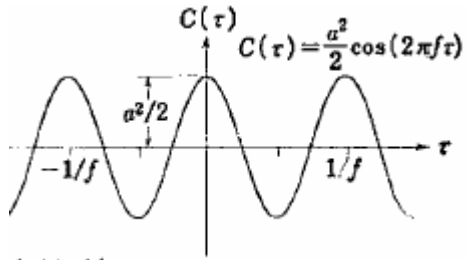
$$x(t)x(t+\tau) = \frac{a^2}{2}[cos2\pi f\tau - cos(2\pi f(2t+\tau))]$$

$$\int_0^{nT} x(t)x(t+\tau)dt = \frac{nTa^2}{2}cos2\pi f\tau$$

$$R(\tau) = \int_0^{nT} x(t)x(t+\tau)dt / \int_0^{nT} x(t)x(t)dt$$

$$= cos2\pi f\tau$$





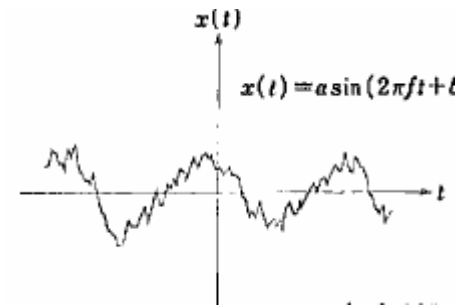
2) 正弦波+ノイズの自己相関関数

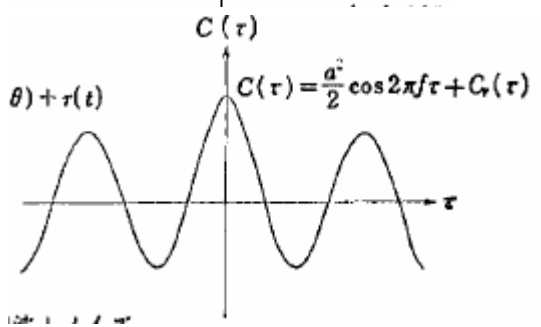
$$x(t) = asin2\pi ft + n(t)$$

$$C(\tau) = \frac{a^2}{2}cos2\pi f\tau + \phi(\tau)$$

$$\phi(\tau) = E[n(t)n(t+\tau)]$$

ここでは $E[x(t)n(t+\tau)] = 0$ を用いている。





3) ノイズ

3-i) 白色雑音（white noise）の自己相関関数

$$x(t) = n(t)$$

$$E[n(t)n(t+\tau)] = n_0^2\delta(\tau)$$

$\tau = 0$ の場合以外は、自己相関が 0 となる。

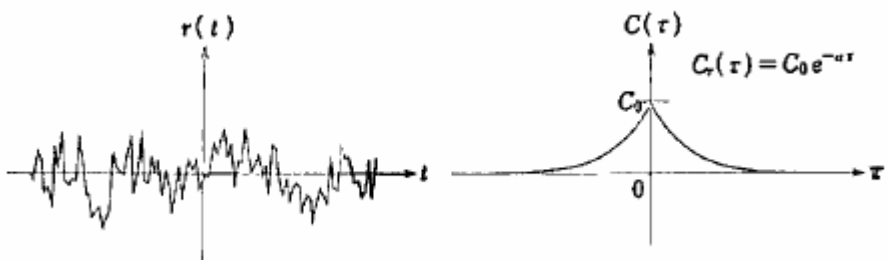
3-ii)AR(1) 過程の自己相関

white noise

$$x(t+\delta t) = rx(t) + n(t) \qquad (|r| < 1)$$

このような過程を自己回帰過程（AutoRegressive process: AR(1)）または一次のマルコフ過程あるいは赤色ノイズ（red noise）と呼ばれる[2]。

$$C(\tau) = C(0)e^{-\alpha\tau}$$



$$C(\tau) = E[x(t)x(t+\tau)]$$

$$C(\tau+\delta t) = E[x(t)x(t+\tau+\delta t)]$$

$$C(\tau+\delta t) = E[x(t)(rx(t+\tau)+n(t+\tau))]$$

$$= rE[x(t)x(t+\tau)] + E[x(t)n(t+\tau))]$$

$$= rC(\tau)$$

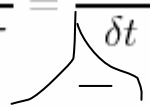
$$C(\tau+\delta t) = C(\tau) + \delta t\frac{dC}{d\tau} + O(\delta t^2)$$

$$r = 1 + \delta t \cdot \frac{1}{C}\frac{dC}{d\tau} + \frac{1}{C}O(\delta t^2)$$

2 次以上の項を無視して

$$\frac{1}{C}\frac{dC}{d\tau} = \frac{1-r}{\delta t} = -\alpha$$

上式を積分してもとまる。



4.3 相互相関関数 （cross-correlation function）

異なる変数間でのラグ相関を求める

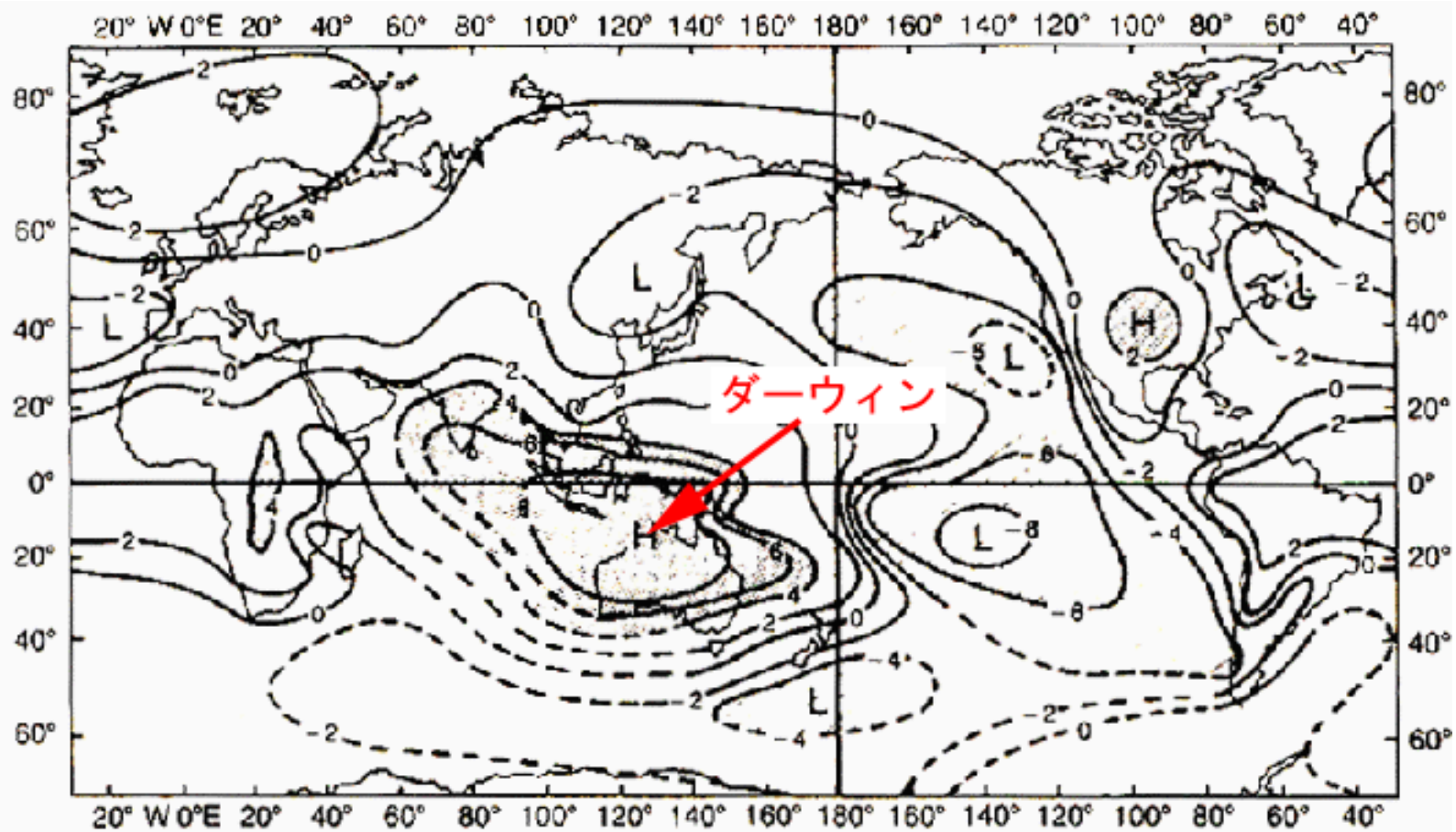
$$C_{xy}(\tau) = \overline{x(t)y(t+\tau)}$$

$$R_{xy}(\tau) = \frac{\overline{x(t)y(t+\tau)}}{\sqrt{\overline{x^2}}\sqrt{\overline{y^2}}}$$

$$= \frac{C_{xy}(\tau)}{\sqrt{C_x(0)C_y(0)}}$$

Rxy(0)=1にはならない

4.4 相関解析の実例　その1.南方振動

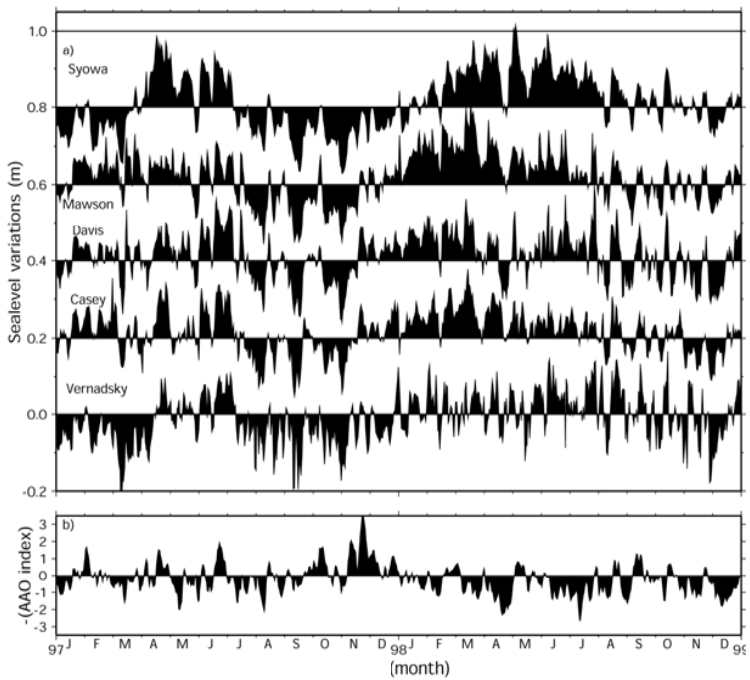


季節変化は落ちている

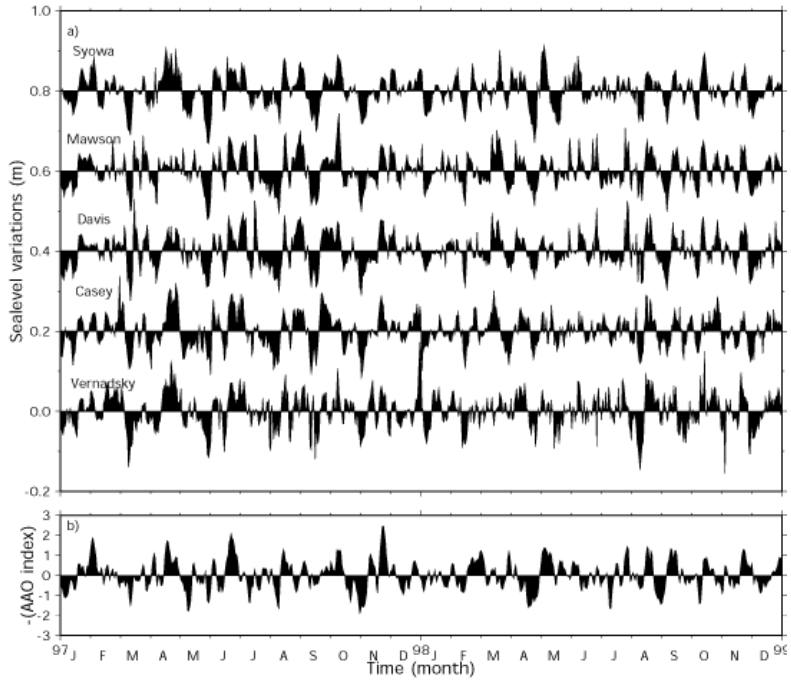
図c ダーウィンと世界各地の年平均海面気圧偏差の相関係数(x10)。
係数が正の値のところはダーウィンの気圧が通常より高いときにその場所の気圧も通常より高い傾向にあり、係数が負の値のところはダーウィンの気圧が通常より高いとき、逆に通常より低い傾向にある。数字の大きさがその傾向の程度を示す。(Trenberth and Shea,1987,Mon. Weather Rev.)

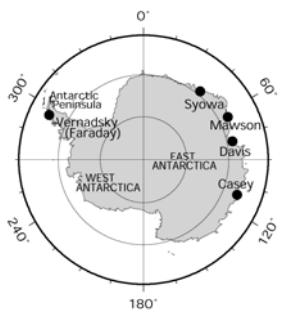
相関係数の例　その２．南極の水位の相関関係

Original



High-passed





	Mawson	Davis	Casey	Vernadsky
original	0.669	0.679	0.648	0.699
high-passed	0.685	0.638	0.634	0.611
high-ln.tide	0.618	0.564	0.557	0.548

ラグ相関解析の実例　ー　その3.擾乱の位相伝播

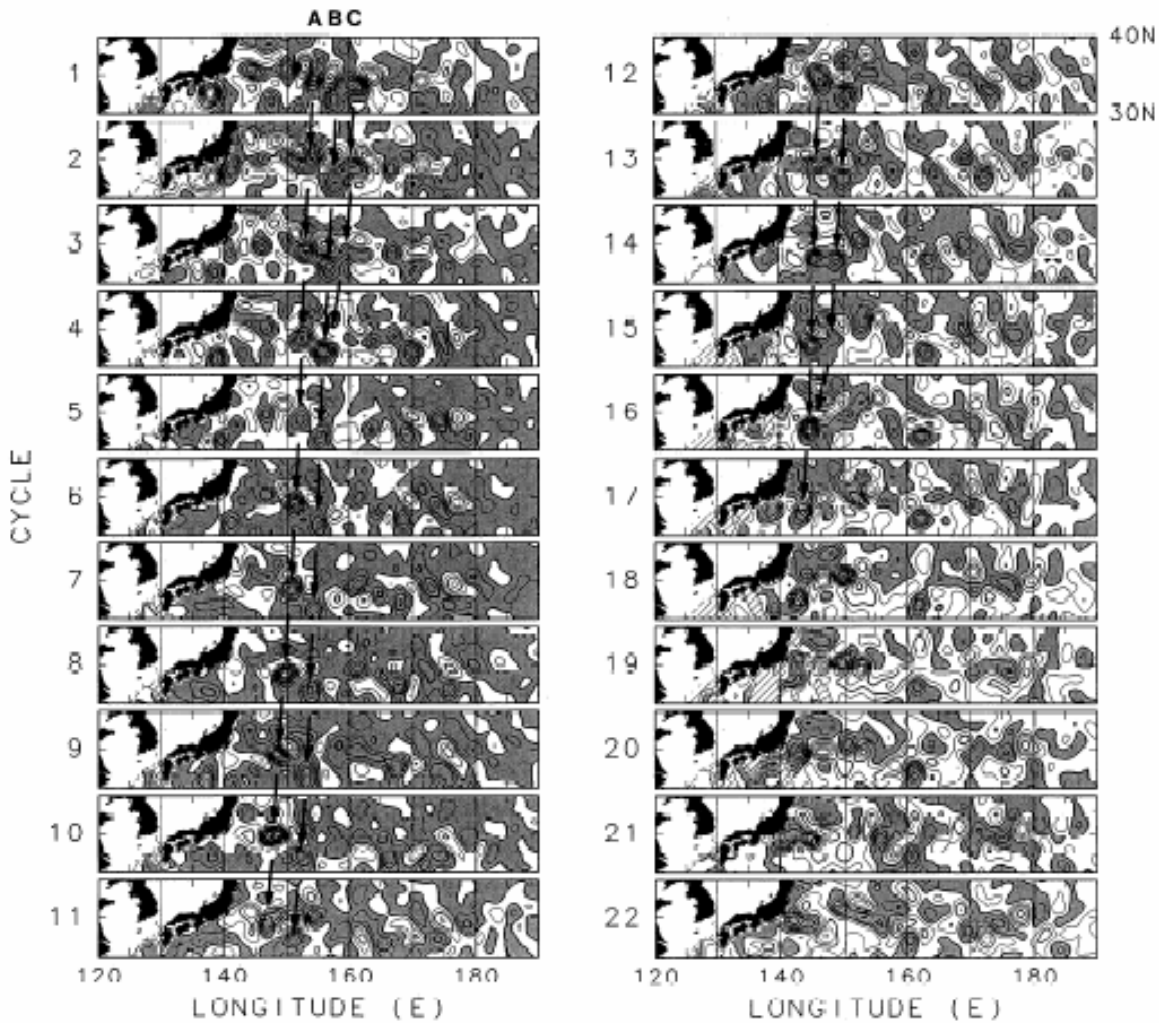


Figure 7. Time series of SSDT anomaly field in the Kuroshio and Kuroshio Extension regions (30°–40°N, 120°E–170°W) derived from Geosat altimetry data. Contour interval is 10 cm. Negative anomalies are shaded. Estimated error is large in hatched regions. The number on the left-hand side denotes the Geosat 17-day repeat cycle. Arrows show the movement of anomalies *A*, *B*, and *C* to guide the reader's eye.

空間構造
時間構造
伝播特性

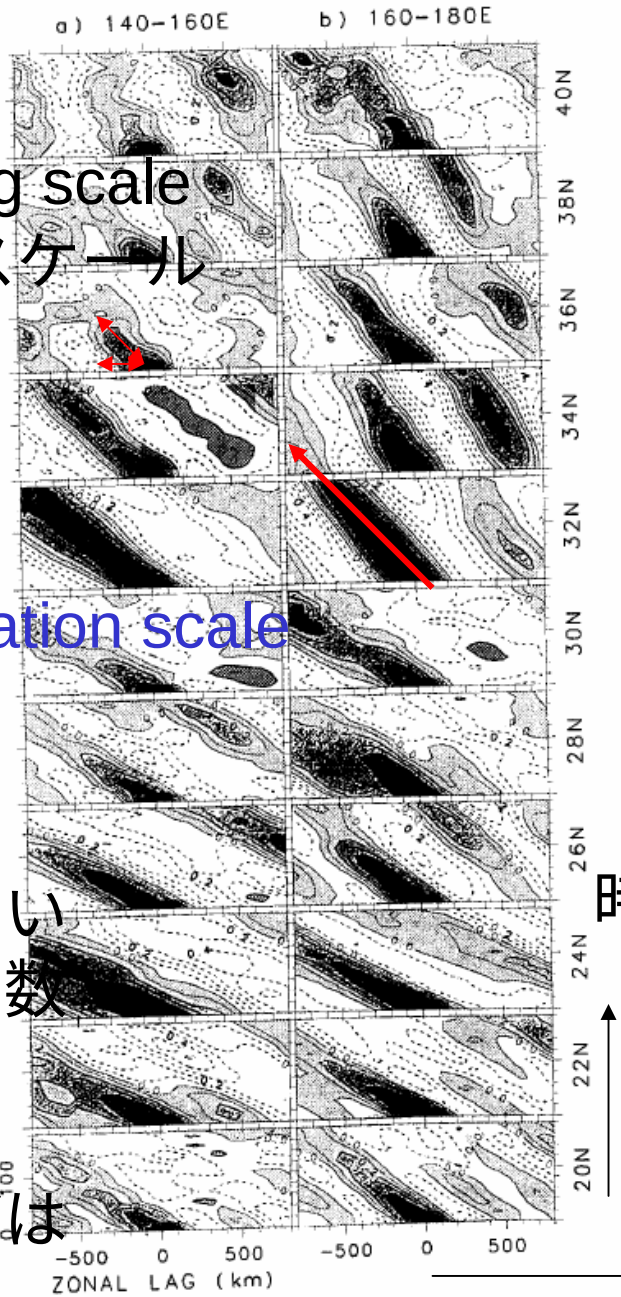
SSH

e-holding scale
無相関スケール

de-correlation scale

ラグが大きい
ときには個数
が少ない

大きなラグは
とれない



時間

空間

SST

周期性

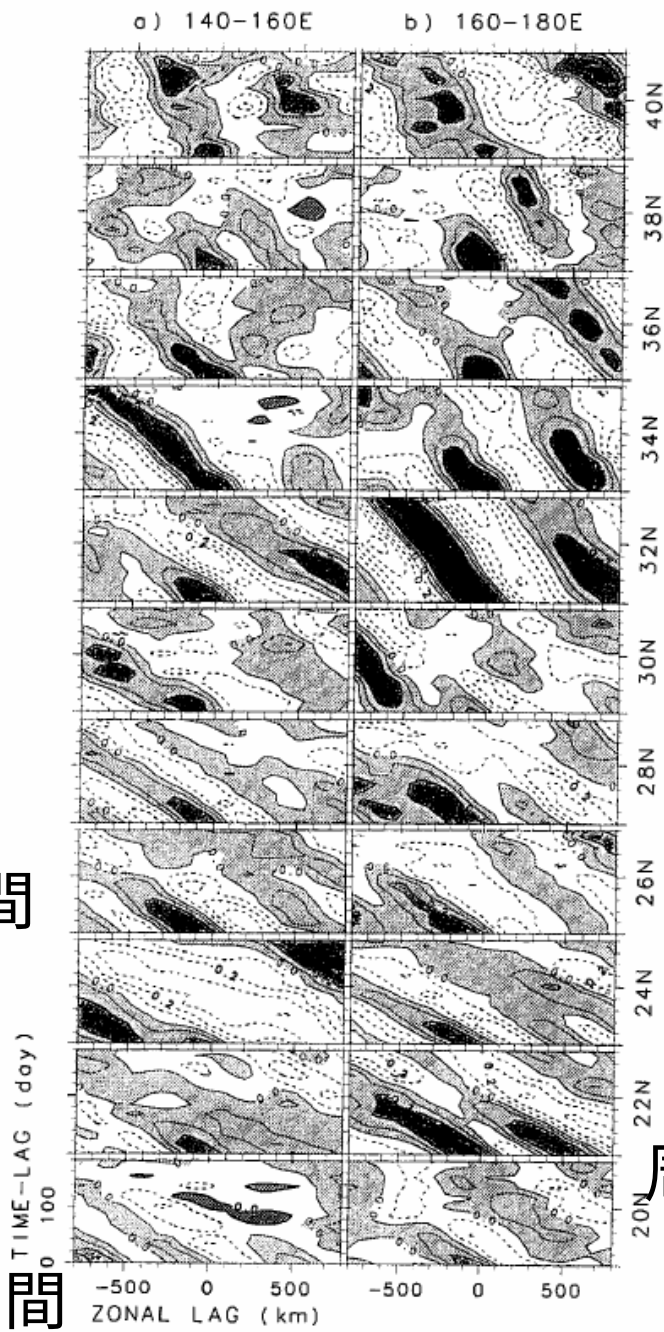
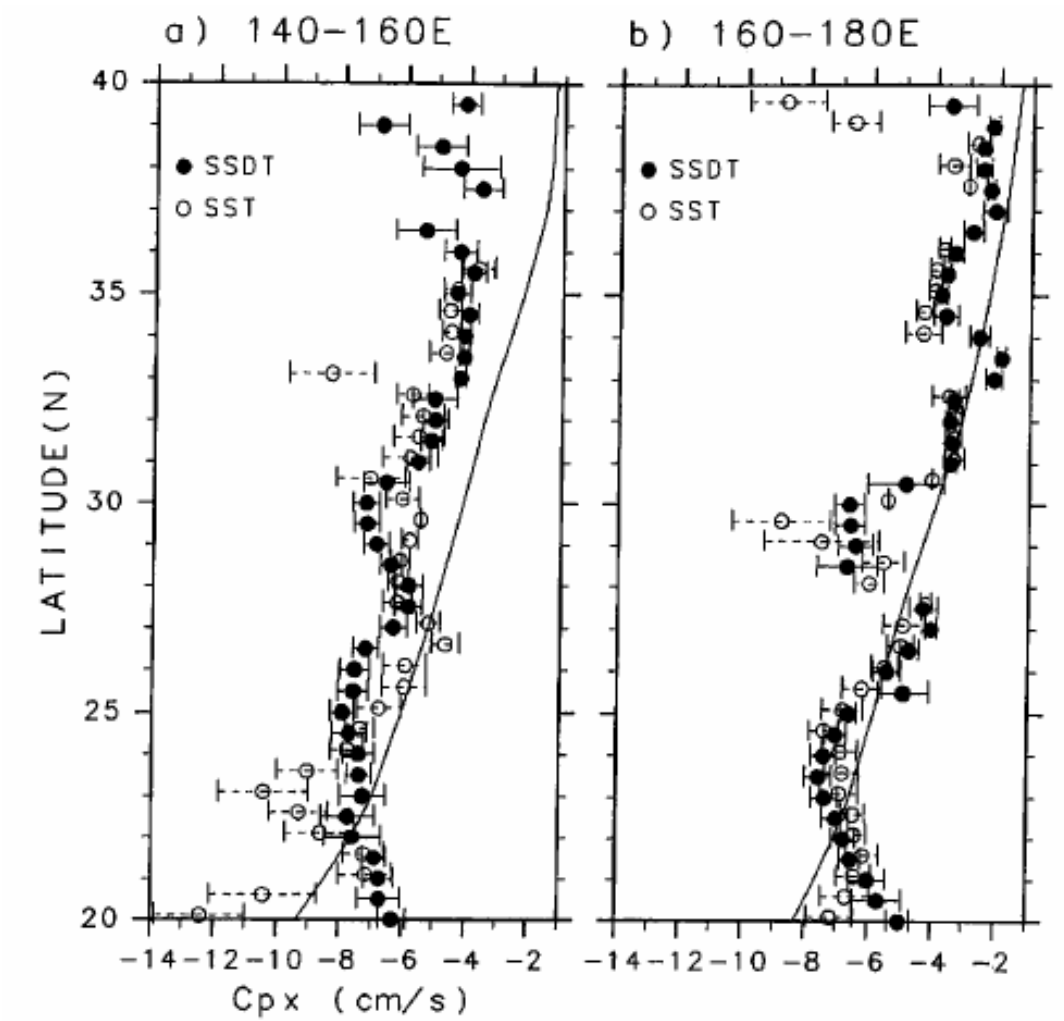


図4　中緯度域（20°-40°N）でのSSDTの擾乱の経度-時間ラグ相関図．a）は西海域（140°-160°E），b）は東海域（160°-180°E）

図6　図4に同じ．ただし，SSTとSSDTの相互ラグ相関のもの

位相速度 phase speed



4.5　相関の有意性

4.5.1 相関係数の検定　test of correlation coefficient

母相関係数の検定

　母相関係数ρ＝０のときは、標本数ｎの相関係数ｒは次のtについて、(近似的に)自由度n-2のｔ分布に従うことが知られている。

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

母相関係数に関する検定は一般に母相関係数ρ＝０という帰無仮説を検定する。したがって、上の式のｔを求めてｔ検定すればよい。

（面倒な計算をしなくてもよいように検定の表がある）。

$$r = \frac{T}{\sqrt{N-2+T^2}}$$

APPENDIX E: CORRELATION COEFFICENTS AT THE 5% AND 1% LEVELS OF SIGNIFICANCE FOR VARIOUS DEGREES OF FREEDOM ν

Degrees of freedom	5%	1%	Degrees of freedom	5%	1%
1	0.997	1.000	24	0.388	0.496
2	0.950	0.990	25	0.381	0.487
3	0.878	0.959	26	0.374	0.478
4	0.811	0.917	27	0.367	0.470
5	0.754	0.874	28	0.361	0.463
6	0.707	0.834	29	0.355	0.456
7	0.666	0.798	30	0.349	0.449
8	0.632	0.765	35	0.325	0.418
9	0.602	0.735	40	0.304	0.393
10	0.576	0.708	45	0.288	0.372
11	0.553	0.684	50	0.273	0.354
12	0.532	0.661	60	0.250	0.325
13	0.514	0.641	70	0.232	0.302
14	0.497	0.623	80	0.217	0.283
15	0.482	0.606	90	0.205	0.267
16	0.468	0.590	100	0.195	0.254
17	0.456	0.576	125	0.174	0.228
18	0.444	0.561	150	0.159	0.208
19	0.433	0.549	200	0.138	0.181
20	0.423	0.537	300	0.113	0.148
21	0.413	0.526	400	0.098	0.128
22	0.404	0.515	500	0.088	0.115
23	0.396	0.505	1000	0.062	0.081

両側確率 (two-sided probability)

サンプル数ｎ（自由度ｆ＝ｎ-２）のときに標本の相関係数が表の値よりも大きければ、母相関係数ρ＝０という帰無仮説が棄却され、有意な相関があるといえる。

例）サンプル数10（自由度８）だと、標本の相関係数が0.632以上ならば5%の有意水準で母相関係数は０でなく、0.765以上ならば1%の有意水準で母相関係数は０ではない。

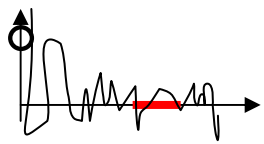
サンプル数ｎ	自由度ｆ	両側確率.05	両側確率.01
10	8	.63190	.76459

◯相関係数の検定はあくまでも母相関係数が０でない（すなわち相関が弱いとしてもある）ことを判断するだけで、帰無仮説が棄却されたからといって「相関が強い」わけではない。一方、相関係数が大きくても、サンプル数が少なければ、検定の結果、相関があるとはいえないこともある。

4.5.2 等価自由度
effective degree of freedom

- 大気海洋データは、時・空間的に相関をもっているため
「ν(自由度)＝N（データ数）」
にはならない。
- 時系列がランダムである場合は自由度ν＝Nでよいが、特定の狭帯域波や長周期波が含まれている場合には自由度は著しく下がる。

- 例えば三角関数は振幅と位相で決まってしまうので、自由度は２しかない。



- 等価自由度の推定
  - （ある狭帯域シグナルがある場合)その５－６倍の間のラグでのラグ相関のRMSをとり、その二乗の逆数をもって等価自由度とする（Davis 1976, 77; Chelton,1982）。データの長さを対象とする現象のスケールで割る（松山・谷本,2005）。

低い係数
でも有意

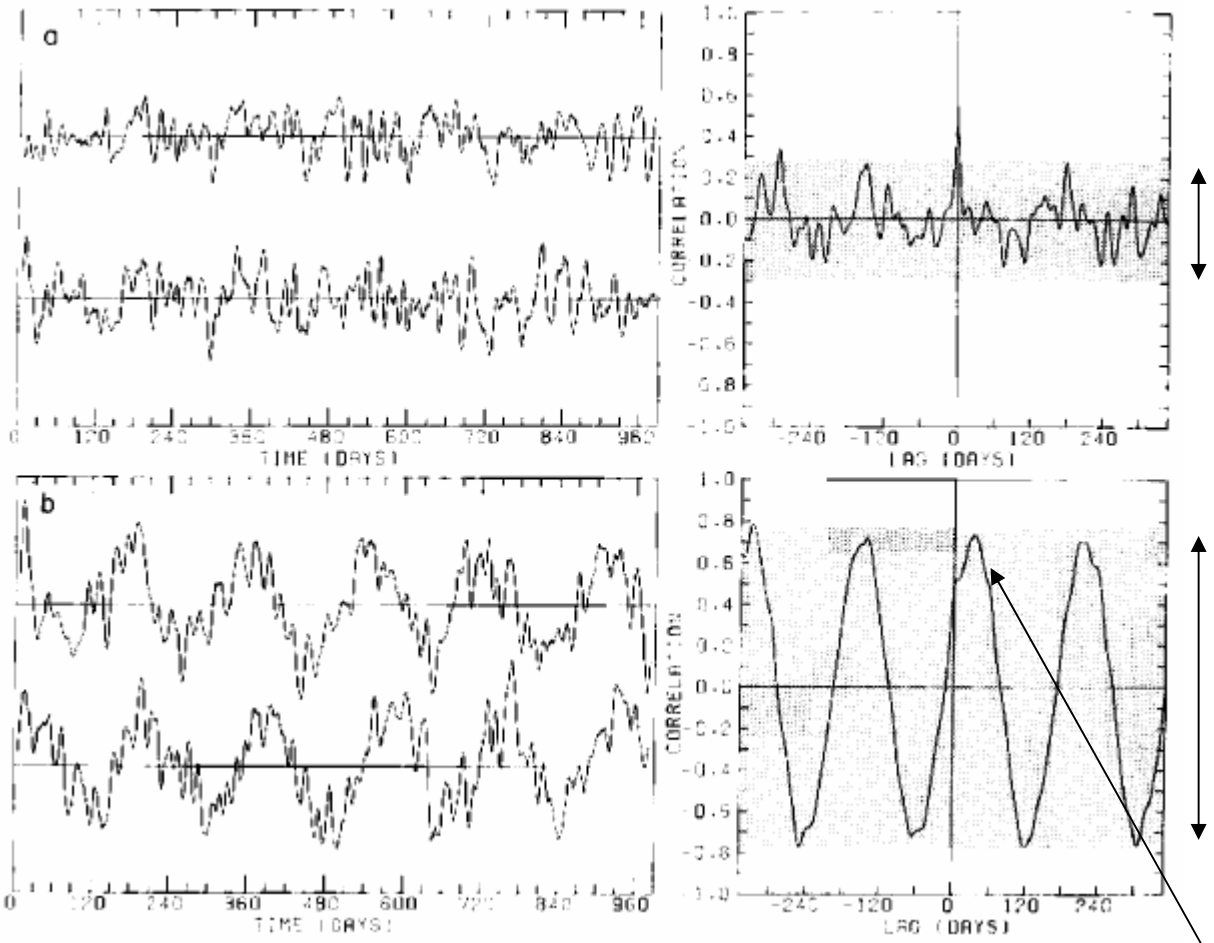
有効自由度

５０

serial
correlation
で有意相関
係数高く

６

高い係数
でも有意
ではない



$$x_1(t) = A \cos 2\pi f t + N_1(t).$$

$$x_2(t) = CN_1(t + \tau) + B \cos 2\pi f(t - t') + N_2(t)$$

$\langle N_2^2 \rangle = \langle N_1^2 \rangle$

$A = B$ (= 0 for case a. $2.5 \langle N_1^2 \rangle^{1/2}$ for the case b)

$C = 0.6$

$\tau = 0$ days

$t' = 30$ days

$f = 2$ cycles year^{-1}.

おまけ　「相関」の注意点

擬似相関

表 21.1　50 m 走のタイムと年収のデータ

No.	x：50m 走のタイム(秒)	y：年収(万円)	z：年齢(歳)
1	7.7	342	23
2	8.2	923	43
3	8.5	985	50
4	7.8	581	35
5	8.0	627	33
6	7.8	388	25
7	7.7	290	20
8	8.2	860	44
9	8.5	787	48
10	8.1	654	37
11	8.4	788	39
12	7.7	334	22
13	7.9	412	29
14	8.3	915	46
15	8.2	648	43
16	7.9	761	33
17	7.8	589	30
18	8.4	946	47
19	7.8	477	28
20	7.7	412	25

Rxy=0.8781
足の遅いひとほど
年収が高い？

Rzx=0.9407
Rzy=0.9400

永田(1996)より

4.6 回帰

誤差を最小にする $y = a + bx$ を与える a, b は次のように書ける。

$$\hat{b} = r \cdot \frac{s_y}{s_x} = \frac{\sum x'y'}{\sum x'^2}$$

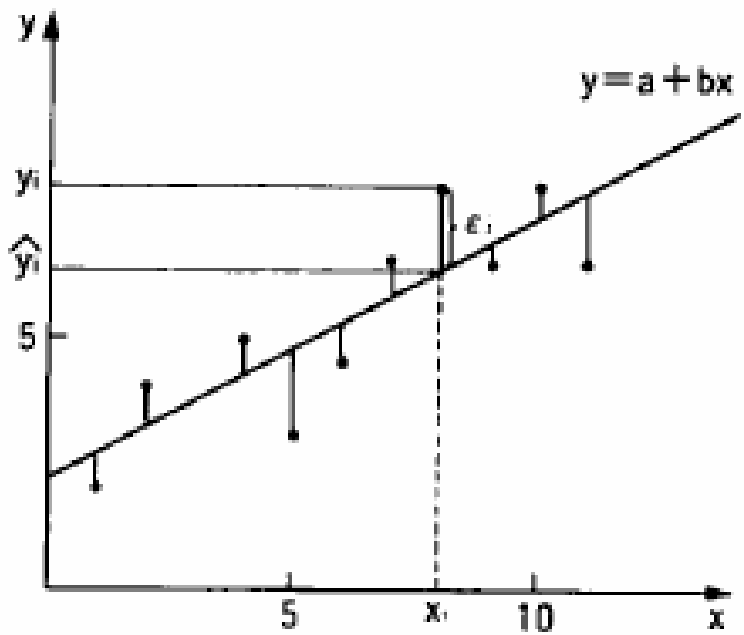
$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

ここで、$x' = x - \bar{x}$, $y' = y - \bar{y}$ として

$$r = \frac{\sum x'y'}{\sqrt{\sum x'^2 \sum y'^2}} = \frac{\sum x'y'}{n s_x s_y}$$

$$s_x^2 = \frac{\sum x'^2}{n}$$

$$s_y^2 = \frac{\sum y'^2}{n}$$



r は相関係数

$\hat{y} = a + bx$ と書けるとする。

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

$$= \sum_{i=1}^{n} \{y_i - (a + bx_i)\}^2$$

これを最小にしたい。そのため、偏微分を取って極値を求める。

$$\frac{\partial Q}{\partial a} = -2 \sum (y_i - a - bx_i) = 0$$

$$na + b \sum x_i - \sum y_i = 0$$

$$\frac{\partial Q}{\partial b} = -2 \sum (y_i - a - bx_i)(x_i) = 0$$

$$n \sum x_i + b \sum x_i^2 - \sum x_i y_i = 0$$

よって

$$a = \frac{1}{n} \sum y_i - b \frac{1}{n} \sum y_i = \bar{x} - b\bar{y}$$

$$b = \frac{\sum x_i y_i - \bar{y} \sum x_i}{\sum x_i^2 - \bar{x} \sum x_i} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum x' y'}{\sum x'^2}$$

決定係数

$$y' = y - \bar{y}$$

$$= (y - \hat{y}) + (\hat{y} - \bar{y})$$

$$\text{総変動} = \text{残差変動} + \text{回帰変動}$$

$$\text{決定係数} = \frac{\text{回帰変動}}{\text{総変動}} = r^2$$

回帰係数の区間推定

$$\hat{b} - t_{\alpha/2,\nu} \frac{s_\varepsilon}{\sqrt{(N-1)s_x}} < b < \hat{b} + t_{\alpha/2,\nu} \frac{s_\varepsilon}{\sqrt{(N-1)s_x}}$$

$$s_\varepsilon = \sqrt{\frac{1}{N-2} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}$$

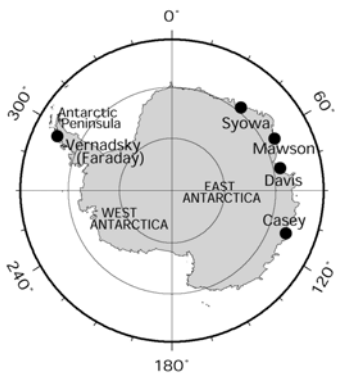
$$s_x = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})}$$

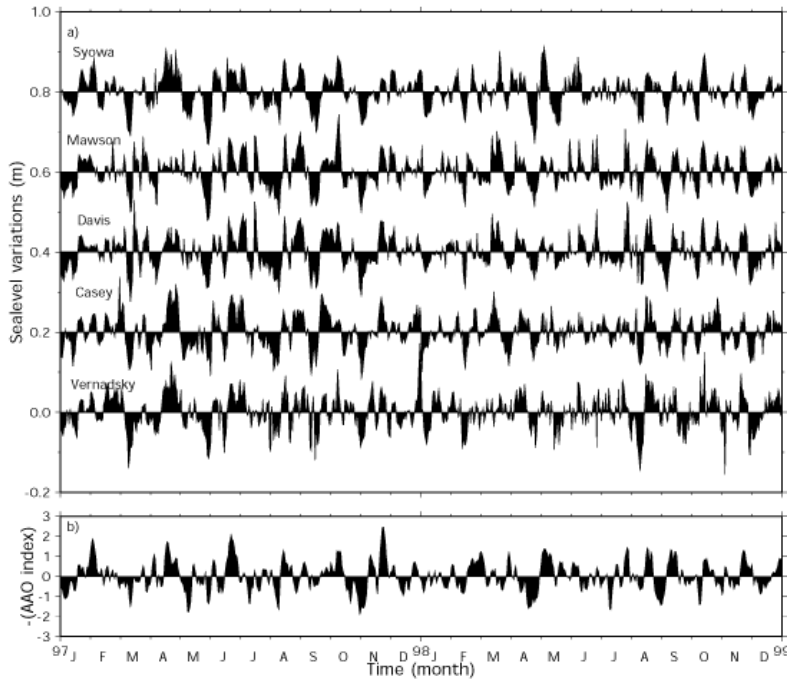
回帰係数の例 その１

High-passed

Table 2. Correlation and Regression Coefficients Between Sea Level and AAO Index for the High-Passed Signals and for the Signals After Being High-Passed and Having Long-Period Tides Removed

	Syowa	Mawson	Davis	Casey	Vernadsky
c.c.(high)	−0.457	−0.497	−0.458	−0.484	−0.414
Var.(%)	20.9	24.7	21.0	23.4	17.2
ratio	−2.31	−2.71	−2.56	−2.57	−2.44
c.c.(h-ltide)	−0.532	−0.574	−0.533	−0.519	−0.457
Var.(%)	28.3	32.9	28.4	27.0	20.8
ratio	−2.41	−2.81	−2.69	−2.53	−2.52





回帰係数の例　その２

データの分布は不規則

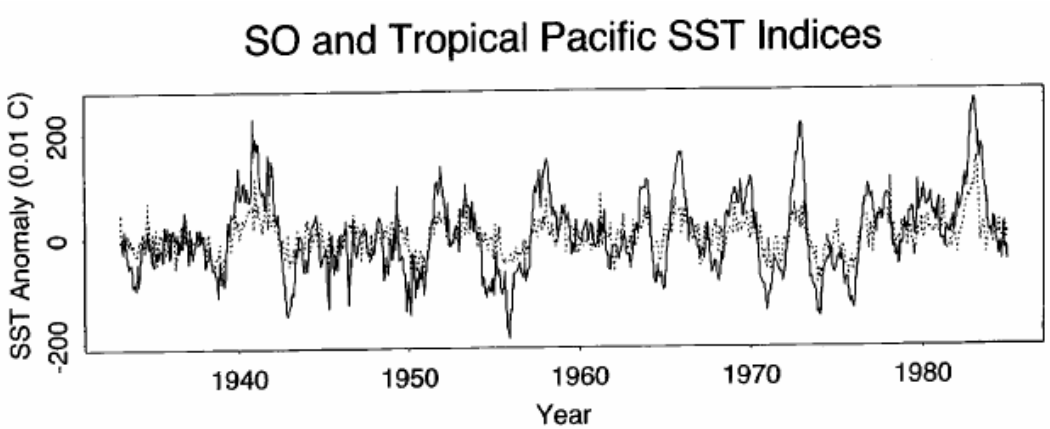


Figure 1.4: *The conventional Southern Oscillation Index (SOI = pressure difference between Darwin and Tahiti; dashed curve) and a sea-surface temperature (SST) index of the Southern Oscillation (solid curve) plotted as a function of time. The conventional SOI has been doubled in this figure.*

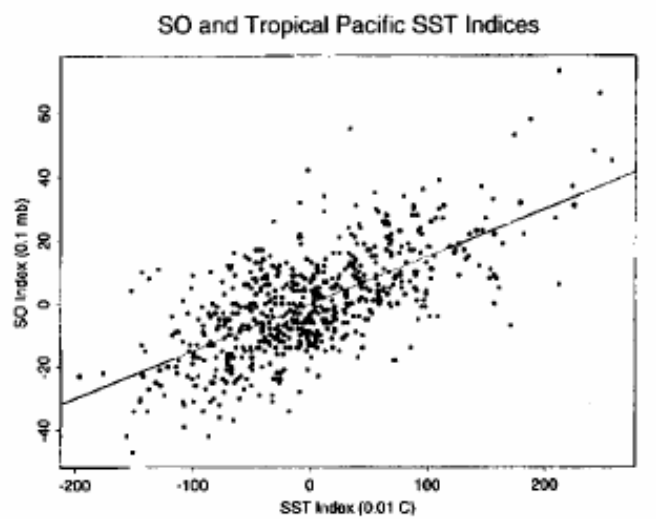


Figure 8.1: *Scatter plot of monthly values of the SO index versus the SST index for 1933–84 inclusive. Units: 0.1 mb (SOI), 0.01 °C (SST Index).*

SST Index
[180-90W, 6S-6N]

von Storch and Zwiers 1999
Wright 1984

回帰係数の例　その３

データ個数を標準化

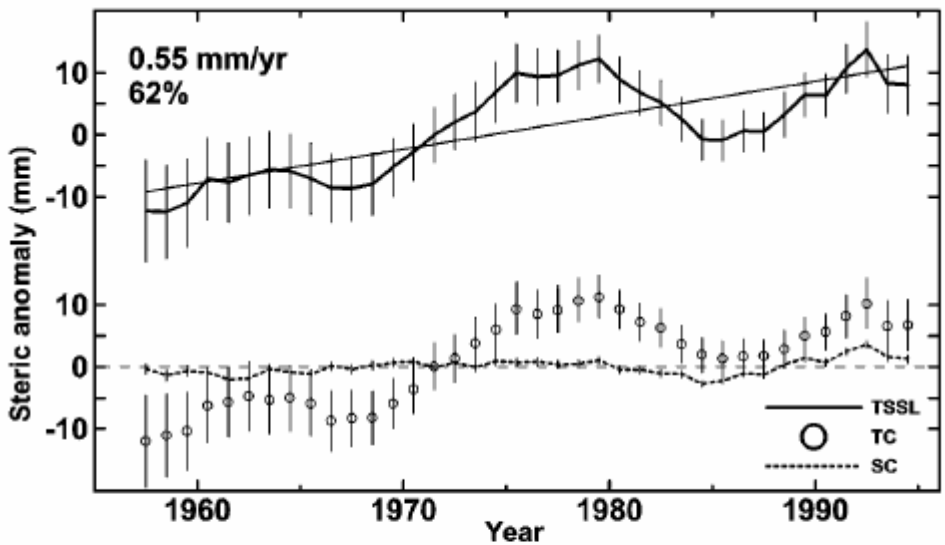


Figure 3. Time series of spatially averaged (50°S–65°N) 5-year running composites of thermosteric (open circles), halosteric (dashed line), and total steric (solid line) anomalies (in millimeters) of the 0–3000 m layer for the 1957–1994 period. Vertical lines represent ±1 standard error of the 5-year mean estimates of steric components. The linear trend is plotted for the TSSL anomaly time series. The trend and the percent variance accounted for by this trend are given in the top left corner.

説明変数は時間

Antonov et al.2002 JGR

まとめ

- 相関係数は変数同士の関連の強さを示す指標
- 変数の周期性を調べたい場合、相関関数をもちいることがある
- 無相関の検定はｔ検定により行うことができる
- 相関関係と因果関係は別物である
  → 擬似相関 spurious correlation
- （単）回帰係数は被説明変数を直線であてはめたときの傾きを示す。
- 相関係数に説明変数・被説明変数の分散の比をかけたものが回帰係数になる。