Universitas Gadjah Mada
Fakultas Teknik
Departemen Teknik Sipil dan Lingkungan
Prodi Magister Teknik Pengelolaan Bencana Alam

Teknik Pengolahan Data

Korelasi

Korelasi

- Acuan
  - Haan, C.T., 1982, *Statistical Methods in Hydrology*, 1st Ed., 3rd Printing, The Iowa State Univ. Press, Ames, Iowa, USA
    - Chapter 11, pp 222-235

Korelasi

- Koefisien korelasi antara dua variabel random *X* dan *Y*

$$r_{X,Y} = \sqrt{1 - \frac{S_r}{S_t}} = \sqrt{1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{Y})^2}}, \quad \hat{y} = y_{reg} \Rightarrow r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y}$$

koefisien korelasi sampel

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

koefisien korelasi populasi

$$s_{X,Y} = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{n-1}$$

kovarian *X* dan *Y*

$$s_X = \sqrt{\frac{\sum(x_i - \bar{X})^2}{n-1}}$$

simpangan baku *X*

$$s_Y = \sqrt{\frac{\sum(y_i - \bar{Y})^2}{n-1}}$$

simpangan baku *Y*

Korelasi

- Koefisien korelasi antara dua variabel random *X* dan *Y*

$$r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y}$$

$$r_{X,Y} = \frac{=\text{COVARIANCE.S(X,Y)}}{=\text{STDEV.S(X)}*\text{STDEV.S(Y)}}$$

$$r_{X,Y} = =\text{CORREL(X,Y)}$$

MSExcel

Koefisien Korelasi

- Pengertian koefisien korelasi
  - Koefisien korelasi menunjukkan tingkat keeratan hubungan linear antara suatu variabel random *Y* dan suatu variabel kedua yang merupakan fungsi linear dari satu atau lebih variabel(-variabel) *X*
    - Setiap variabel *X* dapat berupa variabel random atau bukan variabel random

Koefisien Korelasi

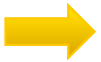
- Nilai koefisien korelasi adalah $-1 \leq r_{X,Y} \leq 1$
  - $r_{X,Y} = \pm 1$ menunjukkan hubungan linear sempurna antara *X* dan *Y*
  - $r_{X,Y} = 0$ menunjukkan independensi (ketidak-tergantungan) linear, namun dapat saja keduanya memiliki hubungan (ketergantungan) yang lain, yang tidak linear
  - Jika *X* dan *Y* tidak saling tergantung (*independent*), maka $r_{X,Y} = 0$

Inferensi terhadap Koefisien Korelasi Populasi

- Dua variabel random
  - tak berkorelasi, $\rho_{X,Y} = 0$
  - berkorelasi, $\rho_{X,Y} \neq 0$
- Situasi
  - Sampel yang diperoleh dari variabel random yang tidak berkorelasi
    - jarang menunjukkan nilai $r_{X,Y} = 0$
    - koefisien korelasi sering $r_{X,Y} \neq 0$, karena kebetulan
  - Oleh karena itu perlu pengujian
    - untuk mengetahui penyimpangan koefisien korelasi dari nol tersebut benar disebabkan oleh kebetulan, atau
    - penyimpangan tersebut terlalu besar untuk dikatakan sebagai akibat kebetulan

Inferensi terhadap ρ

- Uji hipotesis
  - H_0: $\rho_{X,Y} = 0$
  - H_a: $\rho_{X,Y} \neq 0$

statistik uji $T = r\left[\frac{n-2}{1-r^2}\right]^{\frac{1}{2}}$  $|T| > t_{1-\alpha/2,n-2}$ H_0 ditolak

Inferensi terhadap ρ

- Uji hipotesis
  - H_0: $\rho_{X,Y} = \rho^*$ (ρ^* konstanta, diketahui)
  - H_a: $\rho_{X,Y} \neq \rho^*$

  ukuran sampel $n > 25$

  statistik uji: $z = (W - \omega)(n-3)^{\frac{1}{2}}$ → $|z| > z_{1-\alpha/2}$ H_0 ditolak

$$W = \frac{1}{2}\ln\left[\frac{1+r}{1-r}\right] = \operatorname{arctanh} r \qquad \omega = \frac{1}{2}\ln\left[\frac{1+\rho^*}{1-\rho^*}\right] = \operatorname{arctanh} \rho^*$$

- Rentang keyakinan ρ:

$$l = \tanh\left[W - \frac{z_{1-\alpha/2}}{(n-3)^{1/2}}\right] \qquad u = \tanh\left[W + \frac{z_{1-\alpha/2}}{(n-3)^{1/2}}\right]$$

Inferensi terhadap ρ

- Ada sejumlah *k* populasi *bivariate*, distribusi normal, memiliki
  - koefisien korelasi populasi $\rho_1, \rho_2, \ldots, \rho_k$
  - koefisien korelasi sampel $r_1, r_2, \ldots, r_k$
  - ukuran sampel $n_1, n_2, \ldots, n_k$
- Hipotesis
  - H_0: $\rho_1 = \rho_2 = \ldots = \rho_k = \rho^*$ (ρ^* konstanta)
  - H_a: $\rho_1 \neq \rho_2 \neq \ldots \neq \rho_k \neq \rho^*$

Statistik uji

$$\chi^2 = \sum_{i=1}^{k} (n_i - 3)(\operatorname{arctanh} r_i - \operatorname{arctanh} \rho^*)^2$$

$\chi^2 > \chi^2_{1-\alpha,k}$ H_0 ditolak

Inferensi terhadap ρ

- Hipotesis
  - H_0: $\rho_1 = \rho_2 = \ldots = \rho_k$ (semua koefisien korelasi sama)
  - H_a: $\rho_1 \neq \rho_2 \neq \ldots \neq \rho_k$

Statistik uji

$$\chi^2 = \sum_{i=1}^{k} (n_i - 3)\left(W_i - \bar{W}\right)^2 \quad \Rightarrow \quad \chi^2 > \chi^2_{1-\alpha,k-1}$$ H_0 ditolak

$$W_i = \operatorname{arctanh} r_i$$

$$\bar{W} = \sum_{i=1}^{k} (n_i - 3) W_i \Big/ \sum_{i=1}^{k} (n_i - 3)$$

Inferensi terhadap ρ

- Jika
  - hipotesis bahwa semua koefisien korelasi tidak ditolak, maka
  - perlu diketahui nilai koefisien korelasi *r* yang mewakili nilai koefisien korelasi populasi ρ

$$\bar{r} = \tanh\left(\bar{W} - m\,\rho^{*}/2\right)$$

$$\bar{W} = \sum_{i=1}^{k}(n_i - 3)W_i \Big/ \sum_{i=1}^{k}(n_i - 3)$$

$$\rho^{*} = \sum_{i=1}^{k} r_i/k$$

$$m = \sum_{i=1}^{k}\left(\frac{n_i - 3}{n_i - 1}\right) \Big/ \sum_{i=1}^{k}(n_i - 3)$$

Korelasi Serial

- Korelasi serial (*serial correlation*)
  - dikenal pula sebagai autokorelasi (*autocorrelation*)
  - yaitu korelasi antara data hasil pengukuran pada suatu waktu dengan data hasil pengukuran pada waktu sebelumnya
  - elemen dalam sampel yang memiliki korelasi serial **bukan** elemen random (ingat definisi variabel random)

Korelasi Serial

- Dalam korelasi serial, dengan demikian
  - sampel berukuran *n* yang memiliki korelasi serial akan memberikan informasi yang lebih sedikit dibandingkan dengan informasi yang dimiliki oleh sampel random berukuran *n*
  - sebagian informasi pada sampel yang memiliki korelasi serial dapat diperoleh dari atau telah diketahui dalam data hasil pengukuran pada waktu sebelumnya

Korelasi Serial

- Korelasi serial (*serial correlation*)
  - dapat pula dijumpai antara suatu pengukuran pada waktu tertentu dengan pengukuran pada waktu *k* periode waktu sebelumnya (terdahulu), *k* = 1,2, …
  - asumsi
    - selang waktu antar pengukuran adalah sama (seragam)
    - sifat-sifat statistis proses atau peristiwa yang diukur tidak berubah terhadap waktu (bersifat permanen)
    - ρ(*k*) --- koefisien korelasi serial populasi
      r(*k*) --- koefisien korelasi serial sampel

Korelasi Serial

$$r(k) = \frac{\sum_{i=1}^{n-k} x_i x_{i+k} - \sum_{i=1}^{n-k} x_i \sum_{i=1}^{n-k} x_{i+k}/(n-k)}{\left[\sum_{i=1}^{n-k} {x_i}^2 - \left(\sum_{i=1}^{n-k} x_i\right)^2/(n-k)\right]^{1/2} \left[\sum_{i=1}^{n-k} {x_{i+k}}^2 - \left(\sum_{i=1}^{n-k} x_{i+k}\right)^2/(n-k)\right]^{1/2}}$$

- $r(0) = 1$ → korelasi suatu elemen data dengan dirinya sendiri adalah sama dengan satu
- semakin besar k, jumlah pasangan data untuk menghitung $r(k)$ semakin sedikit; $r(k)$ adalah nilai estimasi $\rho(k)$
- oleh karena itu, $k << n$
- jika $\rho(k) = 0$ untuk semua k, maka proses atau peristiwa atau populasi tersebut bersifat random murni

Korelasi Serial

- *Time series* sirkular
  - yaitu *time series* yang berulang, X_n diikuti oleh X_1
  - untuk *time series* sirkular, normal, permanen
    - asumsi

$$r(k) = \frac{\sum_{i=1}^{n} x_i x_{i+k} - n\bar{X}^2}{(n-1){s_X}^2}$$

persamaan $r(k)$ ini memberikan hasil hitungan yang mirip dengan persamaan yang panjang pada hlm sebelumnya apabila $n >>$ dan $k << n$

  - dengan asumsi di atas, maka $r(k)$ berdistribusi normal dengan
    - nilai rerata $-1/(n-1)$
    - simpangan baku $(n-2)/(n-1)^2$

jika $\rho(k) = 0$

Korelasi Serial

- Rentang keyakinan ρ(*k*)

$$l = \frac{1}{n-1}\left(-1 - z_{1-\alpha/2}\sqrt{n-2}\right) \qquad u = \frac{1}{n-1}\left(-1 + z_{1-\alpha/2}\sqrt{n-2}\right)$$

- Uji hipotesis

H_0: ρ(*k*) = 0
H_a: ρ(*k*) ≠ 0

➡ hipotesis ditolak jika *r*(*k*) berada di luar rentang keyakinan

Korelasi dan Analisis Regional

- Informasi yang dikandung dalam data dari sejumlah *n* stasiun di suatu wilayah yang memiliki korelasi (rerata) interstasiun $\bar{\rho}$ sama dengan informasi yang dikandung dalam data dari sejumlah *n*' stasiun yang tidak saling berkorelasi

$$n' = \frac{n}{1 + (n-1)\bar{\rho}}$$

- Seiring dengan *n* yang besar, maka *n*' mendekati $1/\bar{\rho}$

$$n \gg \implies n' \approx 1/\bar{\rho}$$

- Melihat hubungan antara *n*' dan *n*, maka menempatkan sedikit stasiun yang tidak saling bergantung (*independent*) lebih baik daripada menempatkan banyak stasiun yang saling berkorelasi

Korelasi dan Analisis Regional

- Korelasi dalam satu wilayah
  - dapat dipakai untuk memperoleh nilai estimasi yang lebih baik terhadap suatu variabel hidrologik di suatu titik lokasi melalui korelasi dengan variabel hidrologik lain di titik lokasi tersebut atau dengan suatu karakteristika yang mirip di titik lokasi yang lain
  - sebagai contoh
    - *Y* dan *X* adalah dua variabel random hidrologik yang tidak berkorelasi
    - jumlah sampel *Y* adalah n_1
    - jumlah sampel *X* adalah $n_1 + n_2$
    - *Y* dan *X* memiliki koefisien korelasi $\rho_{Y,X}$

Korelasi dan Analisis Regional

- Korelasi dalam satu wilayah
  - Data *Y* dapat diperpanjang dengan memakai korelasi antara *Y* dan *X*
    - pada dasarnya ini sama dengan regresi
    - persamaan hubungan *Y* dan *X* dibentuk dari n_1 pasangan data

$$y = r_{Y,X}\frac{s_Y}{s_X}x$$

y dan *x* adalah deviasi (selisih) data terhadap nilai rerata masing-masing variabel

    - persamaan di atas dipakai untuk memperoleh sejumlah n_2 nilai *Y* berdasarkan n_2 data *X*
    - nilai rerata *Y* yang baru adalah: $\bar{Y} = \frac{n_1\bar{Y}_1 + n_2\bar{Y}_2}{n_1 + n_2}$
    - syarat: $r_{X,Y} > \frac{1}{(n_1 - 2)}$

Korelasi dan Sebab-Akibat

- Penting diketahui
  - Korelasi yang tinggi antara dua variabel tidak selalu berarti adanya hubungan kausal, sebab-akibat antar kedua variabel tersebut
    - Adanya korelasi antara debit bulanan di suatu sungai dengan debit bulanan di sungai tetangga tidak berarti bahwa perubahan debit bulanan di sungai yang satu akan mengakibatkan perubahan debit bulanan di sungai tetangga
    - Perubahan debit bulanan di kedua sungai mungkin saja disebabkan oleh faktor eksternal
  - Ingat
    - Variabel-variabel yang *independent* pastilah tak saling berkorelasi, tetapi
    - Variabel-variabel yang tak saling berkorelasi tidak selalu *independent*
    - Kebergantungan antar variabel yang saling berkorelasi adalah kebergantungan yang bersifat stokastik
      - bukan kebergantungan dalam arti fisik, dan
      - bukan kebergantungan dalam arti sebab-akibat

Korelasi *Spurious*

- *Spurious correlation* (korelasi "palsu")
  - Tampak seperti ada korelasi antar variabel-variabel, tetapi variabel-variabel tersebut sebenarnya tidak berkorelasi
  - Dapat terjadi karena adanya data yang mengelompok

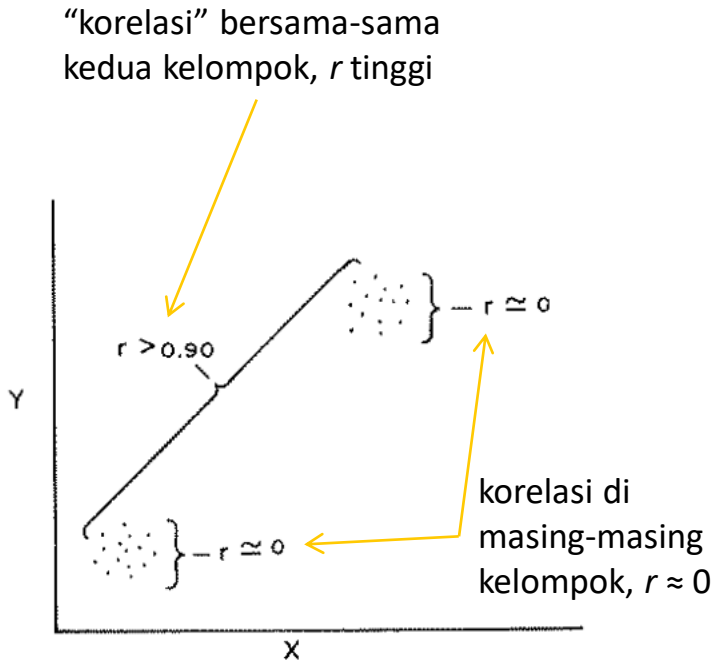


Fig. 11.2. Spurious correlation due to data clustering.

Korelasi *Spurious*

Y dan *X* tidak berkorelasi
(*Y* dan *X independent*)

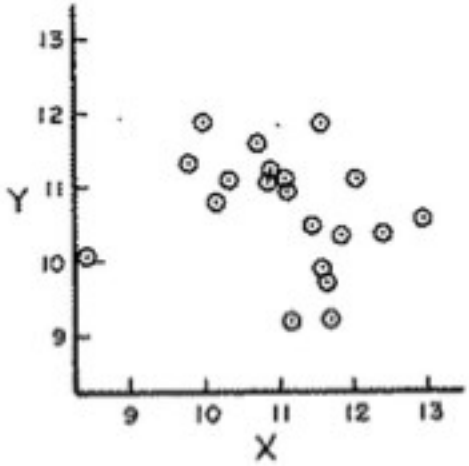


Fig. 11.3. Absence of correlation between two random variables.

Y dan *X* dibagi *Z* → *Y* dan *X* memiliki koefisien korelasi *r* >> 0

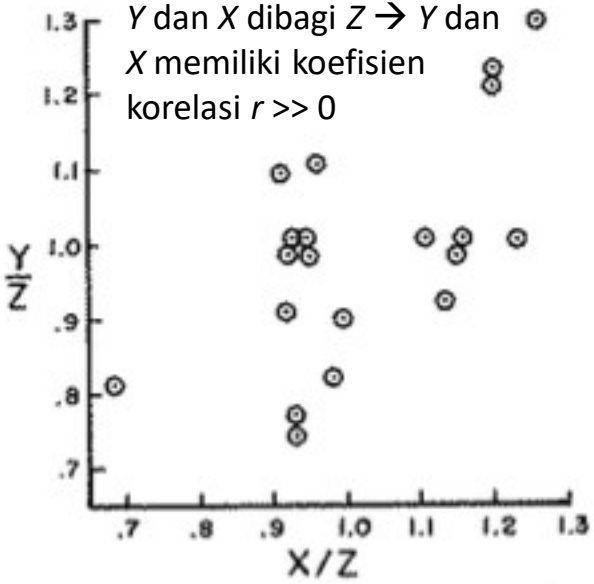


Fig. 11.4. Spurious correlation introduced by dividing 2 random variables by a common third random variable.

Terima kasih