Cleaning and Analysis of Special Codes in Automatic Station Rainfall Data

YUNG-CHUN LAN

Outline

1. PP01 (資料探索、生成機制、分析、處理)
2. 12 月累積雨量網格製作
3. Simple Kriging & Spherical Variogram 數學推導概述
4. Simple Kriging & Spherical Variogram 手寫推導

PP01資料探索

PP01 欄位內容:

- 總列數: 469170
- -999.1(儀器故障待修): 4464
- -999.6(資料累積於後): 13715
- -999.5(因故障而無資料): 189
- None(未觀測而無資料): 1176

```
df_rain.shape
```

```
(469170, 3)
```

```
irregular = [-999.1, -9.6, -999.6, -9.5, -99.5, -999.5, -
for i in irregular:
    print(f"{i}: ",df_rain["PP01"].isin([i]).sum())
print("na: ",df_rain["PP01"].isna().sum())
```

```
-999.1:  4464
-9.6:  0
-999.6:  13715
-9.5:  0
-99.5:  0
-999.5:  189
-9999.5:  0
-9.7:  0
-99.7:  0
-999.7:  0
-9999.7:  0
-9.8:  0
na:  1176
```

PP01資料生成機制

雨水過濾斗，進入接水器，累積 0.5mm 後將水傾倒，產生脈衝信號，傳輸到紀錄器上，因此可看到 PP01 資料是以 0.5 mm 為單位

```
df_rain = df[["# stno", "yyyymmddhh", "PP01"]].copy()
df_rain.rename(columns = {"# stno": "stno"}, inplace = True)
df_rain.head(5)
```

	stno	yyyymmddhh	PP01
0	C0A520	2023120100	0.5
1	C0A520	2023120101	0.5
2	C0A520	2023120102	1.0
3	C0A520	2023120103	1.0
4	C0A520	2023120104	3.0

Source: 中央氣象署官網(https://south.cwa.gov.tw/inner/meck1572422009QjcD)

-999.6 分析

工程邏輯上，出現無線電訊號中斷，該時間點會標記為 -999.6，並在恢復信號時間點將中斷期間的雨量觀測累加於後

Reference:國家災害防救科技中心氣象組 – 台灣地區短延時強降雨事件氣候特性分析
二、資料、定義與分析方法 (https://reurl.cc/dqN0x8)

資料分析層面，使用資料累積於後(-999.6)，而不是直接補上斷訊期間脈衝，推測回補機制並非 100% 可靠。

我們也能觀察到 0 → -999.6 → 0的數據，在此情況卻不將 -999.6 設定為 0，顯示其不確定性。

PP01
0.5
0.5
1.0
1.0
3.0
4.0
1.5
1.5
0.5
0.0
0.0
0.0
0.0
-999.6
0.0
0.0
0.0
0.0

-999.6 處理原則

參考 Guide to Climatological Practices (WMO-No.100)
提及之氣象資料缺失值處理原則，-999.6處理方式如下:

1. 不更動原數據
2. 新增 PP01_cleaned 欄位，將各種原因導致之遺失值全部改為 na
3. 新增 PP01_accumulated 欄位，出現 -999.6 及標註為 True
4. 新增 PP01_caaumulated 欄位標註 -999.6 影響範圍
    - -999.6: 標示 True
    - -999.6 且下一列不為 0: 下一列標示 True
    - -999.6 且下一列為 0: 下一列標示 False
5. 分析者可根據不同需求使用mask

```
[14] df_rain.head(3)
✓ 0.0s
```

	stno	yyyymmddhh	PP01	PP01_clean	PP01_accumulated	PP01_accumulated_window
0	C0A520	2023-12-01 00:00:00	0.5	0.5	False	False
1	C0A520	2023-12-01 01:00:00	0.5	0.5	False	False
2	C0A520	2023-12-01 02:00:00	1.0	1.0	False	False

Reference: *Guide to Climatological Practices* (WMO-No.100): 2.4.5 Quality Control & 3.5.8 Data Estimation

12 月累積雨量網格資料製作

目標: 參考氣象署雨量網格化生產履歷，產出全台 12 月月雨量網格資料
資料需求: 測站月雨量、測站經緯度

- 月雨量: 以測站為單位，將同測站逐時雨量加總
- 測站經緯度: 爬取中央氣象署雨量觀測站-雨量資料（O-A0002-001）
- 欄位包含: stno、lon、lat、rain_dec

```
df_rain_month.head(3)
```

[56] ✓ 0.0s

	stno	lat	lon	rain_dec
0	C0A520	24.974944	121.402008	103.0
1	C0A530	24.938183	121.709750	165.5
2	C0A550	24.971197	121.823711	538.0

Source:
中央氣象署雨量網格化生產履歷: https://www.cwa.gov.tw/Data/data_catalog/2-3-6-a.pdf
中央氣象署雨量觀測站資料: https://opendata.cwa.gov.tw/dataset/observation/O-A0002-001?

12 月累積雨量網格資料製作 – 建立全台網格系統

1. 設定網格空間範圍：
   - 左下角經緯度為 (117.43, 20.76)
   - 右上角經緯度為 (123.92, 26.70)
2. 設定網格解析度：
   - 2.5 公里解析度，網格大小為 260 × 260

Construct the Grid (resolution: 2.5 km)

```
nx = 260
ny = 260
lon_min, lon_max = 117.43, 123.92
lat_min, lat_max = 20.76, 26.70
lon = np.linspace(lon_min, lon_max, nx)
lat = np.linspace(lat_min, lat_max, ny)
grid_lon, grid_lat = np.meshgrid(lon, lat)
```

[15] 0.0s

使用Simple Kriging Interpolation
估計網格點雨量

相關設置:
- Mean = 0 (附近無測站地區估計 0)
- Bin = 20
- Variogram model: Spherical
- Nugget 手動調參，不參與訓練

(詳細推導與方法說明見附錄)

```
x = df_rain_month["lon"].to_numpy(dtype=float)
y = df_rain_month["lat"].to_numpy(dtype=float)
z = df_rain_month["rain_dec"].to_numpy(dtype=float)

m = 0.0
z_res = z - m

# bin_center: representative value of each bin (bin=20)
# gamma: estimate variogram of each interval
bin_center, gamma = gs.vario_estimate((x, y), z_res)

# initialize Spherical variogram model unknown parameter
sill_0 = np.nanvar(z_res)
range_0 = 0.25 * (bin_center.max() - bin_center.min())
nug_0 = 0

model = gs.Spherical(dim=2, var = sill_0, len_scale = range_0, nugget = nug_0)

# train model (LSE)
np.random.seed(123)
model.fit_variogram(bin_center, gamma, nugget=False, max_eval=20000)

# compute
sk = gs.krige.Simple(model, cond_pos=(x, y), cond_val=z_res, mean=0)
z_grid_res, var_grid = sk((grid_lon, grid_lat))
z_grid = z_grid_res + m
```
✓ 19.3s

Reference:
gstools 官方 documentation
Guide to Climatological Practices (WMO-No.100): 2.4.5 Quality Control & 3.5.8 Data Estimation
https://geostat-framework.readthedocs.io/projects/gstools/en/stable/examples/03_variogram/00_fit_variogram.html#sphx-glr-examples-03-variogram-00-fit-variogram-py

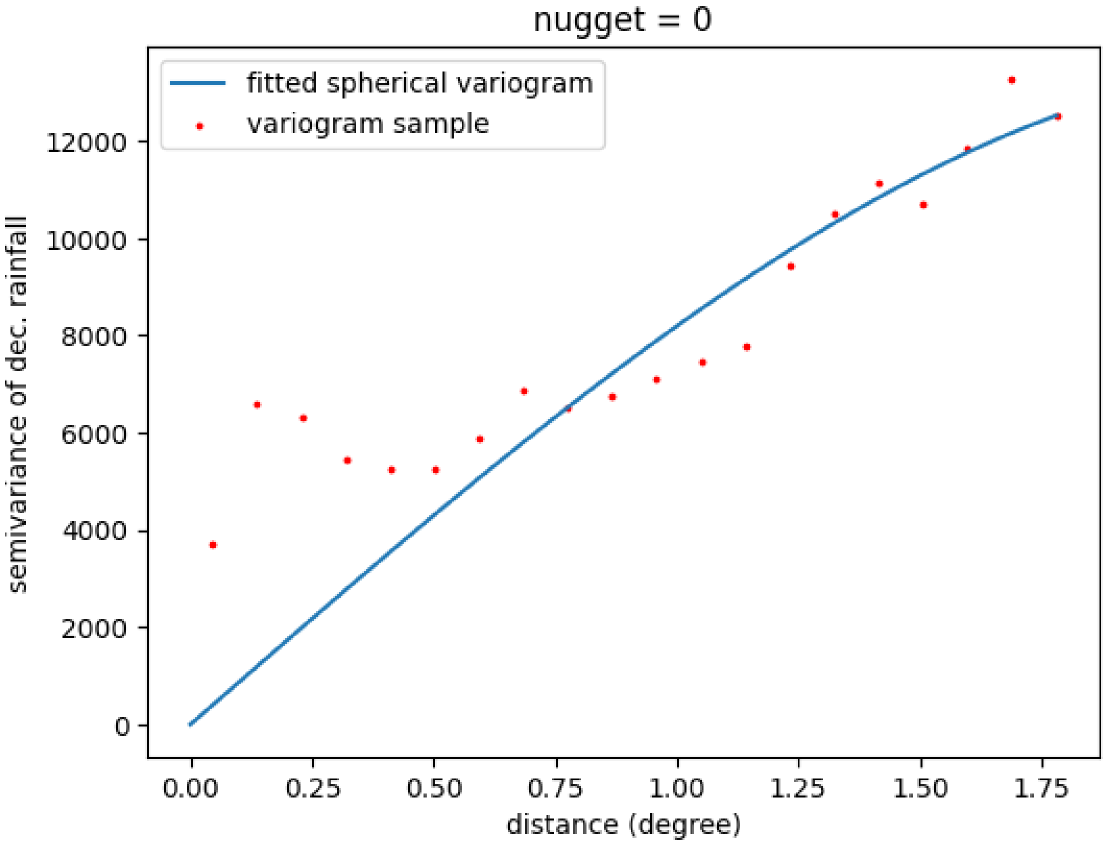
12 月累積雨量網格資料製作 – baseline 模型

Original parameter setting

sill_0 = np.nanvar(z_res)
range_0 = 0.25 * (bin_center.max() - bin_center.min())
nug_0 = 0

Optimal result

sill_opt = 13576.1648
range_opt = 2.3337
nug_opt = 0



- 模型能捕捉 距離增加 →半變異數上升趨勢
- 無法反映近距離觀測誤差
- 下一步: 提高 nugget 至 3000

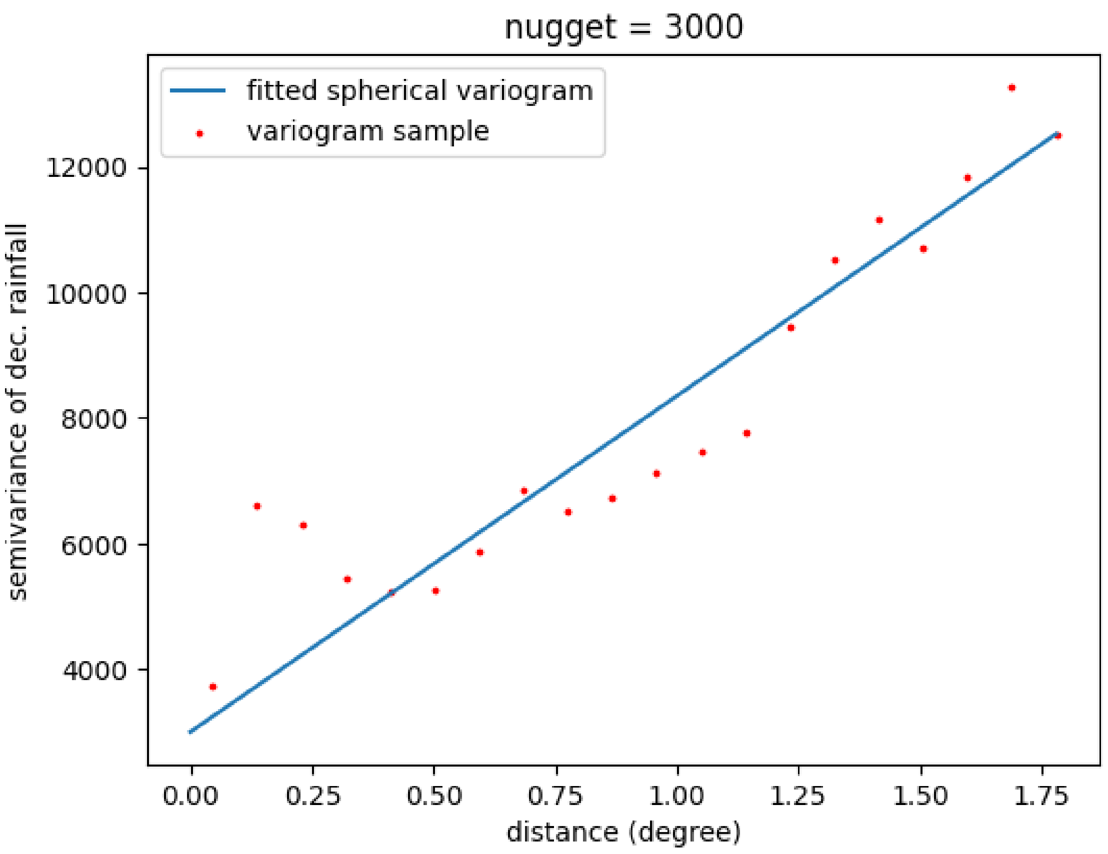
12 月累積雨量網格資料製作 – nugget設置3000

Original parameter setting

sill_0 = np.nanvar(z_res)
range_0 = 0.25 * (bin_center.max() - bin_center.min())
nug_0 = 3000

Optimal result

sill_opt = 2823075.6513
range_opt = 791.9788
nug_opt = 3000



- 模型擬合度比baseline好
- Sill 極高
- Range 不符合自然現象 (此結果相當於全球都具空間相關性
- 模型為貼合 nugget 失去物理意義
- 下一步: 降低 nugget 至 1000

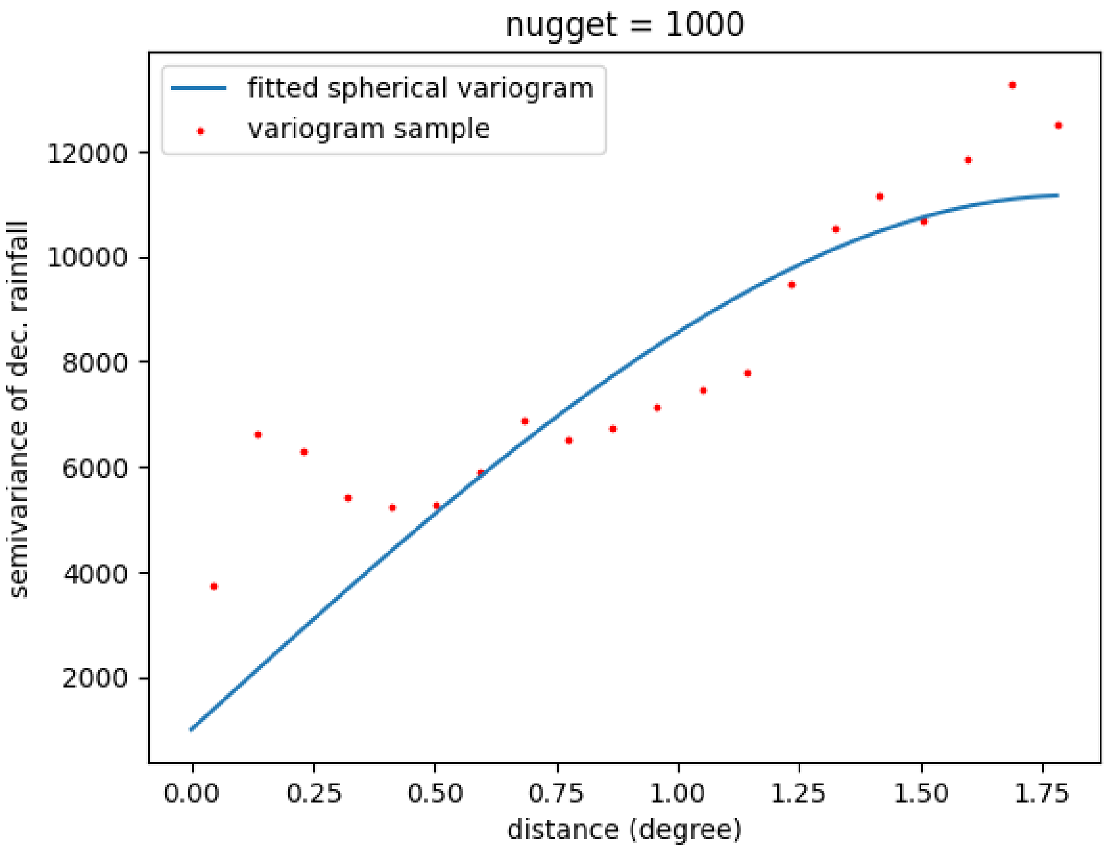
12 月累積雨量網格資料製作 – nugget設置1000

Original parameter setting

sill_0 = np.nanvar(z_res)
range_0 = 0.25 * (bin_center.max() - bin_center.min())
nug_0 = 3000

Optimal result

sill_opt = 10158.9674
range_opt = 1.8155
nug_opt = 1000

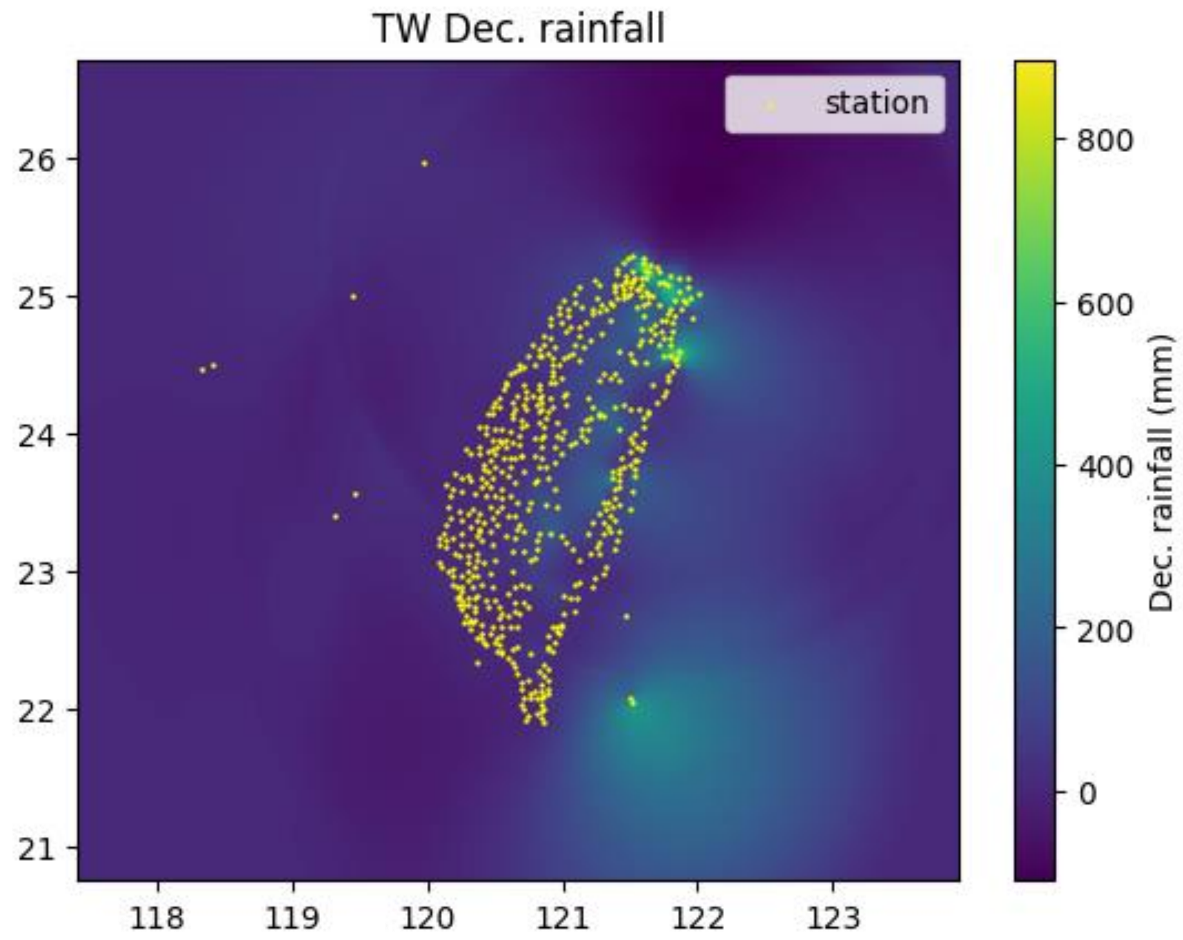


- 模型擬合度在中段貼合良好
- Sill、Range 數值穩定，無發散現象
- Nugget = 1000 合理表示短距離不確定性
- 採用此模型做視覺化

12 月累積雨量網格資料製作 – 視覺化

1. 北部與東北側 (迎東北季風)、蘭嶼雨量較高
2. 未發生高值出現於鄰近無測站情況

- 結果與地理常識大致相同
- 海上數據可透過遮罩方式消除



$u_1, \ldots, u_n$: locations of stations

$Z(u_1), \ldots, Z(u_n)$: rainfall at stations

Target: estimate at unobserved location: $Z(u_0)$

Assumption of Simple Kriging:

$$Z(u) = m + \epsilon(u)$$

where m is the global mean and $\epsilon(u)$ is the random residual.

Let $E[\epsilon(u)] = 0$ and $m = 0$

Kriging estimation form:

$$\hat{Z}(u_0) = m + \sum_{i=1}^{n} \lambda_i \left[Z(u_i) - m \right]$$

Caculation target: weight $\lambda_1 \ldots \lambda_n$

Reference: *Make your own Kriging interpolation algorithm with python*, Jui-Fa Tsai
https://www.dropbox.com/scl/fi/lw19ieu12cv1wlzxy29re/outreach_2019-09-24_kriging_meetup.pdf?rlkey=8wm1diwj3usph7y6cm1o37j45&e=1&dl=0

Estimation condition:

1. $E[\epsilon(u)] = 0 \Rightarrow E[\hat{Z}(u_0)] = m$
2. Find min. $V[\hat{Z}(u_0) - Z(u_0)]$

Expand the variance and take its partial derivative with $\lambda_1, \dots, \lambda_n$, we get:

$$\begin{bmatrix} \mathrm{Cov}[\epsilon(u_1), \epsilon(u_1)] & \cdots & \mathrm{Cov}[\epsilon(u_1), \epsilon(u_n)] \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}[\epsilon(u_n), \epsilon(u_1)] & \cdots & \mathrm{Cov}[\epsilon(u_n), \epsilon(u_n)] \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} \mathrm{Cov}[\epsilon(u_1), \epsilon(u_0)] \\ \vdots \\ \mathrm{Cov}[\epsilon(u_0), \epsilon(u_n)] \end{bmatrix}$$

Solving $\lambda_1 \dots \lambda_n$ through linear equations.

However, $\mathrm{Cov}[\epsilon(u_i), \epsilon(u_j)]$ is unknown.

Next, we will fit them with a Spherical variogram model.

Reference: *Make your own Kriging interpolation algorithm with python*, Jui-Fa Tsai https://www.dropbox.com/scl/fi/lw19ieu12cv1wlzxy29re/outreach_2019-09-24_kriging_meetup.pdf?rlkey=8wm1diwj3usph7y6cm1o37j45&e=1&dl=0

Spatial related assumptions:

1. Correlation: close > far
2. The mean value is a constant
3. Covariance depends only on distance h and is independent with location.

Let $C(h) = Cov[Z(u), Z(u+h)]$

Define Variogram:

$$\begin{aligned}\gamma(h) &= \frac{1}{2}E[Z(u) - Z(u+h)]^2 \\ &= \frac{1}{2}[2V(Z) - 2C(h)] \\ &= C(0) - C(h)\end{aligned}$$

Reference: *Make your own Kriging interpolation algorithm with python*, Jui-Fa Tsai https://www.dropbox.com/scl/fi/lw19ieu12cv1wlzxy29re/outreach_2019-09-24_kriging_meetup.pdf?rlkey=8wm1diwj3usph7y6cm1o37j45&e=1&dl=0

1. Sill

The value that the semivariogram approaches as the distance nears infinity. The greater the distance, the lower the correlation:

$$Let \lim_{h\to\infty} C(h) = 0$$
$$\Rightarrow sill = \lim_{h\to\infty} \gamma(h) = \lim_{h\to\infty} [C(0) - C(h)] = C(0)$$

2. Range

The distance at which the semivariogram reaches the sill. Beyond this distance, spatial correlation between observations becomes negligible:

$$h > range \Rightarrow Z(u) \perp Z(u+h)$$

3. Nugget

The semivariance at an infinitesimally small separation distance. It reflects measurement error or spatial variability occurring at distances smaller than the sampling distance:

$$\gamma(0^+) = C_0 > 0$$

Reference: *Make your own Kriging interpolation algorithm with python*, Jui-Fa Tsai https://www.dropbox.com/scl/fi/lw19ieu12cv1wlzxy29re/outreach_2019-09-24_kriging_meetup.pdf?rlkey=8wm1diwj3usph7y6cm1o37j45&e=1&dl=0

Spherical Variogram

$$\gamma(h) = \begin{cases} c_0 + c[\frac{3}{2}\frac{h}{a} - \frac{1}{2}(\frac{h}{a})^3] & , 0 \leq h \leq a \\ c_0 + c & , h > a \end{cases}$$,where c: sill, a: range, c_0: nugget

Properties

1. $h = 0 \Rightarrow \gamma = 0$ ($c(h) = c(0)$)
2. $h = a \Rightarrow \gamma = c$ ($c(h) = 0$)
3. $h > a \Rightarrow \gamma = c$ ($c(h) = 0$)

Reference: gstools 官方 documentation
https://geostat-framework.readthedocs.io/projects/gstools/en/stable/api/gstools.covmodel.Spherical.html#gstools.covmodel.Spherical

Simple Kriging + Spherical Variogram Summary

1. Simple Kriging:

$$\begin{bmatrix} \mathrm{Cov}[\epsilon(u_1),\epsilon(u_1)] & \cdots & \mathrm{Cov}[\epsilon(u_1),\epsilon(u_n)] \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}[\epsilon(u_n),\epsilon(u_1)] & \cdots & \mathrm{Cov}[\epsilon(u_n),\epsilon(u_n)] \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} \mathrm{Cov}[\epsilon(u_1),\epsilon(u_0)] \\ \vdots \\ \mathrm{Cov}[\epsilon(u_0),\epsilon(u_n)] \end{bmatrix}$$

$$\hat{Z}(u_0) = \sum \lambda_i Z(u_i)$$

2. Unknown $\mathrm{Cov}[\epsilon(u_i),\epsilon(u_j)] \Rightarrow$ estimated by Spherical Variogram model $\gamma(h)$
3. Because $\gamma(h) = C(0) - C(h)$:

$$\begin{aligned} \mathrm{Cov}[\epsilon(u_i),\epsilon(u_j)] &= \mathrm{Cov}[Z(u_i),Z(u_j)] \\ &= C(h_{ij}) \\ &= C(0) - \gamma(h_{ij}) \end{aligned}$$

4. Substituting the result of (3.) into (1.) to obtain $\lambda_1,\ldots,\lambda_n$

Simple Kriging + Spherical Variogram 手寫推導

Let: $u_1, u_2, \dots, u_n$: n 個測站位置

$z(u_1), z(u_2), \dots, z(u_n)$ 在位置觀測到的雨量

欲估計：非觀測位置 u_0 之雨量 $z(u_0)$

Simple Kriging 假設：

$z(u) = m + \varepsilon(u)$, where m: 全域 mean, $\varepsilon(u)$: 隨機殘差

corrected variances withins neighbors.

Let. $E[\varepsilon(u)] = 0$ and $m = 0$

Kriging 估計形式.

$$\hat{z}(u_0) = m + \sum_{i=1}^{n} \lambda_i [z(u_i) - m]$$

$$= m + \sum \lambda_i \varepsilon(u_i)$$

$$= m + \hat{\varepsilon}(u_0)$$

計算目標：$\lambda_1, \lambda_2, \dots, \lambda_n$.

估計條件：① $E[\varepsilon(u)] = 0 \Rightarrow E[\hat{z}(u_0)] = m$.

② Find min. $V[\hat{z}(u_0) - z(u_0)]$

$\equiv$ Find min. $V[\hat{\varepsilon}(u_0) - \varepsilon(u_0)]$

展開 var. of error:

Define. $e(u_0) = \hat{\varepsilon}(u_0) - \varepsilon(u_0) = \sum \lambda_i \varepsilon(u_i) - \varepsilon(u_0)$

$$V[e(u_0)] = E[e(u_0)^2]$$

$$= E\left\{ \left[\sum \lambda_i \varepsilon(u_i)\right]^2 - 2\varepsilon(u_0) \sum \lambda_i \varepsilon(u_i) + \varepsilon(u_0)^2 \right\}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \mathrm{Cov}(\varepsilon(u_i), \varepsilon(u_j)) - 2\sum \lambda_i \mathrm{Cov}[\varepsilon(u_i), \varepsilon(u_0)] + \mathrm{Cov}(u_0, u_0)$$

對 $V[e(u_0)]$ 取 partial

$$\forall \lambda_k, \quad \frac{\partial}{\partial \lambda_k} V[e(u_0)] = 0$$

$$\Rightarrow 2\sum_{j=1}^{n} \lambda_j \mathrm{Cov}[\varepsilon(u_k), \varepsilon(u_j)] - 2\mathrm{Cov}[\varepsilon(u_k), \varepsilon(u_0)] = 0$$

$$\Rightarrow \sum_{j=1}^{n} \lambda_j \mathrm{Cov}[\varepsilon(u_k), \varepsilon(u_j)] = \mathrm{Cov}[\varepsilon(u_k), \varepsilon(u_0)]$$

$$\Rightarrow \begin{bmatrix} \mathrm{Cov}[\varepsilon(u_1), \varepsilon(u_1)] & \cdots & \mathrm{Cov}[\varepsilon(u_1), \varepsilon(u_n)] \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}[\varepsilon(u_n), \varepsilon(u_1)] & \cdots & \mathrm{Cov}[\varepsilon(u_n), \varepsilon(u_n)] \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} \mathrm{Cov}[\varepsilon(u_1), \varepsilon(u_0)] \\ \vdots \\ \mathrm{Cov}[\varepsilon(u_n), \varepsilon(u_0)] \end{bmatrix}$$

透過方程組解 $\lambda_1, \dots, \lambda_n$,

便可得到 $\hat{z}(u_0) = \underset{0}{m} + \sum \lambda_i [\underset{\text{測站 rain}}{z(u_i)} - \underset{0}{m}]$

$$= \sum \lambda_i z(u_i)$$

* why $m = 0$ ① 無測站地区不能算長雨量，需保守估計

② 若 $m = 0$，遠離測站地区会 $\to 0$

* $\mathrm{Cov}[\varepsilon(u_i), \varepsilon(u_j)]$ 計算：由 variogram 模型算 $\gamma(h_{ij})$（半變異數）

後轉成 covariance.

Simple Kriging + Spherical Variogram 手寫推導

$Cov[\varepsilon(u_i), \varepsilon(u_j)]$ 計算 Spherical Variogram

* 指定 $m=0$ $(E(Z(u))=0)$ $\Rightarrow Z(u) = \varepsilon(u)$

* 空間相關假設：
  ① 相關性：近 > 遠
  ② 平均值為常數
  ③ Covariance 只與距離 h 有關，與位置無關

$\Rightarrow$ Let $Cov[Z(u), Z(u+h)] = C(h)$

* Define Variogram：

$$\gamma(h) = \frac{1}{2} E[Z(u) - Z(u+h)]^2 = \frac{1}{2}[2V(Z) - 2C(h)] = C(0) - C(h)$$

($V(Z)$ let $= C(0)$)

* Sill（總變異），Range（相關距離），Nugget（0距離不確定性）

- $\boxed{\text{Sill}}$

距離遠，不相關 $\Rightarrow \lim_{h\to\infty} C(h) = 0$

$\Rightarrow \text{sill} = \lim_{h\to\infty} \gamma(h) = \lim_{h\to\infty}[C(0) - C(h)] = C(0)$

- $\boxed{\text{Range}}$：當 $C(h) \approx 0$ $(\gamma(h) \approx \text{sill})$ 時的距離 (h)

$h >$ range $\Rightarrow Z(u) \perp Z(u+h)$

- $\boxed{\text{Nugget}}$ 理論：$\gamma(0) = 0$

實務：① 量測誤差 ② 微尺度變化

$\gamma(0^+) = C_0 > 0$

* Variogram $\gamma(h)$ 未知，使用 Spherical Variogram model 擬合

- Spherical Variogram (Let nugget not exist)

$$\gamma(h) = \begin{cases} c\left[\frac{3}{2}\frac{h}{a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right], & 0 \le h \le a \\ c, & h > a \end{cases}$$

where c: sill, a: range

性質
① $h = 0 \Rightarrow \gamma = 0$ $(C(h) = C(0))$
② $h = a \Rightarrow \gamma = c$ $(C(h) = 0)$
③ $h > a \Rightarrow \gamma = c$（不再增加） $(C(h) = 0)$

* Summary：

① Simple Kriging：

$$\begin{bmatrix} Cov( ) & \cdots & Cov( ) \\ & \ddots & \\ Cov( ) & & Cov( ) \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} Cov( ) \\ \vdots \\ Cov( ) \end{bmatrix}$$

求出後得 $\hat{Z}(u_0) = \sum \lambda_i Z(u_i)$

② $Cov[\varepsilon(u_i), \varepsilon(u_j)]$ 未知 $\Rightarrow$ 用 Spherical Variogram $\gamma(h)$ 估

③ $\because \gamma(h) = C(0) - C(h)$

$$\Rightarrow Cov[\varepsilon(u_i), \varepsilon(u_j)] = Cov[Z(u_i), Z(u_i + h_{ij})] = C(h_{ij}) = V(Z(u)) - \gamma(h_{ij})$$

④ result of ③ 反代回 ① 求解 #