平成27年度　土木学会北海道支部　論文報告集　第72号

距離の重みを考えた補間・推定法

A Method of Interpolation and Estimation by Distance Weighting

北海学園大学工学部社会環境工学科　○学生員　水上裕平 (Yuuhei Mizukami)
北海学園大学工学部社会環境工学科　学生員　井田博也 (Hiroya Ida)
北海学園大学工学部社会環境工学科　フェロー　許士達広 (Tatsuhiro Kyoshi)

1. はじめに

水文確率値を求める場合、既往観測の最大値や２番目の値が飛びぬけて大きく外れ値になっている場合や、小さい順に並べると小さいデータと大きいデータで傾向が変わり、確率分布に乗らないことがしばしばあり、計画時の判断に苦慮することがある。外れ値を異常値として棄却することもあるが、近年のように異常と考えられていた自然現象が頻繁に発生している現状では、大きなデータを除外して計画することは防災上危険であり、好ましくない。

外れ値を入れて計算する場合には、対応として閾値を設けることやリサンプリング法の適用などが検討されているが、最適な方法は定まっていない。従って実務上はいろいろな方法で確率値を計算して、比較し最終的には人間が最適値を判断しているのが現状である。

この問題は縦軸に観測値、横軸にプロッティングポジションを用いて標準変量を取った時に、回帰曲線やＢスプライン等における補間や外挿・平滑化と共通の問題となる。しかしこれら多くの研究によっても、結局どのように補間推定するのが最適なのか明確になっていない。

２．距離による精度の重み

データから推定値を求める時に一般的に条件の近いデータから求めたほうが，遠い条件のデータから推定するより正確である．例えば気温と相関がある商品の売上高について，データの無い 35℃の場合を推定するのに，－5～14℃のデータから回帰を取るのと 15℃～34℃のデータの回帰を用いるのとでは，当然後者のほうの信頼性が高いと感じるであろうし，高い温度のみから求めることはあっても低い温度のみから求めることはない．時間的には「最近の情報のほうが昔の情報より信頼性が高い．」空間的には「対象となる地点に近いデータのほうが遠くのデータより実用性がある．」質的には「似た性質のものから類推したほうが違うものより正確である．」などの現象は感覚的に自明のことである．しかし通常は回帰分析においては，外挿あるいは内挿して値を求める場合，説明変数 x 方向の距離が遠く条件の大きく違うものも，条件の近いものも同等に扱われ，全てのデータの重みは等しく，全体の誤差を最小とするように推定線が定められる．

３．回帰における推定誤差と予測誤差

一般にデータ（ｘ，ｙ）に対する直線の回帰式は平均値を通る直線として以下のように表される．

$$\hat{y} = ax + b = \bar{y} + a(x - \bar{x}) \qquad \cdots 1)$$

（$\hat{y}$ ：推定値　a，b ：定数　$\bar{x}$，$\bar{y}$ ：平均値）

「推定値」の信頼性はその分散で表され，

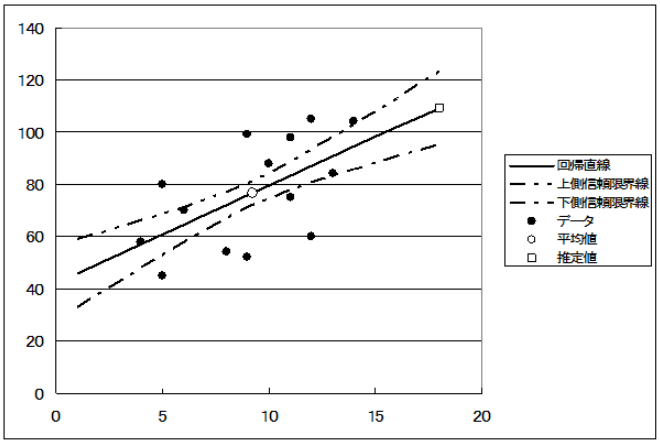
$$S_h{}^2 = \left\{ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right\} s_e{}^2 \qquad \cdots 2)$$

この時の直線の信頼区間 $\tilde{y}$ は，以下のように表される．

$$\tilde{y}_0 = \hat{y}_0 \pm t(n-2, \alpha) s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \qquad \cdots 3)$$

$t(n-2,\alpha)$ ：自由度ｎ－２，片側有意水準$\alpha/2$に対する臨界値　$\hat{y}_0$ ：回帰線上の推定値，x_0 ：$\hat{y}_0$ に対応する説明変数

$$s_e = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n-2}} \qquad S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$



図―１　信頼区間の図

図-１に示す信頼区間はそれが狭いほど推定値のばらつきが小さいことを表し，線上の点の信頼性が高くなる．当然 2）式の $S_h{}^2$ が小さい方が推定値の信頼性が高く、$S_h{}^2$ の広がりに反比例して点 $x = x_0$ における推定値の信頼性が小さくなる．これは平均値から遠くなるほど信頼性が低いという一般の認識と一致する。

４．重み付回帰による解法

２）式は以下のように書き換えられる。

$$S_h{}^2 = \frac{s_e{}^2}{n}\left\{1+\frac{(x_0-\bar{x})^2}{\sigma_x{}^2}\right\} \quad \cdots 4)$$

$\sigma_x{}^2$ ：ｘの標本分散

かっこ内の逆数は以下のように表される。

$$w = \Big/\left(1+\frac{(x_0-\bar{x})^2}{\sigma_x{}^2}\right) \quad \cdots 5)$$

これは距離により小さくなりｘ０の値の$\bar{x}$に対する重要性あるいは信頼度を表すことになる。これを個々のデータに置き換えて考えると，下から ｉ 番目の $x = x_i$ のデータによる $x = x_0$ における推定値の信頼分散 $S_{hi}{}^2$ および信頼度 w_i は，

$$w_i = \Big/\left(1+\frac{(x_0-x_i)^2}{\sigma_x{}^2}\right) \quad \cdots 6)$$

で、マハラノビスの距離の逆２乗となる。この信頼度 w_i を重みとして、重み付直線回帰式を求めれば、距離の逆２乗による重みを考えた推定値が求まる。

重みつき回帰式の定数 a^*, b^* は下記の（７）式を a^* および b^* で偏微分してその値が ０ となるときの値である．

$$E = \sum_{i=1}^{n} w_i (y_i - a^* x_i - b^*)^2 \quad \cdots 7)$$

$$a^* = \frac{\sum w_i \left(x_i - \overline{wx}\right)\left(y_i - \overline{wy}\right)}{\sum w_i \left(x_i - \overline{wx}\right)^2} = \frac{\sum w_i \sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i}{\sum w_i \sum w_i x_i{}^2 - (\sum w_i x_i)^2} \quad \cdots 8)$$

$$b^* = \overline{wy} - a^* \overline{wx} = \frac{1}{\sum w_i}\left(\sum w_i y_i - a^* \sum w_i x_i\right) = \frac{\sum w_i x_i{}^2 \sum w_i y_i - \sum w_i x_i \sum w_i x_i y_i}{\sum w_i \sum w_i x_i{}^2 - (\sum w_i x_i)^2} \quad \cdots 9)$$

なおこの推定値が意味を持つのは重みに対応する一つの推定点 $\left(x_0, \hat{y_0}^*\right)$ のみである。

５．重み付回帰曲線

この方法を用いていろいろなデータで回帰推定線を求めた。図−２に示すように重みの種類を６種類変えて比較している。重みは

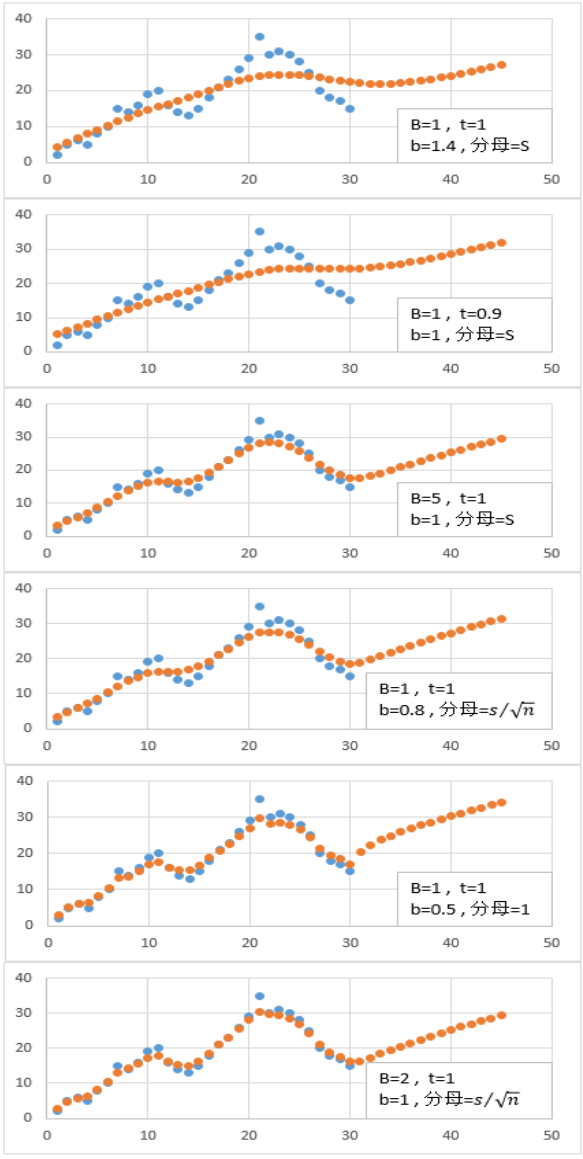
$$w = \Big/\left(1+\left(\frac{x-x_0}{s}\times B\right)^{2\times t}\right)^b \quad \cdots 10)$$

の形として変数を変動させたものである。ｂ、Ｂ、ｔ の３つの変数のうちケース毎に１つを変化させ、他は１としている。分母の分数は

$$s = \sqrt{\frac{1}{n-1}\sum (x-\bar{x})^2} \quad \cdots 11)$$

の場合と$s/\sqrt{n}$の場合、１の場合を用いている。

図-2 におけるｂ、ｔあるいはＢの１以外の値は、重み付の予測誤差を最小化した時のものである。分母ｓを標準偏差σｘとすると各点の母集団に対する距離を考えたもの、分母をσｘ/√ｎとすると各点間の距離を考えたものとなる。図の下から ２ 番目はσｘ＝１としたものである。重み付予測誤差による最適値はｂあるいはｔが１ を前後して変化し、データ数あるいはデータｘの最大の幅と相関がみられる。



図―2　重み付回帰曲線