

観測値の距離の重みを考えた補間および推定法について

A Method of Interpolation and Estimation by Distance Weighting of Observation Data

北海学園大学工学部社会環境工学科	○学生員	東龍弥 (Ryuya Azuma)
北海学園大学工学部社会環境工学科	学生員	清瀬祐一 (Yuichi Kiyose)
北海学園大学工学部社会環境工学科	フェロー	許士達広 (Tatsuhiko Kyoshi)

1. はじめに

水文確率値算定には、その確率分布の選定や算定方法などの課題があるが、それ以外にも外挿や補間に関する問題が従来から議論されている。水文確率値を求めるとき、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たり、データを大きい順にプロットすると途中でデータの勾配の傾向が変わるなど、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。本研究ではこの解決のためデータと推定値の間に距離の重みを考えて、推定値を求める方法を検討する。

2. 推定の問題点

図-1に示すように確率値は確率分布の超過確率を示す端部の形状で決まる。確率分布の形状は通常観測されたデータ全体の分布で決めているが、端部のデータの傾向が他の確率の低い部分と異なっていれば、求める確率値付近の推定値と合わないことになる。したがってデータ全体に合わせるのではなく端部のデータに合わせればよいという考え方が生ずるが、どこで区分するのが良いかは理論がなく、現状では分からない。

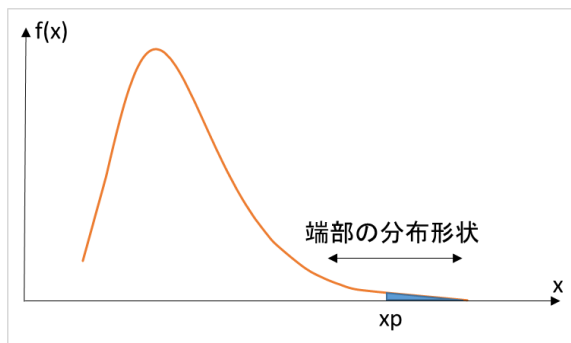


図-1 分布と推定値の関係

同様に分布の縦軸に観測値、横軸にプロットングポジションによる標準変量を取れば図-2のようになる。

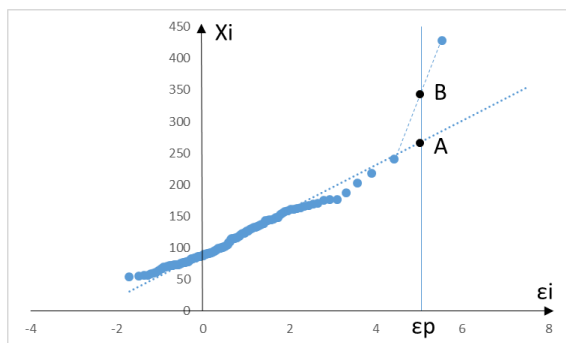


図-2 データが多い場合の確率値

近年は過去百何十年間の降雨データが得られている観測所が多くあり、その場合を図2で考えれば150年確率値は通常は全体のデータからA点となるが、上から第1位と2位の間のB点で求めて良いのではないかと、あるいはプロットの並びが直線に近い上位何点かで決めて良いのではないかとという考え方ができる。このためにデータを区分し閾値を考えその範囲でリサンプリング法を用いることなどが検討されているが、部分的に点列の傾向が変わっても、少し離れると同じ傾向に戻るなど、明確に区分できないことが多く生ずる。閾値の設定法はいろいろな分野で考えられているが、良い方法は見出されていない。

また分布に合わせるために大きな値を異常値（外れ値）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状況では、大きなデータを人為的に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。

3. 通常座標での研究の現状

この問題は確率分布ばかりでなく通常のデータによるX-Y座標上の推定でも同様に存在する。関数を当てはめるパラメトリックな方法では解消できないため、データに距離の重みをつけて推定するノンパラメトリック回帰といわれる補法が考えられている。これはカーネル密度関数やBスプライン等における補間推定などであり、回帰における頑健性・外れ値問題にも通ずる内容で、多次元に拡張すれば地理学や、計量経済学における空間統計の問題にも共通する。それぞれの分野で手法が独自の見地から研究され発展しているが、現状では以下のような課題がある。

- ・データから推定値への距離による影響の形状が分からないため、正規分布などが便宜的に使われている。
- ・影響を最適化するためにクロスバリデーションや情報量基準などが用いられるが、データ分布により当てはまりが悪い場合が多い。
- ・一般的に計算が複雑になり計算時間がかかる。

これらの状況から、結局どのようにデータから最適値を推定するのかはまだ各分野で研究段階にあり、一般の技術的解析では分からないこととなっている。

4. 距離による重みの導入

ここでは距離の重みを以下のように考える。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。

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

($\hat{y}$: 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以下のようになる。

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

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

2) 式は以下のように書き換えられる。

$$S_h^2 = \frac{s_e^2}{n} \left\{ 1 + \frac{(x_0 - \bar{x})^2}{\sigma_x^2} \right\} \quad \cdots 3)$$

σ_x^2 : x の標本分散

かつこの逆数は以下のように表される。

$$w = \frac{1}{1 + \frac{(x_0 - \bar{x})^2}{\sigma_x^2}} \quad \cdots 4)$$

これは距離により小さくなり x_0 の値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 $x = x_0$ における推定値の重み (信頼度) w_{hi} は、

$$w_{hi} = \frac{1}{1 + \frac{(x_0 - x_i)^2}{\sigma_x^2}} \quad \cdots 5)$$

これは「標準化距離の2乗プラス1」の逆数である。距離による影響は y 方向にも存在し重みは以下のように表される。

$$w_{vi} = \frac{1}{1 + \frac{(\hat{y}_i - y_i)^2}{s_e^2}} \quad \cdots 6)$$

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

x, y 両方向に重みをかける場合は合成して

$$w_i = w_{hi} * w_{vi} \quad \cdots 7)$$

この信頼度 w_i を重みとして、各推定点 x_0 ごとに重み付直線回帰式を求めれば、距離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 8)

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。

$$E = \sum_{i=1}^n w_i (y_i - a^* x_i - b^*)^2 \quad \cdots 8)$$

$$a^* = \frac{\sum w_i (x_i - \bar{wx}) (y_i - \bar{wy})}{\sum w_i (x_i - \bar{wx})^2} \quad \cdots 9)$$

$$= \frac{\sum w_i \sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i}{\sum w_i \sum w_i x_i^2 - (\sum w_i x_i)^2}$$

$$b^* = \bar{wy} - a^* \bar{wx} = \frac{1}{\sum w_i} (\sum w_i y_i - a^* \sum w_i x_i)$$

$$= \frac{\sum w_i x_i^2 \sum w_i y_i - \sum w_i x_i \sum w_i x_i y_i}{\sum w_i \sum w_i x_i^2 - (\sum w_i x_i)^2} \quad \cdots 10)$$

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 $(x_0, \hat{y}_0^*)$ のみである。

この方法で $X-Y$ データに曲線を当てはめたものを示す。赤がデータ、青が推定値である。人間の感覚に近い線が描けている。確率データに対しては、横軸の標準変量の間隔が変化するなどの違いがあり、修正する。

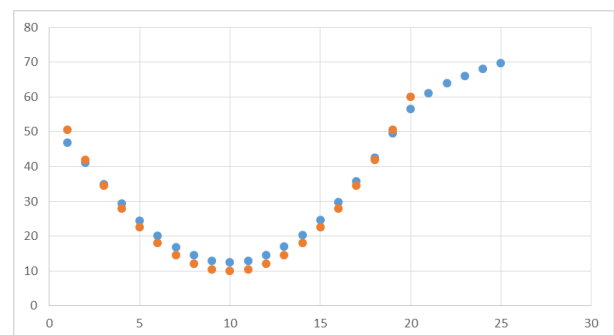
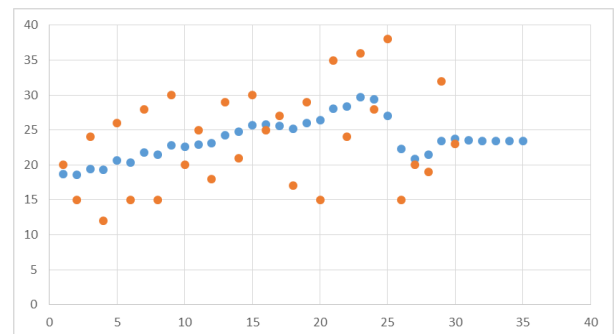
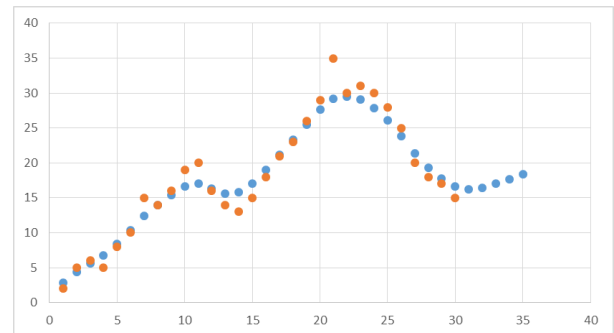


図 - 3～5 距離の重みを用いた推定