[研究新領域報導]

空間統計簡介

中央研究院統計研究所　黃信誠

前　言

空間統計顧名思義為分析空間資料的統計方法，因其與各類實際問題之緊密關係，是一個快速發展的領域。主要想法在於空間中鄰近的資料通常比相離較遠的資料具有較高的相似性，空間統計乃透過位置建立資料間的統計關係。其應用的範圍包羅萬象，包括地質、大氣、水文、生態、天文、遙測、地震、環境監測、流行病以及影像處理等。此外，任何其他領域如所收集的資料與位置有關，亦可為空間統計研究之範疇。除了極少數的例子，真實世界的空間資料大多不能為物理及化學機制以簡單的公式描述。為解決資料中所隱含的空間不確定因素，空間統計模型乃嘗試從凌亂的空間資料中，以統計方法發掘空間變動 (spatial variation) 之規律。

空間資料之分析與傳統的統計分析主要有兩大差異：(1)空間資料間並非獨立，而是在 d 維空間中具有某種空間相關性，且在不同的空間解析度下呈現不同之相關程度；(2)因地球只有一個，大多數空間問題僅有一組（不規則在空間分佈的）觀測值，而無重複觀測的資料。因此，空間現象的了解與描述是極為複雜的，而傳統的統計分析技巧，尤其是建立在獨立樣本的統計方法，並不適合用來分析空間資料。其與時間序列最大的差異在於空間中並無過去、未來之次序，因而不易透過某種因果關係的描述來建構空間模型。目前空間統計模型大致可分為三類：地理統計 (geostatistics)、格點空間模型 (spatial lattice model)以及空間點分佈型態 (spatial point pattern)。以下就各類分別敘述。

地理統計

地理統計主要用於分析地質、大氣、水文等與地理有關之空間資料。例如，一礦區中礦物的含量、空氣中懸浮微粒之濃度等。因所描述的變量大多在空間中呈現連續變化，通常假設其由一個連續空間的 d 維隨機過程（隨機域）$\{Z(\boldsymbol{s}):\boldsymbol{s}\in D\subset\Re^d\}$ 產生。地理統計方法乃透過已知（且通常是不規則分佈的）n 個位置 $\{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_n\}$ 的觀測資料 $\{Z(\boldsymbol{s}_1),\ldots,Z(\boldsymbol{s}_n)\}$，建構適切的隨機過程模型，從而做有效合理的統計推論。其中一個主要之問題為空間預測 (spatial prediction)，即藉由 n 個位置的資料建構空間關係，以預測一區域中任意地點（或區塊）的變數值。此一預測方法在地理統計文獻中通常叫做 kriging。

地理統計的主要模型為：

$$Z(\boldsymbol{s})=m(\boldsymbol{s})+h(\boldsymbol{s})+e(\boldsymbol{s});\ \ \boldsymbol{s}\in D,$$

其中 $m(\cdot)$ 為一非隨機 (deterministic) 之平均函數，用以表示大尺度的空間變化趨勢，$h(\cdot)$ 為一平均值為零之隨機過程，用以表示較小尺度的空間變化趨勢，亦為建構空間相關性之主要結構，而 $e(\cdot)$ 為一白噪 (white noise) 隨機過程，代表雜訊。Kriging 方法乃透過 n 個位置的觀測值 $\{Z(\boldsymbol{s}_1),\ldots,Z(\boldsymbol{s}_n)\}$ 預測

$$S(\boldsymbol{s})\equiv m(\boldsymbol{s})+h(\boldsymbol{s});\ \ \boldsymbol{s}\in D.$$

然而我們所觀測的資料僅是連續空間之隨機過程中的一種可能的一個極小部分（即 n 個點），除非對此模型做進一步的假設，否則無法做任何有效的統計推論。通常 $m(\cdot)$ 假設為常數或一些已知函數（多項式或其他解釋變數之函數）的線性組合，$h(\cdot)$ 則假設為一個內在平穩的隨機過程 (intrinsically stationary process)，即

$$E(h(\boldsymbol{s}+\boldsymbol{h})-h(\boldsymbol{s}))=0,$$

$$\mathrm{var}(h(\boldsymbol{s}+\boldsymbol{h})-h(\boldsymbol{s}))=2g(\boldsymbol{h}).$$

其中 $2g(\cdot)$ 稱做 variogram，為地理統計中描述空間相關性的主要因子。此函數通常隨空間中兩點的距離有一遞增的趨勢，且在數學上必須滿足條件負定 (conditional negative definiteness)，也就是對任意有限位置 $\{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_m\}$ 及滿足 $\sum_{i=1}^m a_i=0$ 的任意實數 $\{a_1,\ldots,a_m\}$，皆有以下關係：

$$\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j 2g(\boldsymbol{s}_i-\boldsymbol{s}_j)\le 0\,.$$

模型之參數可用概似函數法 (likelihood method) 或樣本動差法推估。

近年來時空資料分析的需求與日遽增，時空模型之建構因而成為空間統計的一個主要發展方向。如何結合空間統計與時間序列方法以描述大氣資料、衛星資料、及環境監測等資料所呈現複雜的時間與空間關係，及如何分析由衛星或其他自動監測方式所帶來極大量時空資料，並從中發掘出有用的資訊，皆將對空間統計之理論及應用方面帶來新的發展。其他如空間模型診斷、空間模型選擇、非平穩 (nonstationary) 空間模型之建構、及由一空間解析度的資料去推論另一解析度的空間關係等問題，皆有待進一步研究發展。

格點空間模型

格點空間模型用以描述分佈於有限（或無窮離散）空間點（或區域）上資料的空間關係。例如，在流行病學中欲以地理區域（如縣市、鄉鎮）為單位之發病個數資料，研究疾病發生率與地理位置的關係，及在影像處理中欲從扭曲或帶有雜訊之數位影像（如醫學或衛星影像）資料，重建背後的真實影像，皆為此類空間統計方法研究之範疇。主要之統計模型為馬可夫隨機域 (Markov random field)，以下簡稱 MRF。MRF 乃一條件空間模型 (conditionally specified spatial model)，其空間隨機變數之聯合機率分佈並非直接建構，而是間接地透過一組條件機率建構；這組條件機率描述空間任一位置之隨機變數給定其鄰近區域隨機變數之機率分佈。此一建構方法的主要優點在於很多實際空間問題，不易描述整體之空間關係，但卻很容易透過局部空間關係的描述，建構適當的條件空間模型。

假設 $\{q(\boldsymbol{s}_1),\ldots,q(\boldsymbol{s}_n)\}$ 為空間中定義在 n 個位置 $\{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_n\}$ 上之隨機變數。MRF 依據以下條件機率建構其局部空間關係：

$$p(q(\boldsymbol{s}_i)\mid\{q(\boldsymbol{s}_j):j\ne i\})=p(q(\boldsymbol{s}_i)\mid\{q(\boldsymbol{s}_j):j\in N_i\});\ i=1,\ldots,n,$$

其中 $N_i\subset\{1,\ldots,n\}$，稱為 $\mathbf{s}_i$ 之鄰域 (neighborhood) 集。例如，定義在二維長方形格點上之 MRF，$\mathbf{s}_i$ 之一鄰域選擇方式為最接近 $\mathbf{s}_i$ 的上下左右四點，空間關係則透過 $\mathbf{s}_i$ 與上下左右四點的條件機率關係建立。建構 MRF 之困難在於任意給定之一組條件機率並不保證背後存在一隨機過程有如此的條件機率分佈，欲構成一有效的聯合機率分佈，這些條件機率間必須滿足某種複雜的一致關係。此一困難因 Hammersley 和 Clifford [6] 發現了 MRF 與以下統計模型的關係而解決：

$$p(q(\boldsymbol{s}_1),\ldots,q(\boldsymbol{s}_n))\propto \exp\left\{-\sum_{i=1}^{n}V_i(q(\boldsymbol{s}_i))-\sum_{1\le i<j\le n}V_{\{i,j\}}(q(\boldsymbol{s}_i),q(\boldsymbol{s}_j))-\cdots -V_{\{1,\ldots,n\}}(q(\boldsymbol{s}_1),\ldots q(\boldsymbol{s}_n))\right\}.$$

此模型稱為 Gibbs field，他們證明了其與 MRF 的等價關係。Besag [1] 因而據此建構了一系列的 MRF 模型。

MRF 可用以分析數位影像。貝氏影像分析 (Bayesian image analysis) 方法即根據所欲重建影像 $\mathbf{q}$ 之特性，先以某類 MRF 模型所建構之空間關係做為其先驗 (prior) 分佈 $p(\mathbf{q})\equiv p(\{q(\boldsymbol{s}_i):i=1,\ldots,n\})$。再依循所觀測的資料 $\mathbf{Z}\equiv(Z(\boldsymbol{s}_1),\ldots,Z(\boldsymbol{s}_n))'$ 與真實影像 $\mathbf{q}$ 的條件機率關係 $f(\mathbf{Z}\mid\mathbf{q})$，得出其事後 (posterior) 分佈：

$$p(\mathbf{q}\mid\mathbf{Z})=\frac{p(\mathbf{q})f(\mathbf{Z}\mid\mathbf{q})}{\int p(\mathbf{t})f(\mathbf{Z}\mid\mathbf{t})d\mathbf{t}}.$$

最後依據適當的準則，得出影像 $\mathbf{q}$ 的貝氏估計量，例如事後分佈之眾數 (posterior mode)。此方法之主要困難在於計算，因為一數位影像通常為 $n_1\times n_2=64\times 64$（或 512×512）或更大之畫素 (picture element) 構成，亦即 $\mathbf{q}$ 乃一數千至數十萬維度之隨機過程，如每一畫素有 $K=16$（或 256）種不同灰階，則隨機過程 $\mathbf{q}$ 即有 $K^{n_1\times n_2}$ 種不同可能之影像。此一極高維度之計算問題本為一不可能之任務，因近代計算機的發展及 Markov chain Monte Carlo (MCMC) 法的提出而得以進行，其想法在於透過一簡單機制建構一 Markov chain 使其極限分佈為所欲生成之分佈。例如，Geman and Geman [5]提出模擬冷卻 (simulated annealing) 法，用以找尋事後分佈之眾數 $\hat{\mathbf{q}}$。他們考慮以下的分佈：

$$p_T(\mathbf{q}\mid \mathbf{Z}) = \frac{\left(p(\mathbf{q})f(\mathbf{Z}\mid \mathbf{q})\right)^{1/T}}{\int \left(p(\mathbf{t})f(\mathbf{Z}\mid \mathbf{t})\right)^{1/T} d\mathbf{t}},$$

其中 T 代表此一系統之溫度。當溫度$T=1$時，$p_T(\mathbf{q}\mid \mathbf{Z})$即為其事後分佈；當溫度$T\to\infty$時，$p_T(\mathbf{q}\mid \mathbf{Z})$趨近一均勻分配；當溫度$T\to 0$時，$p_T(\mathbf{q}\mid \mathbf{Z})$逐漸集中在一點。模擬冷卻法乃結合 MCMC 法及一降溫之過程$T(t)\to 0$，在不同時間 t 以不同溫度$T(t)$藉由 Gibbs sampler 從$p_{T(t)}(\mathbf{q}\mid \mathbf{Z})$中抽樣取得$\hat{\mathbf{q}}_\tau$。在一適度的降溫條件下，Geman and Geman 證明$\hat{\mathbf{q}}_\tau \to \mathbf{q}_*$。

目前 MRF 仍有許多問題有待解決。例如，Gibbs field 中不易處理之標準化積分常數造成參數估計之困難、邊界效應 (boundary effect) 的處理、及 MRF 之極限理論等。此外，MRF 雖對影像重建、影像分類、及紋理分割等問題皆有極好的表現，然而其計算卻常需倚賴耗時的 MCMC 法。如何同時兼顧計算效率與統計精度，將是未來一項極大的挑戰。

空間點分佈型態

在自然科學中，許多資料為點（或小區域）所構成的集合。例如，地震發生地點之分佈、樹木在森林中之分佈、某種鳥類鳥巢之分佈、生物組織中細胞核之分佈、及太空中星球之分佈等。我們稱此類資料為空間點分佈型態，並稱其中點之位置為事件。空間點分佈型態因背後形成的機制不同而造成隨機、叢聚或規則等不同分佈型態。藉由空間點分佈型態的研究，我們可以找尋叢聚之所在，並瞭解其背後形成之原因及其可能產生的影響。

空間點分佈型態通常由一個 d 維的空間點過程 (spatial point process) 描述。此類模型之隨機機制在於位置本身，其中最基本的空間點過程為 homogeneous Poisson 點過程，主要假設有二：(1)在空間中任一集合所包含事件之個數為一 Poisson 分佈，且其分佈之平均值與其集合面積（體積）成正比；(2)不相交之集合所包含事件之個數呈統計獨立關係。此類點過程通常用以定義所謂完全空間隨機 (complete spatial randomness) 之點分佈型態，以別於叢聚或規則之分佈。初步的統計分析在於檢定一組點資料的分佈型態是否為完全空間隨機，此檢定可透過數個不相交區域事件個數之分佈、或事件與事件最短距離之分佈等，構造適當的檢定量。當此檢定被拒絕時，我們可以進一步配適其他的空間點過程模型。例如，叢聚型態可用 inhomogeneous Poisson、Cox、或 Poisson cluster 等點過程建構，規則型態可用 inhibition 點過程建構，至於在小尺度呈現規則型態而在大尺度呈現叢聚型態則可用 Markov 點過程建構。

假設 $N(A)$ 代表在 A 區域內發生事件的個數。點過程之第一階性質可用以下之強度函數 (intensity function) 描述：

$$l(\mathbf{s}) = \lim_{|\mathrm{d}\mathbf{s}|\to 0} \frac{E(N(\mathrm{d}\mathbf{s}))}{|\mathrm{d}\mathbf{s}|},$$

其中 $\mathrm{d}\mathbf{s}$ 為包含 $\mathbf{s}$ 點之一極小區域，$|\mathrm{d}\mathbf{s}|$ 為 $\mathrm{d}\mathbf{s}$ 之面積(體積)。通常假設點過程為平穩 (stationary) 且無方向性 (isotropic)，即其機率結構不隨任意平移和旋轉而改變，在此一假設下 $l(\mathbf{s})\equiv l$，為一常數。一平穩且無方向性之點過程的第二階性質可用以下之 K 函數描述：

$$K(d) = \frac{1}{l}E\,(\text{與任一事件距離}\,d\,\text{以內之其他事件的總數})。$$

此函數可描述點過程之空間相關性，其角色及重要性如同地理統計中之 variogram，故在空間點過程中佔有極重要的地位。參數估計主要以最大概似法或透過 K 函數以最小平方法為主。其統計性質之結果尚少，有待更進一步研究。

結　語

本人僅對空間統計做一極簡略的介紹，範圍局限於單變量之空間統計問題。空間資料可說隨處可見，其發展起因於各類實際問題，其內容也因此而日漸豐富，而其所蘊含有趣的理論及應用問題，則有待更多研究學者及學生投入解決。有興趣的讀者可參閱 Cressie [3]，此書對空間統計各方面均有一完整的介紹。其他在地理統計方面可參閱 Matern [7], Wackernagel [10], Stein [9] 及 Chiles and Delfiner [2]，其中 Stein 在極限理論方面有較深入的探討；在格點空間模型於影像處理方面可參閱 Geman and Geman [5] 和 Winkler [11]；在空間點分佈型態方面可參閱 Ripley [8] 和 Diggle [4]。

參考文獻

[1] J. E. Besag, *Journal of the Royal Statistical Society B*, **36**, 192-225 (1974).

[2] J.-P. Chiles and P. Delfiner, *Geostatistics: Modeling Spatial Uncertainty*. Wiley, New York (1999).

[3] N. Cressie, *Statistics for Spatial Data, revised edition*. Wiley, New York (1993).

[4] P. J. Diggle, *Statistical Analysis of Spatial Point Patterns*, Academic Press, New York (1983).

[5] S. Geman and D. Geman, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-6**, 721-741 (1984).

[6] J. M. Hammersley and P. Clifford, Markov fields on finite graphs and lattices. Unpublished manuscript, Oxford University (1971).

[7] B. Matern, *Meddelanden fran Statens Skogsforskningsinstitut,* **49**, No. 5 (1960). [Second edition (1986), *Lecture Notes in Statistics*, **No. 36**, Springer, New York.]

[8] B. D. Ripley, *Spatial Statistics*. Wiely, New York (1981).

[9] M. L. Stein, *Interpolation of Spatial Data*, Springer-Verlag, New York (1999).

[10] H. Wackernagel, *Multivariate Geostatistics*, revised edition, Springer-Verlag, Berlin (1998).

[11] G. Winkler, Image Analysis, Random Fields and Dynamic Monte Carlo Methods. Springer-Verlag, Berlin (1995).