《人文及社會科學集刊》
第十九卷第四期（96/12），pp. 529-562

區位推論的統計模型：發展、比較與評估*

黃信豪
國立政治大學政治學系博士候選人

所謂「區位推論」，爲當研究者可得資料爲集體層次，而研究旨趣與目的是個體層次關聯性，因而產生研究目的與資料分析單位層次不相符之情形。如何運用適當的統計方法與程序，從已知的總體集合資訊推估未知個體層次關聯性，即是區位推論首要關注的問題。

本文的主要目的，在於回顧區位推論模型的發展歷程，同時比較幾個主要區位推論模型在估計上的表現。筆者首先回顧了區位推論模型的發展，以Goodman 迴歸、King EI 模型以及階層模型爲例，說明這三種主要的區位推論模型於統計元素上的特點。接著，運用模擬實驗的研究設計，筆者嘗試評估三者模型在「不偏性」與「有效性」統計性質的表現。在比較不同已知資訊分佈型態後，本文的模擬實驗結果顯示，在放寬估計係數的機率分配假定、置入較精確的函數形式以及更爲彈性的階層結構後，Wakefield 的階層模型具有較佳的估計性質，同時也能有效的矯正極端值的估計。

關鍵字：計量方法、區位推論、貝式統計

＊本論文使用資料全部係採自「行政院主計處」進行之「人力資源調查」94 年 12 月資料。該資料由中央研究院調查研究專題中心釋出，感謝上述機構及相關人員提供資料協助。筆者也感謝政大政治系黃紀講座教授對本文在研究設計上的諸多建議，以及承蒙兩位匿名審查人所提出的精闢見解，讓論文修改後更具可讀性，當然，本論文內容由作者自行負責。
收稿日期：96 年 2 月 26 日；接受刊登日期：96 年 5 月 8 日

1、前言

自 Robinson（1950: 351-357）說明不同分析單位相關係數產生的集合（aggregation）差異以來，學界已廣泛認知到從集體資料來推論個體層次關聯性的困難。當研究者「錯誤地」以集體資料來指涉個體（人）特性，並進行關聯性分析時，將會犯了區位推論謬誤（ecological fallacy）。然而，對某些個體資料無法可得的研究議題（如無法追溯的歷史事件），集體資料卻是這些議題唯一的資料來源。因此，如何運用適當方法與程序以集體資料來探求個體層次變項間關聯性，即是「區位推論」方法上的關切焦點。

所謂「區位推論」（ecological inference），爲當研究者可得資料爲集體層次，而研究旨趣與目的在個體層次關聯性時，所產生層次不相符的情形。例如，我們欲探索個別選民在同時（期）選舉裡，對不同職位產生的一致與分裂投票行爲。即便從民調個體資料我們可探索選民的心理動機，但以全國性調查樣本規模而言，就特定地域上樣本數不足所產生資料代表性之問題，將成爲運用個體資料時所必須面臨的難題；另一方面，若以政府部門依各級地理區域（如投開票所、村里、鄉鎮市等）所出版的選舉實錄來進行數據分析，則僅可得依特定區域所加總之結果。也就是說，這些集體資料雖然客觀地呈現各黨在不同職位選舉結果的得票情況，但由於資料匯集僅至投開票所的地域單位，在喪失數據之間個體選民的聯繫下，將使依集體資料推論至個體行爲顯得困難。但若因此偏廢客觀的集體數據而僅著重個體民調資料，則又顯得本末倒置。因而如何有效整合不同資料所呈現之訊息，實爲方法論學者不得不去審愼面對之課題（黃紀，2001: 548）。

不論方法或應用上，「區位推論」在國內外皆屬快速發展中之議題。在國內的相關研究裡，黃紀（2001: 554-560）首先說明「區位推論」的概念，並闡述如何適當地運用總體資料於個體層次因果假設的驗證與推論。就區位推論的統計模型的應用上，黃紀與張益超（2001: 183-218）則開始採用 Gary King 的 EI 模型，分析 1997 年嘉義市長與立委選舉裡選民的一致與分裂投票之行爲；徐永明（2000: 167-196）與駱明慶（2006: 639-669），也分別以該

模型對於「宋楚瑜現象」與「廢票來源」等重要議題，運用集體資料所推論出之資訊，作爲其實證研究之基礎。就方法上，King 的 EI 模型可說是當前國內學者在處理區位推論問題時，最主要的估計模型。

在區位推論模型的發展過程中，不可否認地，EI 模型對估計「係數」開啓了另一個新紀元。[1] 而近年來階層模型的快速發展，也成爲方法論學者在區位推論模型中熱切關注的焦點。爲了更系統性地瞭解區位推論模型的發展歷程與不同模型的特性，在本文中，筆者將以「區位推論模型」爲研究焦點，從模型設定（model specification）、參數估計（parameter estimation）與機率分配（probability distribution）等統計元素，來介紹區位推論統計模型的發展。[2] 其次，經由模擬實驗設計來評估 Goodman 迴歸、King EI 模型與 Wakefield 階層模型的估計表現。在不同已知資訊分佈的研究設計下，我們將從不偏性（unbiasedness）與有效性（efficiency）的估計元特性作爲評斷標準。模擬實驗結果顯示，階層模型的估計結果在總體區域與個別區域觀察值上皆較爲準確，具有較佳的估計特性，同時模型也擁有較好的統計性質來估計極端值。因此，若我們嘗試以總體資料來對個體層次關聯性從事區位推論時，相較於 King EI 模型來說，筆者認爲階層模型應是較好的選擇。

以下，我們將首先說明區位推論的問題本質與特性。第 3 節的部分，則以 Goodman 迴歸模型、King EI 模型以及階層模型三者爲討論的主軸，除了探討區位推論統計模型之發展歷程外，也將說明這些模型的特點。在回顧學界對不同模型的模擬實驗結果後，我們將於第 4 節介紹本研究的模擬實驗設計與資料來源；第 5 節是區位推論模型模擬的實證結果；在結論中筆者除簡述研究發現外，也嘗試探討區位推論模型未來可能發展方向。

---

1 在本文中，筆者所指涉的「係數」（coefficient）爲研究者欲估計個體層次關聯性之數值，不同於描述機率分配性質的「參數」（parameter）。

2 總的來說，區位推論的估計可分爲判定性途徑（deterministic approach）與統計途徑（statistical approach）兩者（King et al., 2004: 1-12）。所謂判定性途徑，自 Duncan 與 Davis (1953: 665-666) 說明各估計參數可能的上下限範圍以降，學界則透過數值方法的運用，來對各估計參數找尋出唯一解（Johnston and Pattie, 2000: 333-345），但由於最大亂度法（maximize entropy）的估計特性，使得推估結果往往與眞實狀況不盡相符。

2、「區位推論」的問題本質與特性

總的來說，「區位推論」所關注的是研究者所得資料與研究課題兩者層次不相符的問題。若我們探索之研究主題爲個體層次變項之關係，但所得資料卻是變項各自的聚集資料時，則總體資料將使我們無法確切得知變項裡各項次聯合分佈（joint distribution）之情況。因此，區位推論是在於如何就已知總體資訊，推導出未知聯合分佈或條件機率之數值。在本節裡，我們將首先討論區位推論的問題本質以及資料特性。

從資料屬性而論，總體資料是依地理區域或特定單位聚集而成之集體數據。在層次高於個體的匯集下，從個體來看將必然出現資訊漏失（information loss）的結果，造成集體偏誤（aggregation bias）。例如我們運用集體數據來推估不同省籍選民的投票率差異，同時觀察到本省籍選民比例較高的區域中投票率也較高的結果。從資料上來看，或許我們會直觀地得出「本省籍民眾投票率較高」的結論。但在某些特定省籍聚集程度較高的區域中，則很可能讓弱勢族群產生受威脅的感受，因而使「本省籍、投票」的估計係數過度高估。換句話說，在區位推論中我們所擁有的資料畢竟仍以地區爲單位，故無法完整地還原至個體之結果。King（1997: 46-53）便認爲，除非我們能夠基於變數類別的分佈進行編組，或將其作爲蒐集總體資料的依據，來得到完全同質性（homogeneous）的分佈數據，否則「集體偏誤」的問題將無可避免。

因此，正如 Freedman 等人（1998: 1518-1522）所強調的：「若能夠避免由總體資料來對個人特性從事區位推論，則應避免。」在區位推論的過程中，研究者永遠無法將總體資料還原成個體狀態之限制。然而，此非必然表示研究者對總體資料應束諸高閣，甚至避之唯恐不及；反之，在一些個體資料缺漏、或擁有個體資料但有代表性之虞的研究主題上，總體資料仍具理論與實務上之優勢與重要意義。因此，如何適當地運用統計程序透過總體資訊來進行推估，進而呈現總體資料之可靠訊息，將是區位推論課題之核心意義。

接下來，我們將說明「區位推論」之資料特性。下表 1 是典型區位推論

表 1　典型的「區位推論」問題：已知資訊與未知資訊

	$Y_i=1$	$Y_i=0$	
$X_i=1$	$n_{\beta_i}(\beta_i)$	$n_{X_i}-n_{\beta_i}(1-\beta_i)$	$n_{X_i}(P_{X_i})$
$X_i=0$	$n_{W_i}(W_i)$	$n_i-n_{X_i}-n_{W_i}(1-W_i)$	$n_i-n_{X_i}(1-P_{X_i})$
	$n_{T_i}(P_{Y_i})$	$n_i-n_{T_i}(1-P_{Y_i})$	$n_i(1)$

說明 1：細格內的符號爲次數（frequencies），跨弧內的則爲比例（proportions），β_i 與 W_i 爲條件機率。

說明 2：細格內灰底者爲未知資訊。在本文中的說明，以及隨後不同區位推論模型的介紹裡，筆者將以此表符號系統爲基準。

問題裡，針對已知與未知資訊的具體表達方式。[3] 對許多社會科學議題來說，在最簡要的狀況裡 X 表示的是如性別或種族變項，Y 爲是否投票、犯罪等二分變項。假定我們擁有 i 個聚集單位的總體資料（如投開票所、鄉鎮市、選區等），那麼，針對所有已知資訊與未知資訊，我們皆可用「次數」與「比例」來表示。[4]

在我們所得的聚集資料裡，每一個總體觀察值皆可以如此的交叉表來表示。在此，已知資訊是 X_i 與 Y_i 各類別的分佈或比例，而細格內部如 $P(Y_i=1|X_i=1)$ 的條件機率或聯合機率次數之數值則是未知資訊，同時也是研究者所感興趣之標的。因此，區位推論的目的即在於如何從邊際總和（marginal total），推導出細格內的條件或聯合機率數值。在限定的 $n_i(1)$下，顯然地我們只要估算出係數 $n_{\beta_i}(\beta_i)$ 或 $n_{W_i}(W_i)$，其他未知係數即可換算。更重要的是，上表說明以下恆等式的存在（accounting identity, Goodman, 1953: 663-664）：

$$n_{T_i}=n_{\beta_i}+n_{W_i}$$

$$\text{故 } P_{Y_i}=P_{X_i}\beta_i+(1-P_{X_i})W_i \qquad (1)$$

3 本表部分構想來自於 Mattos 與 Veiga（2004: 353），並由筆者進一步整理之。

4 在區位推論中，每一個觀察值的邊際總和與總個數皆是母體資訊，故理論上我們應以 N 來表示如此的母體數據。但在全體 i 個觀察值裡，每一個交叉表卻通常具有各自獨特的數值，爲了便利往後在抽樣分配與統計模型上數學符號的表示，在這裡筆者將以 n_i 表示之。如此也說明了區位推論的資訊特質與一般個體層次資料相當不同，筆者感謝審查人的細心提醒。

恆等式(1)可說是區位推論統計模型發展之起點。這表示當研究者嘗試運用任何的區位推論模型來估算係數 $n_{\beta_i}(\beta_i)$ 與 $n_{W_i}(W_i)$ 時，估計結果將必須符合此恆等式。除此之外，由於邊際總和是已知眞實資訊，因此細格內的係數必定要符合邊際總和之限制，以此我們可以計算出未知係數最大值與最小值之上下限（Duncan and Davis, 1953: 665-666）：[5]

$$\max\left(0, \frac{P_{Y_i}-(1-P_{X_i})}{P_{X_i}}\right) \leq \beta_i \leq \min\left(\frac{P_{Y_i}}{P_{X_i}}, 1\right)$$
$$\max\left(0, \frac{P_{Y_i}-P_{X_i}}{1-P_{X_i}}\right) \leq W_i \leq \min\left(\frac{P_{Y_i}}{1-P_{X_i}}, 1\right) \quad (2)$$

就 β_i 而言，其最小值必然大於 0 與 $P_{Y_i}-(1-P_{X_i})/P_{X_i}$，兩者之中數值較大者，最大值則必然小於 1 與 P_{Y_i}/P_{X_i} 其中數值較小者，W_i 亦然。除了以上兩者特性外，由於 P_{Y_i} 爲 β_i 與 W_i 的線性組合，因此也具以下的恆等關係（King, 1997: 80）：

$$W_i=\frac{P_{Y_i}}{1-P_{X_i}}-\frac{P_{X_i}}{1-P_{X_i}}\beta_i \quad (3)$$

此三者已知資訊與未知資訊間之恆等關係，將是區位推論模型中，進行估計所必須依循的限制條件，也成爲往後區位推論統計模型發展的重要基礎。在說明區位推論的問題本質與特性後。接下來筆者將透過模型設定、機率分配與參數估計等重要統計元素，來說明區位推論統計模型的發展。

3、區位推論統計模型的發展

回顧區位推論統計模型發展的歷程，[6] Goodman（1953: 663-664, 1959:

5 有關上下限的數學證明，可見 King（1997: 301-303）。

6 在此，筆者主要從模型發展的角度來進行探討。事實上在方法學界尙有其他關於區位推論之估計模型，如 neighbourhood model（Freedman et al., 1991: 673-711）、地理學界所關注的空間共變模型（Anselin and Cho, 2002: 276-297）等。但大體來說，Goodman 的迴歸模型與 King 的 EI 模型在區位推論的發展歷程中，可說占了相當重要的地位。值得一提的是，

610-625）首先將區位推論問題的特性形式化（formalization），將恆等式運用線性迴歸來進行估計。而爲了克服「估計係數」與「個數」的問題，King (1997: 91-122) 將估計係數置入隨機係數模型（random coefficient model）結構，至此區位推論估計進入一個新的估計途徑。近年來模型結構的階層化，將貝式機率運用在區位推論的估計裡，也讓區位推論成爲方法學界相當受重視的議題。在本節中，筆者將介紹此三者模型在模型設定、機率分配與參數估計上的特點。

3.1 Goodman 迴歸模型：恆等式的建立與應用

在恆等式(1)中，可以發現未知資訊與已知資訊所呈現的線性關係。基本上 Goodman 迴歸模型相當直觀，他假定所有 i 個觀察值中未知係數皆相同（constancy assumption），即 $\beta_i=\beta$、$W_i=W$，因此模型設定爲：

$$(P_{Yi}|P_{Xi})=P_{Yi}=P_{Xi}\beta+(1-P_{Xi})W+\varepsilon_i$$

此即是常見的線性迴歸關係式，可以透過普通最小平方法（OLS）來求出區位推論的係數估計值。其中 P_{Xi} 與 P_{Yi} 爲所有總體資料觀察值裡 $X_i=1$ 與 $Y_i=1$ 的機率值。若在資料上出現變異異質性（heteroscedasticity）的現象，Goodman 也主張運用加權最小平方法（WLS）來矯正估計係數之標準誤（Goodman, 1959: 612-616），最小平方法係數估計的方程式如下：

$$\begin{bmatrix}\hat{\beta}_i\\ \hat{W}_i\end{bmatrix}=\begin{bmatrix}\beta\\ W\end{bmatrix}=(P_{X_i}{}'P_{X_i})^{-1}P_{X_i}{}'P_{Y_i}$$

Goodman 可說是將區位推論議題形式化於統計模型中的先行者。但也可發現他對於係數估計的問題過於簡化。第一，雖然他發展恆等式(1)爲主要

在 King 提出 EI 模型後，方法學界則開始重新省思區位推論的相關議題。在著名的方法論期刊—Political Analysis 在 2003 年 11 卷第 1 期中，也特刊探討區位推論模型。其中主要關注的焦點爲 King EI 模型與相關延伸的設定，以及如何運用模型估計值，來作爲下一個階段進行實證研究估計的依變項（Adolph et al., 2003: 86-94; Adolph and King, 2003: 65-76; Herron and Shotts, 2003a: 44-64, 2003b: 77-85）。

的模型設定依據，但不同觀察值之間係數的固定假設卻很難出現在現實世界當中。換句話說，我們很難預期每一個觀察值中細格的條件機率皆是相同數值；其次，他對於估計係數 β、W 無任何受限，使得估計係數往往出現大於 1 或小於 0 等機率上不可能出現的數值，也讓人對於該模型之實用性產生質疑（Achen and Shively, 1995: 73-94; King, 1997: 56-68; Freedman et al., 1991: 673-711）。但不可否認地，Goodman 迴歸模型的確開啓區位推論議題與統計模型的對話空間。

3.2 King 的 EI 模型：估計係數參數化的突破

在 Goodman 運用線性迴歸於區位推論的估計後，直到 90 年代末期區位推論的統計估計模型方有突破性的進展。由於區位推論關切的就是如何以已知資訊來推論未知資訊的過程，而上節所述的問題本質裡，我們也可發現欲估計的係數是多於觀察值個數的，這也讓研究者無法運用傳統的統計程序，對個別觀察值係數進行估計。

爲了克服估計上無法認定（under-identified）的問題，[7] King 以恆等式(1)爲基礎，將估計係數「參數化」（parameterization）。於是，原先兩倍於觀察值的估計係數數目（β_i、W_i），在其雙變數常態分配（bivariate normal distribution）預設下，我們可先就常態分配的參數進行推估。也就是說，全體估計係數的分配將成爲估計模型置入機率分配之「抽樣分配」。[8]

$$E(P_{Yi}|P_{Xi})=P_{Yi}=P_{Xi}\beta_i+(1-P_{Xi})W$$
$$(\beta_i, W_i|P_{Xi})\sim TBN_U(\check{\psi})$$

7 所謂「無法認定」（under-identified），在於未知數多於方程式數（指個數），使得無法運用統計模型求出可能解，在區位推論問題中，若可得 i 個聚集單位的觀察值，則一般來說欲估計係數爲 $2i$ 個。

8 從資料型態來看，區位推論的觀察樣本多爲母體資料，故理論上應無「抽樣分配」的問題。然而，爲了要定位個別觀察值在機率分配的位置，則我們必須運用抽樣分配的統計原理來進行定位的區辨。其次，有時我們獲得的資料，也不見得全然是特性區域的所有觀察值。舉例而言，如果我們以台灣地區爲研究範圍，並透過機率抽樣的方法獲得一定數目代表性的村里，但在區位推論的過程中，推論標的仍是以「全台灣」爲主，而非僅限於進行調查研究之地區。因此，「抽樣分配」的概念從此開始也被帶進區位推論估計的議題中。

$$\text{而 } \breve{\psi}' = [\beta, W, \sigma_B^2, \sigma_W^2, \rho]$$

向量 $\breve{\psi}$ 表示了 β_i 與 W_i 在聯合機率分配上的特性與樣態。我們可經由平均數（β, W）、標準誤（σ_B, σ_W）與相關係數（ρ），來定位每一個觀察值係數於常態分配的位置。King 對於區位推論估計係數雙變數常態分配的假定，可說是區位推論問題在方法論的一大突破，初步克服了統計上「無法認定」的問題。同時，King 也對機率分配進行設限（TBN_U），由於 β_i 與 W_i 不可能是 1 與 0 間以外的其他值，因此值域外的部分，可視爲被截斷之機率（truncated）；同時他假定 β_i 與 W_i 兩者與 P_{X_i} 是互爲獨立的，也就是無聚集偏誤的存在，在統計的意義上，則代表著無設定偏誤與漏失變數的問題。

在參數估計上，不同於 Goodman 的最小平方法，由於區位推論的聚集資料多以地理區域爲單位，King 假定觀察值間沒有空間相關（spatial autocorrelation）存在，因此可採最大概似法（MLE）以機率密度函數的連乘對參數向量 $\breve{\psi}$ 求解：

$$L(\breve{\psi}) = \prod_{i=1}^{I} p(P_{Yi} | \breve{\psi})$$

由於 King 對機率分配的預設，使得我們可以定位個別觀察值 β_i 與 W_i 的參數位置，並透過模擬的技術求解。在上下限的限定下，也讓 $\hat{b}_i$ 與 $\hat{w}_i$ 將限於已知資訊，也就是邊際總和值域內。針對個別觀察值係數的估計結果，可以下列方程式表示（U、L 表示爲上下限，且兩式具有恆等式(3)之線性關係）：

$$\hat{b}_i = E(\beta_i | P_{Y_i} = P_{y_i};\ \hat{\breve{\psi}}) = \int_{L_i^b}^{U_i^b} b_i p(b_i | P_{y_i};\ \hat{\breve{\psi}})\, db_i$$

$$\hat{w}_i = E(W_i | P_{Y_i} = P_{y_i};\ \hat{\breve{\psi}}) = \int_{L_i^w}^{U_i^w} w_i p(w_i | P_{y_i};\ \hat{\breve{\psi}})\, dw_i$$

King 將貝式機率事前機率的概念，運用在對於個別觀察值係數推估的事後預測上。在預設參數分配的隨機係數模型置入下，至此有關區位推論統計模型的發展進入另一個嶄新的階段。近年的統計模型發展趨勢上，爲了能夠運用最大概似法進行估計，他假定觀察值間沒有空間相關，此一假設與以地

理區域單位爲主的總體資料特性似乎有所矛盾，因此學者也開始致力發展空間效應相關的估計方法，他們主要透過設定上的矯正來捕捉觀察值間之共變效果（Anselin and Cho, 2002: 276-297）；另一方面，對於係數的機率分配預設，學者則開始將機率分配階層化，運用貝式機率與階層模型的概念來放寬對於區位推論中估計係數的機率分配預設。[9]

3.3 區位推論的階層模型：階層結構的置入

除了各項次的邊際總和外，對於欲估計係數 β_i 與 W_i 我們可說是一無所知。雖然，King EI 模型對於兩者呈現雙變數常態分配的預設已帶入貝式機率所主張的事前分配（prior distribution）的概念。[10] 然而，在對估計係數樣態一無所知的情況下，預設任何的直觀機率分配則顯得過於武斷（Cho, 1998: 143-163; Cho and Gaines, 2004: 152-171）。因此，學者開始嘗試對於係數的機率分配置入階層結構（hierarchical structure），來放寬分配假設對於估計係數的限定。如此階層式機率分配便是貝式統計（Bayesian statistics）的核心概念。[11] 在區位推論的估計上，階層模型的應用尚於快速之發展階段中。在此，我們主要將就幾位學者所發展出來的不同階層模型，就模型設定與機率分配預設上進行探討與比較。大體來說這些模型的差異在於個別觀察值內係數的預設限定、係數以及最終層（hyperprior）機率分配的預設。

在觀察值內的係數限定上，如先前表 1 所示，對於欲估計係數，我們可將 n_{β_i} 與 n_{W_i} 視爲是白努力事件（bernoulli trials），β_i 與 W_i 則爲背後決定

9 有關 King 的 EI 模型的適用性，可見 Freedman 等人（1998: 1518-1522）與 King（1999: 352-355）的辯論。

10 在 Wakefield (2004: 404-412) 對於區位推論模型的回顧裡，其將 King EI 模型歸類爲「階層模型」的一種。而在本文裡，筆者所指的「階層模型」是對於參數分配置入階層結構，而非僅止對於估計係數置入參數分配預設的階層化。

11 貝式統計與傳統頻率（frequencist）統計的最大差異，在於傳統頻率統計針對依變項進行機率分配的假定，在估計上概似函數可以直接構成以進行疊代與逼近的估計程序，對於所有觀察到的樣本，可以求出唯一最可能係數解；至於貝式統計，則是進一步的對於估計係數進行分配預設，如 $Y_i|\beta \sim p_1(y|\beta)$，而 $\beta \sim p_2(\beta|\theta)$（King et al., 1999: 69），則估計時將結合兩層的機率分配來求取同一維度之係數機率：爲 $p(y|\theta)=\int_{-\infty}^{\infty} p_1(y|\beta)p_2(\beta|\theta)d\beta$，進一步有關貝式機率的介紹，可見 Gill（2002: 65-85）。

事件發生的機率，在假定兩者互爲獨立的情況下，針對估計係數可預設爲（King et al., 1999: 61-90）：

$$n_{Ti}|\beta_i, W_i \sim Bin(N_i, \beta_i P_{Xi} + W_i(1-P_{Xi}))$$

故個別觀察值的概似函數可表達爲：

$$L(\beta_i, W_i) = (P_{Xi}\beta_i + (1-P_{Xi})W_i)^{n_{Ti}}(1-P_{Xi}\beta_i - (1-P_{Xi})W_i)^{N_i - n_{Ti}}$$

以上的概似函數，表示以條件機率作爲限定基礎的形式。另外，Wakefield (2004: 385-445) 則提出另一個係數分配的預設方式 (binomial convolution model)，若我們進一步將每個觀察值的細格次數 $n_{\beta i}$ 與 n_{Wi} 皆視爲獨立，以細格爲限定基礎的話，則參數分配將爲：

$$n_{\beta i}|\beta_i \sim Bin(n_{Xi}, \beta_i)、n_{Wi}|W_i \sim Bin(N_i - n_{Xi}, W_i)$$
$$即\ n_{Ti}|\beta_i, W_i \sim ABin(n_{Xi}, N_i, \beta_i, W_i)$$

此時的概似函數是：

$$L(\beta_i, W_i) = \beta_i^{n_{\beta i}}(1-\beta_i)^{n_{Xi}-n_{\beta i}} W_i^{n_{Ti}-n_{\beta i}}(1-W_i)^{N_i - n_{Xi} - n_{Ti} + n_{\beta i}}$$

雖然以上兩者皆運用最大概似法估計程序，Wakefield 與 King 的設定也會產生相同期望值（P_{yi}）之估計結果，但在標準誤則將產生些許差異。[12] 一般來說，在得到機率分配之參數值、瞭解各觀察值之定位後，可運用事後模擬抽樣方法，求取各觀察值之係數解。從概似函數的比較中，也可發現在每個細格的「限定」下，Wakefield（2004: 388-392）的設定會讓係數解在細格個數上完全地符合原先已知的邊際總和（n_{Xi}）。

除了函數形式外，在置入階層結構時，「參數分配」（第二層與最終層）的預設，則是學界在探討以階層模型進行區位推論估計時之另一個焦點。特別重要的是，參數分配預設對於各觀察值係數估計（事後模擬）具有決定性的影響。由於 β_i 與 W_i 是恆爲正值之機率值，King 等人（1999: 61-90）以及

12 有關兩者係數分配預設的標準誤差異，請見 Wakefield（2004: 391）的推導。

Mattos、Veiga（2004: 351-382）皆主張估計係數可預設依循 beta 分配（binomial-beta hierarchical model）。但 Wakefield（2004: 406-409）則認爲，雖然 beta 分配可以直觀地運用機率方式來進行詮釋，但是由於最終層（hyperprior）的參數分配設定，King 等人（1999: 61-90）、Mattos、Veiga（2004: 351-382）皆設定了無任何預設資訊的均等分配（uniform distribution），加上 beta 分配過於彈性的特性，將使事後模擬結果會過度受到 beta 分配裡極端參數之影響，也就是說極端數值出現的機率，在他們的設定下將會遠遠被高估，讓事後係數估計產生偏誤。

爲了更符合區位推論的資料系絡與結構，除了各觀察值細格內的獨立假定外，Wakefield（2004: 398-399）將恆等式(1)加入 logit 連結函數（link function），故原先介於 0 與 1 的機率值（β_i、W_i），轉換後的 θ_i^k 將介於 $-\infty$ 與 ∞，在 2×2 的交叉表裡，Wakefield 假定 θ_i^β 與 θ_i^W 依循雙變數常態分配。他主張如此分配預設使得研究者可以合理地將區域共變（area level covariate）列入第二層模型設定的考量（如 $\theta_i^k=u_k+\gamma_0 Z_i+\delta_{ki}$），在最終層參數分配則也分別依循著常態分佈的結構，以避免極端值出現的機率被過度高估，大體來說，我們可以由以下的式子，來說明 Wakefield 的階層模型：

$$P_{Y_i}=\text{logit}^{-1}(\theta_i^\beta)P_{X_i}+\text{logit}^{-1}(\theta_i^W)(1-P_{X_i}),\ \beta_i=\text{logit}^{-1}(\theta_i^\beta)$$
$$\text{故 } -\infty\leq\theta_i^k\leq\infty,\ k=\beta,\ W$$

第二層的事前機率分配：

$$\text{令 } \theta_i^k=u_k+\delta_{ki} \text{ 而 } \delta_{ki}\sim N(\mu_k,\Sigma),\ \mu_k=\begin{bmatrix}\mu_\beta\\ \mu_W\end{bmatrix}、\Sigma=\begin{bmatrix}\sigma_\beta^2 & \sigma_{\beta W}\\ \sigma_{\beta W} & \sigma_W^2\end{bmatrix}$$

最終層參數分配：[13]

$$\mu_k\sim N(m_k,M_k)、\text{而 } \sigma_k^2\sim Inverse\ Gamma\left(\frac{a_k}{2},\frac{b_k}{2}\right)$$

在第二層與最終層機率分配的結合下，Wakefield 階層模型估計的集合

13 這裡的事前機率的標準誤分配，主要來自 Imai、Kosuke、King 與 Lau（2006: 142-147）的設定。

事後機率 $\left(P_A(\lambda|n_i)\right.$，向量 λ' 爲 $[\beta_i', W_i', h']$，$\beta_i'=[\beta_1 \ldots \beta_i]$，$W_i=[W_1 \ldots W_i]$，而 $h'=\left.\left[m_k, M_k, {}^{a_k}\!/_2, {}^{b_k}\!/_2\right]\right)$ 可由下表示：

$$P_A(\lambda|n_i) \propto \prod_{i=1}^{I} L(\beta_i, W_i) N_1(\theta_i^k|\mu_k, \Sigma) N_2\left(\mu_k, \Sigma|m_k, M_k, \frac{a_k}{2}, \frac{b_k}{2}\right)$$

$L(\beta_i, W_i)$ 爲前述的基本概似函數形式。而上述式子表示在考量第二層（N_1）與最終層（N_2）的常態分配預設後的事後機率估計。以此，我們可以估計向量 λ' 中每一個觀察值的參數位置，再透過積分的方式去求取條件機率。透過上述不同階層模型的整理與探討，筆者認爲，階層模型發展至今，以 Wakefield (2004: 385-445) 所提出來的估計設定與機率分配預設模型具有較佳的估計特性，因此在本文隨後的模擬實驗裡，在階層模型上我們將以 Wakefield 的設定爲主。

以上的說明，大致勾勒出區位推論模型發展脈絡與歷程。在確定區位推論問題的本質與資料特性後，可將已知資訊與未知係數之關係形式化爲線性關係式。爲了克服估計係數個數與方程式數目（觀察值）的問題，隨機係數模型的應用，也成爲區位推論模型發展的一大突破；在階層化模型結構後，近年來各種參數分配預設與模型設定的嘗試，則成爲方法論學者探討區位推論模型的重要課題。

4、模擬研究設計與資料來源

在說明區位推論模型的發展，同時介紹主要的區位推論模型—Goodman 迴歸、King EI 模型與階層模型，在本節中，我們首先回顧相關模擬研究的成果，在這些研究成果的累積上，建立研究設計的判定指標；其次將介紹研究資料的來源，以及說明本研究針對不同資訊量所設計模擬實驗的資料屬性。

4.1 相關模擬研究成果的回顧與判定指標的建立

如前所述，區位推論的目的，在於如何由各變數類別的邊際總和，對於

研究者所欲探討的個體層次關聯性（如表 1 細格內資訊），適當地運用統計程序予以推估。在應用上，學界也評估這些模型在估計上的表現。King 等人（1999: 68-70）認爲貝式統計「縮動」（shrinkage）的特性，會較傳統的估計方式來得較佳，原因在於我們對於欲預估係數之分配一無所知，預設任何機率分配，特別是單一群集（single cluster）的常態分佈會過於理想，他們的模擬實驗結果也顯示，對於偏斜程度較大，或是估計係數呈現多峰分佈的係數而言，貝式階層模型將會得到更準確的估計值。Mattos 與 Veiga（2004: 351-382）則針對 Goodman 迴歸模型、King EI 模型、King 等人階層模型與他們自身所發展的階層模型進行比較，結果則與 King 等人的結論相左。他們發現對於個別觀察值係數估計而言，King 的 EI 模型反而是較爲準確的。Imai、Kosuke、Lu、Strauss（2006: 1-33）的研究則控制 Logit 轉換函數，也就是在設定上皆採 Wakefield 的 binomial convolution model，再進一步比較有無置入階層結構後模型的估計結果，實驗結果顯示置入階層結構的非參數模型（nonparametric model）在個別觀察值的估計上較符合眞實值。

前述的模擬研究皆以個別觀察值的個體層次關聯性（β_i、W_i）爲評比模型表現之依據。然而，除了個別區域的個體層次關係是我們欲推估的對象外，全體區域的因果關係也往往是研究者感興趣，甚至是想要推論之標的。雖然在已知個別觀察值上下限的情況下，我們可直接透過個別觀察值中的個數加權來得到全體之值域（King, 1997: 83-85）。然而，既有的模擬研究多把全體區域的係數合理化爲個別觀察值總合之結果。即使我們能夠透過模型推估程序來得到所有觀察值之估計係數，但畢竟「估計值」只可能近似，卻不等同於「眞實值」，在沒有實質檢證程序下，筆者以爲如此的假定仍過於武斷。

因此，除了個別觀察值之係數外，本文也將總體區域係數納入評斷範疇。在研究設計上，在考量前述的相關研究後，我們將系統化以「不偏性」與「有效性」兩者，來具體判定估計元優劣之指標來評估三種區位推論模型的估計表現。具體操作化方式如下：

1. 不偏性：筆者以「總體區域」與「個別觀察值」的係數爲比較基準，以瞭解何者模型能得到更爲準確的估計結果。在此我們將觀察三者模型與眞值

在總體區域係數（即加權平均數）的誤差（error），[14] 以及對個別觀察值係數估計的平均偏誤（bias）。雖然後者的加總會產生誤差正負相消的問題，造成低估「變異量」結果，但在「不偏性」的指標中，我們則主要以估計準確的偏誤問題爲關切焦點，如下所示：

(1)誤差（error）$=\hat{B}-B$

(2)平均偏誤（bias）$=\sum_{i=1}^{n}(\hat{\beta}_i-\beta_i)/n$

2.有效性：以平均絕對誤差（MAE, mean absolute error）與均方根誤差（RMSE, root mean squared error）作爲操作化指標。以「絕對值」來加總所有觀察值偏誤的用意，在於避免「平均偏誤」指標觀察值間偏差正負值相消，導致低估誤差之問題，將焦點著重在偏差的總量上。另外，用均方根誤差則可進一步瞭解估計值對眞值中可能的極端值之矯正結果，以及模型估計的處理特性，如下所示：

(1)均方根誤差 RMSE$=\sqrt{\sum_{i=1}^{n}(\hat{\beta}_i-\beta_i)^2/n}$

(2)平均絕對誤差 MAE$=\sum_{i=1}^{n}|\hat{\beta}_i-\beta_i|/n$

4.2 研究資料與模擬實驗的資料性質控制

在研究資料上，過去有關區位推論模型的模擬實驗多從估計係數（β_i、W_i）的各種分配特性著手。學者扮演「上帝」的角色來創造各種不同形式的係數組合，來得知在不同分配下模型的估計表現。然而，我們畢竟無法得知估計係數的「眞實值」，也無法合理解釋在哪些眞實情況下，係數分配將呈現何種機率分佈。另一種方式則是運用特定事實資料來評估各種模型的估計表現。可預期的是，此類模擬實驗所得到的結論，將會限定於研究者所採用資料之特性。但也由於資料屬性較爲明確，因而在相關模擬研究成果中，不論學者們如何探索各種的係數分配性質，最終仍多採事實資料來總結其研究成果。因此，在本研究裡，筆者將主要運用大量民調資料所重構的事實資料，

14 總體區域「加權平均數」的計算方式，爲 $B=\frac{1}{N}\sum_{i=1}^{I}n_{Xi}\beta_i$。

來比較 Goodman 迴歸、King EI 模型與階層模型的估計表現。爲了具體比較各模型的估計表現，我們將以調查資料爲基礎來進行資料重建，如此將可得各聚集單位實際的個體層次關聯性數值，以作爲比較各模型估計之眞值。

一般來說，大多數橫斷性調查資料的樣本規模約略在 2,000 筆左右。如此的樣本規模將使我們無法依以較小聚集單位（如鄉鎮市、村里）來進行資料建構，換句話說，此將會使得個別觀察值內細格數值過小，無法有效地建立一定規模地觀察值數量。因此，在同時考量集體資料觀察值個數以及個別觀察值中細格分佈下，筆者將採行政院主計處於 94 年 12 月所進行的「人力資源調查」資料。[15] 在此資料規模下，我們可運用較小的地域單位進行資料之群集，同時也顧及觀察值中細格分佈，不會產生過多細格觀察值爲 0 的情況。在資料群集的過程中，我們將以「村里」爲聚集單位進行資料重建。在全數 59,458 筆成功樣本中，以村里區分將可得 507 筆總體觀察值，全體的村里如附錄一所示。值得一提的，雖然該資料在分層抽樣下並非擁有全台灣所有村里的樣本，故從地域分佈來看可能將呈現碎裂的現象。但由於本文所比較的模型皆未處理「空間相關」共變效果，因此純粹從資料性質來看，能夠具有龐大數量的個體資料來建立可比較的眞值進行模擬實驗，方爲筆者選擇研究資料的主要目的。

在大量的調查研究資料下，我們可得知各細格中之眞實值。然而在眞實狀況中，已知的資訊卻僅有 P_{X_i} 與 P_{Y_i} 等變項類別之總和。因此，在 2×2 的交叉表中，就總體的邊際總和分佈上，大致來說可簡化成以下三種類型：

類型(1)：兩者變項中，各自項次的比例大致相等，即 P_{X_i}、$P_{Y_i} \cong 0.5$。

類型(2)：兩者變項中，其一變項內的項次的比例大致相等，另一則否，即 $P_{X_i} \cong 0.5$，但 $P_{Y_i} > 0.5$。

類型(3)：兩者變項各自項次的比例皆不等，P_{X_i}、$P_{Y_i} < 0.5$。

15 在樣本取得上，「人力資源調查」資料主要透過抽取率與單位大小成比例（probability proportional to size, PPS）的「多階段抽樣」抽取鄉鎮市及村里，在每一個村里中則成功至少 47 個樣本。

在細格分佈未知下，依已知的邊際總和，則我們可以將交叉表的類型區分成以上三者。對區位推論來說，以上三者分類最大意義在於「上下限」值域的範圍大小。可以預期的是，類型(3)交叉表由於變數項次比例分佈不均，因此從已知邊際總和所推算的上下限值域範圍最小。換句話說，就已知資訊而言此類分佈交叉表能帶來最大的資訊量（informative），其次爲類型(2)的分佈。至於類型(1)交叉表，則清楚表示即便我們運用上下限方式排除理論上不會存在的值域，也無法達到限縮值域之目的，也就是已知資訊將無法帶來任何有效的資訊效果（non-informative）。因此，我們將從以上三者不同分佈交叉表的類型，來建立三種不同情況（scenario）的模擬研究，以期盡可能地涵蓋不同的已知資訊分佈概況。而不同分佈類型模擬研究的變項名稱與邊際總和，則如下表 2 所示。

在「教育程度」方面，筆者主要合併專科以上的項目，歸併爲「高教育程度」與「其他」兩者；「婚姻狀況」則是合併未婚、離婚或分居與配偶死亡爲單身，歸併爲「已婚」與「單身」兩者。而由於區位推論模型多以條件機率（$P(y|x)$）爲估計係數，以 P_Y 爲估計標的，因此對估計係數（β_i、W_i）的分佈而言，P_Y 的比例分佈將遠遠較 P_X 來的重要，爲了更能控制眞實值分佈的型態，而僅操控已知資訊所帶來資訊量的限制，在模擬一當中，即便 P_X 並非趨近於 0.5，但尙能展現出已知資訊不具資訊效應（non-informative）的現象。

就進行比較的 Goodman 迴歸、King EI 模型與階層模型而言，Goodman 迴歸假定所有觀察值中有相同的未知係數（$\beta_i=\beta$、$W_i=W$），故估計結

表 2 本文模擬研究交叉表之分佈類型與相關資訊

	模擬一	模擬二	模擬三
變項名稱 X	教育程度	性別	婚姻狀況
變項名稱 Y	性別	教育程度	教育程度
P_X	0.2760	0.5082	0.4295
P_Y	0.5082	0.7240	0.2760

資料來源：行政院主計處，「人力資源調查」94 年 12 月。

果爲定值。至於 King EI 模型與 Wakefield 的階層模型，則是先運用最大概似法求估計係數背後的參數分配解，確定每一個觀察值於特定參數分配的定位後，再透過馬可夫鍊蒙地卡羅法（Markov Chain Monte Carlo，MCMC）的資料擴增程序（data augmentation），求取每一個觀察值係數的期望值解。[16] 本文所運用的統計軟體，分別是 Stata 9.0（Goodman 迴歸）、EzI（King EI 模型）與 R（階層模型）。[17]

5、模擬結果的分析與討論

說明了研究設計與資料來源後，在本節裡，我們首先將呈現模擬實驗的資料特性，接著比較 Goodman 迴歸、King EI 模型與階層模型於區位推論的模擬實驗結果。我們主要透過大量民調資料來建構可供比較的眞值，以「不偏性」與「有效性」等統計估計指標，來評量三者主要區位推論模型的估計特性。

在模擬研究的資料分佈特性方面，在研究設計中筆者依已知資訊所帶來的資訊量，來區分三者不同已知資訊的分佈概況。有關用來比較的眞實值（β_i、W_i），二維分佈如下頁圖 1 左方所示。就此可以發現三者不同分佈類型的交叉表裡，眞實值皆大致呈現單一群集的現象，但隨著不同資訊量的設計下，其單一群集的區域則有所改變。在模擬一中，群集的位置約略在圖的正中央，模擬 2 則是右上方，模擬 3 爲左下方。

當然，在一般估計時我們並無法得知眞實值的分佈。如第 2 節所述的恆等式(3)，在無法得知眞值下我們可將單一個別觀察值的可能值域，進一步運用分佈圖的方式來予以呈現（tomography plots，King, 1997: 81），如圖 1 右方所示。在模擬一的設計裡，可以發現大多的值域線與 X 軸、Y 軸交接在

16 對於未知係數的分配，若我們已得知（或預設）其機率分配，則可在 MCMC 法下，運用隨機過程進行模擬抽樣，在一定次數的累積（積分）下，求取特定觀察值係數的收斂值解，有關 MCMC 法的估計原理與不同的抽樣方式，進一步的介紹可見 Gill（2002: 301-350）。

17 有關 EzI，可見 Benoith 與 King（2003: 1-6）；至於 R 的階層模型估計，筆者主要參考 Zelig 中 ei. Hier 的程式語法，可見 Imai、Kosuke、King 與 Lau（2006: 142-147）。

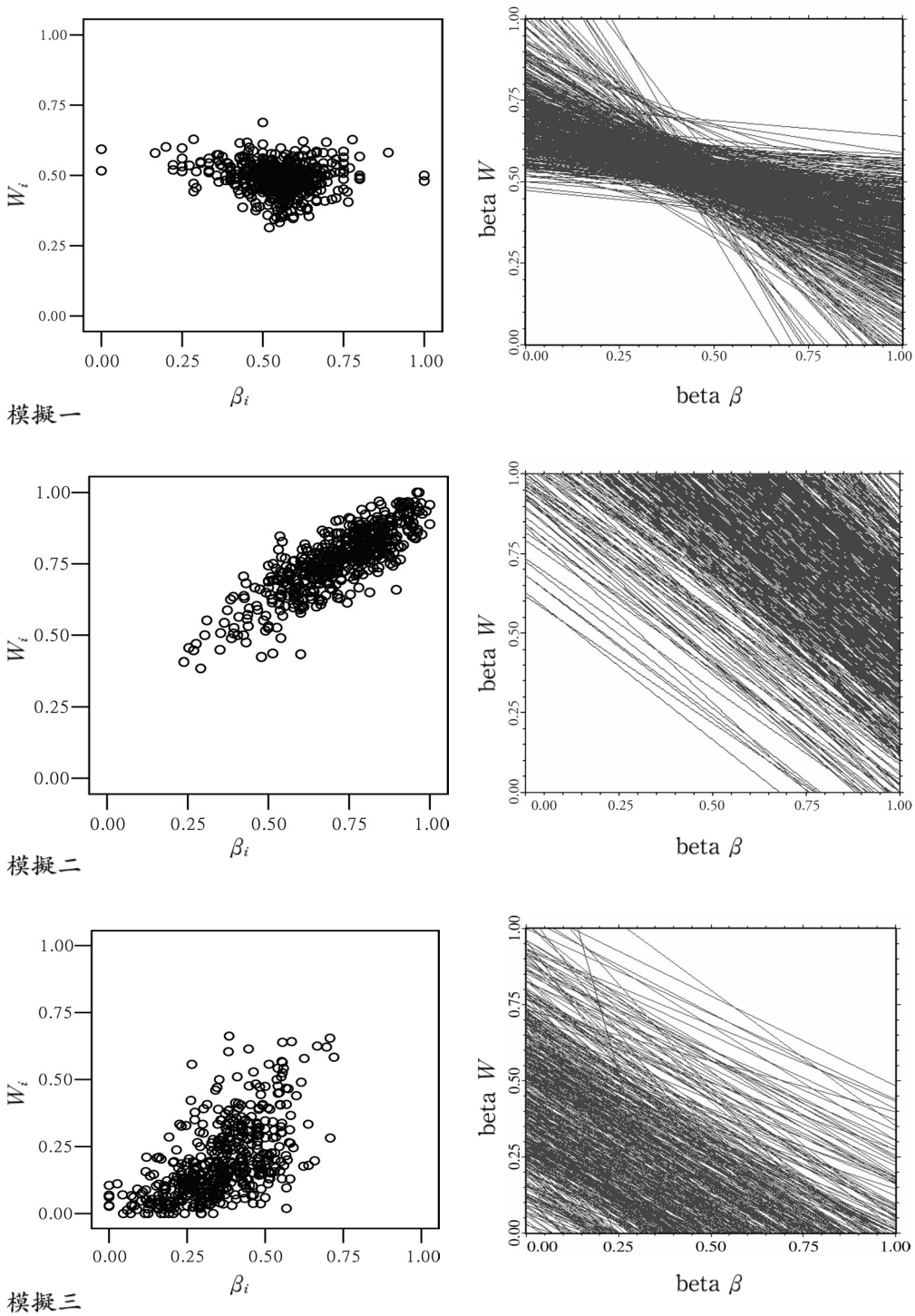

圖 1　係數(β_i、W_i)的眞值分佈與可能值域分佈圖

1.0 附近的位置，這表示即便我們加入上下限的限定，仍無法對未知的 β_i、W_i 帶來更多的有效限定資訊。換句話說，就已知邊際總和裡所限定的合理估計係數組合，將有非常多的可能性。在模擬二上，在筆者設計 $P_Y > 0.5$ 下，則大多觀察值的值域則是集中在右上方，這表示係數 β_i、W_i 多數大於 0.5。模擬三所有觀察值的值域線則是集中在左下方，其中更有部分觀察值值域的斜率相當陡峭，這表示合理的值域組合數，在已知資訊的限定下，將變得相對有限。因此在圖 1 中，三者的分佈呈現了筆者所預期的模擬設計效果，依已知邊際總和呈現資訊量多寡下，來進行值域的限定。而對研究者來說，模擬一的已知訊息分佈將是最糟的狀況（King, 1997: 160），因爲上下界限制所能帶給我們對於估計係數的資訊量相當有限，甚至完全無限縮值域的效果。

在實際估計的情況下，我們並無法知道估計係數的分佈型態，而 King 的 EI 模型假設眞實值爲單一群集（single cluster），預設了雙變數常態分佈。即使如此的機率分配預設，在 Cho 與 Gaines（2004: 157）的研究中，也明確呈現另一種完全極端的可能性（也就是散佈在值域邊緣的四周，非呈現聚集的現象，但兩者的平均數將會相同），但在本研究運用大量調查資料所建構的眞實值裡，從圖 1 則可初步判定，我們用來比較的眞實值則大致符合單一群集的分配型態。

在說明模擬研究的資料性質後，以下的分析中筆者將分別討論三者模型於不同資料型態下的估計性質，包括以不偏性與有效性兩者具體指標來呈現模擬實驗的結果；此外，我們也將呈現眞值與不同模型估計結果的散佈圖，期望能更進一步闡述這些模型的估計特性。

5.1 區位推論模型估計的「不偏性」比較結果

在區位推論中，一般而言個別觀察值中的係數是研究者的估計標的。除此之外，在本文裡我們也進一步將全體的估計係數作爲比較判準，來瞭解不同模型對於係數不偏性的估計結果。下表 3 爲 Goodman 迴歸、King EI 模型、階層模型與眞值在不同類型已知資訊下的估計比較。在「加權平均數」上，我們用來比較依據是爲總體估計係數，「平均偏差」則是就個別觀察值而言，模型估計的平均偏差值。

表 3　區位推論模型的「不偏性」比較列表

	模擬一		模擬二		模擬三	
加權平均數	β	W	β	W	β	W
眞值	0.5487	0.4928	0.7020	0.7467	0.3677	0.2071
Goodman 迴歸(誤差)	−0.1256	0.0482	0.6584	−0.6590	−0.1502	0.0913
King EI 模型(誤差)	−0.1230	0.0469	0.2091	−0.2161	−0.1526	0.1148
階層模型(誤差)	0.0245	0.0458	0.2030	−0.2097	−0.1329	0.1004
平均偏差	β_i	W_i	β_i	W_i	β_i	W_i
Goodman 迴歸	−0.1208	0.0503	0.6474	−0.6729	−0.1365	0.1057
King EI 模型	−0.1184	0.0493	0.2013	−0.2054	−0.1448	0.1120
階層模型	0.0294	0.0479	0.1959	−0.2040	−0.1250	0.0997

資料來源：行政院主計處，「人力資源調查」94 年 12 月。

就「不偏性」的意義上，研究者主要希望能得到最「接近」眞值的估計結果。就總體估計係數而言，模擬一與模擬二的實驗皆顯示階層模型估計結果表現最佳，誤差分別在 0.2030 至−0.2097 之間，其次是 King EI 模型，至於 Goodman 迴歸模型估計誤差則最大。在已知資訊量最大的模擬三中，在 β 方面則仍是階層模型誤差最小，爲−0.1329，而 W 的估計則以 Goodman 迴歸最佳，僅有 0.0913，其次是階層模型的 0.1004，King EI 模型誤差的 0.1148 則是誤差最大者。就個別觀察值的估計係數而言，在平均偏差上與總體係數估計結果所呈現的訊息則相當一致。從表 2 中可以發現，在筆者所設定的不同邊際總和分佈型態的模擬實驗裡，階層模型的估計表現皆是最佳的，數值在 0.1959 至−0.2040 之間。

除了眞實值與估計值的比較外，另一方面，我們也可由不同模擬設計間估計差異的比較，來探討已知資訊量多寡與模型估計的關係。在階層模型上，可以發現在已知資訊量最少的模擬一中，對加權平均數與平均偏差估計最爲準確，其次是資訊量最大的模擬三，誤差與偏差最大的是模擬二。當然，這與筆者所採用的資料型態有直接的關聯性，從 King EI 模型估計中，我們也可看出相同的現象。與過去相關研究不同的是，筆者認爲當邊際總和所能提供的已知資訊愈少，假定眞實值分佈在 0.5 周圍時，這表示雖然我們無法透過

上下限來對限制估計係數的值域。但相反地在係數分配型態呈現單一群集的情況下，這也表示我們不需要透過機率分配的截斷（truncated）來武斷地調整參數值，反而在估計上的表現較佳。

雖然在模擬三中，W 的總體係數估計以 Goodman 迴歸最爲接近眞實值，但綜合來說，在「不偏性」的估計性質裡，即便估計係數的分配型態大致符合 King EI 模型「單一群集、雙變數常態分配」的核心假設，階層模型的估計表現，相對來說仍是較爲準確的。

5.2 區位推論模型估計的「有效性」比較結果

在有效性上，筆者以「平均絕對誤差」與「均方根誤差」作爲評判標準。兩者的差異在於，前者爲估計值與眞值之偏差取絕對值後加總平均之結果；後者則是加入平方的運算。換句話說，後者將可進一步觀察模型在矯正極端值上是否有著較好的性質。

在估計上，有效性的意涵即爲最小變異（minimum variance）。一般來說，對估計模型的要求除了能盡可能接近眞值外，我們也期望能夠得到最爲穩健（robust）的估計結果。換句話說，透過估計值與眞實值差異量的總計，我們可以瞭解每一個估計係數與眞實值的平均差異量。下頁表 4 爲筆者的模擬結果，就進行比較的 Goodman 迴歸、King EI 模型與階層模型三種區位推論模型，在平均絕對誤差上，模擬一中顯示階層模型對 β_i 的估計最佳，爲 0.0833，而 W_i 則是以 King EI 模型的 0.0518 最佳；除此之外，在模擬二與模擬三中，不論是 β_i 或 W_i，階層模型估計的平均絕對誤差皆爲最小者，在 0.1007 至 0.2040 之間，其次是 King EI 模型，至於 Goodman 迴歸模型所估計的平均絕對誤差則最大。在均方根誤差上則呈現相同的現象，除了模擬一中 W_i 以 King EI 模型的 0.0662 表現最好外，其餘則仍是階層模型表現最佳。這表示階層模型的估計有效性比其他兩者區位推論模型來的較好。

除了分別說明不同模型在平均絕對誤差與均方根誤差的估計表現外，我們更可透過兩者的比較，來探討何者對極端值具矯正效果。在定義上，若我們欲對兩者模型在極端值的估計進行評價，可以理解的是，在均方根誤差放大估計值與眞實值差距下，則兩者模型在「均方根誤差」值上的差異，將會

表 4　區位推論模型的「有效性」比較列表

	模擬一		模擬二		模擬三	
	β_i	W_i	β_i	W_i	β_i	W_i
平均絕對誤差						
Goodman 迴歸	0.1408	0.0595	0.6474	0.6729	0.1607	0.1536
King EI 模型(1)	0.1377	0.0518	0.2016	0.2057	0.1482	0.1141
階層模型(2)	0.0833	0.0575	0.1961	0.2040	0.1315	0.1007
(1)-(2)	0.0544	−0.0057	0.0055	0.0017	0.0167	0.0134
均方根誤差						
Goodman 迴歸	0.1648	0.0742	0.6642	0.6821	0.1924	0.1754
King EI 模型(1)	0.1604	0.0662	0.2267	0.2284	0.1734	0.1334
階層模型(2)	0.1164	0.0722	0.2177	0.2177	0.1539	0.1145
(1)-(2)	0.0440	−0.0060	0.0090	0.0107	0.0195	0.0189

資料來源：行政院主計處，「人力資源調查」94 年 12 月。

大於「平均絕對誤差」上之差值。

如此的趨勢也具體呈現在表 4 中。在模擬一裡，由於係數相當趨中群集(0.5)，因此兩者模型對極端值矯正之差距並不明顯。在 β_i 上，King EI 模型在考量極端值的估計後，甚至更能縮小與階層模型對眞值估計的差異，從 0.0544 降爲 0.0440；但在已知邊際總和帶來相對較多資訊量的模擬二與模擬三上，King EI 模型與階層模型的均方根誤差與平均絕對誤差之差距則有明顯擴大的趨勢。模擬二中的 β_i 從 0.0055 增加至 0.0090，W_i 則是從 0.0017 提升至 0.0107；模擬三中的 β_i 與 W_i，則是分別從 0.0167 提高到 0.0195，以及 0.0134 至 0.0189。這具體顯現了階層模型對於極端值的估計具有較好的估計性質。

整體來說，對於區位推論中個別觀察值的係數估計，階層模型有著較佳的不偏性特質，這表示相較於其他區位推論模型而言，階層模型的估計表現不但最接近眞實值，同時也最爲穩健，具有最佳的估計元特性。

在上述的分析中，我們主要針對 Goodman 迴歸、King EI 模型與階層模型三者區位推論模型進行估計特性的比較，在不同已知邊際總和分佈下，不論已知資訊所帶來的資訊量多寡，總體來說 Wakefield 的階層模型在「不偏

性」與「有效性」具有最佳的估計元性質。如前所述，Goodman 迴歸模型假定不同觀察值皆爲定值，至於參數化的 King EI 模型與階層模型，則可透過 MCMC 法來估計出各個觀察值係數期望值解，爲了更具體瞭解此兩者模型的估計性質，筆者將進一步呈現此兩者模型估計結果與眞值的散佈圖。

圖 2 至圖 4 爲筆者所設定不同型態已知資訊分佈中，β_i、W_i 眞值與 King EI 模型、階層模型估計結果的分佈圖，其中 X 軸爲估計值，Y 軸則爲眞實值的分佈。一般來說，若估計結果與眞值結果愈接近，則散佈的結果應越近於圖中的左下至右上的對角線上。就不同已知資訊量的模擬設計裡，透過圖 2 我們可以發現當邊際總和資訊量越小，也就是 $P_{Y_i} \cong 0.5$ 時，則不論是 King EI 模型或階層模型，估計結果的分佈與眞值的分佈差距甚大，將會過度集中在

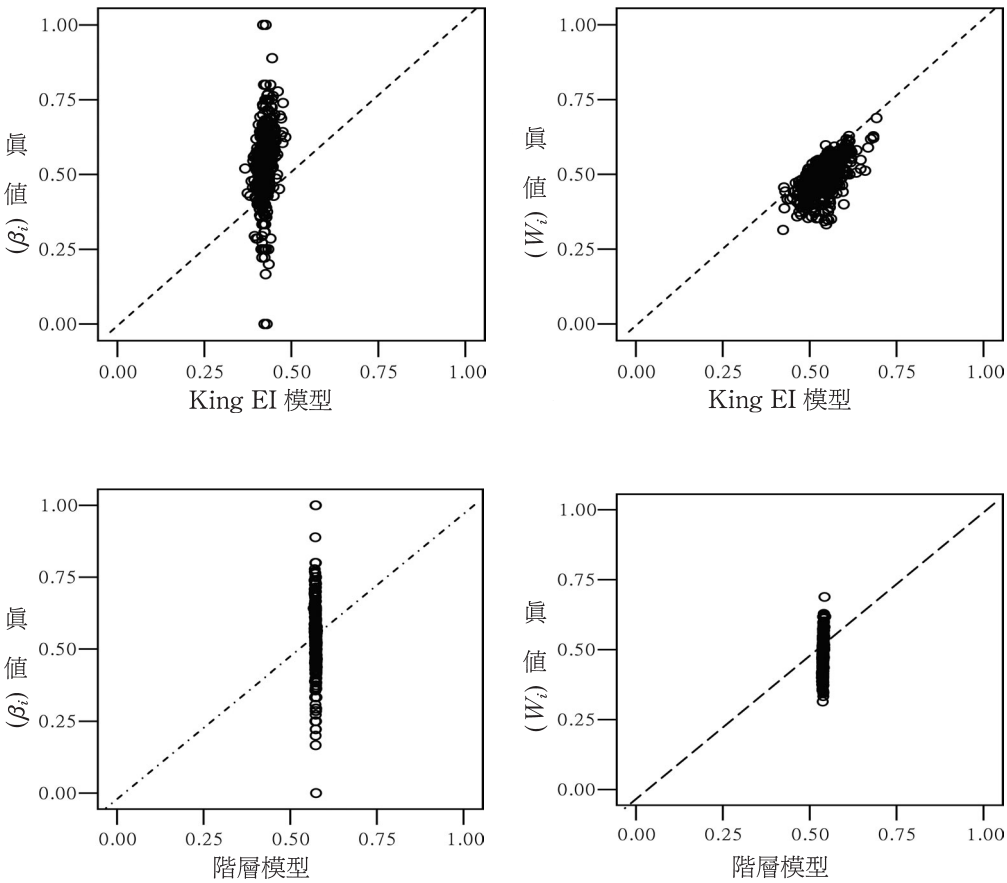

圖 2　眞值與 King EI 模型、階層模型的散佈圖(模擬一)

平均數上。大體來看，即便兩者模型對總體估計係數誤差或各觀察值平均偏差在此類分佈交叉表上表現最好，但對於極端值的估計在缺乏已知資訊進一步限定下，則兩者的估計表現皆有改善的空間。

另一方面，資訊量逐漸增加時，如圖 3 至圖 4 所示，則顯現了 King EI 模型與階層模型在模型預設上之特性。雖然 King EI 模型與 Wakefield 的階層模型都對於各觀察值係數置入參數分配，但從估計值與眞實的散佈便可發現兩者差異。在控制已知邊際總和分佈的情況下，階層模型的估計結果都是更嵌於對角線的，King EI 模型的估計係數則較爲集中。如此的現象在圖 3(模擬二) 最爲明顯。如箭頭所示，對於眞實值單一群集外的極端值而言，在 King EI 模型估計中，隨著眞實值愈小，也就是離群集愈遠時，則估計值的偏離程

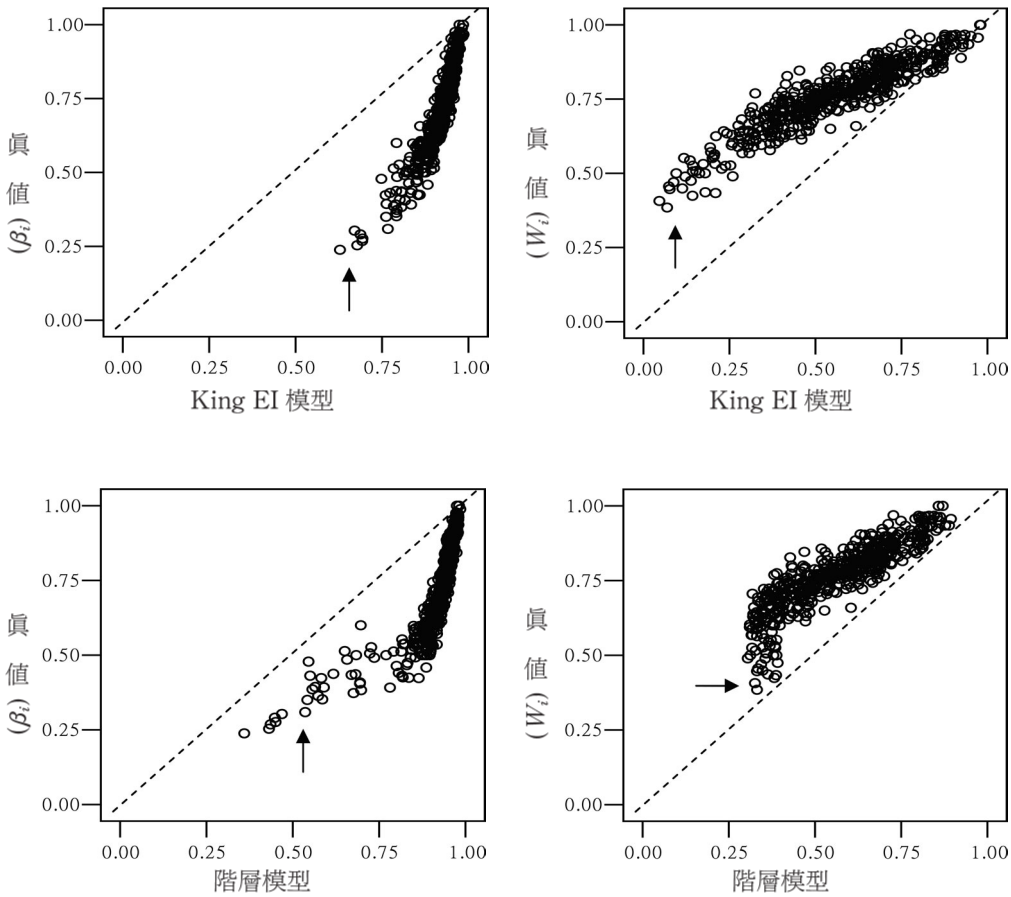

圖 3　眞值與 King EI 模型、階層模型的散佈圖(模擬二)

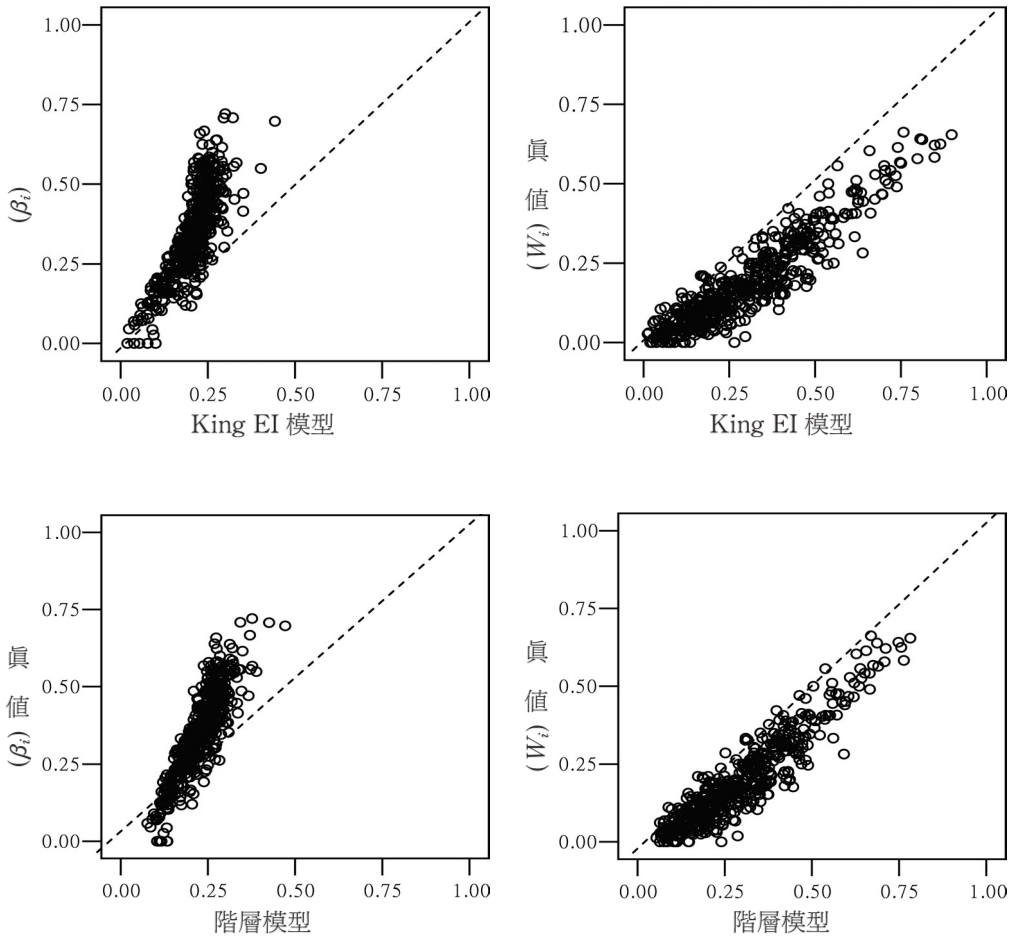

圖 4　眞值與 King EI 模型、階層模型的散佈圖(模擬三)

度愈大；相反地階層模型對於這些離群集較遠的極端值，則有相當顯著的校正效果。在圖 4 的估計中，雖然如此校正極端值的效果不如圖 3 明顯，但也顯示階層模型對於估計係數機率分配單一群集的估計，是較 King EI 模型更接近對角線的。

在本節中，我們首先介紹了模擬研究的資料型態，透過眞值的分佈圖大致確定 β_i 與 W_i 符合單一群集的特性後，筆者認爲，即便在資料上尚符合 King 的核心假設，但階層模型在不同已知邊際總和分佈下，在估計上皆比 King EI 模型擁有較好的「不偏性」與「有效性」；此外，對於極端值的估計也有較好的校正效果。因此，我們認爲而兩者模型對於極端值的矯正差異不

見得僅在於 King EI 模型最受批評的「單一群集雙變數常態分配」預設，而是更在於模型設定與求取參數解的函數形式上。在階層模型置入貝式機率階層結構後，雖然求取出的估計係數解分佈不如 King EI 模型來得集中，但若區位推論的目的在於以各觀察值的已知邊際總和，來求取細格內的未知資訊上，總體來說，則筆者認爲階層模型將較 Goodman 迴歸模型與 King EI 模型具有更佳的估計性質，如此的特性也是先前相關模擬實驗中未被系統性提及的。

6、結論

所謂「區位推論」，爲當研究者所可得資料爲集體層次，而研究旨趣與目的在個體層次的關聯性時，所產生可得資料與研究目的分析單位層次不相符之情形。因此，如何運用適當的統計方法與程序，從已知的總體集合資訊推估未知個體層次關聯性，便是區位推論模型首要關注的問題所在。

在本文中，筆者回顧了區位推論模型的發展，並以 Goodman 迴歸、King EI 模型以及階層模型爲例，說明這三種主要的區位推論模型於統計元素上的特點。在區位推論模型的發展過程中，Goodman 首先釐清問題本質，建立已知資訊與未知資訊的恆等式；而 King 將估計係數參數化，更讓研究者能夠得到針對個別區域（觀察值）的估計結果；至於近年來階層模型的發展，則在模型設定與參數預設上有熱切的討論，試圖能得到更貼近眞實情況的估計結果。在經過相關文獻的檢視後，筆者以 Goodman 迴歸、King EI 模型與 Wakefield 的階層模型爲例，運用模擬實驗方式來評估三者模型在總體與個別區域觀察值估計上之表現。結果顯示，在不同分佈型態的交叉表中，在不偏性與有效性上，階層模型具有最佳的估計元特性。再者，由於階層模型的貝式機率結構，以及更受已知邊際總和限定的函數形式設定，讓事後模擬所求取出的期望值解，將會較 King EI 模型更貼近眞值分佈，特別在於極端值的矯正與估計上。

就現今所發展的區位推論模型而言，在方法上未來仍有許多發展空間。首先，就本文所設定的不同類型的交叉表上，可以發現不論是 King 的 EI 模

型，或是 Wakefield 的階層模型，在上下限無法有效限制值域，也就是已知邊際總和帶來資訊量較小的情況下，對於極端值的估計皆有改善的空間。其次，當今主要的區位推論模型，在便利運用最大概似法求取預設參數分配解的情況下，皆存在著無空間聚集偏誤之假定，或運用特定的模型設定方式來捕捉區域效應，但如此之基本假定與區位推論中基於地區聚集的資料特性又有所矛盾，因此，如何在函數形式中適當且合理的對於區域特性進行設定，相信將成爲未來區位推論模型發展必須受重視之主要課題。

附錄一　模擬實驗之觀察值的縣市與村里列表

縣市	村　　里
台北縣	板橋市國光里、板橋市百壽里、板橋市振興里、板橋市埤墘里、板橋市華貴里、板橋市新生里、三重市光陽里、三重市錦通里、三重市慈生里、中和市福祥里、中和市明德里、中和市興南里、中和市碧河里、永和市文化里、永和市民族里、新莊市自信里、新莊市中隆里、新店市百和里、樹林市保安里、樹林市彭厝里、樹林市樂山里、樹林市南園里、樹林市柑園里、鶯歌鎮大湖里、鶯歌鎮鳳祥里、淡水鎮水碓里、淡水鎮清文里、汐止市白雲里、汐止市大同里、汐止市湖興里、汐止市長青里、土城市員仁里、土城市廣福里、土城市中正里、蘆洲市中路里、蘆洲市九芎里、蘆洲市得仁里、蘆洲市保新里、泰山鄉貴子村、林口鄉麗林村、深坑鄉昇高村、坪林鄉水德村、坪林鄉漁光村、平溪鄉薯榔村、雙溪鄉共和村、貢寮鄉吉林村、金山鄉三界村、萬里鄉大鵬村
宜蘭縣	宜蘭市孝廉里、宜蘭市中山里、羅東鎮大新里、蘇澳鎮永榮里、蘇澳鎮南正里、頭城鎮外澳里、頭城鎮港口里、礁溪鄉白鵝村、礁溪鄉玉光村、壯圍鄉東港村、員山鄉惠好村、員山鄉永和村、冬山鄉群英村、五結鄉三興村、五結鄉二結村、三星鄉大隱村
桃園縣	桃園市永興里、桃園市西埔里、桃園市莊敬里、中壢市自立里、中壢市中正里、中壢市興國里、中壢市龍東里、中壢市正義里、大溪鎮一心里、大溪鎮美華里、楊梅鎮大同里、楊梅鎮金溪里、蘆竹鄉錦興村、蘆竹鄉長興村、大園鄉和平村、大園鄉三石村、龜山鄉迴龍村、龜山鄉新興村、八德市瑞祥里、八德市霄裡里、八德市高明里、龍潭鄉三林村、龍潭鄉富林村、平鎮市北勢里、平鎮市平鎮里、平鎮市東社里、平鎮市龍興里、新屋鄉後庄村、觀音鄉保生村
新竹縣	竹北市竹義里、竹東鎮員山里、新埔鎮寶石里、新埔鎮新埔里、關西鎮錦山里、新豐鄉松林村、橫山鄉橫山村、北埔鄉埔尾村、寶山鄉大崎村、寶山鄉寶山村、峨眉鄉石井村
苗栗縣	苗栗市建功里、苗栗市恭敬里、苗栗市福麗里、苗栗市福安里、苑裡鎮西勢里、苑裡鎮舊社里、通霄鎮通東里、通霄鎮楓樹里、後龍鎮埔頂里、後龍鎮福寧里、卓蘭鎮上新里、大湖鄉富興村、公館鄉大坑村、公館鄉福基村、三義鄉雙湖村、獅潭鄉豐林村
台中縣	豐原市西安里、豐原市東勢里、豐原市民生里、豐原市南嵩里、東勢鎮新盛里、東勢鎮慶東里、東勢鎮茂興里、大甲鎮平安里、清水鎮棟榔里、沙鹿鎮犁分里、沙鹿鎮北勢里、梧棲鎮下寮里、梧棲鎮興農里、后里鄉墩北村、神岡鄉北庄村、大安鄉海墘村、烏日鄉榮泉村、烏日鄉五光村、大肚鄉山陽村、大肚鄉蔗部村、霧峰鄉甲寅村、霧峰鄉南勢村、太平市太平里、太平市新福里、大里市新里里、大里市瑞城里
彰化縣	彰化市延和里、彰化市牛埔里、彰化市阿夷里、鹿港鎮洋厝里、和美鎮鐵山里、線西鄉頂犁村、福興鄉萬豐村、秀水鄉陝西村、花壇鄉南口村、員林鎮黎明里、員林鎮大饒里、員林鎮東北里、溪湖鎮湖西里、田中鎮平和里、埔心鄉經口村、永靖鄉瑚璉村、永靖鄉福興村、社頭鄉松竹村、北斗鎮新政里、北斗鎮重慶里、二林鎮南光里、二林鎮趙甲里、埤頭鄉興農村、芳苑鄉新街村、芳苑鄉漢寶村、竹塘鄉小西村、溪州鄉榮光村

縣市	村　　里
南投縣	南投市龍泉里、南投市平和里、南投市福興里、埔里鎮水頭里、埔里鎮桃米里、草屯鎮和平里、草屯鎮碧洲里、草屯鎮明正里、竹山鎮田子里、集集鎮林尾里、鹿谷鄉秀峰村、鹿谷鄉清水村、中寮鄉清水村、中寮鄉永和村、國姓鄉福龜村、國姓鄉柑林村、水里鄉水里村、水里鄉頂崁村
雲林縣	斗六市榴北里、斗南鎮阿丹里、斗南鎮明昌里、虎尾鎮立仁里、西螺鎮大新里、土庫鎮大荖里、北港鎮南安里、北港鎮共榮里、古坑鄉古坑村、大埤鄉三結村、大埤鄉豐岡村、莿桐鄉饒平村、莿桐鄉四合村、二崙鄉定安村、崙背鄉西榮村、東勢鄉新坤村、臺西鄉臺西村、元長鄉頂寮村、四湖鄉溪尾村、水林鄉萬興村
嘉義縣	太保市埤鄉里、朴子市竹圍里、布袋鎮龍江里、布袋鎮樹林里、大林鎮東林里、大林鎮明華里、民雄鄉北斗村、新港鄉宮前村、新港鄉中庄村、東石鄉永屯村、東石鄉蔦松村、義竹鄉義竹村、鹿草鄉後堀村、水上鄉溪洲村、水上鄉柳林村、竹崎鄉坑頭村、梅山鄉大南村、番路鄉觸口村、大埔鄉永樂村
台南縣	新營市大宏里、新營市南興里、鹽水鎮水仙里、鹽水鎮孫厝里、白河鎮崎內里、後壁鄉頂安村、麻豆鎮涑江里、麻豆鎮油車里、麻豆鎮中民里、下營鄉新興村、下營鄉大埤村、大內鄉大內村、佳里鎮海澄里、學甲鎮一秀里、學甲鎮宜民里、西港鄉劉厝村、七股鄉頂山村、將軍鄉仁和村、將軍鄉長沙村、新化鎮竹林里、新市鄉三舍村、仁德鄉後壁村、關廟鄉松腳村、永康市大橋里、永康市烏竹里、永康市光復里
高雄縣	鳳山市興中里、鳳山市文英里、鳳山市鎮東里、鳳山市天興里、鳳山市興仁里、鳳山市富甲里、林園鄉北汕村、大樹鄉井腳村、大社鄉神農村、仁武鄉大灣村、鳥松鄉夢裡村、岡山鎮後協里、岡山鎮灣裡里、岡山鎮爲隨里、橋頭鄉仕隆村、燕巢鄉尖山村、燕巢鄉鳳雄村、阿蓮鄉清蓮村、湖內鄉葉厝村、茄萣鄉白雲村、茄萣鄉光定村、彌陀鄉文安村、彌陀鄉南寮村、旗山鎮大德里、美濃鎮泰安里、美濃鎮吉東里、六龜鄉寶來村、內門鄉內東村、三民鄉民生村
屏東縣	屏東市大同里、屏東市扶風里、屏東市長春里、屏東市豐源里、屏東市歸心里、屏東市潭墘里、潮州鎮五魁里、潮州鎮檨子里、潮州鎮四春里、東港鎮朝安里、東港鎮嘉蓮里、東港鎮興和里、恆春鎮城南里、恆春鎮鵝鑾里、萬丹鄉竹林村、萬丹鄉水泉村、麟洛鄉麟趾村、內埔鄉東寧村、竹田鄉鳳明村、新埤鄉建功村、枋寮鄉枋寮村、枋寮鄉中寮村、林邊鄉光林村、佳冬鄉萬建村、琉球鄉大福村、車城鄉福安村、枋山鄉楓港村、霧臺鄉霧臺村、泰武鄉佳平村
台東縣	臺東市新生里、臺東市大同里、臺東市興國里、臺東市豐榮里、臺東市建業里、成功鎮忠仁里、鹿野鄉瑞隆村、池上鄉新興村、池上鄉振興村、長濱鄉寧埔村、延平鄉武陵村、達仁鄉台板村
花蓮縣	花蓮市民樂里、花蓮市主權里、花蓮市國華里、花蓮市國裕里、玉里鎮泰昌里、新城鄉康樂村、壽豐鄉池南村、壽豐鄉水璉村、光復鄉大馬村、光復鄉東富村、富里鄉羅山村、萬榮鄉西林村、萬榮鄉見晴村
澎湖縣	馬公市中央里、馬公市光明里、馬公市東文里、馬公市井垵里、湖西鄉隘門村、湖西鄉林投村、白沙鄉中屯村、西嶼鄉池東村、西嶼鄉外垵村
基隆市	中正區正義里、中正區入船里、中正區八斗里、中正區正濱里、七堵區泰安里、七堵區友一里、暖暖區八堵里、仁愛區育仁里、仁愛區書院里、中山區安平里、中山區中和里、中山區和平里、安樂區定邦里、信義區信綠里

縣市	村　　里
新竹市	東區中正里、東區南大里、東區復中里、東區龍山里、東區仙水里、東區新光里、東區錦華里、北區仁德里、北區福林里、北區古賢里、北區康樂里、北區新雅里、香山區東香里、香山區茄苳里
台中市	東區旱溪里、東區東信里、西區民龍里、北區文莊里、北區健行里、北區賴興里、北區六合里、北區大湖里、北區梅川里、西屯區西墩里、西屯區上石里、西屯區何厝里、西屯區永安里、西屯區何仁里、南屯區楓樹里、南屯區鎮平里、南屯區三和里、北屯區北興里、北屯區仁美里、北屯區新平里、北屯區大德里
嘉義市	東區後湖里、東區東山里、東區東噴里、東區後庄里、東區東平里、東區芳草里、西區大溪里、西區西平里、西區磚磘里、西區義昌里、西區民生里、西區湖內里
台南市	東區和平里、東區崇成里、東區東門里、東區成大里、東區大同里、南區荔宅里、南區再興里、南區府南里、北區華興里、北區大道里、北區安民里、北區中樓里、北區成德里、安南區淵東里、安南區淵西里、安南區公塭里、安南區城北里、安南區溪北里、安南區理想里、安南區國安里、安平區育平里、中西區開山里、中西區三合里
台北市	松山區吉仁里、松山區敦化里、信義區新仁里、信義區富台里、信義區景新里、大安區民炤里、大安區義村里、大安區龍安里、大安區錦安里、大安區福住里、大安區通化里、中山區新庄里、中正區忠勤里、大同區玉泉里、大同區蓬萊里、大同區國順里、萬華區雙園里、萬華區新忠里、萬華區日善里、萬華區銘德里、萬華區華中里、萬華區忠貞里、文山區景東里、文山區萬隆里、文山區興福里、文山區興昌里、文山區木新里、文山區樟林里、文山區萬興里、南港區合成里、南港區仁福里、內湖區康寧里、內湖區明湖里、士林區福中里、士林區永倫里、士林區岩山里、士林區永福里、士林區菁山里、北投區福興里、北投區中庸里
高雄市	鹽埕區新豐里、鼓山區雄峰里、鼓山區建國里、鼓山區龍井里、左營區路東里、左營區新下里、楠梓區惠豐里、楠梓區玉屏里、楠梓區瑞屏里、楠梓區隆昌里、楠梓區中和里、三民區寶龍里、三民區灣愛里、三民區達明里、三民區港北里、三民區德西里、新興區中東里、前金區長城里、前金區復元里、前金區社西里、苓雅區苓中里、苓雅區普照里、苓雅區林西里、苓雅區五權里、苓雅區民主里、前鎮區明禮里、前鎮區鎮榮里、前鎮區鎮中里、前鎮區瑞竹里、前鎮區瑞和里、前鎮區瑞平里、旗津區慈愛里、小港區港南里、小港區松金里、小港區鳳鳴里

參考資料

行政院主計處

2005 〈人力資源調查——94 年 12 月〉，中央研究院調查研究專題中心「學術調查研究資料庫」。

徐永明

2001 〈南方政治的形成？台灣政黨支持的地域差別，1994-2000〉，《中山大學社會科學季刊》2(4): 167-196。

黃　紀

2001 〈一致與分裂投票：方法論之探討〉，《人文及社會科學集刊》13(5): 541-574。

黃　紀、張益超
2001 〈一致與分裂投票：嘉義市一九九七年市長與立委選舉之分析〉，見黃紀與徐永明（編），《政治分析的層次》，頁 183-218。台北：韋伯文化事業出版社。

駱明慶
2006 〈廢票哪裡來？無效票定義範圍過大對 2004 年總統選舉的影響〉，《人文及社會科學集刊》18(4): 639-669。

Achen, Christopher H. and W. Phillips Shively
1995 *Cross-Level Inference.* Chicago: The University of Chicago Press.

Adolph, Christopher and Gary King
2003 "Analyzing Second-Stage Ecological Regression: Comment on Herron and Shotts," *Political Analysis* 11(1): 65-76.

Adolph, Christopher, Gary King, Michael C. Herron, and Kenneth W. Shotts
2003 "A Consensus on Second-Stage Analyses in Ecological Inference Models," *Political Analysis* 11(1): 86-94.

Anselin, Luc and Wendy K. Tam Cho
2002 "Spatial Effects and Ecological Inference," *Political Analysis* 10(3): 276-297.

Benoith, Kenneth and Gary King
2003 "EzI: An Easy Program for Ecological Inference," Manuscript. http://gking.harvard.edu/stats.shtml (November 18, 2006)

Cho, Wendy Tam
1998 "Iff the Assumption Fits. . .: A Comment on the King Ecological Inference Solution," *Political Analysis* 7(1): 143-163.

Cho, Wendy Tam and Brian J. Gaines
2004 "The Limits of Ecological Inference: The Case of Split-Ticket Voting," *American Journal of Political Science* 48(1): 152-171.

Duncan, Dudley and Beverly Davis
1953 "An Alternatives to Ecological Correlation," *American Sociological Review* 18(6): 665-666.

Freedman , D. A., Kevin, S. P., Ostland, M., and Roberts, M. R.
1998 "Review of A Solution to the Ecological Inference Problem," *Journal of the American Statistical Association* 93(444): 1518-1522.

Freedman, D. A., Kevin, S. P., Sacks, J., Smyth, C. A., and Everett, C. G.
1991 "Ecological Regression and Voting Rights," (with discussion) *Evaluation Review* 15: 673-711.

Gill, Jeff
2002 *Bayesian Methods: A Social and Behavioral Sciences Approach.* Boca Raton: Chapman & Hall/CRC.

Goodman, Leo A.
1953 "Ecological Regressions and Behavior of Individuals," *American Sociological Review* 18(6): 663-664.
1959 "Some Alternatives to Ecological Correlation," *American Journal of Sociology* 64(6): 610-625.

Herron, Michael C. and Kenneth W. Shotts
2003a "Using Ecological Inference Point Estimates as Dependent Variables in Second-Stage Linear Regressions," *Political Analysis* 11(1): 44-64.
2003b "Cross-Contamination in EI-R: Reply," *Political Analysis* 11(1): 77-85.
Imai, Kosuke, Gary King, and Olivia Lau
2006 "Zelig: Everyone's Statistical Software," *The Comprehensive R Archive Network* (CRAN). http://cran.r-project.org/ (December 24, 2006)
Imai, Kosuke, Ying Lu, and Aaron Strauss
2006 *Bayesian and Likelihood Inference for Models for 2×2 Ecological Tables: An Incomplete Data Approach.* Tech. rep., Department of Politics. Princeton: Princeton University.
Johnston, Ron and Charles Pattie
2000 "Ecological Inference and Entropy-Maximizing: An Alternative Estimation Procedure for Split-ticket Voting," *Political Analysis* 8(4): 333-345.
King, Gary
1997 *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data.* Princeton: Princeton University Press.
1999 "The Future of Ecological Inference Research: A Comment on Freedman et al.," *Journal of the American Statistical Association* 94(445): 352-355.
King, Gary, Ori Rosen, and Martin A. Tanner
1999 "Binomial-Beta Hierarchical Models for Ecological Inference," *Sociological Methods & Research* 28(1): 61-90.
King, Gary, Ori Rosen, and Martin A. Tanner (eds.)
2004 *Ecological Inference: New Methodological Strategies.* Cambridge: Cambridge University Press.
Mattos, Rogerio S. and Alvaro Veiga
2004 "A Structured Comparison of Goodman Regression, the Truncated Normal, and the Binomial-Beta Hierarchical Methods for Ecoligical Inference," in Gary King, Ori Rosen and Martin A. Tanner (eds.), *Ecological Inference: New Methodological Strategies*, pp. 351-382. Cambridge: Cambridge University Press.
Robinson, W. S.
1950 "Ecological Correlations and the Behavior of Individuals," *American Sociological Review* 15(3): 351-357.
Wakefield, Jon
2004 "Ecological Inference for 2×2 Tables," (with discuss) *Journal of the Royal Statistical Society, Series A* 167(3): 385-445.

The Statistical Approaches of Ecological Inference: A Brief Review, Comparison and Evaluation of Three Main Models

Hsin-hao Huang
Ph.D. Candidate, Department of Political Science,
National Chengchi University

ABSTRACT

The ecological inference problem arises when people make inferences about individual behavior from aggregate data. While this approach proves useful when individual information is inadequate, scholars should make inferences for unknown individual associations based upon observable aggregate information. Therefore, how to generate unknown individual associations accurately is what methodologists need to focus on making ecological inference.

To address the concerns addressed above, this paper intends to evaluate the major models used for ecological inference, by constructing a series of simulation experiments based on a real case and testing the model performances upon there. Firstly, the author briefly reviews the development of statistical approaches for making ecological inference, and discusses the statistical properties of three main models (Goodman's regression, King's EI model and Wakefield's hierarchical model). Secondly, the author tests the performance of different models through simulations, and compares the model estimations with the true value generated from plenty of survey data. Finally, the author concludes that the hierarchical model performs best in predicting quantities of interest about different individual associations, by relaxing the distributional assumption, and imposing a more exact functional form as well as a more flexible hierarchical structure.

Key Words: quantitative method, ecological inference, bayesian statistics