智慧科技與應用統計學報
第 11 卷 第 1 期
2013 年 6 月
25-34 頁

台灣地區二氧化硫之空間分佈

張雅梅*

淡江大學統計系

摘 要

二氧化硫一直是台灣地區空氣污染中主要污染物之一，本研究將二氧化硫的時空資料，表示為數個基底函數與平穩過程的線性組合，可把平均數與變異數估計問題視為迴歸分析，再利用 Tibshirani (1996)所提出的「最小絕對壓縮挑選機制」 (Least absolute shrinkage and selection operator，Lasso)，來選取基底函數及估計參數，分析結果以圖的方式呈現二氧化硫之空間分佈。總體而言，西半部地區在春季(五月)時，空氣中二氧化硫含量會上升偏高，而整個東半部及離島則全年皆低；至於各地區標準差在桃竹、雲林及高屏等三大工業區皆偏高。

關鍵詞：二氧化硫；空氣汙染；最小角度迴歸法；正最小絕對壓縮挑選機制；空間預測

*通訊作者聯絡方式：yamei628@gmail.com

Spatial Distribution of SO_2 in Taiwan Area

Ya-Mei Chang

Department of Statistics, Tamkang University, Taiwan

Abstract

SO_2 has been one of the major air pollutions in Taiwan area. In this research, the spatial-temporal data of SO_2 is decomposed as a linear combination of basis functions and stationary processes. The problem of mean and variance estimations can be considered as a regression. A subset selection method, least absolute shrinkage and selection operator (Lasso), proposed by Tibshirani (1996) is used to choose a suitable subset of the basis functions and estimate the parameters. The analysis results are demonstrated in graphs which can be easily observed the spatial distribution of SO_2 and the seasonal variation of it. The monthly mean of SO_2 is increasing in May over western Taiwan. As in eastern Taiwan and off-shore islands, the monthly mean of SO_2 is very low through the year. High values of the standard deviation in SO_2 occur in the surrounding area of three major industrial zones.

Key words: SO_2; air pollution; least angle regression; positive Lasso; spatial prediction

1. 簡介

二氧化硫一直是台灣地區空氣污染中主要污染物之一，它會造成呼吸器官的疾病，且提高某些癌症的致癌風險(陳美華，1984；林茂盛等，1984；郭耀綸，2008)，對人體健康影響甚鉅。不僅如此，二氧化硫會傷害農作物及行道樹(葛應欽，1996；姜伯穎與謝文斌，2000；陳姿汶與謝伸裕，2009)，且腐蝕建築物招致損失(魏維新，1975)。本研究利用空間統計方法，來分析二氧化硫的空間分佈情形，對政府單位或決策者在政策上，提供佐證的統計依據，可幫助健全機制或法令。

空間統計在許多領域上非常重要，文獻上最被廣泛使用的平穩模式爲 Matern Class (1986)，其假設兩點間的共變異數爲兩點距離的函數。但平穩的假設與許多實際收集到的空間資料相抵觸，因此建立非平穩(non-stationary)模式顯得格外重要。非平穩模式方面，有 Cohen and Jones (1969)、Creutin and Obled (1982)、Obled and Creutin (1986)、Wikle and Cressie (1999)等諸位學者，將非平穩模式建立在經驗正交函數(empirical orthogonal functions)上；Sampson and Guttorp (1992)、Guttorp and Sampson (1994)、Meiring et al. (1997)則是利用應變方法(deformation approach)；Nott and Dunsmuir (2002)利用核估計方法 (the kernel-based approach)；而 Nychka et al. (2002)提出小波收縮法(a wavelet shrinkage approach)來近似非平穩的共變異函數。以上這些非平穩模式的估計方法需要樣本共變異數，資料需重複量測。此外，Higdon et al. (1999)、Fuentes (2001)、Fuentes and Smith (2002)等人提出的迴旋方法(the convolution approach)，這些迴旋方法中，資料無需重複量測。Chang et al. (2010)將空間訊號表示爲數個基底函數與平穩過程的線性組合，把變異數估計問題視爲迴歸分析，再利用 Tibshirani (1996)所提出的「最小絕對壓縮挑選機制」(Least absolute shrinkage and selection operator，Lasso)，來選取基底函數及估計參數。此種方法可建構廣泛的模式，可用以描述平穩與非平品穩過程，且提供快而有效率的估計方法。

本研究參考 Chang et al. (2010)的作法來估計平均數與變異數，更進一步將每月的月平均表示爲數個基底函數的線性組合，此種方式可顯示二氧化硫的月平均在空間上不平穩的分佈情形。

本文共分爲五個章節，第二章是有關統計模式之建構，第三章介紹在該統計模式之下所採用的估計方法，第四章是二氧化硫資料的分析結果，第五章爲總結。

2. 統計模式

假設 s 爲時空訊號發生之地點，t 爲時空訊號發生之時間，將有興趣的時空訊號 $\{Y(s,t): s \in D \subset R^2, t \in T\}$ 表示如下：

$$Y(s,t) = \mu_m(s) + \eta(s,t), \tag{1}$$

$\mu_m(s)$ 爲時空訊號在地點 s 及月份 m 之平均值；而 $\eta(s,t)$ 爲期望值是 0，在空間上有相依性(spatially dependent)，但在時間上是獨立(temporally independent)的時空訊號。如此假設之下，$Y(s,t)$ 亦同樣在空間上有相依性，但在時間上是獨立的時空訊號。

時空訊號資料是藉由幾個觀測點站 $\{s_1, s_2, \ldots, s_n\}$ 所搜集得來的，搜集到的觀測資料

$\{Z(s,t): s \in D \subset R^2, t \in T\}$ 可表示爲：

$$Z(s,t) = Y(s,t) + \varepsilon(s,t),$$

$\varepsilon(s,t) \sim (0, \sigma_\varepsilon^2)$ 是一白噪音過程(white-noise process)，假設 $\varepsilon(s,t)$ 彼此互相獨立且具相同分配(independently and identically distributed，i.i.d.)，其可視爲觀測誤差。

將地點 s 及月份 m 之平均值 $\mu_m(s)$ ，表示爲數個基底函數 ϕ_j 的線性組合：

$$\mu_m(s) = \sum_{j=1}^{J} b_{j,m} \phi_j(s). \tag{2}$$

時空訊號 $\eta(s,t)$ 則視爲數個基底函數 ψ_k ，再加入數個平穩過程 W_l 的線性組合，如下：

$$\eta(s,t) = \sum_{k=1}^{K} \xi_k(t) \psi_k(s) + \sum_{l=1}^{L} W_l(s), \tag{3}$$

$\xi_k(t)$ 爲隨機係數(random coefficient)，其互相獨立且具相同分配 $\xi_k(t) \sim (0, \upsilon_k^2)$ ；$W_l(s)$ 的變異數爲 σ_l^2 ，且其相關係數函數爲

$$\rho_l(h) = cor(W_l(s), W_l(s+h)).$$

假設 $\{\phi_j(\cdot)\}$、$\{\psi_k(\cdot)\}$ 與 $\{\rho_l(\cdot)\}$ 皆爲已知，隨機係數 $\{\xi_k(\cdot)\}$、平穩過程 $\{W_l(\cdot)\}$ 與 $\{\varepsilon(\cdot)\}$ 間均爲獨立。

在以上的統計模式假設之下，對任何時間點 t ， $Z(s,t)$ 與 $Z(s',t)$ 之空間共變異數(spatial covariance)滿足下式：

$$C(s,s') = \sum_{k=1}^{K} \upsilon_k^2 \psi_k(s) \psi_k(s') + \sum_{l=1}^{L} \sigma_l^2 \rho_l(\|s-s'\|) + \sigma_\varepsilon^2 I(s=s'), \tag{4}$$

$\|\cdot\|$ 表歐氏距離(Euclidean distance)， $I(\cdot)$ 表指示函數(indicator function)。假設 $\{\phi_j(\cdot)\}$ 、 $\{\psi_k(\cdot)\}$ 與 $\{\rho_l(\cdot)\}$ 爲已知之下，有 $J \times 12 + K + L + 1$ 個參數要估計。在此研究中，將估計程序分成兩個步驟，先估計 $J \times 12$ 個參數 $b_{j,m} \geq 0$， $j = 1, \cdots, J$； $m = 1, \cdots, 12$ ，再估計剩餘的 $K + L + 1$ 個參數 $\upsilon_1^2, \cdots, \upsilon_k^2$, $\sigma_1^2, \cdots, \sigma_L^2, \sigma_\varepsilon^2$ ，其估計方法在下節詳述。

3. 估計方法

時空訊號 $\{Z(\cdot)\}$ 在地點 s 且在月份 m 之期望值爲 $\mu_m(s)$ ，如果我們在觀測站 $s_i \in \{s_1, s_2, \cdots, s_n\}$ 且在月份 m 有 T_m 筆觀測資料，可以求得樣本平均數

$$E_{i,m} = \frac{1}{T_m} \sum_{t \in \{Month\ m\}} Z(s_i, t),$$

$E_{i,m}$ 則爲 $\mu_m(s_i)$ 的動差估計式(moment estimator)，利用第(2)式的假設與動差法(method of moment estimator)可得：

$$E_{i,m} = b_{j,m}\phi_j(s_i).$$

由於月平均資料皆爲正值，因此需要 $b_{j,m} \geq 0$， $j = 1,\cdots,J$； $m = 1,\cdots,12$ 的限制，以保證 $\mu_m(s)$ 的估計值皆爲正的。在此研究中，我們利用限制最小平方法(constrained least squares, constrained LS)來估計 $b_{j,m}$ ：

$$\sum_{i=1}^{n}(E_{i,m} - \sum_{j=1}^{J} b_{j,m}\phi_j(s_i))^2 \text{ ，限制於 } b_{j,m} \geq 0,\ \ j = 1,\cdots,J;\ \ m = 1,\cdots,12, \tag{5}$$

也就是將第(5)式最小化，以求得 $b_{j,m}$ 的估計值 $\hat{b}_{j,m}$ 。求得 $\hat{b}_{j,m}$ 之後，對於空間上任何一點 s 其在月份 m 之期望值，皆可利用 $\hat{\mu}_m(s) = \sum_{j=1}^{J}\hat{b}_{j,m}\phi_j(s)$ ，求得該期望值之估計值。

我們預計參數個數 J 會很大，估計誤差也將會很大，爲了克服這個問題，本研究利用 Tibshirani (1996)提出之 Lasso 法，來選取與估計參數，在第(5)式中加入 Lasso 限制式：

$$\sum_{j=1}^{J}\left|b_{j,m}\right| < \alpha,$$

以 CV (cross validation)法選取參數 α 。加入 Lasso 限制式後，某些 $\hat{b}_{j,m}$ 的估計值爲 0，其被視爲未選取之參數，因此 Lasso 法可將選取與估計參數兩步驟同時進行，增加運算速度。Efron 等人(2004)所提的 Lars 演算法(least angle regression algorithm)，讓選取的變數與當前殘差的角度都相同，此舉更可有效運算求得 Lasso 估計值。

求得月平均的估計值 $\hat{\mu}_m(s) = \sum_{j=1}^{J}\hat{b}_{j,m}\phi_j(s)$ 之後，可以將觀測值減掉月平均的估計值，來算樣本共變異矩陣 $\Gamma = [\gamma_{ij}]$ 。令 $\mathbf{Z}_t = (Z(s_1,t),\ldots,Z(s_n,t))'$ 與 $\hat{\boldsymbol{\mu}}_m = (\hat{\mu}_m(s_1),\ldots,\hat{\mu}_m(s_n))'$ ，可得

$$\Gamma = \frac{1}{365}\sum_{m=1}^{12}\sum_{t\in\{Month\ \ m\}}(\mathbf{Z}_t - \hat{\boldsymbol{\mu}}_m)(\mathbf{Z}_t - \hat{\boldsymbol{\mu}}_m)'.$$

設

$$\boldsymbol{\beta} = (\beta_1,\cdots,\beta_{K+L+1})' = (\upsilon_1^2,\cdots,\upsilon_k^2,\sigma_1^2,\cdots,\sigma_L^2,\sigma_\varepsilon^2)',$$

與

$$\mathbf{x}_{ij} = (\psi_1(s_i)\psi_1(s_j),\cdots,\psi_K(s_i)\psi_K(s_j),\rho_1(\left|s_i - s_j\right|),\cdots,\rho_L(\left|s_i - s_j\right|),I(s_i = s_j))',$$

從第(4)式可知 $C(s_i,s_j) = \mathbf{x}_{ij}'\boldsymbol{\beta}$ 。將樣本共變異數 γ_{ij} 取代 $C(s_i,s_j)$ ，在 $\beta_K \geq 0$， $k = 1,\cdots,K+L+1$ 的限制下，同樣利用限制最小平方法來估計 $\boldsymbol{\beta}$ ：

$$\sum_{i=1}^{n}\sum_{j=i}^{n}(\gamma_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta})^2 \text{ ，限制於 } \beta_K \geq 0,\ \ k = 1,\cdots,K+L+1, \tag{6}$$

也就是將第(6)式最小化。在此，同樣利用 Lars 演算法，來挑選與估計參數 β_K 。

至於基底函數ϕ_j與ψ_K，本研究在此選用高斯核函數(Gaussian kernels)，核中心(kernel centers)選擇在觀測站位置$\{s_1, s_2, \cdots, s_n\}$，而核刻度(kernel scales)則選擇核中心間最小距離的倍數(Chang et al.，2010)，如下：

$$\phi_j(s), \psi_k(s) \in \left\{ \exp\left\{ -\frac{\|s-u\|}{2v^2} \right\}, (u,v) \in \bigcup_{i=1}^{n} \Omega_i \right\};$$

$$\Omega_i \equiv \{(s_i, \tau d_i) : \tau = 0.75, 1, 1.25\},\ d_i = \min_{j:j\neq i}\{\|s_i - s_j\|;\ j = 1, \ldots, n\}.$$

其中τ的值是控制帶寬參數(bandwidth parameter)的大小。Chang et al. (2010)建議τ的選取範圍爲$0.75 \le \tau \le 1.25$，因此本研究選取$\tau = 0.75, 1, 1.25$。對於基底函數ϕ_j與ψ_K，雖然一開始皆置入一組相同的高斯核函數，但經由 Lars 演算法選取與估計參數後，最後選取的基底函數將不盡相同。而平穩過程W_l的相關係數函數則爲$\rho_l(h) = \exp\{-d/\alpha_l\}$，$\alpha_l$的選取則是參考 Chang et al. (2010)。

4. 二氧化硫實例分析結果

本研究分析 2009 年台灣地區二氧化硫(SO_2)日平均資料，資料來源爲行政院環保署空氣品質監測站。全台共有 76 個監測站，監測站之位置分佈請見圖 1，監測站位置集中在台灣本島西半部，共有 68 個監測站，東半部有 5 個監測站，離島地區僅有 3 個監測站。

台灣地區二氧化硫在空間上的分佈情形，會隨著季節而改變，此可藉圖 2 一窺端倪。圖 2 爲原始資料的月平均折線圖，因監測站個數不少，在此僅展示在永和、新竹、仁武、花蓮、林園及小港等六個監測站的月平均值。永和、新竹、花蓮等三個監測站之二氧化硫月平均，不會受季節影響而有太大的波動；但仁武、林園及小港等三地則受季節的影響很大。仁武與林園的二氧化硫月平均會在春夏時下降許多，秋冬時期則上升；而小港地區的二氧化硫則在春季時下降，夏初時略升，夏末又會稍降，秋冬時期上揚。因此要描述台灣地區的二氧化硫平均值在空間上分佈情形，需逐月探討。

本研究根據上節描述之方法估計各地的月平均$\hat{\mu}_m(s) = \sum_{j=1}^{J} \hat{b}_{j,m}\phi_j(s)$，圖 3 爲$\hat{\mu}_m(s)$在春(五月)、夏(八月)、秋(十一月)、冬(二月)的空間分佈情形。整體而言，高屏工業區空氣中的二氧化硫含量全年都偏高，北部桃園地區亦是，但桃園地區的二氧化硫含量會秋季(十一月)時下降許多。其餘西半部地區會在春季(五月)時，空氣中二氧化硫含量會上升偏高，而整個東半部及離島則全年皆低。

至於各地區標準差$\sqrt{\hat{C}(s_i, s_j)} = \sqrt{\mathbf{x}_{ij}\hat{\boldsymbol{\beta}}}$的空間分佈狀況可見圖 4，在桃竹、雲林及高屏等三大工業區皆偏高。由圖 3 與圖 4 可見空氣中二氧化硫含量，在平均值與標準差方面皆顯示空間上的不平穩(spatially nonstationary)狀況。

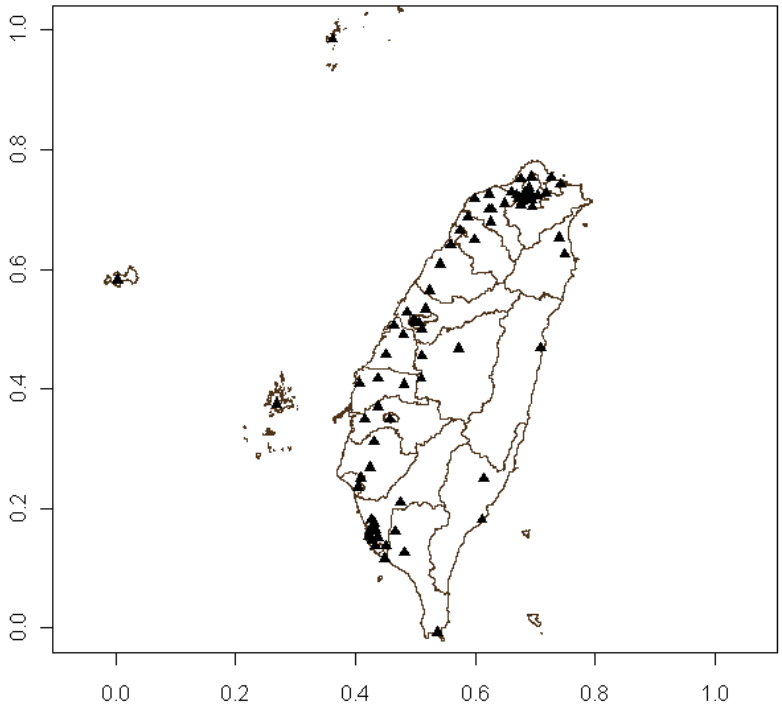

圖 1. 監測站之位置分佈

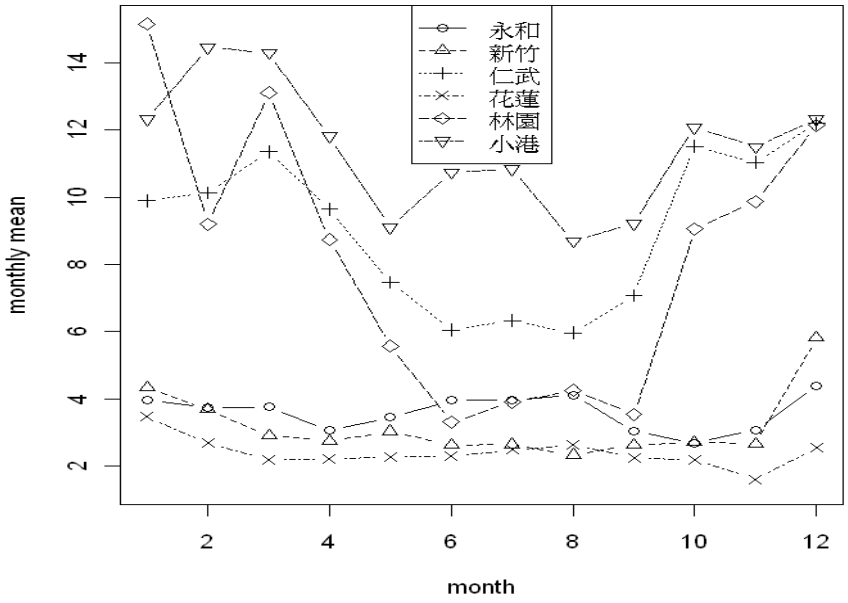

圖 2. 永和、新竹、仁武、花蓮、林園及小港等六個監測站的月平均折線圖

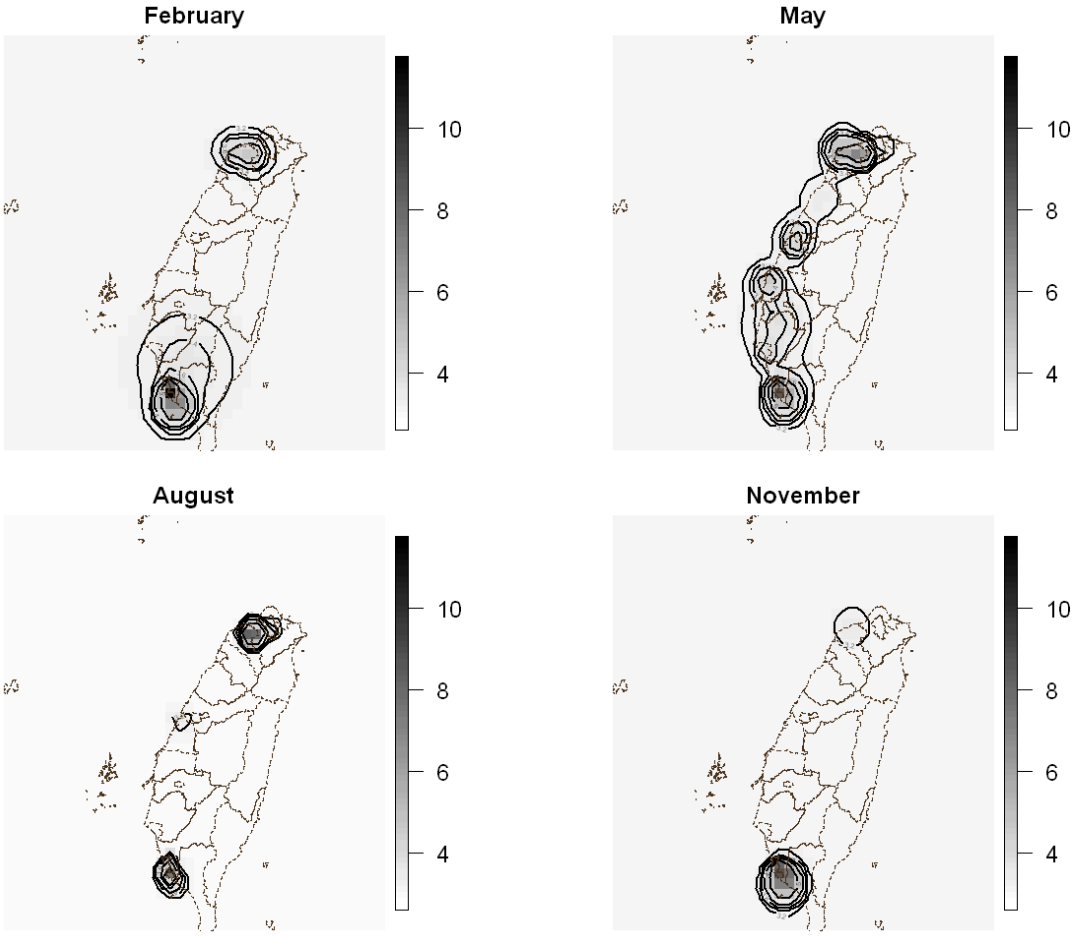

圖 3. 二氧化硫之月平均空間分佈圖

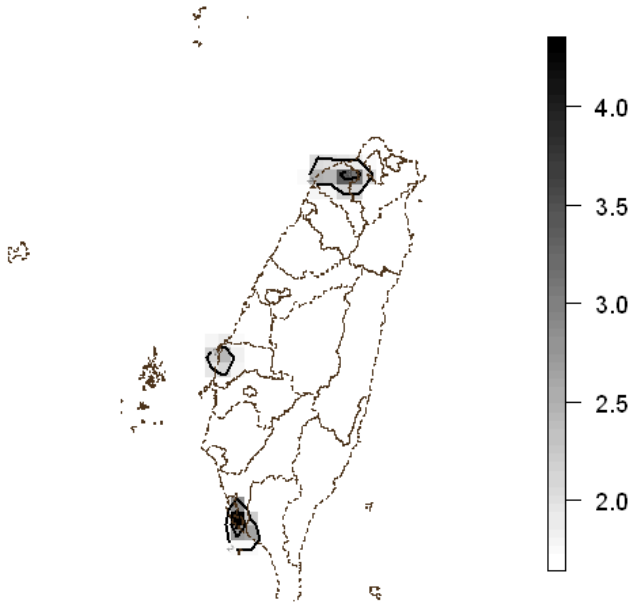

圖 4. 各地區標準差之空間分佈

5. 總結

二氧化硫對人體健康與生態環境的傷害不容小覷，長久以來一直爲台灣地區的主要汙染物之一。環保署長期對空氣品質作監測，觀察空氣品質的變化，這些空汙監測資料得知不易，如何有效運用該資料成爲一主要課題。

本研究利用空間統計方法，將二氧化硫的空間分佈依季節以圖方式呈現，更可簡易觀測出其分佈狀況的改變。由月平均可知，空氣中二氧化硫的含量會隨季節不同而有所增減，各地區的增減幅度不盡相同。整體而言，春天的二氧化硫的含量偏高，空氣品質較差。桃竹、雲林及高屏等三大工業區的標準差皆偏高，顯示其周遭地區二氧化硫的變化較大。

本文僅單純研究二氧化硫的空間分佈，未來可考慮在統計模式中加入一些解釋變數，對預測空氣中二氧化硫含量將有所幫助，這些建議也可作爲台灣空氣汙染研究的重要參考依據。

感謝

本文研究成果由國科會專題研究計畫 NSC 100-2118-M-032-001 之經費補助，特此感謝。

參考文獻

[1] 魏維新 (1975)，二氧化硫的污染及其控制，*科學月刊*，62。

[2] 陳美華 (1984)，二氧化硫對作物細胞構造之影響，*中華農業研究*，33(4)，387-394。

[3] 林茂盛、陳美華、張瑞明 (1984)，二氧化硫爲害臺灣重要作物徵狀診斷研究，*中華農業研究*，33(2)，153-158。

[4] 葛應欽 (1996)，台灣空氣污染與社區居民健康效應，*高雄醫學科學雜誌*，12(12)，657-699。

[5] 姜伯穎、謝文斌 (2000)，急性吸入性傷害，*當代醫學*，324，789-802。

[6] 郭耀綸、陳家儀、李芳胤、仲崇毅 (2008)，二氧化硫短期燻氣對十七種樹苗葉部傷害及生理活性之影響，*作物、環境與生物資訊*，5(2)，113-126。

[7] 陳姿汶、謝伸裕 (2009)，都市空氣污染環境中居住與運動探討，*輔仁大學體育學刊*，8，278-288。

[8] 李嘉塗、余伍洲、仲崇毅、呂召宏、黃怡萍、葉軒丞、吳福霖 (2011)，臺灣五種行道樹對二氧化硫耐受性之分析，*臺灣農業化學與食品科學*，49(1)，11-18。

[9] Chang, Y.-M., Hsu, N.-J. and Huang, H.-C. (2010). Semiparametric estimation and selection for nonstationary spatial covariance functions. *Journal of Computational and Graphical Statistics*, 19, 117-139.

[10] Cohen, A. and Jones, R. H. (1969). Regression on a random field. *Journal of the American Statistical Association*, 64, 1172-1182.

[11] Creutin, J. and Obled, C. (1982). Objective analysis and mapping techniques for rainfall fields: an objective comparison. *Water Resources Research*, 18, 413-431.

[12] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals*

of Statistics, 32, 407-499.

[13] Fuentes, M. (2001). A high frequency kriging approach for non-stationary environmental processes. *Environmetics*, 12, 469-483.

[14] Fuentes, M. and Smith, R. L. (2002). A new class of nonstationary spatial models. *Technical Report*, North Carolina State University.

[15] Higdon, D., Swall, J. and Kern, J. (1999). Non-stationary spatial modeling. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds.) *Bayesian Statistics*, Oxford University Press, Oxford, 6, 761-768.

[16] Matern, B. (1986). *Spatial Variation*, Second Edition, Lecture Notes in Statistics, Springer, New York.

[17] Meiring, W., Monestiez, P., Sampson, P. D. and Guttorp, P. (1997). Developments in the modeling of nonstationary spatial covariance structure from space–time monitoring data. *Geostatistics Wollongong '96*, In E. Y. Baafi and N. A. Schofield (eds.) Kluwer Academic, Dordrecht, 162-173.

[18] Nott, D. J. and Dunsmuir, W. T. M. (2002). Estimation of nonstationary spatial covariance structure. *Biometrika*, 89, 819-829.

[19] Nychka, D., Wikle, C. and Royle, J. A. (2002). Multiresolution models for nonstationary spatial covariance function. *Statistical Modeling*, 2, 315-331.

[20] Obled, C. and Creutin, J. D. (1986). Some developments in the use of empirical orthogonal functions for mapping meteorological fields. *Journal of Applied Meteorology*, 25, 1189-1204.

[21] Sampson, P. and Guttorp, P. D. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87, 108-119.

[22] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267-288.

[23] Wikle, C. K. and Cressie, N. (1999). A dimension-reduced approach to space–time kalman filtering. *Biometrika*, 86, 815-829.