INTERPOLACIÓN ESPACIAL DE CONCENTRACIONES DE OZONO EN LA ZONA METROPOLITANA DEL VALLE DE MÉXICO, BASADA EN MÉTODOS DE KRIGING Y COKRIGING

Óscar BORREGO HERNÁNDEZ[1], Mario Miguel OJEDA RAMÍREZ[1],
José Agustín GARCÍA REYNOSO[2], Claudio Rafael CASTRO LÓPEZ[1]
[1]*Facultad de Matemática, Universidad Veracruzana, México*
[2]*Centro de Ciencias de la Atmósfera, Universidad Nacional Autónoma de México, México*
oborrego@uv.mx, mojeda@uv.mx, agustin@atmosfera.unam.mx, ccastro@uv.mx

RESUMEN

La contaminación atmosférica es el principal problema ambiental que enfrenta la Ciudad de México, donde el ozono, particularmente, es uno de los contaminantes de mayor impacto para la salud humana. En el presente trabajo se evalúan tres procedimientos de interpolación espacial para estimar la concentración de ozono en la tropósfera de la Ciudad de México, en localizaciones donde esta no es medida, a partir de datos provenientes de estaciones de monitoreo. Los procedimientos evaluados son el kriging, el cokriging y el cokriging colocado. Como variables secundarias para los métodos basados en cokriging, se consideran las concentraciones de dióxido de nitrógeno (NO_2), la temperatura y la humedad relativa. La relación entre las variables secundarias y el ozono se analiza previamente a partir de un método estadístico de partición recursiva con estructura de árbol. Los tres procedimientos se comparan mediante validación cruzada, considerando datos del año 2010.

Palabras clave: interpolación espacial, kriging, cokriging, contaminación atmosférica, ozono, detección automática de interacciones.

ABSTRACT

Air pollution is the main environmentl issue in Mexico City, where ozone is one of the most damaging pollutant for human health. In this work they are compared three spatial interpolation procedures to estimate the tropospheric ozone concentration for non-observed locations in Mexico City, considering data obtained in monitoring stations. The compared procedures are kriging, cokriging and collocated cokriging. Nitrogen dioxid (NO_2) concentration, temperature and relative humidity are considered as secondary variables for the cokriging methods. The relationship among the secondary variables and ozone is previously analyzed with a tree based recursive partitioning statistical method. The interpolation procedures are compared by cross-validation, using data for the year 2010.

Keywords: spatial interpolation, kriging, cokriging, air pollution, ozone, automatic interaction detection.

1. INTRODUCCIÓN

La contaminación atmosférica es el principal problema ambiental que enfrenta la Zona Metropolitana del Valle de México, donde la salud de la población está en riesgo o es afectada por

diversos contaminantes, los cuales pueden rebasar los límites normales de concentración ambiental ocasional o sistemáticamente. El Sistema de Monitoreo Atmosférico del Valle de México (SIMAT) se encarga de vigilar y evaluar la calidad del aire en dicha región, a partir de la información proveniente de estaciones de monitoreo. En esta zona, el ozono es uno de los contaminantes de mayor impacto en la salud humana y en los ecosistemas.

En la presente investigación se evalúan modelos y procedimientos de interpolación espacial para estimar la concentración de ozono en la tropósfera de la Ciudad de México, en localizaciones donde esta no es medida, a partir de datos provenientes de estaciones de monitoreo. Los procedimientos evaluados son el kriging ordinario, el cokriging ordinario y el cokriging colocado. Como variables secundarias para los métodos basados en cokriging, se consideran las concentraciones de dióxido de nitrógeno (NO_2), la temperatura y la humedad relativa. La relación entre las variables secundarias y el ozono se analiza previamente a partir de un método estadístico de partición recursiva con estructura de árbol. Los modelos se comparan mediante validación cruzada, considerando datos del año 2010.

2. MODELACIÓN DE PROCESOS ESPACIALES

Un proceso estocástico en el espacio se define como una familia de variables aleatorias que pueden ser indexadas en el espacio: $\{Z(s): s \in D\}$. En esta investigación se considera el caso del plano, o sea, $D \; R^2$.

2.1. Estacionariedad e isotropía

En el caso la modelación espacial, es muy frecuente considerar como hipótesis la estacionariedad del proceso. Cuando $Z(s)$ presenta una media invariante en el espacio: $E[Z(s)] =$, y se cumple:

(1) $$Cov[Z(s), Z(s+h)] \equiv C(h)$$

para cualesquiera s,h, o sea, esta expresión solo depende del vector separación h y no de la localización s, entonces se dice que el proceso es **débilmente estacionario**. Puede demostrarse, ver (Cressie, 1993) o (Schabenberger et al., 2005), que como consecuencia se cumplirá

(2) $$\frac{1}{2} Var[Z(s) - Z(s+h)] = \gamma(h)$$

también es invariante en el espacio. La función $\gamma(h)$ es llamada **variograma**.

También resulta interesante la situación en que la función de variograma no depende de la dirección del vector de diferencias h, sino solo de su valor absoluto, o sea, $\gamma(h) = \gamma^*(|h|)$, donde $|h|$ es la norma euclidiana de h. En lo sucesivo, abusando un poco de la notación, se usará la expresión $\gamma(h)$ (donde h se refiere a la distancia y no al vector separación) en lugar de $\gamma^*(|h|)$.

Las estimaciones empíricas de la función de variograma no siempre son consistentes con las condiciones de estacionariedad , ver (Cressie, 1993), (Schabenberger et al., 2005) o (Sherman, 2011). Por otra parte, solo se obtienen estimaciones para los valores de distancia correspondientes a las localizaciones observadas. Es por ello que en la práctica se adoptan modelos paramétricos, y se buscan estimados para los parámetros basados en los datos. En el presente trabajo se consideraron los modelos: esférico, exponencial, gaussiano y lineal, ver (Cressie, 1993), (Schabenberger et al., 2005) o (Sherman, 2011).

2.1.a. KRIGING

La predicción espacial clásica, comúnmente llamada **kriging**, tiene por objetivo estimar el valor $Z(s_0)$ de no observado, a partir de las observaciones $z_i = Z(s_i)$, $i = 1, \ldots, n$. El objetivo es encontrar el Z_0 mejor predictor lineal insesgado de $Z(s_0)$, el cual estará dado por una combinación lineal de la forma

$Z_0=\lambda^T\mathbf{Z(s)}$ donde $\lambda\ R^n$ es el vector de coeficientes y **Z(s)** es el vector de variables $Z(S_i)$. El proceso $Z(s)$ suele descomponerse según el modelo: $Z(s)=(s)+e(s)$, donde (s)=E[Z(s)] es la media del proceso y $e(s)$ es un proceso de errores con $E[e(s)]=0$. Particularizando la estructura de (s) y exigiendo propiedades como la estacionariedad para $e(s)$ se obtienen diferentes modelos de kriging.

2.1.b. Kriging ordinario

Si el proceso $e(s)$ es débilmente estacionario y $(s)=$ (constante), el modelo es conocido como kriging ordinario, y se plantea de la siguiente forma:

$$Z(s),=,1_n\ +e(s),Z(s_0),=,\ +e(s0),$$

Donde $e(s)$ es el vector de variables $e(s_i)$, y $1_n\ R^n$ es el vector de coordenadas unitarias. La solución se obtiene a partir del sistema de ecuaciones, ver (Cressie, 1993), (Schabenberger et al., 2005) o (Sherman, 2011):

$$\Gamma\lambda+1_n m,=\gamma_0,1_n{}^T\lambda,=1$$

Donde $\Gamma=[\gamma(|s_i-s_j|)]\ R^{nxn}$, $\gamma_0=[\gamma(|s_i-s_0|)]$, R^n y m es un multiplicador de Lagrange. En este punto $\gamma(h)$ se entiende como la función de variograma del proceso $e(s)$.

2.1.c. Kriging universal

Cuando la media no es constante, también puede expresarse como combinación lineal: $M(s)=\sum_{i=1}^{p}\beta_i f_i(s)$ Es conveniente fijar $f_1(s)=1$ para incluir el modelo de media constante.

Con esa estructura para la media se obtiene el modelo de kriging universal:

$$Z(s),=,X(s)\beta+e(s),Z(s_0),=,x(s_0)^T\beta+e(s_0),$$

Donde $X(s)=\left[f_j(s_i)\right]_{ij}\ R^{nxp}$ y $x(s_0)=\left[f_i(s_0)\right]_i\ R^p$ son respectivamente una matriz y un vector de valores explicatorios, $\beta=\left[\beta_i\right]_i$. Aplicando el método de los multiplicadores de Lagrange se llega al sistema de ecuaciones:

$$\Gamma\lambda+X(s)m,=\gamma_0,X(s)^T\lambda,=x(s_0)$$

Donde $m\ R^p$ es el vector de multiplicadores. Frecuentemente se modela (s) con estructura polinomial sobre x e y.

2.1.d. Cokriging

Los modelos de kriging pueden extenderse para casos multivariados, recibiendo la denominación de *cokriging*. La idea es la predicción espacial en nuevas localizaciones que utiliza tanto la información de mediciones directas del proceso, como las medidas de otros procesos componentes. De manera similar a los modelos de kriging se transforma el problema de optimización en un sistema de ecuaciones lineales, ver detalles en (Cressie, 1993), (Schabenberger et al., 2005) o (Sherman, 2011).

Cokriging colocado

El modelo de cokriging colocado es un caso particular de cokriging donde para las variables secundarias no se consideran todas las observaciones, sino solo aquellas correspondientes a la localización a predecir (S_0).

3 METODOLOGÍA

En la presente investigación se consideraron los siguientes modelos de (co)kriging:

Abreviatura	Tipo de (co)kriging	Variables secundarias
OK	kriging ordinario	-
UK	kriging universal	-
COK.TMP	cokriging ordinario	temperatura
COK.ALL	cokriging ordinario	temperatura, humedad y concentración de NO_2
CUK.TMP	cokriging universal	temperatura
CUK.ALL	cokriging universal	temperatura, humedad y concentración de NO_2
CUK.col	cokriging ordinario colocado	temperatura

Para OK y UK se ajustan y comparan los diferentes modelos de variograma antes referidos. La estructura considerada para la media en el caso de UK es:

$$M(x, \gamma)=\beta_6 x^2+\beta_5 \gamma^2+\beta_4 x\gamma+\beta_2 \gamma+\beta_1$$

Los siete modelos de kriging son ajustados y comparados mediante validación cruzada, siguiendo la regla *dejar uno afuera* (*leave one out*), según la cual se elige un registro del conjunto de datos para conformar un conjunto de prueba unitario y el resto se utiliza como conjunto de entrenamiento, repitiendo este proceso para cada uno de los registros. La comparación de los modelos se realiza teniendo en cuenta la raíz del error cuadrático medio (RMSE) en la predicción.

4. DETECCIÓN DE INTERACCIONES

Es conocido que la concentración de ozono está relacionada con variables meteorológicas y otros contaminantes como el NO_2, ver por ejemplo (Rojas-Avellaneda y Martínez-Cervantes, 2011). Para explorar la existencia de tales interacciones se consideraron 6236 mediciones horarias (1559 para cada variable, entre las 10 y las 17 horas inclusive) de O_3, NO_2, humedad relativa (RH) y temperatura (TMP) tomadas en la estación Merced (MER) durante el año 2010. Se aplicó el algoritmo TAID (*Tau Automatic Interaction Detection*) tomando como variable respuesta el O_3 y las restantes como predictoras. Este es un algoritmo de partición recursiva que genera un modelo de árbol ternario para la clasificación o segmentación de las observaciones, basado en el análisis de correspondencias no simétrico sobre tablas de contingencia.

Pero el TAID fue diseñado para la detección de interacciones entre variables nominales y no continuas, por lo que fue necesario discretizar los datos. Para ello se aplicó una división en subintervalos del intervalo de valores de cada variable, codificando cada valor según el subintervalo correspondiente. Los subintervalos se determinaron a partir de los cuartiles de cada variable como muestra la siguiente tabla:

Subintervalos				
O3	NO2	TMP	RH	Codificación
[0.001,0.020]	[0.004,0.020]	[7.0,17.4]	[1,18]	1
(0.020,0.040]	(0.020,0.030]	(17.4,20.9]	(18,29]	2
(0.040,0.060]	(0.030,0.050]	(20.9,23.5]	(29,45]	3
(0.060,0.150]	(0.050,0.150]	(23.5,29.3]	(45,100]	4

En la figura 1 aparece el árbol obtenido, mostrando para cada nodo la siguiente información:

- La variable de mayor poder predictivo en ese nivel de la partición
- Los valores de dicha variable correspondientes al nodo
- **N**: la cantidad de observaciones (registros) incluidos en el nodo
- Las clases que predominan en el nodo
- La precisión en esa clasificación, o sea, la porción de registros que se encuentran en tales clases

En este caso cada clase se corresponde con uno de los subintervalos de concentración de O3. La clasificación solo se muestra para nodos terminales en los que el índice de entropía fue menor o igual a 0.8 (40% del máximo posible: log_2 4=2), y en el rango de clases se incluyeron aquellas que tuvieron más del 25% de representación en el nodo.

Como puede observarse en la figura 1, la temperatura es la variable de mayor poder predictivo para el 1er y 2do nivel de la partición, seguida por la humedad relativa en el 3ro y el dióxido de nitrógeno en el 4to. Las siguientes reglas pueden extraerse del árbol obtenido:

Premisa	Conclusión	Precisión
TMP ≤17.4	O3 ≤0.020	0.75
17.4< TMP ≤20.9, RH >45	O3 ≤0.040	0.87
TMP >20.9, RH >45	0.020< O3 ≤0.060	0.85
TMP >20.9, 18< RH ≤29, 0.020< NO2 <0.050	O3 >0.060	0.66
TMP >20.9, 29< RH ≤45, 0.020< NO2 <0.050	O3 >0.040	0.89
20.9< TMP ≤23.5, RH ≤18, (NO2 ≤0.020 ó NO2 >0.050)	O3 >0.020	1.00

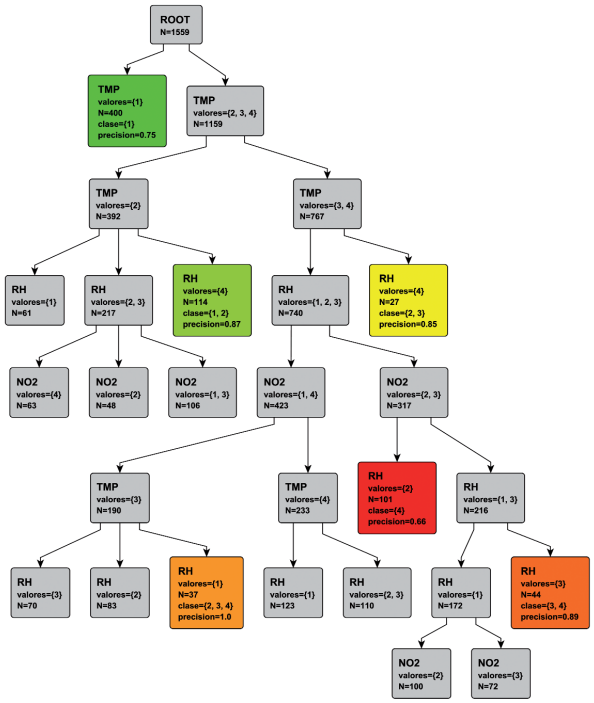

FIG. 1: *Árbol de partición recursiva*

En sentido general puede observarse que las concentraciones de O3 son mayores en situaciones de mayor temperatura, menor humedad y menor concentración de NO2.

6. RESULTADOS Y DISCUSIÓN

Para comparar el desempeño de los diferentes modelos se tomaron en cuenta mediciones de 22 estaciones para el día 12 de junio de 2010 a las 15 horas (se eligió esa fecha y esa hora por presentar una mayor disponibilidad de mediciones).

6.1 Ajuste de modelos de variograma

6.1.a. Media constante

Bajo la hipótesis de estacionariedad e isotropía, se ajustaron los cuatro modelos de variograma mencionados anteriormente. En la siguiente tabla se presentan los valores de los parámetros ajustados así como la raíz del error cuadrático medio (RMSE).

Modelo	Nugget	Sill	Range	RMSE
esférico	4.402503e-05	0.0002839510	32.37717	2.387804e-05
exponencial	4.029569e-05	0.0005555363	41.00488	2.387632e-05
gaussiano	6.731472e-05	0.0002273423	11.07599	2.381687e-05
lineal	7.500000e-05	0.0002250000	20.00000	2.433976e-05

El modelo elegido para aplicar el kriging ordinario fue el gaussiano, por ser el mejor evaluado según el RMSE, aunque los errores son relativamente cercanos para todos.

6.1.b. Media polinomial

Para ninguno de los modelos de variograma se alcanzó un nivel de convergencia aceptable, debido a que el ajuste degeneró en casos singulares. No obstante, esto no quiere decir que los ajustes no sean útiles (ver Pebesma, 2004), pero sí puede ser una sugerencia de no estacionariedad o no isotropía. Los parámetros ajustados y los errores fueron los siguientes:

Modelo	Nugget	Sill	Range	RMSE
esférico	8.239013e-05	8.239013e-05	50.00	1.677730e-05
exponencial	3.522604e-05	3.227092e-04	86.40	1.927492e-05
gaussiano	9.619965e-05	6.181583e-01	2956.04	2.255259e-05
lineal	2.000000e-05	2.700000e-04	50.00	2.281464e-05

A pesar de lo anterior se procedió con la aplicación del kriging universal, en aras de la comparación, eligiendo el ajuste de menor RMSE. En este punto, se desechó el modelo esférico dado que el ajuste devino en un efecto de *nugget* puro, lo que condujo a la elección del modelo exponencial.

6.2. Comparación de los modelos de kriging

Los resultados de la comparación mediante validación cruzada son los siguientes:

Modelo	RMSE
CUK.TMP	0.009367621
UK	0.009726906
CUK.ALL	0.010415382
COK.TMP	0.010904973
OK	0.011239855
COK.col	0.013927397
COK.ALL	0.029120853

El modelo CUK.TMP resultó ser el mejor evaluado, en sentido general los modelos de media polinomial mostraron mejor evaluación que sus homólogos de media constante. En el caso del cokriging, se obtuvieron mejores resultados considerando solo la temperatura como variable secundaria que al considerar de manera conjunta la temperatura, la humedad y el NO_2.

En la figura 2 se muestra el mapa resultante de la interpolación hacia una malla, mediante cuatro modelos de (co)kriging (OK, UK, CUK.TMP, COK.col), donde se observa la similitud entre los mapas de los modelos de (co)kriging universal. La varianza de kriging para los cuatro modelos aparece en la figura 3, notándose en general un mejor ajuste en las regiones de mayor concentración de estaciones, en particular en el caso de CUK.TMP es conexa y más extensa la región de menor varianza, con respecto a los otros modelos.

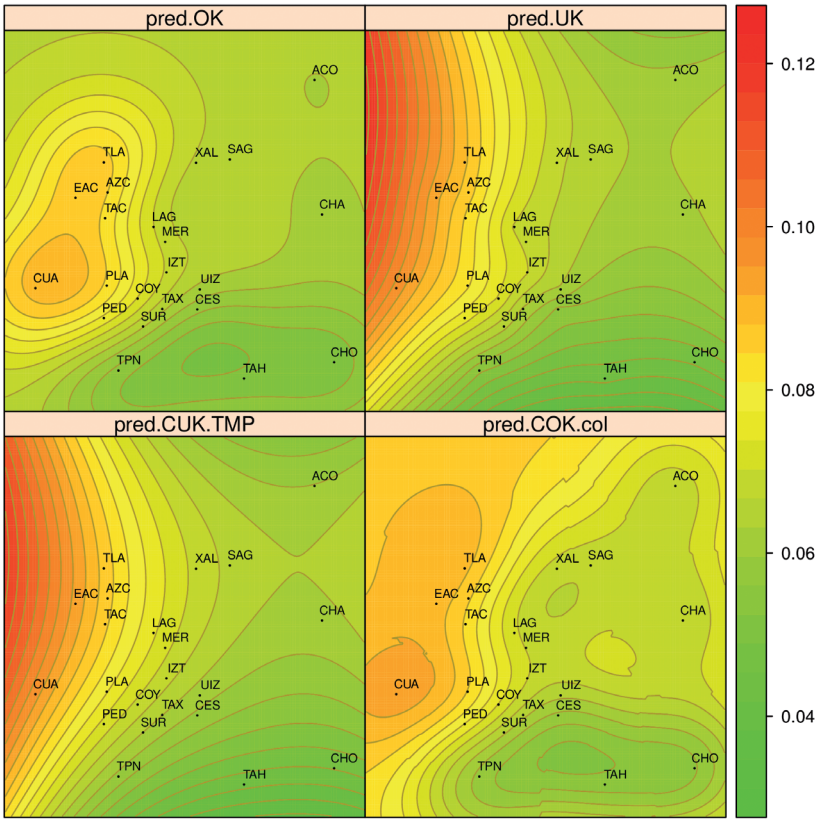

FIG. 2: *Predicción de los modelos de Kriging*

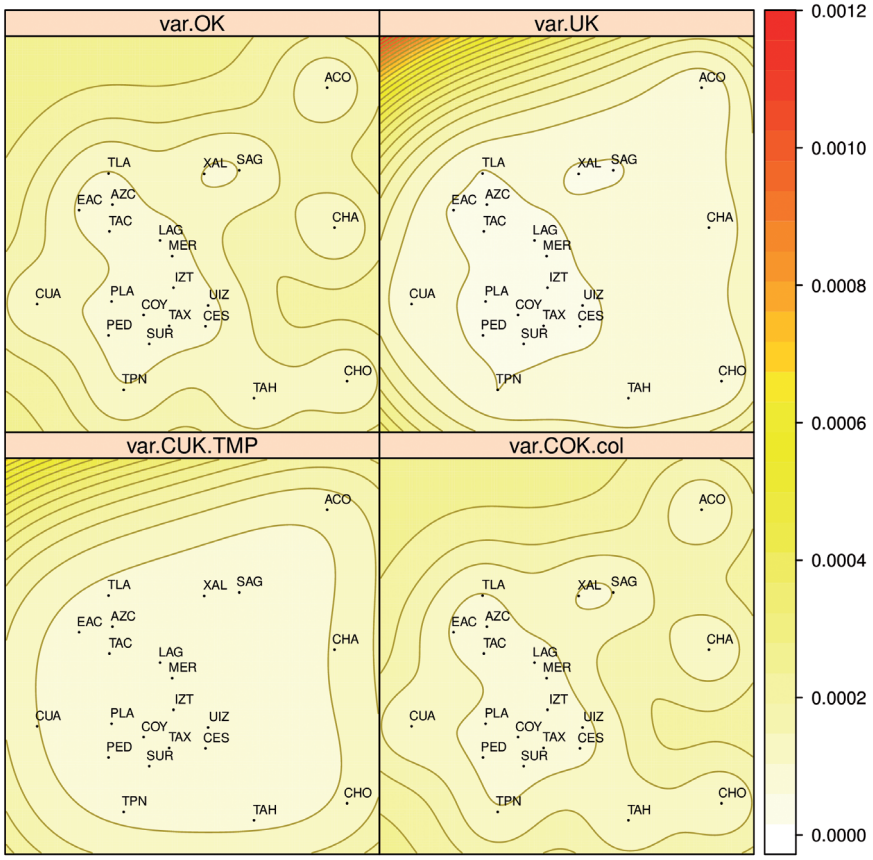

Fig. 3: *Varianza de Kriging*

7. CONCLUSIONES

Se ha confirmado la hipótesis de la influencia que tienen otras variables en la concentración de ozono, fundamentalmente la temperatura, y con ello, la validez del enfoque multivariado del problema. A diferencia de otras investigaciones, como (Rojas-Avellaneda y Martínez-Cervantes, 2011), los modelos de (co)kriging universal han mostrado un mejor desempeño que los de (co)kriging ordinario. Aunque curiosamente, en el ajuste de modelos de variograma con media polinomial el proceso se detuvo por singularidades en la optimización.

REFERENCIAS BIBLIOGRÁFICAS

Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V. (2008). *Applied spatial data analysis with R* Use R! Series, Springer.

Castro López, Claudio Rafael y Galindo Villardón, Purificación (2005). *Contribuciones al Análisis de Segmentación Basada en Árboles Ternarios* Universidad Veracruzana, Universidad de Salamanca.

Choi, Jungsoon (2008). *Multivariate spatio-temporal modeling of environmental-health processes.* Ph.D., North Carolina State University.

Choi, J., Fuentes, M., Reich, B.J. y Davis, J.M. (2009). *Multivariate spatial-temporal modeling and prediction of speciated fine particles.* J. Stat. Theory Pract.

(Cressie, 1993) Cressie, Noel A. C. (1993). *Statistics for spatial data*. Wiley series in probability and mathematical statistics.

Deligiorgi, D. and Philippopoulos, K. (2011). *Spatial Interpolation Methodologies in Urban Air Pollution Modeling: Application for the Greater Area of Metropolitan Athens, Greece*. Advanced Air Pollution. National and Kapodistrian University of Athens, Greece.

Dutilleul, P. y Binel-Alloul B. (1996). *A double multivariate model for statistical analysis of spatio-temporal environmental data* Wiley Environmetrics Journal, Volume 7, Issue 6.

Gelfand Alan E. et al. (2010). *Handbook of Spatial Statistics.* Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Taylon & Francis Group. Boca Raton.

Glover, D., Jenkins, W. and Doney, S. (2008). *Objective Mapping and Kriging.* Chapter 7 in *Modeling Methods for Marine Science.* Woods Hole Oceanographic Institution.

Hoffman, F. M., Hargrove, W. W., Mills, R. R., Mahajan, S., Erickson, D. J., y Oglesby, R. J. (2008) *Multivariate Spatio-Temporal Clustering (MSTC) as a Data Mining Tool for Environmental Applications*. iEMSs: International Congress on Environmental Modelling and Software.

Hooyberghs J., Mensink C., Dumont G., Fierens F. (2006). *Spatial interpolation of ambient ozone concentratios from sparse monitoring points in Belgium*. J. Environ. Monit., 8 1129-1135.

Pebesma, E.J. (2004). *Multivariable geostatistics in S: the gstat package. Computers & Geosciences*, 30: 683-691.

Pebesma, E. (2011). *Classes and methods for spatio-temporal data in R: the spacetime package*. Institute for Geoinformatics, University of Münster. R package version 0.4-0.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rojas-Avellaneda, D. y Martínez-Cervantes, J. (2011). *Using the bivariate approach to spatial estimation of air pollution by ozone*. Procedia Environmental Sciencies.

Sherman, M. (2011). *Spatial Statistics and Spatio-Temporal Data. Covariance Functions and Directional Properties*. Wiley Series in Probability and Statistics.

Sistema de Monitoreo Atmosférico de la Ciudad de Mexico(SIMAT) (2011). Secretaría del Medio Ambiente. http://www.calidadaire.df.gob.mx/.

Sistema de Monitoreo Atmosférico de la Ciudad de Mexico (2010). *Calidad del aire en la Ciudad de México. Informe 2010*. Secretaría del Medio Ambiente, Distrito Federal. http://www.sma.df.gob.mx/sma/links/download/biblioteca/flippingbooks/informe_anual_calidad_aire_2010.

Schabenberger, Oliver y Gotway, Carol A. (2005). *Statistical Methods for Spatial Data Analysis.* Chapman & Hall/CRC Texts in Statistical Science. Boca Raton.

Singh, V., Carnevale, C., Finzi, G., Pisoni, E. and Volta, M. (2011). *A cokriging based approach to reconstruct air pollution maps, processing measuremente station concentratios and deterministic model simulations*. Environmetnal Modelling and Software, 26, pp. 778-786.