MODELACIÓN ESPACIAL DE CONCENTRACIONES DE CONTAMINANTES ATMOSFÉRICOS

Oscar Borrego Hernández*Mario M. Ojeda Ramírez*[1], Jose A. García Reynoso**, Claudio R. Castro López
*Universidad Veracruzana, México
**Universidad Nacional Autónoma de México, México

ABSTRACT

The study of air quality, not only in urban zones but also in rural ones, has grown in relevance among international institutions, governments, scientific communities and people in general. Any air quality model should consider the spatial and/or temporal nature of this phenomenon. Frequently, the pollutants concentrations are measured in monitoring stations, but it is necessary to estimate such concentrations in non-observed locations. This contribution deals with different approaches to solve that estimation problem, and the spatial models that support them. It is also presented and analyzed a case study about ozone (O_3) concentrations, considering its interaction with nitrogen dioxide (NO_2), temperature and relative humidity. The relationship among pollutants and meteorological variables is analyzed using TAID, a method for automatic interaction detection. Finally, some of the presented methods are compared.

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

RESUMEN

El estudio de la calidad del aire, tanto en zonas urbanas e industrializadas como en zonas rurales, ha cobrado una relevancia creciente entre instituciones internacionales, gobiernos, comunidades científicas y población en general. Cualquier modelo de calidad del aire debe tener en cuenta la naturaleza espacial y/o temporal de este fenómeno. Generalmente las concentraciones de contaminantes son medidas en estaciones de monitoreo, pero es necesario estimar estas concentraciones en localizaciones no observadas. En la presente contribución se abordan diferentes enfoques para resolver ese problema de estimación, así como los modelos espaciales en los que se basan. Se presenta y analiza un caso de estudio de concentraciones ozono (O_3) considerando su interacción con el dióxido de nitrógeno (NO_2), la temperatura y la humedad relativa. La relación entre los contaminantes y las variables meteorológicas se analiza previamente utilizando el método de detección automática de interacciones TAID. Finalmente se evalúan y comparan algunos de los métodos presentados.

PALABRAS CLAVE: interpolación espacial, kriging, cokriging, contaminación atmosférica, ozono, detección automática de interacciones.

1. INTRODUCCIÓN

El aire es un componente esencial para la vida en la Tierra. Consiste en una mezcla de gases que permanecen alrededor del planeta gracias a la fuerza de gravedad dando lugar a la atmósfera terrestre. Los pulmones humanos filtran unos 15 kg de aire diariamente, mientras que solo absorbemos unos 2.5 kg de agua y menos de 1.5 kg de alimentos como promedio. El problema de la contaminación atmosférica constituye sin dudas un asunto de vida o muerte para la raza humana, los animales, las plantas y muchísimas especies de organismos vivos en general.

El estudio de la calidad del aire, tanto en zonas urbanas e industrializadas como en zonas rurales, ha cobrado una relevancia creciente entre instituciones internacionales, gobiernos, comunidades científicas y población en general.

Este problema tiene una evidente naturaleza espacio-temporal y multivariada. En la compleja dinámica de la física y la química de la atmósfera, en la que confluyen disímiles factores y fenómenos, la cercanía en el espacio y/o en el tiempo implica cierta inter-relación además de los patrones (en ocasiones cíclicos) que pueden manifestarse en estos dominios.

A lo largo de los años, los métodos estadísticos para el análisis de datos espaciales y/o temporales han devenido en una disciplina que continúa creciendo y desarrollándose. Dado que esta disciplina tiene varios orígenes, se caracteriza por una gran diversidad metodológica. Algunos métodos parten del desarrollo de la estadística orientada a ciencias como la geología, la geografía o la meteorología; otros tienen su origen en áreas de la estadística clásica como los modelos lineales; y otros se derivan de enfoques para series de tiempo o la teoría de procesos estocásticos.

En el presente capítulo se aborda la modelación espacial de las concentraciones de contaminantes atmosféricos (ozono y dióxido de nitrógeno), considerando la interacción entre sí y su relación con variables meteorológicas (temperatura y humedad).

En la sección 2 se describen brevemente los modelos espaciales en que se basa este enfoque, haciendo énfasis en las técnicas de geoestadística clásica conocidas como (co)kriging. En la sección 3 se presenta un caso de estudio de concentraciones de ozono en la Ciudad de México, se analizan los resultados del ajuste de los modelos y se exponen las conclusiones.

[1] ojeda@uv.mx

2. MODELACIÓN DE PROCESOS ESPACIALES

De manera similar a una serie de tiempo, un proceso estocástico en el espacio se define como una familia de variables aleatorias que pueden ser indexadas en el espacio:

$$\{Z(s): s \in D \subseteq \mathbf{R}^n\}$$

En esta investigación se considera el caso del plano, o sea, $D \subseteq \mathbf{R}^2$.

2.1. Estacionaridad[2] e isotropía

En la modelación matemática suelen considerarse simplificaciones e idealizaciones de los fenómenos para hacerlos más tratables matemática y computacionalmente. En el caso de la modelación espacial, es muy frecuente considerar como hipótesis la estacionaridad del proceso.

Cuando el proceso $Z(s)$ presenta una media invariante en el espacio: $E[Z(s)]=\mu$, y se cumple:

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

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 [3], [4] o [16]) que como consecuencia se cumplirá que

$$\tfrac{1}{2} Var[Z(s)-Z(s+h)] \equiv \gamma(h) \quad (2)$$

también es invariante en el espacio [3]. La función $\gamma(h)$ es llamada **variograma**[4].

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 $\gamma^*(|h|)$.

Para estimar empíricamente el valor del variograma correspondiente una distancia h, a partir de las observaciones $Z(s_i), i = 1,\ldots,n,$, se propone la fórmula (ver [3]):

$$\hat{\gamma}(h) = \frac{1}{2\,|N(h)|} \sum_{N(h)} [Z(s_i) - Z(s_j)]^2$$

donde $N(h)=\{(s_i,s_j):|s_i-s_j|=h\}$, o sea, se trata del conjunto de observaciones separadas una distancia h.

Pero esta estimación no siempre es consistente con las condiciones de estacionaridad (ver [3], [4], o [13]). 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 para la función de variograma, y se buscan estimados para estos parámetros basados en los datos.

2.2. Modelos de variograma

En sentido general en los modelos de variograma se tienen en cuenta tres elementos:

Nugget: $\eta = \lim_{h\to 0} \gamma(h)$, la altura del salto del variograma por la probable discontinuidad en el origen.

Sill: $\theta = \lim_{h\to \infty} \gamma(h)$, el valor asintótico del variograma cuando la distancia tiende al infinito.

Range: ρ, distancia a partir de la cual la diferencia entre el variograma y θ se considera insignificante [5].

En el presente trabajo se consideran los siguientes modelos de variograma comúnmente utilizados:

- Modelo esférico

[2] El término *estacionaridad* se refiere a la condición de *estacionario* de un proceso, es una traducción del término en inglés *stationarity*.

[3] Cuando la media es invariante en el espacio y se cumple (2), se dice que el proceso es **intrínsecamente estacionario**

[4] Algunos autores también la llaman **semivariograma**, pero en este trabajo se usará la denominación de **variograma**

[5] En algunos modelos se considera ρ como una parametrización del verdadero *range*, ver [1].

$$\gamma(h) = \begin{cases} \eta + (\theta - \eta)\left(\dfrac{3h}{2\rho} - \dfrac{h^3}{2\rho^3}\right) & para \quad 0 < h \le \rho \\ \theta & para \quad h > \rho \end{cases}$$

• Modelo exponencial

$$\gamma(h) = \eta + (\theta - \eta)\left(1 - e^{-\frac{h}{\rho}}\right)$$

• Modelo gaussiano

$$\gamma(h) = \eta + (\theta - \eta)\left(1 - e^{-\frac{h^2}{\rho^2}}\right)$$

• Modelo lineal

$$\gamma(h) = \eta + \frac{\theta - \eta}{\rho} h$$

Para todos los modelos se define $\gamma(0) \equiv 0$.

2.3. Kriging

El problema de estimación en un campo aleatorio $Z(s)$ se plantea de la siguiente forma:

Dadas las observaciones $z_i = Z(s_i), i = 1, \ldots, n,$ se quiere estimar el valor de $Z(s_0)$ no observado.

Lo ideal sería poder calcular E[$Z(s_0)|Z(s_1)\ldots Z(s_n)$]. Pero para lograrlo es necesario conocer la función de distribución conjunta de $Z(s_0)$, $Z(s_1)\ldots Z(s_n)$, lo cual no es factible por lo general.

Una alternativa es la predicción espacial clásica, comúnmente llamada **kriging** en honor a Daniel G. Krige, ingeniero minero que propuso el método de manera empírica en 1951. Se trata de un grupo de técnicas geoestadísticas que persiguen como objetivo interpolar valores desconocidos en un campo aleatorio a partir de observaciones vecinas, mediante una estimación lineal mínimo-cuadrática.

El objetivo del kriging es encontrar $\hat{Z}_0$, el mejor predictor lineal insesgado de $Z(s_0)$, dado por una combinación lineal de la forma:

$$\hat{Z}_0 = \lambda^T \mathbf{Z(s)}$$

Donde $\lambda \in R^n$ es el vector de coeficientes y $\mathbf{Z(s)}$ es el vector de variables $Z(s_i)$.

Sean

- $\boldsymbol{\Sigma} := \mathrm{Var}[\mathbf{Z(s)}]$, la matriz de covarianzas de las $Z(s_i)$ ($\boldsymbol{\Sigma}_{ij} = \mathrm{Cov}[Z(s_i), Z(s_j)]$),
- $\boldsymbol{\omega} := \mathrm{Cov}[\mathbf{Z(s)}, Z(s_0)]$, el vector de covarianzas entre las $Z(s_i)$ y $Z(s_0)$ ($\boldsymbol{\omega}_i = \mathrm{Cov}[Z(s_i), Z(s_0)]$),
- $\sigma_0^2 := \mathrm{Var}[Z(s_0)]$,

Se formula el problema de optimización con restricciones:

$$\min_\lambda \quad \sigma_k^2 := \mathrm{Var}[\lambda^T \mathbf{Z(s)} - Z(s_0)]$$
$$= \lambda \boldsymbol{\Sigma} \lambda - 2\lambda \boldsymbol{\omega} + \sigma_0^2$$

s.a.

$$E[\lambda^T \mathbf{Z(s)} - Z(s_0)]=0$$

Donde la restricción se refiere a la condición de no sesgo del predictor y la optimalidad está dada a partir de la minimización de la varianza del error en la predicción: σ_k^2, la cual se conoce como *varianza de kriging*.

El proceso $Z(s)$ suele descomponerse según el modelo:

$$Z(s)= \mu(s)+e(s)$$

donde $\mu(s)=E[Z(s)]$es la media del proceso y $e(s)$ es un proceso de errores con $E[e(s)]=0$. Particularizando la estructura de $\mu(s)$ y exigiendo propiedades como la estacionaridad para $e(s)$ se obtienen diferentes modelos de kriging.

2.4. Kriging ordinario

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

$$\begin{aligned} \mathbf{Z(s)} &= \mathbf{1_n}\mu + \mathbf{e(s)}, \\ Z(s_0) &= \mu + e(s_0) \end{aligned}$$

Donde $\mathbf{e(s)}$ es el vector de variables $e(s_i)$, y $1_n \in R^n$ es el vector de coordenadas unitarias.

La solución se obtiene a partir del sistema de ecuaciones (ver [3], [4], [13] o [16]):

$$\begin{aligned} \mathbf{\Gamma}\lambda + \mathbf{1_n} m &= \boldsymbol{\gamma}_0 \\ \mathbf{1_n}^T \lambda &= 1 \end{aligned}$$

Donde $\mathbf{\Gamma} = [\gamma(|s_i - s_j|)]_{i=1\ldots n,\, j=1\ldots n} \in R^{n\times n}$, $\boldsymbol{\gamma}_0 = [\gamma(|s_i - s_0|)]_{i=1\ldots n} \in R^n$ y m es un multiplicador de Lagrange. En lo sucesivo $\gamma(h)$ se entiende como la función de variograma del proceso $e(s)$, al cual se le atribuye la propiedad de estacionaridad.

2.5. Kriging universal

Cuando la media no es constante, también puede expresarse como combinación lineal: $\mu(s) = \sum_{i=1}^{p} \beta_i f_i(s)$. 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:

$$\begin{aligned} \mathbf{Z(s)} &= \mathbf{X(s)}\boldsymbol{\beta} + \mathbf{e(s)}, \\ Z(s_0) &= x(s_0)^T \boldsymbol{\beta} + e(s_0) \end{aligned}$$

Donde $\boldsymbol{X(s)}=[f_j(s_i)]_{i=1\ldots n,j=1\ldots p} \in \boldsymbol{R}^{n\times p}$ y $x(s)=[f_i(s_0)]_{i=1\ldots p} \in \boldsymbol{R}^p$ son respectivamente una matriz y un vector de valores explicatorios, y $\boldsymbol{\beta}=[\beta_i]_{i=1\ldots p} \in \boldsymbol{R}^p$ es un vector de coeficientes.

Aplicando el método de los multiplicadores de Lagrange se llega al sistema de ecuaciones:

$$\begin{aligned} \mathbf{\Gamma}\lambda + \mathbf{X(s)} m &= \boldsymbol{\gamma}_0 \\ \mathbf{X(s)}^T \lambda &= x(s_0) \end{aligned}$$

Donde $m \in \boldsymbol{R}^p$ es el vector de multiplicadores.

Frecuentemente se modela $\mu(s)$ con estructura polinomial sobre x e y.

2.6. 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 general el problema del cokriging se plantea como sigue:

En $s_1 \dots s_n$ se han observado los procesos $Z_1(s) \dots Z_q(s)$, y se quiere predecir $Z_k(s_0)$. La forma del predictor es:

$$\hat{Z}_k^0 = \sum_{i=1}^{n}\sum_{j=1}^{q}\lambda_{ij} Z_j(s_i) = \boldsymbol{\lambda}^{\mathbf{T}}\mathbf{Z(s)}$$

donde $\boldsymbol{\lambda} = [\boldsymbol{\lambda}_j]_{j=1\dots q}$, con $\boldsymbol{\lambda}_j = [\lambda_{ij}]_{i=1\dots n}$, y $\mathbf{Z(s)} = [\mathbf{Z_j(s)}]_{j=1\dots q}$, con $\mathbf{Z_j(s)} = [Z_j(s_i)]_{i=1\dots n}$.

De manera similar a los modelos de kriging se transforma el problema de optimización en un sistema de ecuaciones lineales (ver detalles en [3], [4], [13] o [16]).

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). La estructura del predictor es la siguiente:

$$\hat{Z}_k^0 = \sum_{i=1}^{n}\lambda_i^k Z_k(s_i) + \sum_{j=1, j\neq k}^{q}\lambda_j^0 Z_j(s_0)$$

2.7. Análisis objetivo de Cressman

El análisis objetivo es el proceso de transformar datos de observaciones para puntos irregularmente espaciados en puntos de una malla regular. La propuesta de Cressman (ver [5] y [7]) se basa en correcciones sucesivas a partir de una aproximación inicial. Para determinar las correcciones se consideran los errores entre los datos y la interpolación en los puntos observados.

Dadas las observaciones $z_i = Z(s_i), i = 1, \dots, n$, en las localizaciones s_i ,una interpolación inicial para la malla y un radio de influencia R para las observaciones, la interpolación para una localización s_0 perteneciente a la malla a predecir se realiza de la siguiente forma:

• Sea $d_i \equiv \|s_i - s_0\|_2 < R$ para $i = 1 \dots n$ la distancia euclidiana entre s_i y s_0, se define el conjunto de localizaciones s_i para las que s_0 se encuentra dentro de su radio de influencia: $I_0 = \{i : d_i < R\}$, y para cada $i \in I_0$ se calcula una corrección ponderada:

$$C_i = \frac{R^2 - d_i^2}{R^2 + d_i^2}(z_i - \hat{z}_i),$$

donde $\hat{z}_i$ es el valor estimado para la localización s_i, a partir de la interpolación de la malla. Un enfoque sencillo es calcular $\hat{z}_i$ como el promedio de los valores para los cuatro vértices de la celda a la que pertenece s_i.

• A la interpolación inicial para s_0 se suma la corrección total:

$$C = \frac{1}{|I_0|}\sum_{i\in I_0} C_i.$$

De esa forma se corrigen todos los puntos de la malla. El proceso se repite sucesivamente disminuyendo el valor de R.

A diferencia de los métodos de kriging, el análisis objetivo de Cressman se ha diseñado para realizar la interpolación en una

malla, donde la estimación en un punto depende de las estimaciones de otros, en particular, de los vecinos de las localizaciones observadas. La precisión en la interpolación también depende de los parámetros de entrada del algoritmo: los radios de influencia y la resolución de la malla.

3. CASO DE ESTUDIO: CONCENTRACIONES DE OZONO

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 (ver [14] y [15]). En esta zona, el ozono es uno de los contaminantes de mayor impacto en la salud humana y en los ecosistemas.

En la presente sección se aborda la interpolación espacial de concentraciones de ozono medida en estaciones de monitoreo, hacia una malla regular. Este problema puede atacarse desde diversos enfoques, y empleando diferentes técnicas, véanse, por ejemplo [6], [8], [12] y [17].

3.1. Metodología

Los procedimientos evaluados son el kriging, el cokriging y el análisis objetivo de Cressman. Como variables secundarias para los métodos basados en cokriging se consideran las concentraciones de dióxido de nitrógeno, la temperatura y la humedad relativa. Puntualmente, los modelos evaluados son los siguientes:

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
Cressman	-	-

En los casos de (co)kriging universal el modelo para la media es el siguiente:

$$\mu(x,y) = \beta_6 x^2 + \beta_5 y^2 + \beta_4 xy + \beta_3 x + \beta_2 y + \beta_1.$$

Al respecto, se asume la propiedad de estacionaridad de segundo orden (y por ende intrínseca) para el proceso de errores: $e(s) = Z(s) - \mu(s)$, y no para el proceso $Z(s)$, puesto que trivialmente se violaría E[*Z*(*s*)]=μ constante.

Para el análisis de Cressman se siguió un esquema de radios exponenciales: $e^{\alpha i}$, $i = (n-1),\ldots,0$, donde n es el número de radios, $\alpha = \dfrac{\log(R_M)}{(n-1)}$, y R_M es el radio de influencia máximo, o sea, se trata de una secuencia que comienza en R_M y decrece hasta 1 con una razón de $e^{-\alpha}$.

La malla considerada en la interpolación y la generación de mapas, tiene una dimensión de 54×56 puntos y una resolución de $0.01176°$.

La relación entre las variables secundarias y el ozono se analiza previamente con el método TAID (*Tau Automatic Interaction Detection*) [2], un método estadístico de partición recursiva con estructura de árbol.

Se realiza un análisis de sensibilidad con respecto al radio de influencia de las estaciones, entendido en el caso de los modelos de (co)kriging como la distancia a partir de la cual el valor de la función de covarianza es despreciable $C(h)$. Para esto se aplica una 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 datos registrados el día 12 de junio de 2010 a las 15 horas. Como resultado se elige un valor para el radio de influencia con el cual se continuará el estudio.

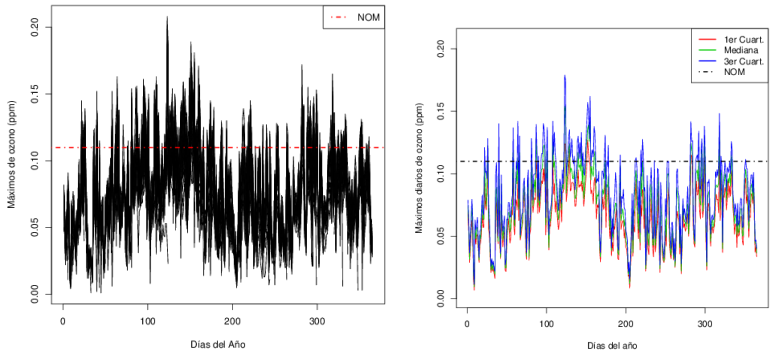
Con el radio de influencia fijado, se procede al refinamiento de los modelos de (co)kriging con respecto a la estructura del variograma, teniendo en cuenta la estructura esférica, la exponencial y la gaussiana.

Finalmente se repite la validación cruzada luego del refinamiento, para comparar los modelos.

El software utilizado en el trabajo es el entorno de computación estadística ***R*** (ver [11]), en particular, los paquetes *gstat* [9] y *spacetime* [10].

3.2. Análisis Exploratorio

En la figura 1.a se grafica el historial de máximos diarios de la concentración de O_3 en cada una de las estaciones de monitoreo, en 1.b se grafican los cuartiles para los máximos diarios de O_3.



1 a 1.b

Figura 1: Máximos diarios de concentración de ozono

Se observó que el día 3 de mayo (día 123 del año) se alcanzó el máximo absoluto de 0.208 ppm en la estación de Cuajimalpa durante la hora 17.

Por otra parte, en el diagrama de cajas de la figura 2 se observa que el mes de mayo presenta la mayor mediana, el tercer cuartil (0.119 ppm) supera el límite NOM (0.110 ppm) y se identificaron valores atípicos para el extremo superior y para el inferior, no obstante el mes de junio es el de mayor máximo (sin considerar los valores atípicos).

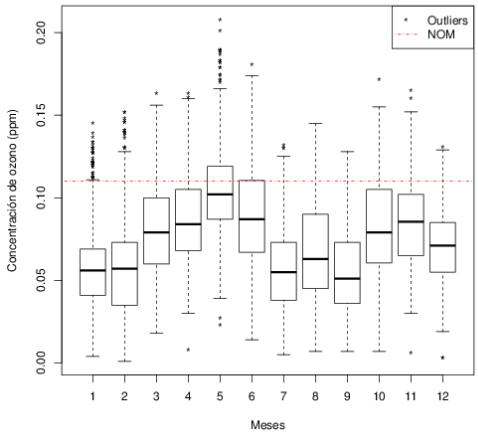


Figura 2: Diagramas de caja por meses para máximos diarios de concentración de ozono

3.3. 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 [12] o [18]). 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*) [2] 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 3 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

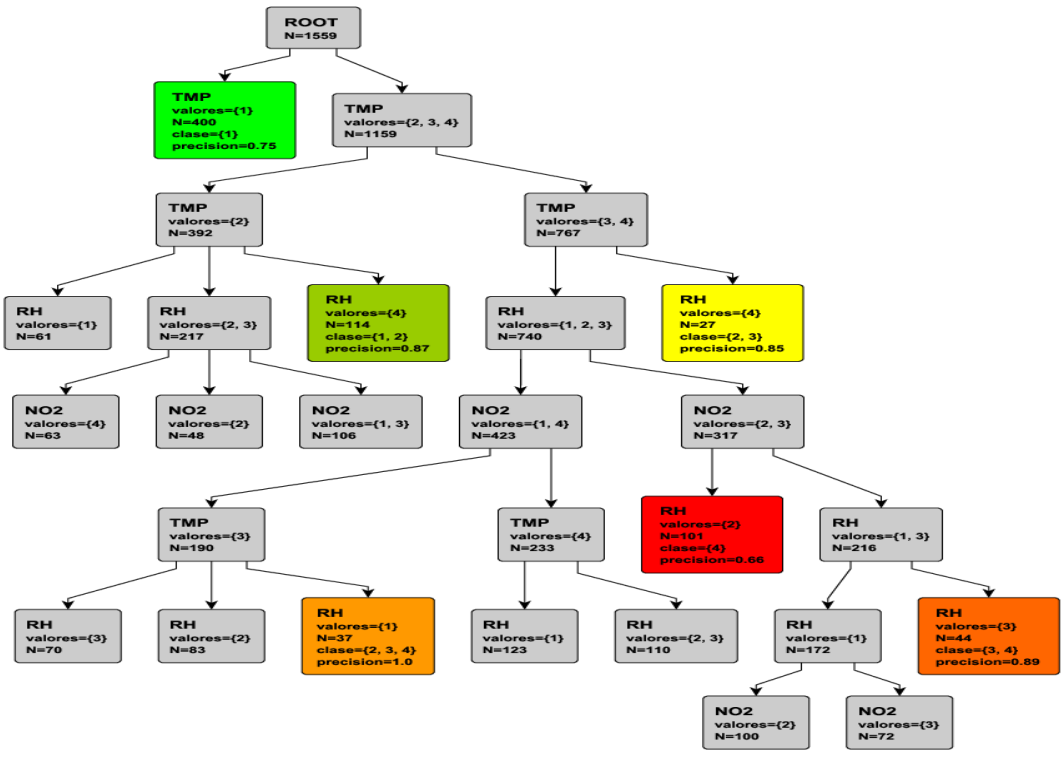


Figura 3: Árbol de partición recursiva

En este caso cada clase se corresponde con uno de los subintervalos de concentración de O_3. 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 3, 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

En sentido general puede observarse que las concentraciones de O_3 son mayores en situaciones de mayor temperatura, menor humedad y menor concentración de NO_2.

3.4. Resultados y discusión

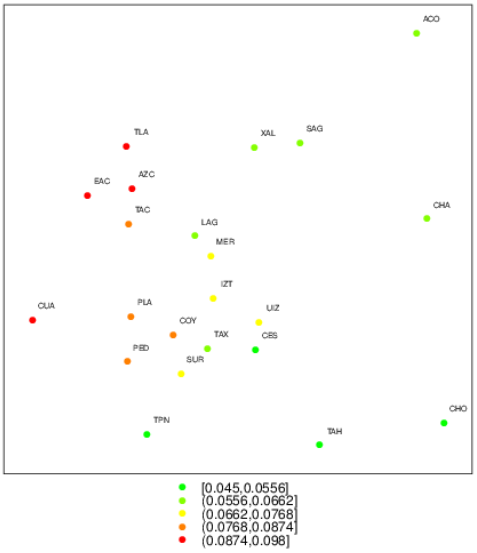


Figura 4: Mapa de O_3, día 12/06/2010 a las 15:00

Los datos considerados en la evaluación de los modelos están conformados por mediciones de la concentración de O_3, NO_2, SO_2, temperatura y humedad en 22 estaciones para el día 12 de junio de 2010 a las 15 horas[6]. En la figura 4 se muestra la ubicación geográfica de las estaciones coloreadas según la concentración de O_3, el mapa sugiere mayores concentraciones en la zona noroeste con respecto a las del sur y el este.

3.4.1. Análisis de sensibilidad para radios de influencia

Para analizar la sensibilidad de los métodos de interpolación con respecto al radio de influencia de las estaciones, se ejecutó el procedimiento de validación cruzada según la regla *dejar uno afuera*, para cada uno de los modelos evaluados, y los siguientes valores para los radios de influencia (Km): 10, 12, 15, 20, 25, 27.5, 30, 35, 40, 45, 48, 50, 52 (la distancia entre estaciones varía

[6] Se eligió esa fecha y esa hora por presentar una mayor disponibilidad de mediciones.

aproximadamente de 3 a 50 Km). En este punto de la evaluación, para la estructura de las funciones de variograma $\gamma(h)$ se empleó el modelo esférico, fijando como rango el radio de influencia.

En la figura 5 se muestran las curvas (suavizadas mediante splines) de RMSE contra radios de influencia por modelo. La mayoría de las curvas tienen un comportamiento oscilatorio y una tendencia a la disminución con el aumento del radio de influencia. Puede notarse un mejor desempeño para los modelos de media polinomial (CUK.ALL, CUK.TMP y UK), aunque esto podría ser

consecuencia de la sobre-adaptación a los datos, dada la flexibilidad de la estructura polinomial (más adelante se volverá este tema).

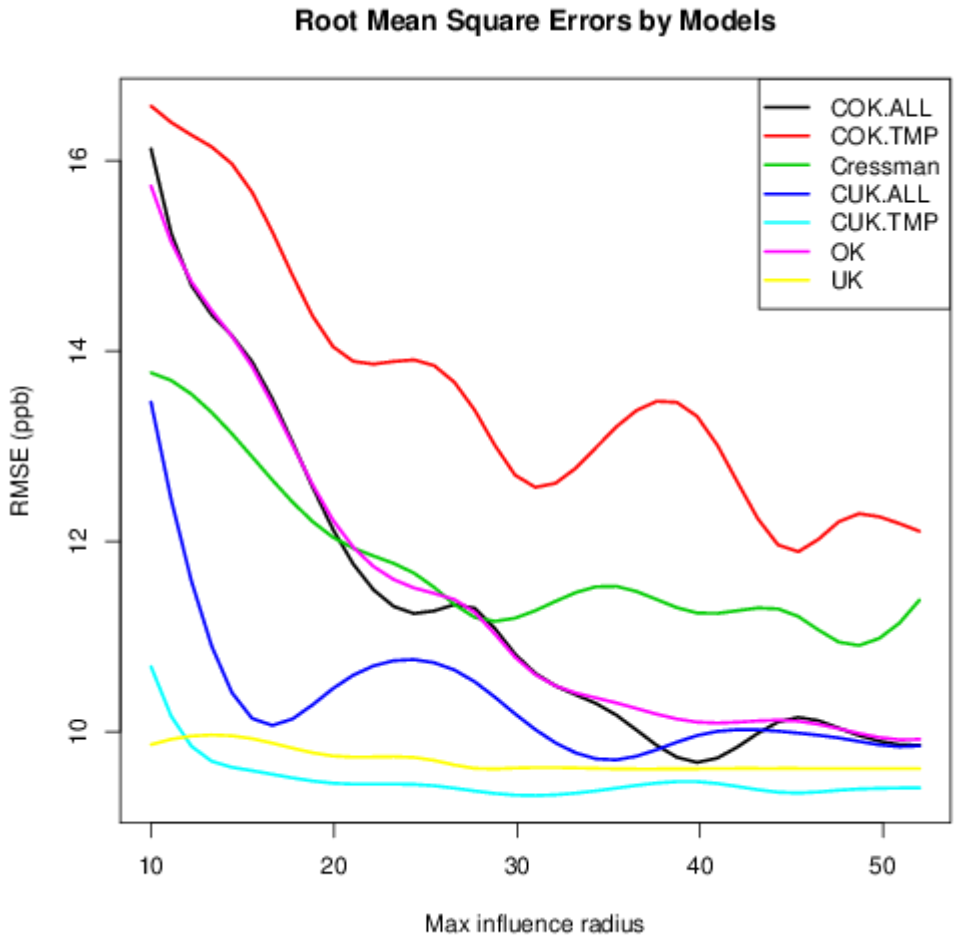


Figura 5: RMSE por modelos

También puede observarse a partir de los 35 Km cierta *convergencia* y estabilidad para la mayoría de los modelos. Se fijó en 37 Km el radio de influencia para proseguir con el análisis, puesto que a partir de esa distancia cinco de los siete modelos mantienen el RMSE por debajo o muy cerca de 10 ppb. Por otra parte, no tiene mucho sentido considerar radios de influencia muy grandes en este dominio.

3.4.2. Ajuste de modelos de variograma: Media constante

La nube de variograma, así como el variograma muestral para el modelo de media constante se grafican en la figura 6. Con estas gráficas se sustenta visualmente la hipótesis considerada de estacionaridad de segundo orden.

Bajo la hipótesis de estacionaridad e isotropía, se ajustaron tres de los modelos de variograma mencionados en el apartado 2.2: esférico, exponencial y gaussiano. En la figura 7 se muestran los resultados del ajuste para cada modelo. En la siguiente tabla se presentan los valores de los parámetros ajustados así como la raíz del error cuadrático medio (RMSE).

a) Nube b) Muestral

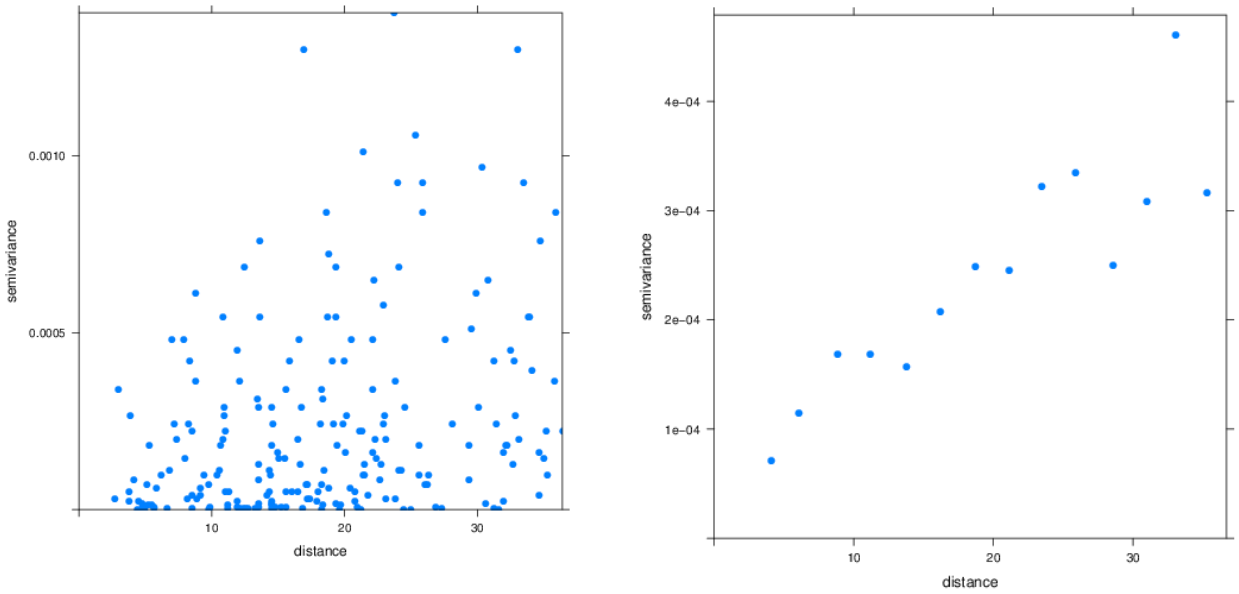


Figura 6: Variograma: modelo de media constante.

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

modelo	Nugget	sill	Range	RMSE
esférico	2.535313e-05	0.0003451827	37.00000	7.58687e-06
exponencial	0.000000e+00	0.0002974244	12.33333	8.93329e-06
gaussiano	7.549592e-05	0.0003943375	21.36196	8.81377e-06

Esférico Exponencial Gaussaino

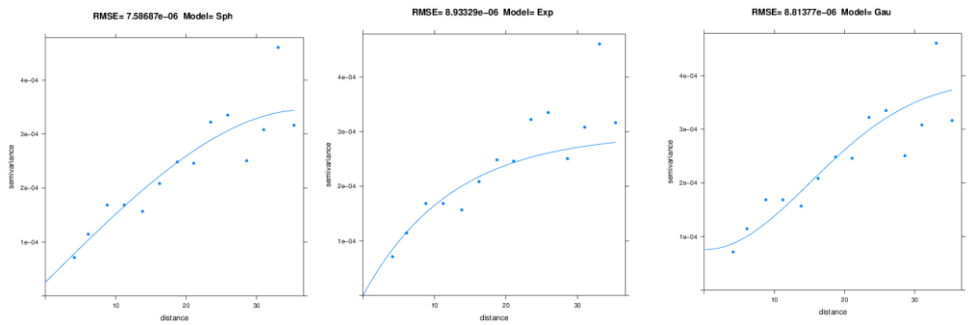


Figura 7: Ajustes de modelos de variograma. Media constante

3.4.3. Ajuste de modelos de variograma: Media polinomial

De manera similar al modelo de media constante, en la figura 8 se muestra la nube de variograma y las estimaciones muestrales para el modelo de media polinomial. Tanto la nube como el variograma muestral hacen dudar de la certeza de la hipótesis de estacionaridad. Esa sospecha se incrementa con el ajuste de cada uno de los modelos de variograma, al resultar en efectos de nugget puro, todos coincidentes en el valor de ese parámetro. Los parámetros ajustados y los errores fueron los siguientes:

modelo	nugget	sill	range	RMSE
esférico	6.28036e-05	6.28036e-05	37.00000	5.50012e-06
exponencial	6.28036e-05	6.28036e-05	12.33333	5.50012e-06
gaussiano	6.28036e-05	6.28036e-05	21.36196	5.50012e-06

Bajo la hipótesis de estacionaridad de segundo orden, esto quiere decir que la función de covarianza $C(h)$ es constante y nula, por lo tanto, los errores $e(s)$ en las diferentes localizaciones son independientes, o sea, $e(s)$ es un proceso de ruido blanco con media nula y varianza 6.28036e-05. El modelo resultante para la concentración de O_3, coincide con un modelo lineal clásico. Este escenario es muy poco realista, y puede ser consecuencia de la sobre-adaptación de la media polinomial, debido a su flexibilidad. A pesar de lo anterior se procedió con la evaluación del modelo, en aras de la comparación.

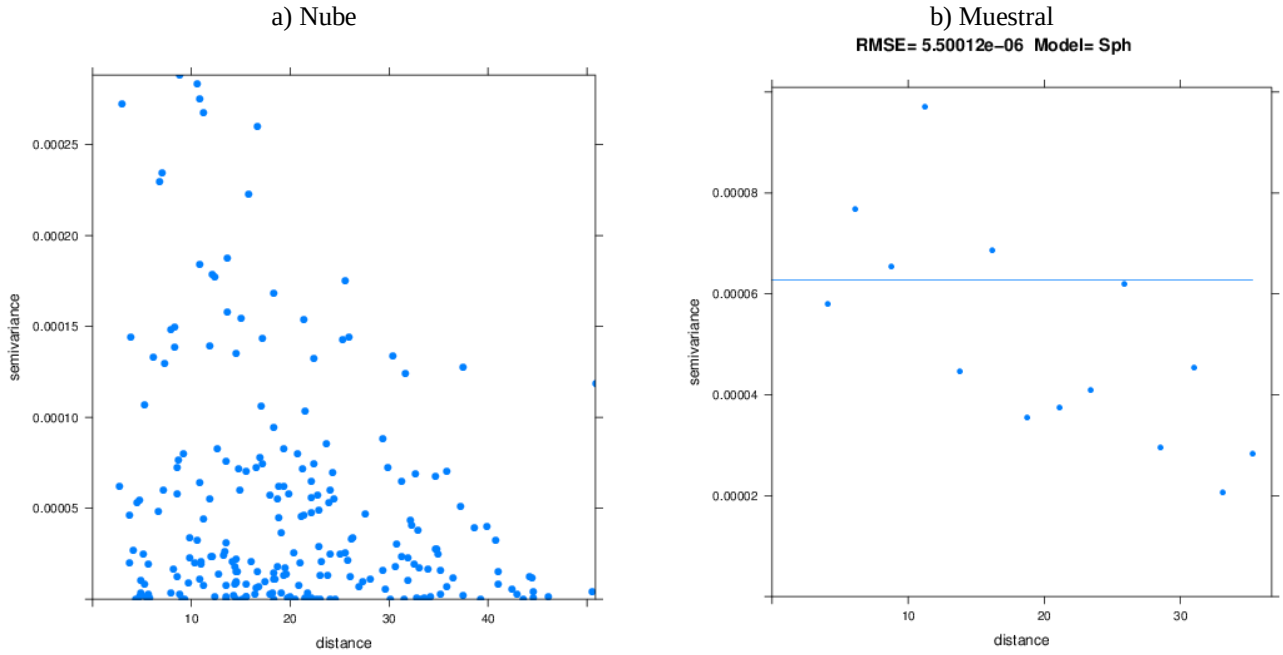


Figura 8: Variograma: modelo de media polinomial.

En el caso b) la línea horizontal representa el efecto de nugget puro.

3.4.4. Comparación de los modelos de kriging

(Co)Kriging ordinario

Con el objetivo de contrastar los resultados, en la figura 9 se grafican los mapas de varianza de kriging para los modelos de media constante. Resalta el caso de COK.TMP con valores elevados para la varianza, más del 82% de los puntos en la malla considerada superan las 25 ppb de desviación estándar, razón por la cual este modelo se desecha y se continúa el análisis con OK y COK.ALL.

A diferencia de resultados como el de [12], la consideración de la temperatura como única variable secundaria en la predicción de concentraciones de O_3 no implicó una mejora con respecto a la modelación univariada de OK. Se conoce que existe relación *parcialmente* directa entre la temperatura y la concentración de O_3. Pero las mediciones de temperatura consideradas varían en un rango de 22 a 28^{o}C, en este caso el aumento de la tempera podría provocar que el aire en la superficie se caliente y suba, disminuyendo la concentración de los contaminantes (al aumentar la altura de mezcla).

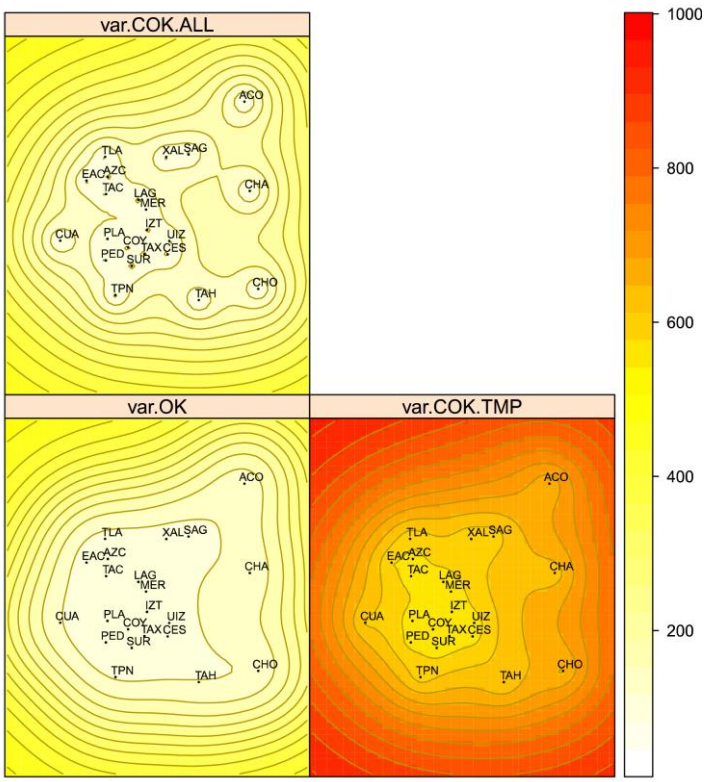


Figura 9: Varianza de Kriging. Modelos de media constante.

(Co)Kriging universal

En la figura 10 se muestran los mapas de varianza de kriging para los modelos de media polinomial. Donde nuevamente resalta el modelo de cokriging con la temperatura como única variable secundaria: CUK.TMP, para el cual el mínimo valor de la desviación estándar sobre toda la malla fue de 33.6 ppb. Por esa razón este modelo también se desecha.

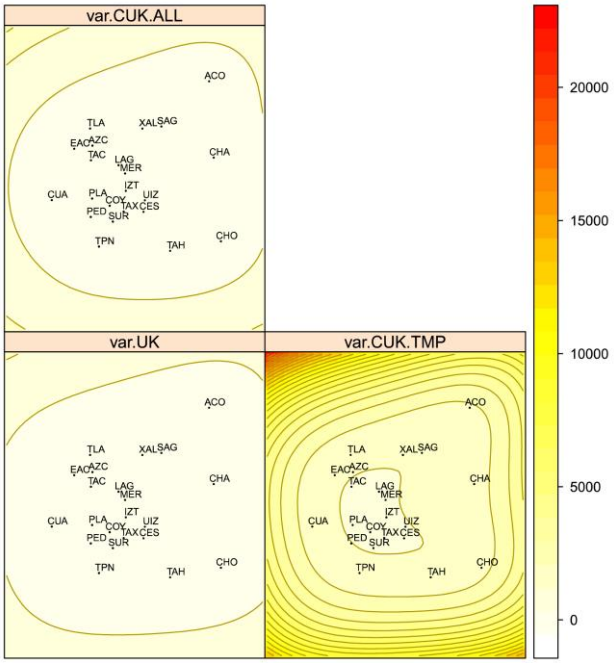


Figura 10: Varianza de Kriging. Modelos de media polinomial.

(Co)Kriging ordinario vs universal

En la figura 11 se contrastan los mapas de varianza de kriging entre los modelos de (co)kriging universal y ordinario. Aquí también pueden notarse regiones de mayor varianza (en especial hacia el noroeste) para los modelos UK y CUK.ALL. Para este último aproximadamente el 12% de los puntos de la malla supera las 25 ppb en desviación estándar, por esa razón no se incluirá en la comparación final.

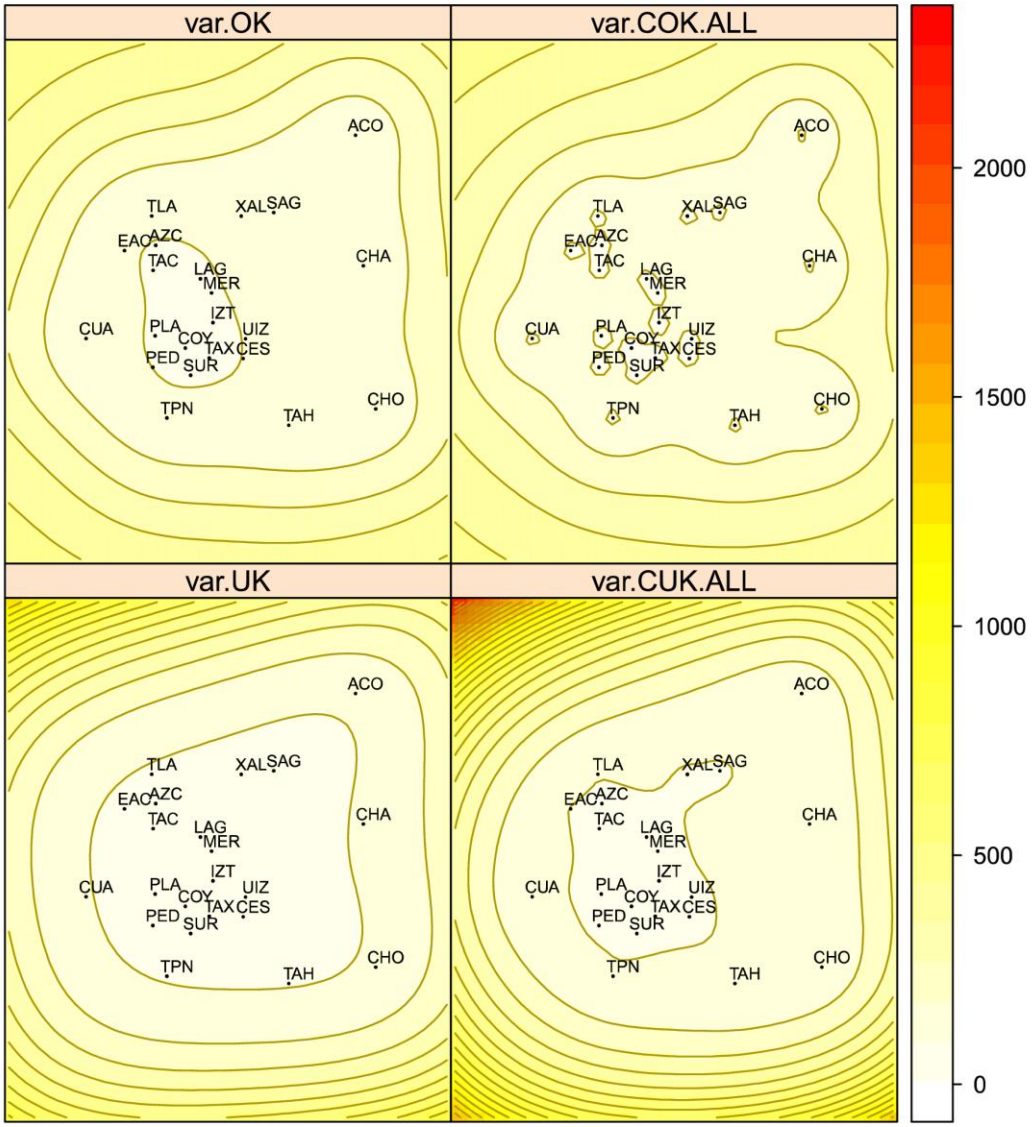


Figura 11: Varianza de Kriging

(Co)Kriging vs Análisis Objetivo de Cressman

En la figura 12 se muestra el mapa resultante de la interpolación hacia la malla, mediante los modelos: OK, COK.ALL, UK y Cressman. Se observa la similitud entre los mapas de los modelos OK y COK.ALL. Para UK se dibujan en blanco las celdas con una desviación estándar superior a las 25 ppb. Como se puede apreciar, el mapa de UK no parece ilustrar la realidad del fenómeno, con curvas de nivel demasiado suaves y demasiado cercanas, en comparación con los otros modelos.

Los errores absolutos y relativos en la validación cruzada para cada modelo se resumen en el cuadro 1. Puede notarse lo siguiente:

• Los cuatro modelos arrojan errores inferiores a 11.5 ppb para RMSE.

• Según RMSE absoluto, el mejor modelo es UK, aunque le siguen COK.ALL con una diferencia de solo 0.29 ppb, OK con 0.57 ppb y Cressman con 1.87 ppb.

• COK.ALL es el mejor según el promedio de errores.
• OK es el único modelo para el que al menos la mitad de los errores absolutos son menores o iguales que 5.73 ppb.

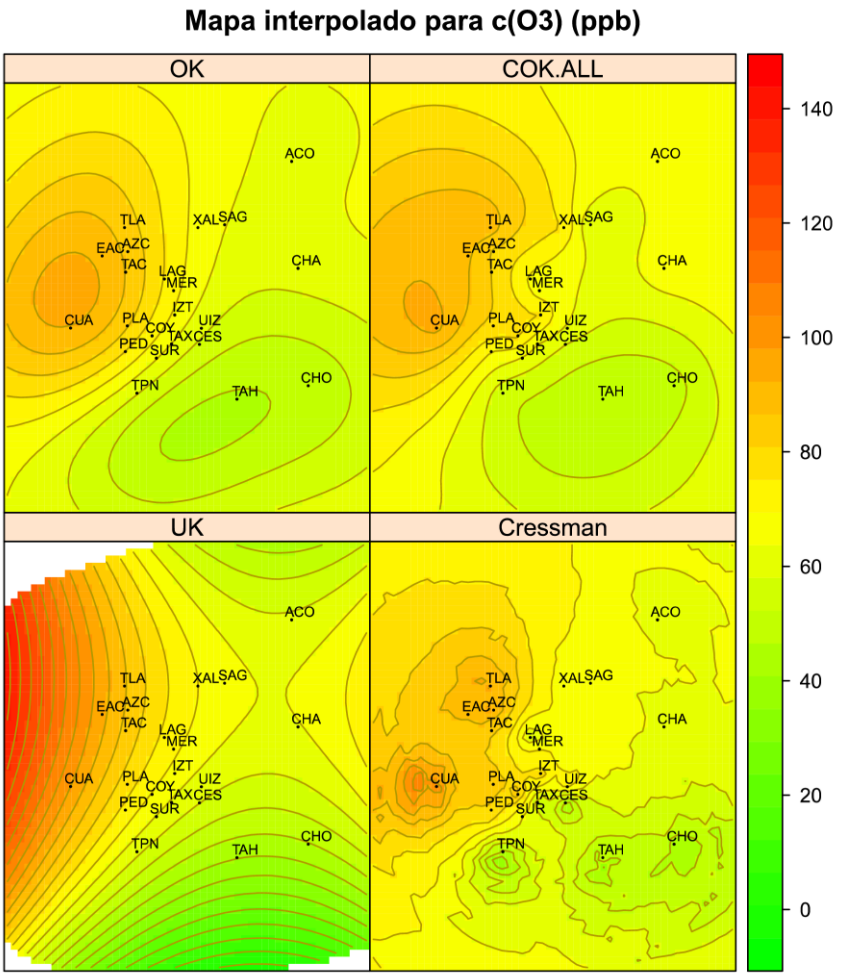


Figura 12: Mapas de interpolación

Aunque el modelo UK tenga el mejor valor para el RMSE absoluto, se ha visto que padece de sobre-adecuación y no incluye para este caso relaciones de covarianza en el espacio. Por ello no parece ser una buena opción ante COK.ALL u OK.

Errores absolutos (ppb)

Modelo	RMSE	Promedio	Mediana	Máximo
COK.ALL	9.890781	7.553256	6.202689	20.87188
Cressman	11.479344	8.853170	6.223485	23.59617
OK	10.174161	7.619835	5.722581	21.29566
UK	9.609665	7.935267	6.390427	20.58493

Errores relativos (%)

Modelo	RMSE	Promedio	Mediana	Máximo
COK.ALL	16.86774	12.03098	8.793137	40.57623
Cressman	20.89773	14.83722	9.396384	50.20463
OK	17.79772	12.30854	7.341439	45.30992
UK	16.91139	12.91406	9.299323	43.79771

Cuadro 1: Errores absolutos y relativos en la validación cruzada de cada modelo

Por otra parte, COK.ALL supera a OK en la mayoría de los indicadores, tanto absolutos como relativos. La modelación

considerando de manera conjunta la temperatura, la humedad y la concentración de NO_2 tuvo mucho mejor desempeño que el caso de la temperatura como única variable secundaria. Pero puede decirse que no se trata de una diferencia considerable con respecto a OK, por ejemplo, es de solo 0.28 ppb en el RMSE absoluto. Esto también puede ser consecuencia del aumento de la altura de mezcla mencionado anteriormente.

3.5. Conclusiones

Con el análisis de TAID sobre mediciones de 2010, 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. No obstante, en el caso particular del día 12 de junio no se ha logrado expresar esta relación de forma aceptable en la modelación multivariada espacial, hecho que debe ser consecuencia de una dinámica más compleja que no se ha tenido en cuenta.

Al igual que en otras investigaciones (ver [12]) los modelos de (co)kriging universal no han mostrado un mejor desempeño que los de (co)kriging ordinario. El ajuste de modelos de variograma con media polinomial arrojó un caso trivial, y lejano a la realidad, de independencia espacial. Esto se atribuye a la sobre-adaptación de la flexible estructura polinomial a las mediciones en las estaciones, con perjuicio del resto del área a predecir. En este sentido mostraron mejor capacidad de generalización los modelos de media constante.

El análisis objetivo de Cressman fue superado por otros modelos, pero tampoco fue amplia la diferencia, mostrando en general buen desempeño, que podría ser mejorado quizás variando los parámetros del algoritmo como la resolución de la malla o el esquema de radios de influencia. Pero también se podría pensar en una modificación o adición al algoritmo, como por ejemplo, una forma de suavizar las fronteras abruptas del *alcance* de las estaciones.

REFERENCIAS

[1] BIVAND, R. S., PEBESMA, E. J. & GÓMEZ-RUBIO, V. (2008): **Applied spatial data analysis with R Use R!** Series, Springer. New York.http://www.asdar-book.org/

[2] CASTRO-LÓPEZ, C. (2005): Contribuciones a la detección y análisis de variables relevantes en tablas de contingencia multivariantes. **Tesis Doctoral Universidad de Salamanca.**

[3] CRESSIE, N. (1993): **Statistics for spatial data** (Wiley series in Probability and Statistics). Wiley-Interscience, New York.

[4] CRESSIE, N. & WIKLE, C. K. (2011): **Statistics for spatio-temporal data** (Wiley Series in Probability and Statistics). Wiley. ISBN-10 0471692743. Hoboken, New Jersey.

[5] CRESSMAN G. P. (1959): An operational objective analysis system. **Monthly Weather Review**, 87, 367-374.

[6] DUTILLEUL, P. & BINEL-ALLOUL B. (1996): A double multivariate model for statistical analysis of spatio-temporal environmental data. **Wiley Environmetrics Journal**, 7: 551-595.

[7] GLOVER, D., JENKINS, W. & DONEY, S. (2008): Objective Mapping and Kriging. Chapter 7 in **Modeling Methods for Marine Science**, 165-191. Cambridge University Press. Massachusetts.

[8] 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.

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

[10] PEBESMA, E. (2011): Classes and methods for spatio-temporal data in R: the **spacetime package** (R package documentation)**. Institute for Geoinformatics, University of Münster**.

[11] R DEVELOPMENT CORE TEAM (2010): **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/.

[12] ROJAS-AVELLANEDA, D. & MARTÍNEZ-CERVANTES, J. (2011): Using the bivariate approach to spatial estimation of air pollution by ozone. **Procedia Environmental Sciencies**, 3: 20-25.

[13] SHERMAN, M. (2011): **Spatial Statistics and Spatio-Temporal Data. Covariance Functions and Directional Properties** (Wiley Series in Probability and Statistics). Wiley. Pondicherry.

[14] SISTEMA DE MONITOREO ATMOSFÉRICO DE LA CIUDAD DE MEXICO (SIMAT) (2011): Secretaría del Medio Ambiente. **http://www.calidadaire.df.gob.mx/**.

[15] SISTEMA DE MONITOREO ATMOSFÉRICO DE LA CIUDAD DE MEXICO (SIMAT) (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/**.

[16] SCHABENBERGER, O. & GOTWAY, C. A. (2005): **Statistical Methods for Spatial Data Analysis**. Chapman & Hall/CRC Texts in Statistical Science. Boca Raton.

[17] SINGH, V., CARNEVALE, C., FINZI, G., PISONI, E. & VOLTA, M. (2011): A cokriging based approach to reconstruct air pollution maps, processing measurement station concentratios and deterministic model simulations. **Environmetnal Modelling and Software**, 26, 778-786.

[18] UHEREK, E. (2008): El smog de ozono. Baja Atmósfera. Bases. Environmental Science Published for Everybody Round the Earth Educational Network on Climate (ESPERE-ENC). **http://www.atmosbhere.mpg.de/enid/3464a3129c75fa0569308e9f2e0a3417,0/3__Ozono_y_oxidos_de_nitrogeno/-_smog_de_ozono_2xb.html**.

ANEXOS

Anexo 1 Árbol TAID

Figura 9: Árbol de partición recursiva