MODELADO ATMOSFÉRICO PARA DETERMINAR NIVELES MÁXIMOS DIARIOS DE OZONO EN LA CIUDAD DE GUADALAJARA

Tesis que presenta:
LORELIE HERNÁNDEZ GALLARDO

Para obtener el grado de:
MAESTRA EN CIENCIAS
(Matemáticas Aplicadas e Industriales)

Asesor de Tesis

Dr. Gabriel Escarela Pérez

Noviembre de 2009

Casa abierta al tiempo

UNIVERSIDAD AUTÓNOMA METROPOLITANA

UNIDAD - IZTAPALAPA

DIVISIÓN DE CIENCIAS BÁSICAS E INGENIERÍA

DEPARTAMENTO DE MATEMÁTICAS

MODELADO ATMOSFÉRICO PARA DETERMINAR NIVELES MÁXIMOS DIARIOS DE OZONO EN LA CIUDAD DE GUADALAJARA

Tesis que presenta
LORELIE HERNÁNDEZ GALLARDO

Asesor
DR. GABRIEL ESCARELA PÉREZ

Sinodales
DR. JUAN RUÍZ DE CHAVEZ SOMOZA
DR. CARLOS CUEVAS COVARRUBIAS

A mi Pa y a Chayo
♡

AGRADECIMIENTOS

Quiero agradecer a todas aquellas personitas que estuvieron conmigo desde el principio de esta etapa. Gabriel gracias por permitirme trabajar contigo. De manera especial agradezco a los sinodales por sus comentarios para que esta tesis tuviera un mejor término. Gracias también al Instituto Nacional de Ecología por proporcionar la base de datos para llevar a cabo este estudio. A la UAM por los amigos que me dió, a mis amigos por su complicidad y apoyo. Carlos A. gracias por todo.

ÍNDICE GENERAL

INTRODUCCIÓN

Panorama General

La exposición de seres humanos y otros seres vivos a niveles altos de contaminación es un problema de salud persistente en áreas urbanas grandes. Por ejemplo, se sabe que si un individuo respira en ambientes con concentraciones mayores o iguales a $80\,\mu g/m^3$ (microgramos de ozono por cada metro cúbico) de aire de gas ozono hay una reducción en su funcionamiento pulmonar. Por consiguiente, exposiciones repetidas a estos altos niveles de contaminación pueden causar daño a los pulmones principalmente. En términos de la Organización Mundial de la Salud (OMS), los niveles de ozono superiores a los $100\,\mu g/m^3$ pueden comenzar a ser riesgosos para la salud (WHO, 1987); concordantemente, la NORMA Oficial Mexicana de la Salud Ambiental establece que la concentración de ozono no debe exceder los $110\,\mu g/m^3$, que equivalen a 0.11 ppm [1] de ozono de aire ambiente a condición estandar de presión a temperatura de 25°C y 1 atmósfera. Una forma efectiva de incrementar el conocimiento del problema para finalmente aumentar la capacidad de predecir la concentración del ozono troposférico y mejorar estrategias de control de emisiones, se obtiene al aplicar una metodología que pueda tomar en

[1] conversión: $1ppm$ (porciones por millón) equivale a $1962\mu g/m^3$

consideración la relación que hay entre los procesos meteorológicos y las reacciones químicas. Esta metodología debe permitir formulaciones que puedan separar los efectos de las tendencias causadas por emisiones precursoras y los efectos debidos a la variabilidad meteorológica (National Research Council, 1991).

Objetivos

El objetivo principal de este estudio es evaluar si las diversas políticas ambientales implementadas en el Área Metropolitana de Guadalajara integrada por los municipios de Guadalajara, Tlaquepaque, Tonalá y Zapopan, en el Estado de Jalisco; han generado una tendencia a la baja en los niveles de contaminación. Para ello, se analizan los niveles máximos diarios de ozono registrados por siete estaciones de monitoreo en el Área Metropolitana de Guadalajara durante el periodo $1997 - 2006$, pues con frecuencia se presentan elevados índices de contaminación en esta ciudad.

Otro objetivo es proponer una metodología para predecir niveles máximos diarios de ozono en presencia de variables atmosféricas. Como el ozono troposférico se crea cuando los óxidos de nitrógeno e hidrocarburos reaccionan químicamente en presencia de la luz solar (Apéndice B), la metodología en este estudio trata de conciliar métodos estadísticos con el conocimiento descriptivo de la química troposférica. Es bien sabido que la variación de los niveles de contaminantes corresponde a varias razones, como la industrialización, grandes concentraciones urbanas, emisión de humos, polvos y gases provenientes de fuentes contaminantes tanto moviles como fijas, por mencionar algunas. También los cambios graduales de las condiciones meteorológicas influyen en esta variación; en particular, temperaturas altas junto con bajas velocidades del viento están asociadas con observaciones altas de ozono (Huang y Smith, 1999).

Un objetivo más, es proponer una técnica estadística que pueda usar tanto información no estacionaria como atmosférica para la predicción de los máximos

diarios de ozono. En la ciencia ambiental, uno de los propósitos principales es analizar datos que corresponden a extremos de un cierto fenómeno durante determinados periodos de tiempo. En este contexto, los valores extremos son los altos índices de contaminación que en verdad causan un grave problema. El objetivo de analizar el comportamiento de estos valores extremos es garantizar, de alguna manera, que la calidad del aire sea ideal para llevar a cabo las actividades de la vida cotidiana, sin que la salud peligre y, para ello, en este trabajo se extienden ideas bien establecidas en la literatura de la Teoría de Valor Extremo. Existe una gran variedad de literatura que trata acerca de la teoría asintótica de los extremos y sus aplicaciones estadísticas. Por mencionar alguno, destaca principalmente el libro de Coles (2001), que proporciona una revisión básica de la Teoría del Valor Extremo y de su aplicación práctica.

El presente trabajo está organizado como sigue: el capitulo I es una introducción a la Teoría de Valor Extremo, que resume los elementos básicos de dicha teoría; en el capitulo II se hace una revisión de los modelos condicionales y también se describe el uso de los polinomios ortogonales. Así, los capitulos I y II se consideran teóricos, pues introducen a la Teoría de Valor Extremo. En el capitulo III, se analizan los datos de la Red Automática de Monitoreo de la Ciudad de Guadalajara, y además, se encuentran el análisis, desarrollo y resultados de la aplicación de la Teoría del Valor Extremo; posteriormente están las conclusiones y los apéndices A, B, C y D; que respectivamente se refieren al bosquejo de una integral, descripción breve de que es el ozono y la programación realizada con el paquete estadístico **R**. [2] Por último, aparecen las referencias bibliográficas citadas a lo largo de la tesis.

[2]@Manual title = R: A Language and Environment for Statistical Computing,
author = R Development Core Team,
organization = R Foundation for Statistical Computing,
address = Vienna, Austria,
year = 2008,
note = ISBN 3-900051-07-0,
url = http://www.R-project.org

CAPÍTULO 1

TEORÍA DE VALOR EXTREMO

La Teoría de Valor Extremo consiste en emplear técnicas estadísticas para identificar y modelar observaciones extremas. Se considera como valor extremo el dato más grande o más pequeño en un conjunto de observaciones que incluso pueden ser variables aleatorias. El objetivo principal de la Teoría de Valor Extremo es, como su nombre lo indica, analizar los extremos registrados de los valores máximos o mínimos de un conjunto de observaciones; en este caso, los datos de la cola derecha son los que ayudan a predecir futuros extremos que son peligrosos para la salud.

1.1. Antecedentes

Las técnicas de Valor Extremo se aplican a variables aleatorias continuas que pertenecen a un espacio muestral que puede ser llamado Ω. El comportamiento de los valores más altos o más bajos es importante en varios campos, los cuales incluyen a la Ingeniería, la Oceanografía, el Medio Ambiente, la Hidrología y la Climatología. En Finanzas, la Teoría de Valor Extremo ha sido aplicada en los valores máximos o mínimos de un portafolio de inversión, incluso en las acciones del

mercado. En las Ciencias Naturales, por ejemplo, se utiliza en los máximos anuales del nivel del mar, en datos diarios de la cantidad de lluvia, en las observaciones de la velocidad del viento (Coles, 2001) o, como en nuestro caso, la cantidad de ozono que se registra en una hora determinada. Como se puede notar, todos estos valores coinciden en que su escala de medición se realiza de manera continua.

1.2. Conceptos Básicos

Puesto que no es posible asignar probabilidades a todos y cada uno de los valores continuos de las variables aleatorias de una manera significativa, para todo número real $z \in \Omega$ se define:

$$F(x) = \Pr\{X \leq x\}$$

En este contexto, a F se le llama *función de distribución de la variable aleatoria* X.

De los axiomas generales de probabilidad, F cumple con las siguientes condiciones:

i. F monótona no-decreciente. Es decir, $z \leq z' \Rightarrow F(z) \leq F(z')$.

ii. F tiene límite 0 en $-\infty$ y 1 en ∞, de manera formal

$$\lim_{z \to -\infty} F(z) = 0 \qquad y \qquad \lim_{z \to \infty} F(z) = 1.$$

1.2.1. Límites

Muchas de las veces nos es complicado realizar cálculos exactos con distribuciones de probabilidad. Esto en gran parte se debe a que la distribución es desconocida; sin embargo, es posible aproximar la verdadera distribución. Esto requiere de una definición de convergencia de variables aleatorias. Hay varias caracterizaciones, pero la más útil para los propósitos del modelado es la convergencia de la distribución.

Definición 1.2.1. *Una sucesión de variables aleatorias $X_1, X_2, \ldots, X_n$ que tienen una función de distribución $F_1, F_2, \ldots, F_n$, respectivamente, se dice que converge en distribución a la variable aleatoria X, que tiene función de distribución F, si*

$$\lim_{n\to\infty} F_n(x) = F(x)$$

para todo $x \in F$.

La utilidad de establecer un límite en la distribución de F para una sucesión de variables aleatorias se justifica por el uso de F como una aproximación a la distribución de X_n para valores grandes de n.

1.3. Formulación

La teoría de valores extremos estudia el comportamiento de los valores máximos y mínimos. Desde el punto de vista estadístico, si $X_1, X_2, \ldots, X_n$ son una sucesión de variables aleatorias independientes que tienen una función de distribución F y

$$M_n = \text{máx}\{X_1, X_2, \ldots, X_n\}$$

los extremos son definidos como el máximo y el mínimo de las n variables aleatorias. En este contexto, sólo se hará referencia a los máximos.

En las aplicaciones, normalmente las X_i representan valores de un proceso medido en un tiempo regular, así que M_n representa el máximo de las observaciones del proceso en n unidades de tiempo. Por ejemplo, si n es el número de observaciones tomadas durante un año, entonces M_n corresponde al máximo anual o, para nuestros fines, si n es el número de observaciones tomadas durante un día, M_n corresponde al máximo diario. La distribución del máximo M_n puede calcularse

para todos los valores de n como sigue:

$$\begin{aligned}
\Pr\{M_n \leq z\} &= Pr\{X_1 \leq z, X_2 \leq z, \ldots, X_n \leq z\} \\
&= \Pr\{X_1 \leq z\} \times \Pr\{X_2 \leq z\} \times \ldots \times \Pr\{X_n \leq z\} \\
&= \underbrace{F(z) \times F(z) \times \ldots \times F(z)}_{n - veces} \\
&= \{F(z)\}^n. \qquad (1.1)
\end{aligned}$$

El comportamiento asintótico del máximo tiene que estar relacionado con la cola derecha de la distribución F, pues es justo aquí donde ocurren los extremos; por lo tanto, la función de distribución estará dada por el producto de probabilidades. Sin embargo, en la práctica esto no es útil desde el momento en que se desconoce cómo es la función de distribución F. Una manera de solucionar este gran inconveniente, es usar técnicas estadísticas para estimar la distribución asintótica F de los extremos M_n, cuando n es grande y entonces sustituir esta estimación en (1.1). Esto es equivalente a la práctica usual de aproximar la distribución de una muestra suponiendo normalidad, lo cual está justificado por el Teorema Central del Limíte. Desafortunadamente, las diferencias, por muy pequeñas que puedan resultar en la estimación de F, podrían conducir a obtener diferencias importantes y significativas para F^n.

Una alternativa para evitar estas diferencias es que, al aceptar que esta distribución es desconocida, no queda más remedio que buscar modelos de familias aproximadas para F^n, que sólo pueden estimarse con base en los datos extremos.

La manera en que se procede es observar el comportamiente de F^n cuando $n \to \infty$. Pero esto en sí no es suficiente, puesto que para cualquier $z < z_+$, donde z_+ es el valor más grande de F, es decir, $z_+ = (z, 0)$, se tiene que $F^n(z) \to 0$ cuando $n \to 0$, tal que la distribución de M_n degenera a un punto de z^+. Esta dificultad se evita mediante la siguiente parametrización lineal de la variable M_n:

$$M_n^* = \frac{M_n - b_n}{a_n}$$

para sucesiones constantes $\{a_n\} > 0$ y $\{b_n\}$ (Coles, 2001). La elección apropiada de $\{a_n\}$ que representa los parámetros de localización, es una medida de tendencia central de la distribución de los valores extremos que no es la media de la distribución, de la misma manera, $\{b_n\}$ representa los parámetros de escala que es una medida de dispersión pero no es la desviación estandar, de M_n^* , evitando así las complicaciones que surgen con la variable original M_n. Por tanto, se buscan las distribuciones del límite para M_n^* cuando n aumenta, con las elecciones apropiadas de $\{a_n\}$ y $\{b_n\}$, en lugar de M_n.

1.4. Clasificación de Extremos

Del desarrollo de las posibles distribuciones del límite para M_n^*, $G(z)$ se puede tomar la forma de una de las tres distribuciones, independientemente de la distribución original de las observaciones (Coles, 2001), las cuales se proporcionan en el siguiente teorema:

Teorema 1.4.1. *Si existen sucesiones constantes $\{a_n > 0\}$ y $\{b_n\}$ tales que*

$$Pr\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \to G(z) \qquad \text{cuando} \qquad n \to \infty$$

dónde G es una función de distribución no-degenerada, entonces, se dice que G pertenece a una de las siguientes familias:

$$I : G(z) = \exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}, \qquad -\infty < z < \infty$$

$$II : G(z) = \begin{cases} 0, & z \leq b \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, & z > b \end{cases}$$

$$III : G(z) = \begin{cases} \exp\left\{-\left[-\left(\frac{z-b}{a}\right)^{\alpha}\right]\right\}, & z < b \\ 1, & z \geq b \end{cases}$$

con parámetros $a > 0$, b y en el caso de las familias II y III $\alpha > 0$.

Generalmente, se dice que estas tres familias pertenecen a distribuciones de valor extremo, con tipos I, II y III que a su vez son conocidos como familias Gumbel, Fréchet y Weibull respectivamente. Cada familia tiene un parámetro de escala a y un parámetro de localización b. Adicionalmente, las familias Fréchet y Weibull están definidas en términos de un parámetro de forma denotado por α, llamado también índice de cola, pues indica el grueso de la misma, es decir, indica el grado de convergencia a cero de la densidad de probabilidad. Cuanto mayor es este parámetro, más se suaviza la cola; en otras palabras, la función converge de manera más rápida a la asíntota, mientras que, si el parámetro se acerca a cero, mayor es el grueso de la cola, es decir, la función converge a cero de manera más lenta.

El teorema 1.4.1 también implica que, cuando M_n puede establecerse con las sucesiones convenientes $\{a_n\}$ y $\{b_n\}$, la correspondiente variable parametrizada M_n^* tiene una distribución restringida a uno de los tres tipos ya mencionados. El rasgo notable de este resultado es que los tres tipos de distribuciones son los únicos límites posibles para las distribuciones de M_n^* sin tener en cuenta la distribución F para la población. En este sentido, el teorema proporciona un valor extremo semejante al Teorema Central del Límite (Coles, 2001).

1.5. Distribución de Valor Extremo Generalizado

Las tres familias que se mencionan en el teorema 1.4.1 tienen diferentes formas de comportamiento que corresponden a diferentes formas de conducta en la cola para la función de distribución F de cada una de las X_i. Esto se debe a la forma de la distribución del límite G a z_+ que es el punto extremo superior.

Para las distribuciones Gumbel y Fréchet $z_+ = \infty$, mientras que para la distribución Weibull, z_+ es finito.

La gráfica 1.1 con parámetros de localización y escala iguales a 0 y 1, respecti-

vamente, muestra que el comportamiento de esta función es, en efecto, de manera exponencial. Recordemos que, para esta familia, z está definida para cualquier valor real.

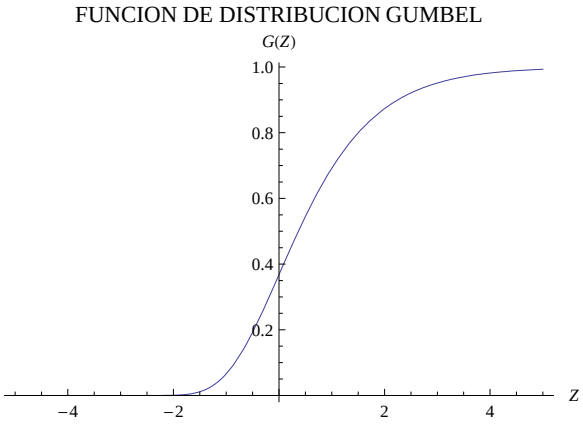


Figura 1.1: Función de Distribución Gumbel.

La gráfica 1.2 ilustra el comportamiento del tipo II o familia Fréchet. El parámetro α indica el comportamiento de la cola; cuanto mayor sea este valor, más se suaviza la cola, mientras que si el parámetro se acerca a cero, el grueso de la cola es mayor, en la línea punteada $\alpha = 1.5$, mientras que en la línea continua $\alpha = 3$, para ambos casos se consideró el parámetro de localización 0 y el paramétro de escala igual a 1. También se puede observar que es acotada por la izquierda, pues según la ecuación 1.2 esta familia acepta sólo valores positivos para z.

Finalmente, la gráfica 1.3 ilustra la función de distribución para la familia Weibull, nuevamente con parámetros de localización y escala iguales a 0 y 1, respectivamente. Como ya se mencionó anteriormente, el parámetro α de la distribución indica el grueso de la cola, la línea punteada es la gráfica para $\alpha = 3$. Se puede notar que, cuanto mayor es este parámetro la cola se suaviza, mientras que cuando el parámetro $\alpha = 1$ que es un valor cercano a cero, puede apreciarse en la línea continua de la misma gráfica, que el grueso de la cola es mayor para esta distribución.

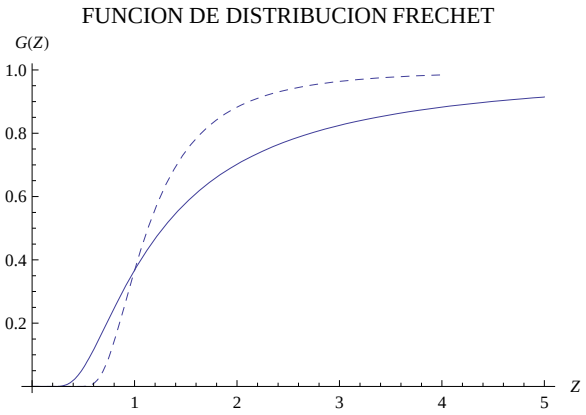


Figura 1.2: Función de Distribución Fréchet.

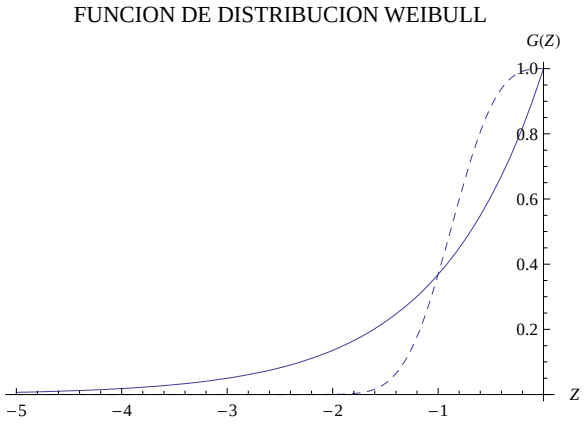


Figura 1.3: Función de Distribución Weibull.

Hace tiempo, en las aplicaciones de ésta teoría, era usual adoptar una de las tres familias para después estimar los parámetros convenientes de dicha distribución. Pero al aplicar este método hay dos desventajas:

- se requiere una técnica para elegir cuál de las tres familias es la más apropiada para los datos y

- una vez tomada la decisión, en las inferencias siguientes ésta opción se supone como correcta y no admite incertidumbre aunque ésta pueda ser importante.

Hasta el momento, se han descrito los tres modelos de distribuciones límite para los máximos de variables aleatorias independientes e idénticamente distribuidas, pero las familias Gumbel, Fréchet y Weibull pueden combinarse en una sola familia de modelos a través de la parametrización $\xi = \frac{1}{\alpha}$ y a la elección apropiada de los parámetros de localización $a = \mu$ y de escala $b = \sigma$ que tienen una función de distribución de la forma:

$$G(z) = \exp\left[-\left\{1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right\}_{+}^{-1/\xi}\right] \tag{1.2}$$

que está definida en el conjunto $\{z : 1+\xi\left(\frac{z-\mu}{\sigma}\right) > 0\}$ y, además, los parámetros cumplen que $-\infty < \mu < \infty$, $\sigma > 0$ y $-\infty < \xi < \infty$.

La ecuación 1.2 es conocida como **Distribución de Valor Extremo Generalizado** ($DVEG$) y contiene tres parámetros:

- el parámetro de localización μ
- el parámetro de escala σ
- el parámetro de forma ξ

Dicha expresión toma diferentes formas en función de cómo sea el parámetro ξ:

Cuando $\xi > 0$ la distribución de valor extremo corresponde a la familia de tipo II o Fréchet.

Mientras que si $\xi < 0$ en la parametrización, se dice que corresponde a la familia de tipo III o Weibull.

Se obtiene $\xi = 0$ cuando $\xi \to 0$ en la ecuación 1.2 que lleva a la familia Gumbel o tipo I cuya función de distribución es:

$$G(z) = \exp\left[-\exp\left\{-\left(\frac{z-\mu}{\sigma}\right)\right\}\right], \qquad -\infty < z < \infty,$$

La unión de las tres familias originales de distribución de valor extremo en una sola familia simplifica la aplicación estadística considerablemente. Por medio de la inferencia de ξ, los mismos datos determinan el tipo más apropiado de la conducta en la cola y no hay ninguna necesidad de hacer un criterio previo sobre cuál es la familia de valor extremo que se debe elegir de manera individual. Por lo que el teorema 1.4.1, se puede reescribir como:

Teorema 1.5.1. *Si existen sucesiones constantes* $\{a_n\}$ *y* $\{b_n\}$ *tales que*

$$Pr\left\{\frac{(M_n - b_n)}{a_n} \leq z\right\} \to G(z) \quad \textit{cuando} \quad n \to \infty$$

para un distribución G no degenerada, entonces, G es miembro de la familia de GEV

$$G(z) = \exp\left[-\left\{1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right\}_+^{-1/\xi}\right]$$

definida en el conjunto $\{z : 1+\xi\left(\frac{z-\mu}{\sigma}\right) > 0\}$ *y, además, los parámetros cumplen que* $-\infty < \mu < \infty$, $\sigma > 0$ *y* $-\infty < \xi < \infty$.

Interpretando el límite en el teorema 1.5.1 como una aproximación para los valores grandes de n, hace pensar en el uso de la familia de Valor Extremo Generalizada para modelar la distribución de máximos de sucesiones (Coles, 2001). La dificultad radica en que constantes normalizadas serán desconocidas, sin embargo, en la práctica eso se resuelve de modo sencillo. Asumiendo

$$Pr\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \approx G(z) \qquad \text{cuando} \qquad n \to \infty$$

De manera equivalente,

$$\begin{aligned} \Pr\{M_n \leq z\} &\approx G(z)\left\{\frac{(z-b_n)}{a_n}\right\} \\ &= G^*(z) \end{aligned}$$

Donde G^* es otro miembro de la familia de GEV. En otras palabras, si el teorema 1.5.1 permite la aproxiamación de la distribución de M_n^* por un miembro de la familia de GEV cuando n es grande, la distribución de M_n también puede aproximarse por un miembro diferente de la misma familia. Así, los parámetros de la distribución tienen que ser estimados, sin embargo, no es conveniente que en la práctica los parámetros de la distribución G sean diferentes de los de G^*.

Este argumento lleva al acercamiento siguiente para los extremos del modelado de una serie de observaciones independientes $X_1, X_2, \ldots$. Los datos forman bloques en las sucesiones de observaciones de longitud n, para algún valor grande de n, mientras se genera una serie de máximos en el bloque $M_{n,1}, M_{n,2}, \ldots, M_{n,m}$, es decir, en cual distribución de GEV puede ajustarse. A menudo los bloques se escogen de tal manera que corresponden a un periodo de tiempo de un año en donde n es el número de observaciones durante un año y los máximos del bloque corresponden a los máximos anuales (Cloes, 2001).

1.5.1. Distribución Gumbel

De las tres distribuciones clásicas, la familia de tipo I conocida como distribución de Gumbel es la que más ha atraído la atención; está ilimitada por ambos extremos y, además, se trata de una función con doble exponencial, cuya función de distribución es

$$G(z) = \exp\left\{-\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\}, \qquad -\infty < z < \infty$$

cuya función de densidad es [1]:

$$g(z) = \exp\left\{-\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\} * \frac{1}{\sigma}\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]$$

La gráfica 1.4 muestra el comportamiento de esta distribución, que está definida para cualquier valor de z. Así mismo, se muestra también la función de densidad, ambas se ilustran con parámetros de localización y escala iguales a 0 y 1, respectivamente, bajo esta distribución la cola en la función de densidad decrece.

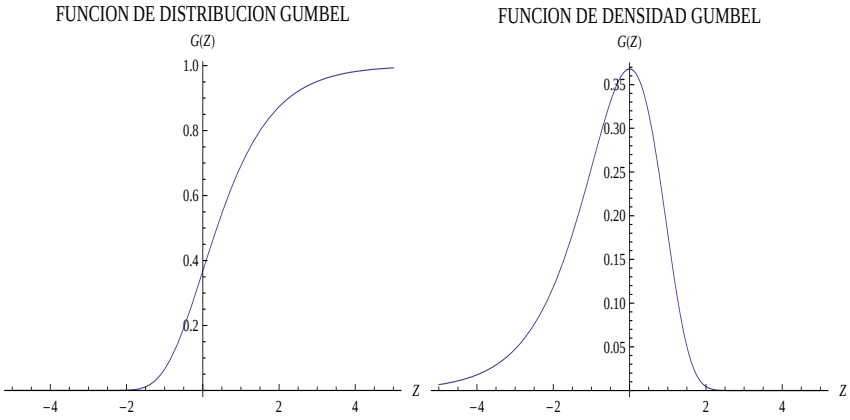


Figura 1.4: Gráficas de la Distribución Gumbel.

Cuando $\mu = 0$ y $\sigma = 1$ la función de densidad es la siguiente

$$g(z) = \exp\{-\exp[-z]\} * \exp[-z]$$

Por lo tanto, el momento de primer orden, es decir, el valor esperado para esta distribución cuando los parámetros toman los valores $\mu = 0$ y $\sigma = 1$ es

$$\begin{aligned} E[Z] &= \int_0^\infty (-\ln z) * \exp(-z)dz \\ &= -\int_0^\infty (\ln z) * \exp(-z)dz \\ &= \lambda \end{aligned} \tag{1.3}$$

[1] $g(z) = G'(z)$

donde $\lambda = -0.577216$ es el valor que toma la primera derivada de la función Gamma $\Gamma(n)$ con respecto a n, cuando $n = 1$, es decir, $\Gamma'(1) = \lambda$. Para un bosquejo de la demostración ver apéndice A.

Introduciendo los parámetros de localización y escala tenemos que

$$E[Z] = \mu + (-0.577216)\sigma$$

El momento de segundo orden para esta distribución conocido también como varianza cuando $\mu = 0$ y $\sigma = 1$ es

$$Var[Z] = \frac{\pi^2}{6}$$

Lo que implica que la Desviación Estándar se define mediante

$$\sigma(Z) = \frac{\pi}{\sqrt{6}}$$

Introduciendo los parámetros de localización y escala tenemos

$$\sigma(Z) = \frac{\sigma\pi}{\sqrt{6}}$$

CAPÍTULO 2

MODELOS CONDICIONALES

Varios procedimientos estadísticos y de química troposférica han sido propuestos para predecir niveles de ozono, ya sea considerando variables topológicas y de periodicidad, o incluso sin usar variables explicativas en lo absoluto. En la literatura estadística, Smith (1989) extiende los modelos del valor extremo al usar la teoría de los procesos puntuales para estudiar excedentes de ozono de alto nivel. Cox y Chu (1992) analizan máximos diarios de ozono de 43 zonas urbanas de Norteamérica, usando un modelo de regresión Weibull, cuyo parámetro de escala se determina en función de variables meteorológicas. Niu (1996) implementa un modelo ARMA de manera tal que incluye funciones suavizadas de las variables explicativas para modelar la media marginal, y permite que las varianzas de las innovaciones puedan ser modeladas a través de una liga log y un componente lineal y, de esta forma, algunas respuestas tienen una varianza más grande que otras. Smith y colaboradores (2002) desarrollaron un modelo autorregresivo bayesiano y semiparamétrico para explorar la forma funcional de la relación del ozono con algunas variables meteorológicas.

La mayoría de estos métodos pueden ser criticados debido a suposiciones poco

fundamentadas, malos ajustes y algunos otros por su gran complejidad. En particular, suponer una familia paramétrica con una cola corta para el ajuste de datos de ozono, tal como la Gaussiana, puede conllevar a inferencias erróneas; de igual forma, en varias circunstancias se ha supuesto independencia entre la sucesión de los máximos (Bloomfield et. al., 1996), lo cual puede ser muy cuestionable pues intuitivamente se puede discernir que los niveles de ozono de un periodo en particular dependen altamente de los niveles de los periodos inmediatos anteriores. Como la meta principal de un estudio de datos de ozono es modelar sus extremos al estimar el comportamiento de la cola superior, una elección más adecuada para el análisis correspondiente es usar una estructura paramétrica que esté compuesta por distribuciones de valor extremo y permita cierto grado de dependencia entre las observaciones.

Este estudio considera el problema de generalizar respuestas de la familia exponencial a un marco de respuestas de valor extremo. Específicamente, se adopta la metodología propuesta por Zeger y Qaqish (1988), la cual consiste en especificar el modelo autorregresivo en forma de distribución condicional cuya parametrización pertenece a la familia exponencial de distribuciones. Al igual que con los modelos lineales generalizados, a esta distribución se le incluyen términos autorregresivos en forma de variables explicativas pasadas y presentes.

2.1. Antecedentes

A los modelos de regresión que tienen como variables independientes a los valores anteriores de la serie de tiempo se les conoce como Modelos Autorregresivos. En este caso, los valores de la serie de tiempo se denotan por y_{t-1}, y_{t-2}, y_{t-3}, $\ldots$, y_{t-p} y la variable independiente que es el ozono, se denota mediante y_t, se trata de hallar una ecuación que relacione y_t con los valores más recientes de la serie de tiempo y_{t-1}, y_{t-2}, y_{t-3}, $\ldots$, y_{t-7}.

2.1.1. Series de Tiempo

Se definen como un conjunto de observaciones de una variable que está medida en puntos sucesivos en el tiempo o en periodos de tiempo (Anderson, 2008), dichos valores pueden ser tomados en intervalos regulares los cuales pueden ser por hora, día, semana, mes o año, aunque se pueden aplicar en cualquier otro periodo. Es conveniente pensar que el comportamiento de los datos de una serie de tiempo se debe escencialmente a cuatro componentes que, combinados, proporcionan los valores de la serie: tendencia, cíclico, estacional e irregular. A continuación, se definen a grandes rasgos cada uno de estos cuatro componentes.

Tendencia: Es un desplazamiento gradual hacia valores relativamente altos o bajos a través de un lapso largo de tiempo, observable a través de varios periodos; estos pueden ser lineales, no lineales e incluso sin tendencia.

Cíclico: La serie de tiempo muestra un comportamiento que consiste en tendencias periódicas que caen por arriba o abajo de la línea de tendencia con una duración de más de un año.

Estacional: Cuando existe un patrón periódico que dura a lo más un año.

Irregular: Corresponde a las variaciones aleatorias que se observan y no son explicadas por los componentes anteriores.

Cabe mencionar que los componentes de las series de tiempo no siempre se presentan solos; pueden presentarse de manera combinada o incluso, pueden presentarse todas juntas.

2.2. Definición del Modelo

De manera análoga al modelo de Zeger y Qaqish (1988), se propone definir un modelo autorregresivo de orden p para respuestas de valor extremo, de tal manera que la distribución condicional pertenece a la familia de distribuciones

de valor extremo cuyo parámetro de localización está ligado a una componente lineal que se forma de variables explicativas que contienen la historia presente y pasada de los últimos p períodos y de un vector de coeficientes de regresión; esto es, la distribución condicional de cada respuesta Y_t dado el conjunto de información presente y pasada

$$H_t = \{x_t, ..., x_{t-p}, y_{t-1}, ..., y_{t-p}\}$$

donde

- $\mathbf{x_t}$ es el vector de variables explicativas en el tiempo t y
- $\mathbf{y_t}$ es la respuesta observada en el tiempo t, está dada por la siguiente distribución de Valor Extremo Generalizado:

$$F(y_t \mid \mathbf{H}_t) = \exp\left[-\left\{1 + \xi\left(\frac{y_t - \mu_t}{\sigma}\right)\right\}_+^{-1/\xi}\right], \qquad \text{para} \quad y_t > \mu_t, \tag{2.1}$$

Notación.

- $\mu_\mathbf{t}$ parámetro de localización
- σ parámetro de escala
- ξ parámetro de forma

con $-\infty < \mu_t < \infty$, $\sigma > 0$, $-\infty < \gamma < \infty$ y $h_+ = \text{máx}(h, 0)$.

Aquí, μ_t está relacionado con la historia presente y pasada a través de una componente lineal de manera tal que

$$\mu_t = \boldsymbol{\beta}^T \mathbf{z}_t$$

donde $\mathbf{z}_t$ es un vector de variables explicativas seleccionadas de $\mathbf{H}_t$, que incluye a la ordenada, y $\boldsymbol{\beta}$ es el vector de coeficientes correspondiente.

La especificación del modelo en la ecuación (2.1) es una generalización del modelo para series de máximos no estacionarios propuesto por Smith (1989), el cual sólo considera a t en H_t. En ésta formulación se propone incluir respuestas y variables explicativas presentes y pasadas que expliquen la dependencia entre las respuestas. Entre los beneficios de la especificación propuesta aquí, se puede mencionar que la función de verosimilitud condicional tiene forma explícita, se pueden comparar modelos y se pueden llevar a cabo los diagnósticos correspondientes a través de los residuales estandarizados propuestos por Dunn y Smyth (1996).

Aunque el modelo propuesto puede estar mal especificado debido a que no hay una distribución multivariada de valor extremo que tenga una distribución condicional de valor extremo, contrario a las series de tiempo Gaussianas donde la condicional es Gaussiana también, Dupuis y Tawn (2001) encontraron que para correlaciones de orden 1 las distribuciones ajustadas del modelo condicional correcto y del mal especificado eran casi idénticas para dependencias relativamente altas.

2.2.1. La Técnica del Estimador de Máxima Verosimilitud

Se han propuesto muchas técnicas para la estimación del parámetro en los modelos de valor extremo. Entre ellas, se incluyen técnicas gráficas basadas en probabilidad; métodos de momento en que se igualan sus funciones con sus equivalentes empíricos; procedimientos en que los parámetros se estiman como las funciones especificas de orden estadístico, así como los métodos basados en máxima verosimilitud. Cada técnica tiene sus contras, pero la utilidad y adaptabilidad giran alrededor del modelo complejo construido por las técnicas de verosimilitud, lo que la hace particularmente atractiva.

Un importante método de estimación es el llamado *Método de Máxima Verosimilitud*. La característica principal del método de máxima verosimilitud es que examina los valores de la muestra para elegir como estimados de los parámetros

desconocidos los valores para los cuales la probabilidad o la densidad de probabilidad de obtener los valores de la muestra es un máximo.

Definición 2.2.1. *Si $x_1, x_2, \ldots, x_n$ son valores de una muestra aleatoria con función de densidad de probabilidad conjunta $f(x;\theta)$, la función de verosimilitud de la muestra está dada por:*

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

para valores de θ dentro de un dominio dado.

Cuando los valores muestrales se conocen, la función de verosimilitud es una función de una sola variable, en este caso θ. Así, el método de máxima verosimilitud consiste en hacer que la función $L(\theta)$ sea lo mayor posible y se refiere al valor de θ que maximiza la función de verosimilitud como estimador de máxima verosimilitud de θ.

Puesto que $\log L$ alcanza su valor máximo para el mismo valor de θ que L, se debe resolver la ecuación de verosimilitud:

$$\frac{\partial \log L}{\partial \theta} = 0$$

con respecto a θ. Cualquier solución de la ecuación de verosimilitud se denomina estimador de máxima verosimilitud de θ. En este trabajo se usa inferencia basada en verosimilitud debido a sus propiedades asintóticas que flexibilizan el modelado.

Una gran dificultad en el uso de métodos de verosimilitud para la distribución GEV es la que involucra a las condiciones de regularidad que se requieren para las propiedades asintoticas asociadas con el estimador de máxima verosimilitud. Tales condiciones no satisfacen al modelo de GEV porque el extremo es funcion del valor del parámetro $\frac{\mu - \sigma}{\xi}$ que es un punto extremo superior de la distribución cuando $\xi < 0$ y un es un punto extremo inferior cuando $\xi > 0$. Esta violación de

la regularidad significa que los resultados de verosimilitud asintótica estandar no son aplicables automáticamente. Según lo escrito en Coles (2001), Smith estudio detalladamente este problema obteniendo los siguientes resultados:

Cuando $\xi > -0.5$, el estimador de máxima verosimilitud es regular, en el sentido de que tiene las propiedades asintóticas usuales.

Cuando $-1 < \xi < -0.5$, el estimador de máxima verosimilitud generalmente se obtiene, pero no tiene propiedades asintóticas.

Cuando$\xi < -1$, es poco probable obtener el estimador de máxima verosimilitud.

El caso $\xi \leq -.05$ corresponde a distribuciones que tienen un pequeño salto en la cola superior. Esta situación casi no aparece en modelos de valor extremo.

2.2.2. Estimador de Máxima Verosimilitud

Sean $Z_1, Z_2, \ldots, Z_n$ variables aleatorias independientes con distribución GEV, la función de log-verosimilitud de los parámetros cuando $\xi \neq 0$ es

$$l(\mu, \sigma, \xi) = -m \log \sigma - (1 + \frac{1}{\xi}) \sum_{i=1}^{m} \log \left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^{m} \log \left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right]^{-1/\xi} \tag{2.2}$$

tal que

$$\left(\frac{z_i - \mu}{\sigma}\right) > 0 \quad i = 1, 2, \ldots, m \tag{2.3}$$

Para el caso $\xi = 0$ se requiere el uso del límite de la distribución Gumbel de GEV. Utilizando log-verosimilitud

$$l(\mu, \sigma) = -m \log \sigma - \sum_{i=1}^{m} \left(\frac{z_i - \mu}{\sigma}\right) - \sum_{i=1}^{m} \exp \left\{ - \left(\frac{z_i - \mu}{\sigma}\right)\right\}. \tag{2.4}$$

La maximización de las ecuaciones 2.2 y 2.4 con respecto al vector de parámetros (μ, σ, ξ) lleva al estimador de máxima verosimilitud con respecto a toda la

familia de GEV. No existe una solución analítica, pero para un conjunto de datos proporcionados, la maximización se obtiene mediante algoritmos numéricos de optimización. Se deben tomar precauciones necesarias para evitar las dificultades numéricas de los logarítmos que aparecen de la evaluación de 2.2 en la vecindad de $\xi = 0$. Este último problema se resuelve utilizando la ecuación 2.4 en lugar de la ecuación 2.2 para valores de ξ que caen dentro de una pequeña vecindad alrededor de cero.

Por consiguiente, las funciones de verosimilitud son relativamente fáciles de expresar; sin embargo, las funciones de log-verosimilitud son altamente no lineales por lo que se debe proponer un método numérico adecuado para optimizar esta función y así obtener tanto los estimadores de máxima verosimilitud como sus errores estándares.

La función de verosimilitud de un modelo de transición de orden p para

$$\{y_{m+1}, ..., y_n\}$$

condicional a las primeras m respuestas puede expresarse como

$$L(\theta) = \prod_{k=m+1}^{n} f(y_k \mid \mathbf{H}_k)$$

donde θ representa al vector de parámetros y f denota la función de densidad correspondiente a F.

En el presente estudio se usó la biblioteca `evd` (*distribución de valor extremo*) descrita por Stephenson y Gilleland (2006) del paquete estadístico **R**, cuya función `fgev` proporciona el estimador de máxima verosimilitud para la distribución de la ecuación (2.1); el proceso optimizador que usa esta función se basa en el método quasi-Newton (también conocido como el algoritmo de métrica variable), el cual usa evaluaciones de la función objetivo y sus gradientes para generar una "fotografía" de la superficie por optimizar y así buscar el punto estacionario donde el

gradiente es 0.

La distribución aproximada de $(\hat{\xi}, \hat{\sigma}, \hat{\xi})$ es normal multivariada con media (μ, σ, ξ) y matriz de varianza-covarianza igual a la inversa de la matriz de información evaluada por el estimador de máxima verosimilitud. Aunque esta matriz puede calcularse analíticamente, es más fácil utilizar técnicas numéricas para evaluar las segundas derivadas y encontrar también la matriz inversa. Los intervalos de confianza y otras inferencias se siguen inmediatamente de la normal aproximada del estimador.

Como en este tipo de datos generalmente se cuenta con un tamaño de muestra grande, las pruebas de cociente de verosimilitud para encontrar un modelo parsimonioso no son muy confiables pues tienden a presentar significancias importantes a variables cuya contribución a la explicación del fenómeno es muy modesta o nula.

2.2.3. Método de Selección hacia Adelante

Para encontrar el modelo más parsimonioso se utilizó el algoritmo de selección hacia adelante, el cual empieza sin ninguna variable independiente y se van agregando variables de una en una para determinar si una variable independiente debe ser ingresada al modelo. El procedimiento finaliza cuando el valor p de cada una de las variables independientes que no están en el modelo es mayor que el nivel de significancia (Anderson, 2008). Se considera que la eliminación hacia atrás es un procedimiento razonable, pues no se ignora a ninguna variable que podría ser importante.

Este método ayuda a determinar cuál de los tres parámetros originales gobierna de mejor manera el ajuste.

2.2.4. Criterio de Información Bayesiano

Para este estudio también, se propone usar el BIC (*Criterio de Información Bayesiano*), que es un procedimiento de selección que determina con mayor grado de precisión el modelo más parsimonioso penalizando tanto al número de parámetros como al tamaño de la muestra.

El criterio del BIC consiste en escoger el modelo para el cual

$$2 \ln L(\hat{\theta}) + n_p \ln n$$

tiene el valor más pequeño, donde n_p es el número de parámetros, $2 \ln L(\hat{\theta})$ se refiere a la Devianza Estadística y n indica el número de observaciones.

CAPÍTULO 3

APLICACIÓN DE VALOR EXTREMO A MÁXIMOS DE OZONO

En múltiples disciplinas científicas, se han hecho diferentes investigaciones que han permitido el desarrollo de métodos para cuantificar estadísticamente eventos extremos y sus consecuencias de un modo óptimo, dando lugar a distribuciones de probabilidad que permiten el modelado de los valores mayores o menores de alguna variable aleatoria.

Por la naturaleza de los datos, y como ya se mencionó anteriormente, el modelado en el presente trabajo, requiere del uso de la Teoría del Valor Extremo y se propone un modelo para el análisis de regresión de máximos correlacionados en serie. Se pretende extender la metodología propuesta por Benjamin et. al. (2003), la cual modela la función de distribución condicional como una distribución de la familia exponencial, en donde se liga a los parámetros asociados a la media con un componente lineal que, a su vez, contiene los términos correspondientes a observaciones pasadas, ya sean respuestas y variables explicativas.

3.1. Antecedentes

En las últimas décadas, se ha experimentado un incremento en la concentración de ozono y el crecimiento de la población en zonas urbanas grandes. De ahí el motivo de elegir a la Ciudad de Guadalajara como muestra de estudio, pues con alta frecuencia se presentan elevados índices de contaminación en esta región del país, debido en gran parte a su evolución industrial y comercial. Es por ello que surge la necesidad de detectar periodos de contaminación peligrosos para la población.

Específicamente, en este estudio se analizan los niveles de ozono de Guadalajara registrados en el periodo que abarca del 6 de enero de 1997 al 31 de diciembre de 2006. Se propone un modelo que en presencia de variables atmosféricas y no estacionarias, tenga la capacidad de predecir en forma efectiva periodos con altos niveles de ozono. La capacidad de un modelo para predecir la ocurrencia de un nivel alto de contaminación es un aspecto relevante para las autoridades ambientales en la ejecución de acciones como medidas preventivas; tales medidas incluyen la reducción de la exposición de grupos vulnerables y la aplicación, por ejemplo, de programas de contingencia ambiental.

Para el objeto de estudio se usa una distribución de la familia de valor extremo similar a los modelos lineales generalizados. La idea principal es ligar a cada uno de los parámetros de la distribución con un componente lineal como lo hace con un parámetro el modelo de Benjamin y otros (2003). Esta especificación ha sido explorada por Dupuis & Tawn (2001) para cuando se tiene una sola variable explicativa. En su estudio, Dupuis & Tawn encontraron que cuando las colas de las dos variables son considerablemente largas el modelo condicional aquí descrito es muy competitivo con una distribución condicional propia. Como lo que se desea predecir son días repetidos de contaminación elevada, esta especificación parece ser muy conveniente para el modelado.

Todos los fenómenos meteorológicos pueden jugar un papel importante en la evolución de los contaminantes en la atmósfera y, por lo tanto, algunos aspectos relacionados con estos fenómenos deben tomarse en cuenta. Las concentraciones de ozono monitoreadas durante el periodo $1997-2006$ fueron utilizadas para hacer comparaciones de tipo estadístico, en el que se involucra tanto información no estacionaria como atmosférica, para tener la capacidad de predecir en forma efectiva los máximos niveles diarios de ozono.

3.2. Los Datos

En la figura 3.1 se ilustra la gráfica de la serie de tiempo de los máximos niveles diarios de ozono y se observa claramente como en el Área Metropolitana de la Ciudad de Guadalajara integrada por los municipios de Guadalajara, Tlaquepaque, Tonalá y Zapopan, en el Estado de Jalisco se presentan con frecuencia elevados índices de contaminación que violan las normas mexicanas al superar los 0.11 ppm en el periodo ya mencionado, pues son excesivos los máximos que diariamente sobrepasaron la línea del umbral que se ubica en el nivel establecido por la NORMA Oficial Mexicana de la Salud Ambiental. Nótese como dicho límite se excedió en repetidas ocasiones, llegando incluso a concentraciones tan altas como la que se tiene registrada el día 13 de mayo del año 1998, que alcanzó los 0.313 ppm, nivel más alto de todo el periodo estudiado.

En esta misma gráfica, también se puede observar una aproximación a la variación estacional en la serie de tiempo debido a que el comportamiento de la variable es regular, lo que sugiere que los efectos producidos por las variables atmosféricas tienen una constante anual. En otras palabras, se puede decir que la máxima cantidad se ozono se concentra siempre en los mismos meses del año y en épocas determinadas.

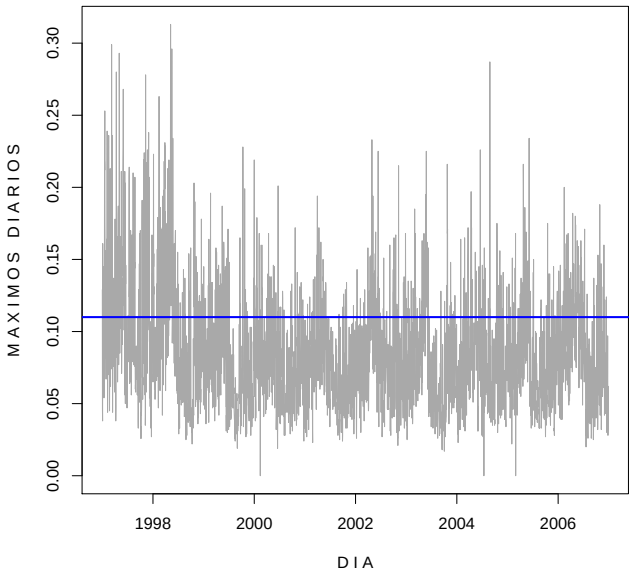


Figura 3.1: Máximos diarios de ozono, el umbral está ubicado en 0.11 ppm

3.2.1. Descripción de los Datos

La base de datos fue proporcionada por el **Instituto Nacional de Ecología** (INE) de la Ciudad de México y consiste en mediciones de diversas variables atmosféricas sobre la superficie del terreno tomadas por varias estaciones de monitoreo las cuales están dispersas en una región y tienen varias altitudes sobre el nivel del mar. Los datos a disposición van del 01 de enero de 1997 al 31 de diciembre de 2006, sin embargo, los datos estudiados sólo se consideran a partir del lunes 06 de enero de 1997 hasta domingo 31 de diciembre de 2006 sumando un total de 3645 días, esto es con el fin de comenzar el estudio en la primera semana completa del año 1997 pues, de otro modo, se tiene la sospecha de que el año anterior influya un poco en el índice de contaminación en estos primeros días del año.

3.2.2. Red Automática de Monitoreo

Según el INE, el monitoreo de la calidad del aire en la Zona Metropolitana de la Ciudad de Guadalajara (ZMG), se inició en el año de 1975, interrumpiéndose cuatro años más tarde. Años después se rehabilitó la red manual y hasta 1991 estuvo constituida por 15 estaciones donde se muestreaban principalmente, partículas suspendidas totales y plomo, para el año de 1993, se puso en marcha la actual red automática de medición continua. Las estaciones de monitoreo realizan mediciones cada 10 minutos y se encuentran operando las 24 horas de los 365 días del año en ocho estaciones que componen la Red Automática de Monitoreo Atmosférico (RAMA) de la ZMG. Estas estaciones reciben los nombres de: Las Aguilas (AGU), Atemajac (ATM), Centro (CEN), Loma Dorada (LDO), Miravalle (MIR), Oblatos (OBL), Tlaquepaque (TLA) y Vallarta (VAL), referenciando el lugar de su ubicación; las abreviaturas entre parentesis indican su clave respectiva, la cual fue proporcionada también por el INE.

De acuerdo con el INE, la ZMG se ubica en el centro del Estado de Jalisco, a una latitud de 20º39´54"N, longitud de 103º18´42"W y una altitud de 1,540 m sobre el nivel del mar, se sitúa en la cuenca del Valle del Río Grande de Santiago, en los Valles de Atemajac y la Planicie de Tonalá, entre las zonas montañosas de la Sierra Madre Occidental y el Eje Neovolcánico. Las montañas que circundan la zona son: al noroeste la Sierra de San Esteban; al sureste, la Serranía de San Nicolás y los conjuntos montañosos Cerro Escondido-San Martín y El Tapatío-La Reyna; al sur, el Cerro del Cuatro-Gachupín-Santa María; y al oeste, la Sierra de la Primavera.

Para tener una mejor idea acerca de su distribución en la Ciudad, la figura 3.2 muestra un mapa del Estado de Guadalajara donde se ilustra la ubicación de estas estaciones. Así mismo, en el cuadro 3.1 se muestran las coordenadas geográficas de la localización de cada una de las 8 estaciones que componen esta red de monitoreo, así como también el domicilio particular de cada estación.

ZONA	ESTACIÓN	CLAVE	LONGITUD	LATITUD	DOMICILIO
NORTE	Atemajac	ATM	103°21′19″	20°43′10″	Calle Hidalgo No. 1 entre Cuauhtémoc y Ramón Corona
	Oblatos	OBL	103°17′48″	20°42′02″	Avelino M. Presa No. 1685, Col. Oblatos
ESTE	Loma Dorada	LDO	103°15′50″	20°37′45″	Calle Loma Plana Norte cruza con Loma Plana Sur
	Tlaquepaque	TLA	103°18′45″	20°38′27″	Calle Constitución 159 esq. con Prisciliano Sánchez
CENTRO	Centro	CEN	103°19′59″	20°40′25″	Calle Churubusco No. 143 entre Dionisio Rodríguez y Javier Mina, Sector Libertad
SUR	Miravalle	MIR	103°20′35″	20°36′51″	Av. Gobernador Curiel cruce con Av. de la Pintura, Col. Miravalle
OESTE	Aguilas	AGU	103°25′01″	20°37′51″	Av. Adolfo López Mateos No. 5250
	Vallarta	VAL	103°23′55″	20°40′48″	Calle Coras entre Lacandones y Rincón del Nardo, residencial Juan Manuel

Cuadro 3.1: Coordenadas geográficas y domicilio de las estaciones de monitoreo

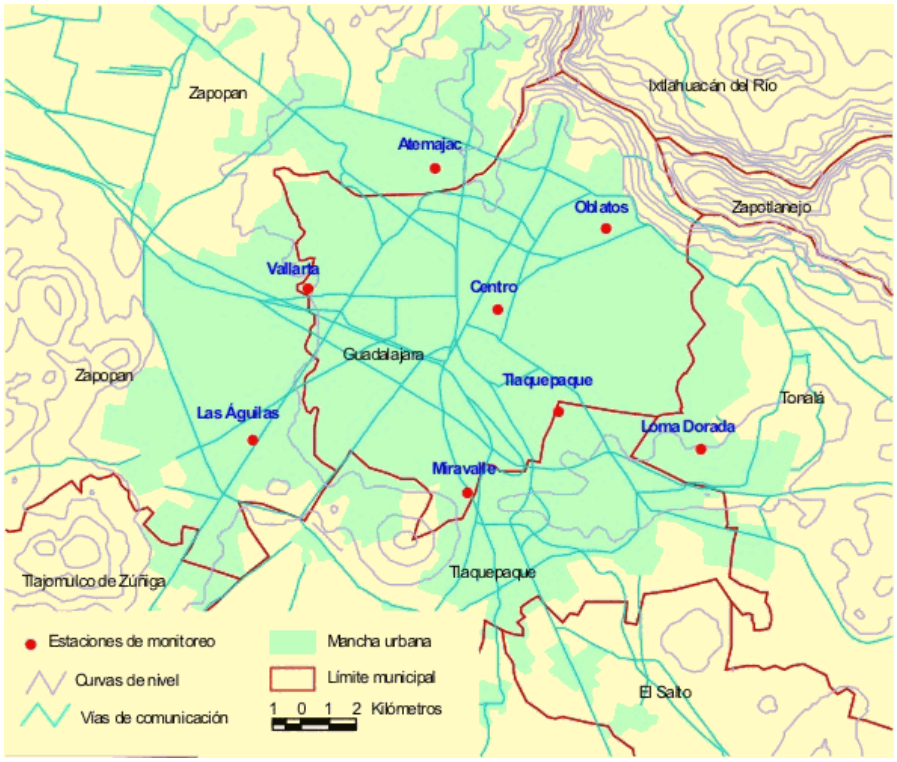


Figura 3.2: Red de monitoreo en la Ciudad de Guadalajara

3.3. Datos atípicos

Es frecuente que en la recolección de datos aparezcan errores de medición, cambios en el modelo de medición o simplemente ocurra un descuido a la hora de la transcripción de los elementos observados. A estas observaciones numéricas que parecen haberse generado de forma desigual del resto de las demás en estadística se les conoce como *outlier* o dicho de otro modo *datos atípicos*. La base de datos, por ejemplo, tenía registradas temperaturas mayores a 100 ^{0} C que sabemos no son posibles. En otro caso, había el registro de una temperatura menor a cero, esto no es grave si hablamos de -2 ^{0} C, pues no es creible que en la Ciudad de Guadalajara entre las 14 y 17 hrs haya un descenso repentino de temperatura para

después recuperarla de una hora a otra, de manera cierta, esto es totalmente atípico. En otras ocasiones, simplemente no existe registro en alguna hora del día. Para solucionar este problema, el primer paso fue detectar las observaciones atípicas que resultaron ser pocas y se optó por reemplazarlas con el valor inmediato anterior registrado para cada una de las variables. Se eligió hacerlo de esta manera debido a que, si se etiquetaban como observaciones no disponibles NA: *Not Available*, había problemas numéricos para llevar a cabo la programación de los modelos en el paquete *R*.

La estación Oblatos, ubicada al Norte de la Ciudad de Guadalajara, presenta observaciones no disponibles en los últimos cuatro años para las 5 variables atmosféricas de interés, así que esta estación no fue tomada en cuenta para llevar a cabo este estudio; por lo tanto, sólo se consideran las 7 estaciones restantes de la RAMA de esta Ciudad.

3.4. Variables Atmosféricas

Cabe mencionar que son diversos los parámetros que se registran en todas y cada una de estas estaciones de monitoreo, como son: el ozono (O_3), óxidos de nitrógeno (NOX, NO y NO_2), bióxido de azufre (SO_2), monóxido de carbono (CO), Partículas menores a 10 micras (PM10), Plomo (Pb), entre otros. En particular, el bioxido de nitrógeno, combinado con el óxido de azufre reacciona contra el agua provocando lluvia acida que afecta sobre todo a las regiones industrializadas. A pesar de esto, sólo cinco de éstas fueron consideradas para este estudio pues no todas son de nuestro interés al no poder clasificarlas como variables atmosféricas. Este es el motivo por el que se eligió trabajar con las variables que se enlistan a continuación en el cuadro 3.2, en el cual se muestran también la notación empleada y las correspondientes unidades de medida:

VARIABLE	NOTACIÓN	UNIDAD DE MEDIDA
ozono	O3	$\mu g/m^3$
dirección del viento	dvel	grados al Norte
humedad relativa	hum	porcentaje
temperatura	tem	$^0\,C$
velocidad del viento	vel	m/s

Cuadro 3.2: Variables atmosféricas observadas

3.4.1. Descripción de las Variables Atmosféricas

La frecuencia de los máximos diarios de ozono por cada hora se muestra en la figura 3.3. Es claro que el máximo nivel se presenta por la tarde, siguiendo un decaimiento de ozono al comienzo de la noche, para finalmente llegar a un mínimo desde la media noche hasta las primeras horas del día.

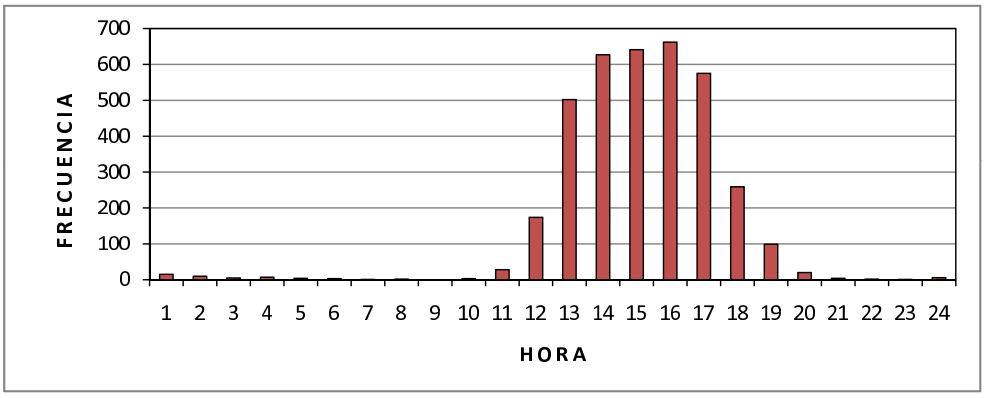


Figura 3.3: Máximos diarios de Ozono

Así pues, y debido a que la concentración de ozono es mayor a media tarde, y para reducir la gran cantidad de datos que se tienen si se consideran 24 datos

diarios durante 10 años de estudio, se toma el valor máximo entre las 12 y 17 horas de cada día, se consideró este intervalo para tener datos de una cuarta parte del día, y esto fue así para cada una de las estaciones de la red de monitoreo (Huang y Smith, 1999). Más adelante se mencionará que las demás variables atmosféricas también se restringieron a este horario.

Se sabe que las temperaturas altas favorecen generalmente la difusión de contaminantes así como el viento ya que desplaza las masas de aire en función de la presión y la temperatura. La humedad juega un papel negativo en el incremento de estos contaminantes, pues ayuda a la acumulación de polvo, por lo que temperaturas altas junto con bajas velocidades de viento favorecen la presencia de altas concentraciones de ozono [15, 3, 11]. Intuitivamente, puede afirmarse que las sierras que rodean a esta ciudad conforman una barrera física natural para la circulación del viento, impidiendo el desalojo del aire contaminado fuera de la ZMG.

En el cuadro 3.3 se resume el número de ocasiones en que las concentraciones de ozono alcanzaron niveles más allá de lo permitido, es decir, rebases mayores o iguales a 0.11 ppm en cada mes durante el periodo estudiado. Este límite de exposición tiene una frecuencia máxima aceptable de una vez cada tres años (NOM-020-SSA-1993). A manera de resumen, en la figura 3.3 se muestran las gráficas de estos datos; nótese como en los meses de marzo a junio, periodo que abarca la primavera, el exceso de ozono es más frecuente debido a que en esta época del año disminuye el movimiento de las masas de aire y las tardes son más cálidas y soleadas, por el contrario, en la temporada de otono-invierno, que son los meses de septiembre en adelante, incluyendo febrero, mes en que la temperatura aún es baja, la presencia del ozono dismunuye un poco.

En la figura 3.5, se ilustra de forma gráfica el número anual de las ocasiones que se rebasó este nivel permitido y se puede observar que en 1997, año en que se puso en marcha el “Programa para el mejoramiento de la calidad del aire de la Zona

Metropolitana de Guadalajara 1997-2001", resultó ser el que excedió más dicho límite, de manera que, a partir de entonces, fue disminuyendo el nivel de la cantidad de ozono hasta el año 2001, cuando dejó de implemantarse dicho programa, lo que sugiere que sí es indispensable implementar un buen sistema que beneficie el modo de vida de la sociedad.

Año	**Ene**	**Feb**	**Mar**	**Abr**	**May**	**Jun**	**Jul**	**Ago**	**Sep**	**Oct**	**Nov**	**Dic**
1997	16	12	15	14	24	23	9	12	4	12	17	12
1998	9	16	19	19	27	12	6	6	3	7	10	9
1999	5	7	3	11	12	7	2	0	3	5	3	0
2000	10	8	4	11	6	3	4	1	1	4	5	7
2001	1	2	7	5	8	2	0	0	2	6	2	1
2002	2	1	3	14	19	8	3	2	1	12	2	6
2003	8	4	7	6	21	10	0	0	1	6	5	3
2004	0	3	8	6	5	6	6	3	0	6	5	2
2005	3	4	2	12	10	14	3	0	3	4	4	6
2006	6	11	12	16	22	8	4	1	3	5	3	3

Cuadro 3.3: Número de veces que las concentraciones de ozono fueron mayores o iguales a 0.11 ppp para cada mes durante 1997-2006

Tomando en cuenta estas afirmaciones, la manera de considerar a las variables atmosféricas empleadas en este trabajo es la siguiente:

tem: Promedio de máximos diarios de temperatura registrados en cada estación entre las 12 y 17 hrs.

rangotem: Diferencia entre los promedios máximos diarios y mínimos diarios de temperatura registrados en cada estación entre las 12 y 17 hrs

vel: Promedio de máximos diarios de velocidad de viento registrados en cada estación entre las 12 y 17 hrs.

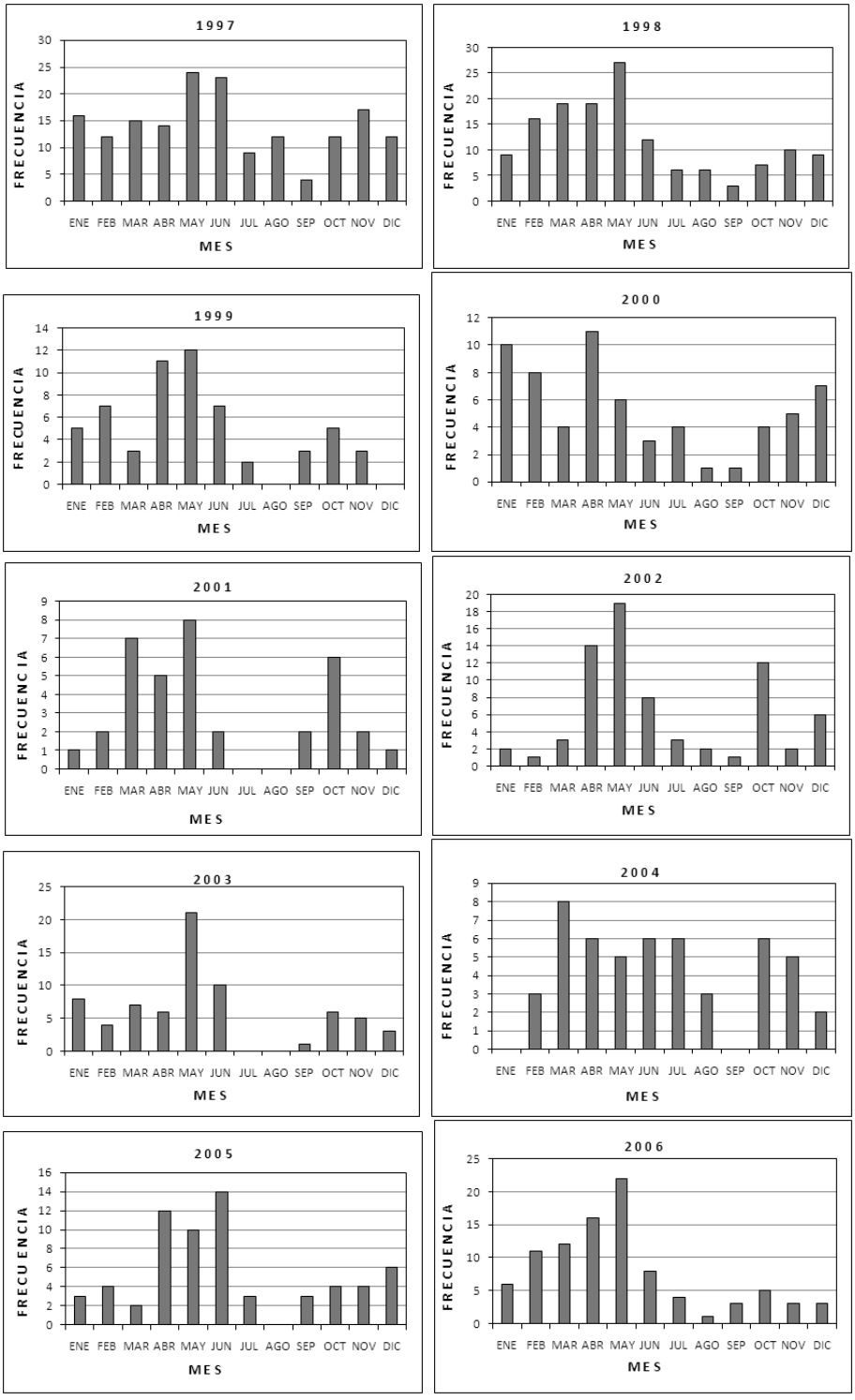


Figura 3.4: Número mensual de ocasiones en que se rebaso la NORMA Oficial Mexicana durante 1997-2001.

rangovel: Diferencia entre los promedios máximos diarios y mínimos diarios de velocidad de viento registrados en cada estación entre las 12 y 17 hrs.

hum: Promedio de máximos diarios de humedad relativa registrados en cada estación entre las 12 y 17 hrs.

rangohum: Diferencia entre los promedios máximos diarios y mínimos diarios de humedad relativa registrados en cada estación entre las 12 y 17 hrs.

La dirección del viento es una variable importante y es un elemento climatológico definido como "el aire en movimiento", el cual, principalmente, se describe por dos características:

- la velocidad y
- la dirección

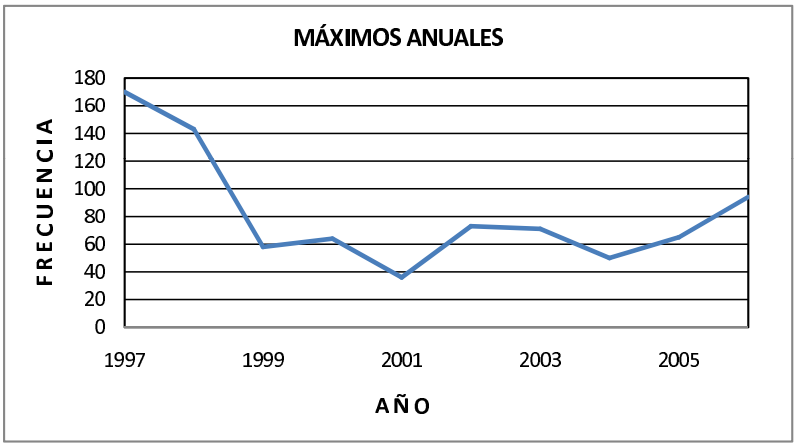


Figura 3.5: Número anual de ocasiones en que se rebaso la NORMA Oficial Mexicana de 1997 a 2001.

Debido a esto, se considera un vector con magnitud (dada por la velocidad) y dirección, entonces, para poder incluirla, es necesario hacer una parametrización,

pues su unidad de medida está en grados con respecto al norte. Para ello, Huang y Smith (1999) proponen crear dos variables adicionales

$$wu \longrightarrow -vv * \frac{\text{sen}(2\pi * dv)}{360}$$
$$wv \longrightarrow -vv * \frac{\cos(2\pi * dv)}{360}$$

La variable `wu` es la componente este-oeste del viento, la cual es positiva cuando el viento proviene del oeste; de manera análoga, `wv` es la componente norte-sur, la cual es positiva cuando el viento proviene del sur.

Estos vectores de viento son los que están registrados los lunes a las 15 hrs, la razón de elegir este día es porque resulta ser el más contaminado. En segundo lugar, se encuentra el sábado y, en tercero el viernes, como puede apreciarse en la figura 3.6, en que se graficó con qué frecuencia presentan los dias mayor contaminación.

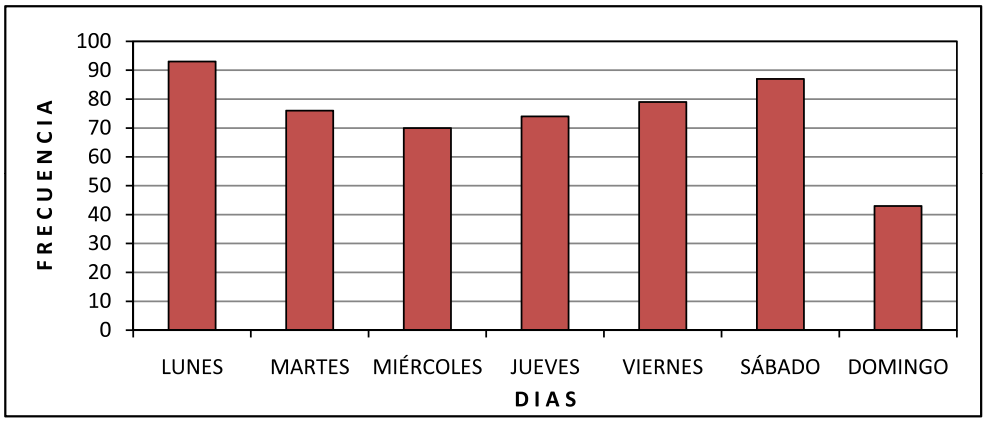


Figura 3.6: Máximos diarios de Ozono

Para ajustar efectos no lineales del tiempo t, se usan funciones polinomiales como t^2, t^3, ...; mientras que para incluir efectos `anuales` se usan los términos

$$\cos\left(\frac{2\pi t}{365}\right) \qquad \text{y} \qquad \text{sen}\left(\frac{2\pi t}{365}\right)$$

y para efectos `semestrales` se agregan los términos

$$\cos\left(\frac{2\pi t}{182.5}\right) \qquad \text{y} \qquad \text{sen}\left(\frac{2\pi t}{182.5}\right)$$

donde 365 y 182.5 equivale a los días que hay en un año y en un semestre, respectivamente.

3.5. Polinomios Ortogonales

Como se mencionó anteriormente, se incluyó a la variable tiempo en términos de bases ortogonales de una regresión polinomial. La misma idea fue implementada para la variable respuesta ozono; sin embargo, ésta sólo fue significativa para la observada en el tiempo $t-1$.

Los polinomios ortogonales se usan para ajustar un modelo polinómico de una variable de cualquier orden, aun cuando esté mal condicionado (Montgomery y Peck, (1982)), todavía se puede tener correlación alta entre ciertos coeficientes de la regresión. Algunas de estas dificultades pueden eliminarse usando polinomios ortogonales para ajustar el modelo. Supongamos que el modelo es:

$$Y = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \ldots + \beta_k z_i^k + \varepsilon_i, \qquad i = 1, 2, 3, \ldots, n \tag{3.1}$$

Notación.

- Y es la variable respuesta
- $\mathbf{z_i}$ es un vector de variables explicativas

Generalmente, las columnas de la matriz $\mathbf{Z}$ que son los vectores z_i pueden no ser ortogonales. Si aumentamos el orden del polinomio agregando el término $\beta_{k+1} z_i^{k+1}$, se debe calcular $(Z'Z)^{-1}$ y el grado de los parámetros estimados $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, cambiará (Draper y Smith, 1981). Al reescribir el modelo (3.1) se tiene

$$Y = \alpha_0 P_0(z_i) + \alpha_1 P_1(z_i) + \alpha_2 P_2(z_i) + \ldots + \alpha_k P_k(z_i) + \varepsilon_i, \qquad i = 1, 2, 3, \ldots, n \tag{3.2}$$

Notación.

- $\mathbf{P_u(z_i)}$ es el polinomio ortogonal de grado u definido como

$$\sum_{i=1}^{n} P_r(z_i)P_s(z_i) = 0, \qquad r \neq s, \qquad r, s = 0, 1, 2, \ldots, n \qquad P_0(z_i) = 1$$

Entonces el modelo se convierte en $y = Z\alpha + \epsilon$; donde la matriz Z es de la siguiente forma

$$Z = \begin{pmatrix} P_0(z_i) & P_1(z_i) & \cdots & P_k(z_i) \\ P_0(z_i) & P_1(z_i) & \cdots & P_k(z_i) \\ \vdots & \vdots & \vdots & \vdots \\ P_0(z_i) & P_1(z_i) & \cdots & P_k(z_i). \end{pmatrix} \tag{3.3}$$

La matriz 3.3 tiene columnas ortogonales, por lo que la matriz $Z'Z$ es igual a

$$Z'Z = \begin{pmatrix} P_0(z_i) & P_0(z_i) & \cdots & P_0(z_i) \\ P_1(z_i) & P_1(z_i) & \cdots & P_1(z_i) \\ \vdots & \vdots & \vdots & \vdots \\ P_k(z_i) & P_k(z_i) & \cdots & P_k(z_i) \end{pmatrix} \begin{pmatrix} P_0(z_i) & P_1(z_i) & \cdots & P_k(z_i) \\ P_0(z_i) & P_1(z_i) & \cdots & P_k(z_i) \\ \vdots & \vdots & \vdots & \vdots \\ P_0(z_i) & P_1(z_i) & \cdots & P_k(z_i) \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^{n} P_0^2(z_i) & 0 & \cdots & 0 \\ 0 & \sum_{i=1}^{n} P_1^2(z_i) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \sum_{i=1}^{n} P_k^2(z_i). \end{pmatrix} \tag{3.4}$$

De nuevo, la inversa de la matriz 3.4, es decir, $(Z'Z)^{-1}$ es también ortogonal y ella se obtiene con el inverso multiplicativo de cada elemento de la diagonal, esto es,

$$(Z'Z)^{-1} = \begin{pmatrix} \dfrac{1}{\sum_{i=1}^{n} P_0^2(z_i)} & 0 & \cdots & 0 \\ 0 & \dfrac{1}{\sum_{i=1}^{n} P_1^2(z_i)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dfrac{1}{\sum_{i=1}^{n} P_k^2(z_i)} \end{pmatrix}$$

El procedimiento de mínimos cuadrados proporciona una estimación de las α_j de 3.2 como

$$\hat{\alpha}_j = \frac{\sum_{1=1}^{n} P_j(z_i) y_i}{\sum_{1=1}^{n} P_j^2(z_i)}, \qquad j = 0, 1, 2, \ldots, k$$

3.5.1. Resultados

En el cuadro 3.4, se muestran los resultados de las devianzas y su correspondiente criterio de Bayes (ver Apéndice C). Se observa que, a partir del tercer modelo, la devianza aumenta y disminuye de manera alternada; nótese que para los modelos 3 y 6 este valor coincide. Estos resultados se obtuvieron utilizando el software **R**; en particular, como ya se dijo anteriormente, se usó el paquete **evd** (Stephenson y Gilleland, 2006), que es un sistema de cálculo estadístico. Cuando se realizó el ajuste de los datos, se encontró que, en presencia de todas las variables explicativas, el parámetro de forma ξ no tiene efectos significativos, por lo que se decidió emplear la distribución marginal Gumbel que está representada por la función de distribución

$$F(y_t; \mathbf{x}_t) = \exp\left[-\exp\left\{-\left(\frac{y_t - \mu(\mathbf{x}_t)}{\sigma}\right)\right\}\right]$$

.

MODELO	DEVIANZA	n_p	BIC
1	−16511.73	17	−16372.3344
2	−16795.39	21	−16623.1955
3	−16817.45	23	−16628.8560
4	−16830.52	26	−16617.3268
5	−16841.33	27	−16619.9370
6	−16817.45	23	−16628.8560
7	−16841.58	26	−16628.3868
8	−16840.29	26	−16627.0968

Cuadro 3.4: Devianza de los modelos propuestos con su respectivo BIC

Un problema que es inminente en este tipo de modelado es el de parametrizar excesivamente a la distribución condicional. Cuando el número de variables explicativas es grande, para encontrar el modelo más parsimonioso se propone usar el algoritmo de eliminación regresiva *(backward selection)*, el cual consiste en evaluar la significancia de un parámetro a la vez y después eliminar al más débil, la manera de proceder ya se explicó en la seccion 2.2.2.

Puesto que en el modelo se incluyen las variables independientes *tem* y *vel*, entonces se puede incluir también a la variable de interacción *tem*vel* de acuerdo con [15, 3, 11], una vez que se eligió al modelo más parsimonioso, se verificó que esta interacción también fuera significativa, utilizando de nuevo el criterio del BIC se determinó que, en efecto, resultan significativas tanto $tem_1 * vel_1$ como $tem_2 * vel_2$.

Así pues, bajo los criterios mencionados anteriormente, en el caso de nuestro estudio, el modelo más parsimonioso resulta un modelo autorregresivo de orden 2 que contiene 23 variables explicativas (modelo 3 en el cuadro 3.4) cuya formulación

resultó ser

$$\mathtt{t}^7 + \mathtt{y_{t-1}}^2 + \mathtt{y_{t-2}} + \mathtt{semestrales} + \mathtt{tem_t} + \mathtt{v_t} + \mathtt{tem_t} * \mathtt{v_t} + \mathtt{rtem_t} + \mathtt{rvv_t} + \mathtt{wu_t} + \mathtt{wv_t} + \mathtt{tem_{t-1}} + \mathtt{v_{t-1}} + \mathtt{tem_{t-1}} * \mathtt{v_{t-1}} + \mathtt{h_{t-1}} + \mathtt{rvv_{t-1}} + \mathtt{wu_{t-2}}$$

donde el superíndice denota el grado del polinomio.

Variable	**Estimador**	**Error Estandar**
μ	0.072870	0.0003473
$\mathtt{y_{t-1}}$		
$\mathtt{y_{t-2}}$	0.073777	0.0244300
$\mathtt{semestral}$	0.001881	0.0005156
$\mathtt{tem_1}$	0.777762	0.0463656
$\mathtt{rangotem_1}$	0.151047	0.0263547
$\mathtt{vel_1}$	-0.852195	0.0344144
$\mathtt{rangovel_1}$	-0.135639	0.0329698
$\mathtt{wu_1}$	0.201069	0.0309544
$\mathtt{wv_1}$	-0.120397	0.0212634
$\mathtt{tem_2}$	-0.371501	0.045416
$\mathtt{vel_2}$	0.264405	0.0286195
$\mathtt{hum_2}$	-0.294538	0.0284432
$\mathtt{rvel_2}$	0.096665	0.0279282
$\mathtt{wu_3}$	-0.081762	0.0255117
$\mathtt{tem_1} * \mathtt{vel_1}$	-15.282099	1.685457
$\mathtt{tem_2} * \mathtt{vel_2}$	3.111811	1.3923151
σ	0.019835	0.0002594

Cuadro 3.5: Modelo con Shape cero

El Cuadro 3.5 muestra el valor de los estimadores de los coeficientes lineales en el modelo así como los estimadores de los parámetros de localización μ y de escala

σ para el modelo elegido. Es posible notar que los coeficientes correspondientes al promedio de máxima temperatura (ver *tem1* y *tem2* en el Cuadro 3.5) son muy significativos. Estos indican que un día caluroso incrementa la severidad de los máximos de ozono. Sin embargo, si el día que precede también es caluroso, los máximos pueden ser aminorados. Como es de esperarse, los vectores de velocidad de viento, también juegan un papel relevante tanto en la dispersión de los contaminantes como en la concentración de ellos, ocasionando efecto inverso. Como es bien sabido, un día con viento dispersa los contaminantes; además, si en el día anterior hubo viento, entonces es posible que el actual sea susceptible de presentar un incremento en el máximo. Un aumento de humedad en el día anterior también contribuye a disminuir la severidad del máximo de ozono.

La Figura 3.7 muestra el ajuste del polinomio de tiempo t ajustado con las bases ortogonales de polinomios de grado 7 que se obtuvieron en el modelo más parsimonioso con bandas de confianza del 99 %.

Es notoria la baja en la severidad de los máximos de ozono los primeros mil días aproximadamente; lo que sugiere que, en el periodo en que se implementó un programa de mejoramiento ambiental, se redujeron los niveles de contaminantes. Sin embargo, en cuanto se dejó de aplicar el programa los niveles se incrementaron de nuevo. También se nota que el comportamiento en la gráfica es generalmente regular, pues aunque se nota un crecimiento en los máximos, éste no alcanza las mismas dimensiones que en los primeros tres años, lo que corrobora la estacionalidad de la serie de tiempo.

La Figura 3.8 muestra el ajuste del polinomio ozono y_{t-1} ajustado con las bases ortogonales de polinomios de grado 2 que se obtuvieron en el modelo más parsimonioso. Al igual que con el polinomio de tiempo, se consideran bandas de confianza del 99 %.

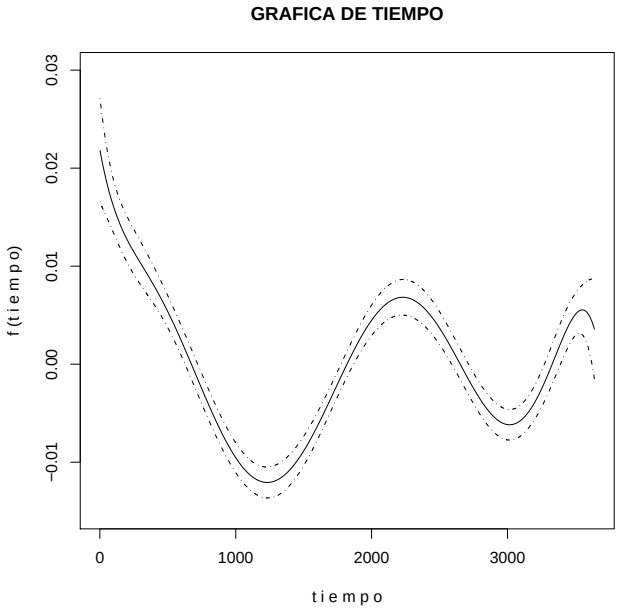


Figura 3.7: Efecto ajustado de tiempo en presencia de variables atmosféricas

3.6. Análisis de Residuales

El análisis de los residuales proporciona excelente información para determinar si las suposiciones sobre el término del error son adecuadas; además, gran parte del análisis residual se justifica en examinar gráficas. La figura 3.9 muestra las gráficas de análisis de residuales, las cuales consisten en: (a) densidad estimada de los residuales, (b) cuantil contra cuantil, donde puede observarse que casi todos los puntos se encuentran situados cerca de la línea recta de 45°, (c) función de autocorrelación, (d) función de autocorrelación parcial y (e) residuales contra tiempo, en la que para cada residual se grafica un punto, la primera coordenada de cada uno de estos puntos está dada por t_i y la segunda coordenada se refiere al valor correspondiente del residual; dicha gráfica ayuda a comprobar la suposición de normalidad puesto que los residuales aparentan la forma de una banda

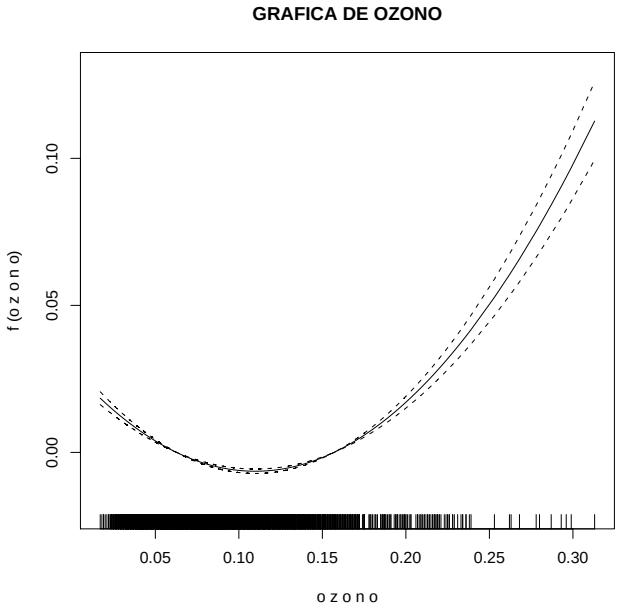


Figura 3.8: Efecto ajustado de la variable ozono

horizontal; por tanto, se concluye que el modelo es admisible. Aquí se muestra también que los residuales estandarizados se distribuyen aproximadamente como una distribución normal como era de esperarse; en general, puede decirse una vez más que el ajuste del modelo propuesto ajusta los datos razonablemente bien, es decir, este es aceptable. El Ápendice D muestra el programa utilizado para realizar estas gráficas en el paquete **R** .

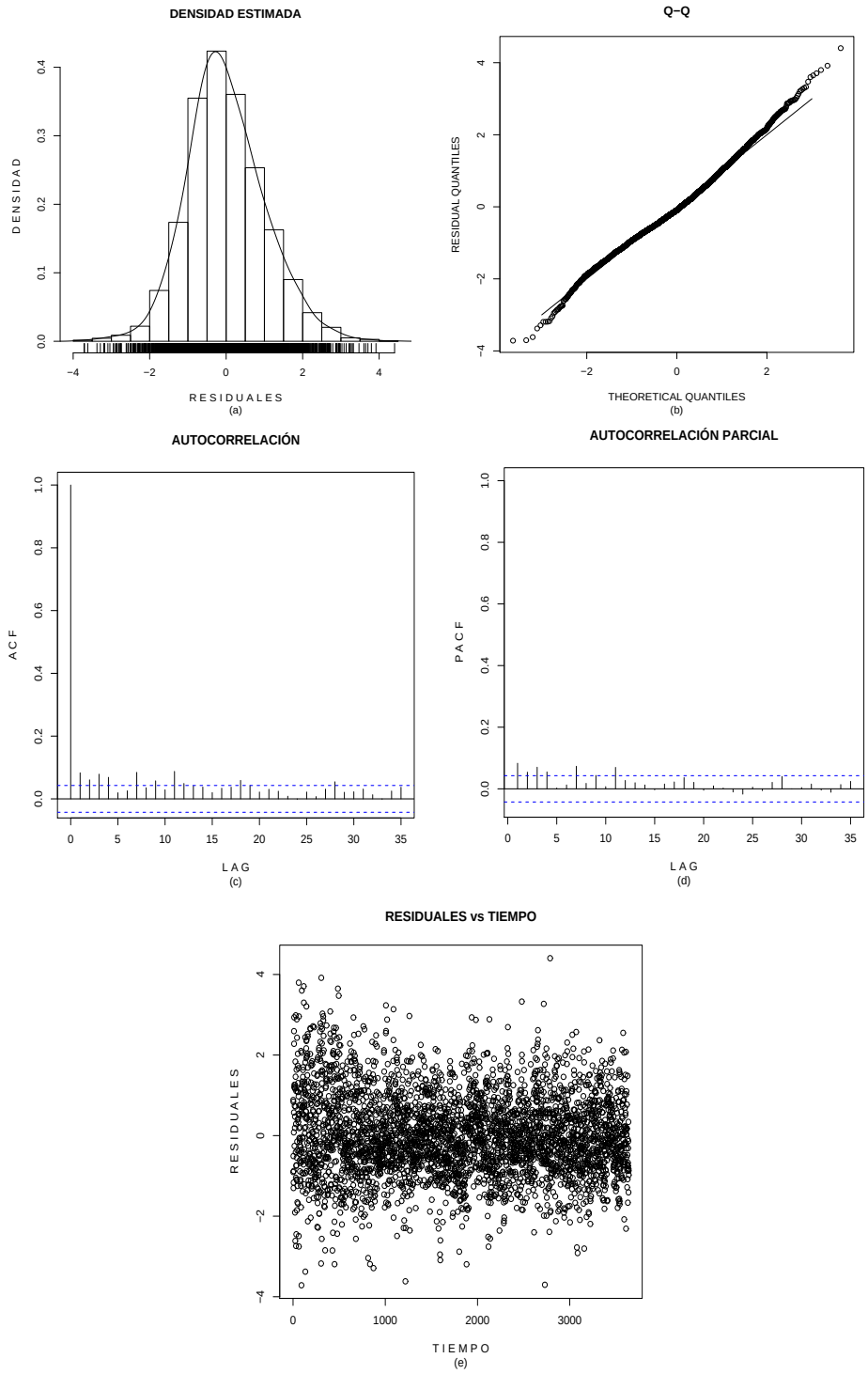


Figura 3.9: Gráficas de diagnósticos de los máximos diarios de ozono.

CONCLUSIONES

La ocurrencia de altos niveles de ozono ocasiona daños a la salud. De ahí es que surge la necesidad de detectar estos niveles para proteger a la población. El tratamiento de eventos extremos es en general complicado dada su baja frecuencia de ocurrencia y los escasos datos con los que se cuenta; sin embargo, esta dificultad no exime de la necesidad de tratar estos valores extremos; como se observó, los primeros cinco años, periodo en que se implementó un programa de mejoramiento ambiental, se redujeron notablemente los altos niveles de contaminación. Sin embargo, en cuanto dejó de aplicarse este programa, los niveles se incrementaron de nuevo, pues es claro que hay un crecimiento en los máximos y que aunque estos no lograron alcanzar las mismas dimensiones que en los primeros años, debe continuarse con este tipo de programas que protegen la salud en contra de contaminantes atmosféricos. Al mismo tiempo, se debe crear conciencia sobre la gravedad del problema de la contaminación. Por lo que, las políticas implementadas para el "Programa para el mejoramiento de la calidad del aire de la Zona Metropolitana de Guadalajara 1997-2001" no son suficientes pues la calidad del aire no es la óptima.

La contaminación troposférica y la constante alteración del clima estan directamente relacionados con una serie de variables atmosféricas. En este estudio,

se comprobó que altas temperaturas influyen para que exista mayor cantidad de ozono. Así mismo, la interacción que hay entre la temperatura y velocidad del viento resultaron significativas; por lo tanto, los resultados indican que el efecto de la temperatura sobre el máximo de ozono depende de la velocidad del viento, con ello se reafirma lo mencionado por Huang y Smith (1999), altas temperaturas junto con bajas velocidades de viento están relacionadas con incrementos en la cantidad de ozono.

Tal como se esperaba, existe estacionalidad, en los máximos diarios de ozono de la Zona Metropolitana de Guadalajara, dado que los efectos semestrales resultaron significativos en el modelo.

También se comprobó que el ajuste del polinomio del modelo más parsimonioso es relativamente bueno. Esto implica que, usando información no estacionaria y atmosférica, se puede, en efecto, proteger a la población de una concentración peligrosa de contaminantes.

En el contexto de los modelos lineales generalizados, Benjamin et. al. (2003) proponen un modelo ARMA(p,q) de manera tal, que los términos autorregresivos y de promedios móviles de la siguiente forma:

$$x_t\beta + \sum_{j=1}^{p} \phi_j \mathcal{A}(y_{t-j}, x_{t-j}, \beta) + \sum_{j=1}^{q} \theta_j \mathcal{M}(y_{t-j}, \tau_{t-j})$$

donde $\mathcal{A}$ y $\mathcal{M}$ son funciones que representan los términos autorregresivos y de promedios móviles respectivamente, $\phi' = (\phi_1, \phi_2, \ldots, \phi_p)$ y $\theta' = (\theta_1, \theta_2, \ldots, \theta_q)$ son los parámetros autorregresivos y de promedios móviles respectivamente. Un revisor de este trabajo, sugiere que éstos términos pueden ser simplemente $\bar{y}_{t-q}$, donde $\bar{y}_{t-q}$ es el promedio de las q observaciones pasadas. De hecho, este modelo ARMA es un caso particular de la especificación sugerida por Benjamin et. al. (2003) y es equivalente a agregar interacciones tomadas de dos en dos entre la respuestas correspondientes a un y más días anteriores, lo cual puede modelarse de manera

fácil con el paquete **R**.

Usando el procedimiento de selección hacia adelante descrito en la sección 2.2.3 de este trabajo, e incluyendo además efectos sexenales, se obtuvo el siguiente modelo:

$$\mathtt{t}^3 + \mathtt{pm}_{t-1} * \mathtt{y}_{t-1} + \mathtt{sexenales} + \mathtt{tem}_t + \mathtt{h}_t + \mathtt{tem}_{t-1} + \mathtt{v}_{t-1} + \mathtt{h}_{t-1} + \mathtt{rh}_{t-1}$$

donde pm_{t-1} denota los promedios móviles de grado 2. Al comparar el Criterio de Información Bayesiano, de los modelos más parsimoniosos, es claro que el menor entre ambos es el que pertenece al modelo que incluye promedios móviles, pues por mucho es el más pequeño, como puede verse en el siguiente cuadro 1.

MODELO	**DEVIANZA**	n_p	**BIC**
original	-16817.45	23	-16628.8560
promedio móvil	-36904.54	16	-36773.3442

Cuadro 1: Comparación de los modelos obtenidos.

Sin embargo, en el cuadro 2, se muestran los estimadores y errores estandar para este modelo, donde todas las variables son significativas a excepción de la interacción entre los promedios móviles y la respuesta y_{t-1}, pues si ésta no se incluye el modelo presenta errores numéricos.

Como en el modelo original, la figura 1 muestra el efecto ajustado de tiempo para el modelo con promedios móviles y puede apreciarse un decaimiento en los niveles de ozono en los primeros mil días aproximadamente, por lo que de nuevo, puede concluirse que durante el periodo en que se implementó el programa de mejoramiento ambiental, los niveles de contaminación se reducen en el mismo periodo del modelo anterior.

Variable	Estimador	Error Estandar
μ	0.084496	0.000035
$\mathtt{y}_{\mathtt{t}-1}$	-2.418270	0.003999
$\mathtt{pm}_{\mathtt{t}-1}$	4.180904	0.004056
$\mathtt{y}_{\mathtt{t}-1} * \mathtt{pm}_{\mathtt{t}-1}$	-0.010779	0.089290
sexenal	-0.000263	0.000064
sexenal	-0.000366	0.000046
$\mathtt{tem}_{\mathtt{t}}$	0.015660	0.004504
$\mathtt{h}_{\mathtt{t}}$	0.020758	0.004813
$\mathtt{tem}_{\mathtt{t}-1}$	0.010134	0.004428
$\mathtt{vel}_{\mathtt{t}-1}$	-0.037370	0.002850
$\mathtt{hum}_{\mathtt{t}-1}$	-0.057397	0.004941
$\mathtt{rhum}_{\mathtt{t}-1}$	0.007980	0.001662
σ	0.001593	0.000002

Cuadro 2: Modelo con promedios móviles.

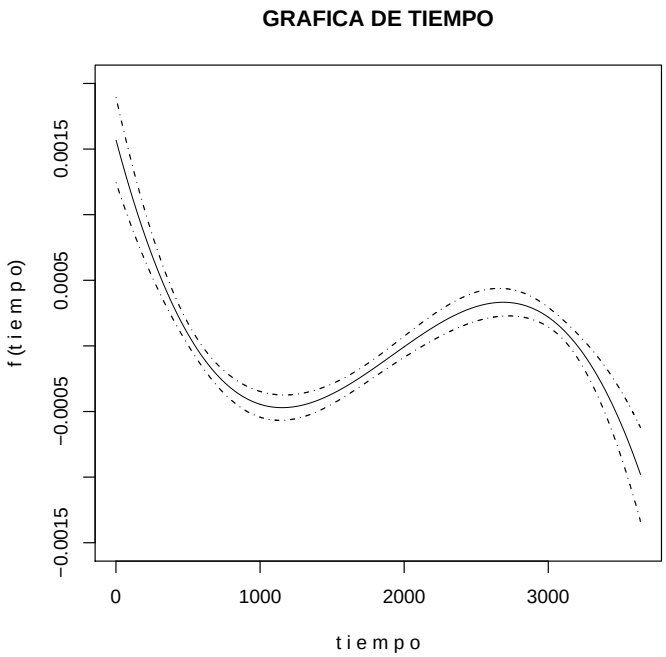


Figura 1: Efecto ajustado de tiempo para el modelo con promedios móviles.

APENDICE A

En el Capítulo 1, se afirma que si en la función de densidad $\mu = 0$ y $\sigma = 1$ y además $z = -\ln z$, entonces,

$$E[Z] = \int_0^\infty (-\ln z) * exp(-z)dz = \lambda$$

Un bosquejo de la demostración es el siguiente

$$\begin{aligned} E[Z] &= \int_0^\infty (-\ln z) * \exp(-z)dz \\ &= -\int_0^\infty (\ln z) * \exp(-z)dz \\ &= -\left[\int_0^1 (\ln z) * \exp(-z)dz + \int_1^\infty (\ln z) * \exp(-z)dz\right] \qquad (1) \\ &= \lambda \end{aligned}$$

La primera integral de (1) se puede escribir como,

$$\int_0^1 (\ln z) * \exp(-z)dz = -\int_0^1 \frac{d}{dz}(\exp(-z) - 1)\ln z dz \qquad (2)$$

Integrando por partes, tenemos que

$$\begin{aligned} u &= \ln z & dv &= \frac{d}{dz}(\exp(-z) - 1)dz \\ du &= \frac{1}{z}dz & v &= (\exp(-z) - 1) \end{aligned}$$

entonces, la integral (2) queda como

$$-\left[(\exp(-z)-1)\ln z\Big|_0^1 - \int_0^1 \frac{\exp(-z)-1}{z}dz\right] = \int_0^1 \frac{\exp(-z)-1}{z}dz \qquad (3)$$

Puesto que

$$\lim_{z\to 1} \exp(-z) - 1 \ln z = 0 \qquad y \qquad \lim_{z\to 0} \exp(-z) - 1 \ln z = 0$$

Para la segunda integral de (1) de nuevo aplicamos integración por partes haciendo

$$\begin{aligned} u &= \ln z & dv &= \exp(-z)dz \\ du &= \frac{1}{z}dz & v &= -\exp(-z) \end{aligned}$$

Entonces

$$\begin{aligned} \int_1^\infty (\ln z) * \exp(-z)dz &= -\exp(-z)\ln z\Big|_1^\infty + \int_1^\infty \frac{\exp(-z)}{z}dz \\ &= \int_1^\infty \frac{\exp(-z)}{z}dz \end{aligned} \qquad (4)$$

Pues

$$\lim_{z\to\infty} -\exp(-z)\ln z = 0 \qquad y \qquad \lim_{z\to 1} \exp(-z)\ln z = 0$$

Por lo tanto, de (3) y (4) concluimos que la ecuacion (1) es igual a

$$\begin{aligned} -\left[\int_0^1 \frac{\exp(-z)-1}{z}dz + \int_1^\infty \frac{\exp(-z)}{z}dz\right] &= \int_0^1 \frac{1-\exp(-z)}{z}dz - \int_1^\infty \frac{\exp(-z)}{z}dz \\ &= \lambda \end{aligned}$$

donde λ se conoce como la constante de Euler-Macheroni. [1] Para ver una demostracion más detallada consultar Boros & Moll (2004).

[1] $\lambda = \lim_{m\to\infty}\left\{\frac{1}{1}+\frac{1}{2}+\ldots+\frac{1}{m}-\ln m\right\} = 0.5772157\ldots$

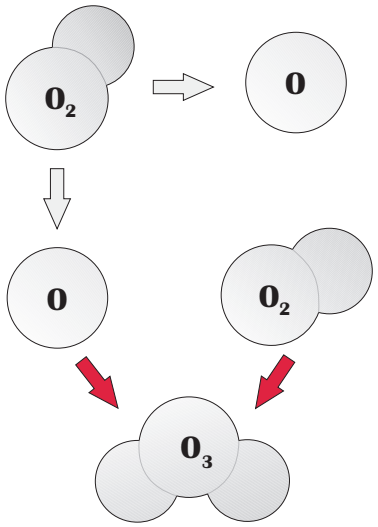
APENDICE B

¿Qué es el ozono?

El oxígeno (O_2) es un elemento químico que tiene la propiedad del estado alotrópico, es decir, por su estructura o número de átomos constituyentes tiene la capacidad de formar moléculas diversas. Así, el ozono (O_3) es una molécula compuesta por tres átomos de Oxígeno, formada al disociarse los 2 átomos que componen dicho gas. Cada átomo (O) liberado se une a otra molécula de Oxígeno (O_2), formando moléculas de Ozono (O_3), con una carga eléctrica negativa, es por esto que resulta ser un gas muy oxidante (Heredia, 1992), ver la siguiente figura.

A temperatura y presión ambientales el ozono es un gas de olor acre e incoloro, en grandes concentraciones puede volverse de color azul pálido. Se encuentra en muy pequeñas proporciones en la atmósfera después de las tempestades y si se respira en grandes cantidades, es tóxico así que puede provocar en el ser humano problemas respiratorios.

La ocurrencia de concentraciones altas de ozono en la parte baja de la tropósfera es considerada como uno de los tópicos más importantes de la química troposférica.



Los mecanismos químicos que controlan la formación del ozono troposférico son complejos y las variables condiciones meteorológicas contribuyen a la dificultad de predecir periodos de altos niveles de ozono con exactitud.

La ***troposfera*** o ***tropósfera*** es la primera capa de la atmósfera. En esta capa es donde generalmente vuelan los aviones, aunque algunos, para evitar los problemas climáticos lo hacen en la capa superior que es la estratósfera, por ser ésta más estable. Llega hasta un límite superior llamado tropopausa (frontera entre ambas) situado a 9 Km. de altura en los polos y a 18 Km. en el ecuado; en ella se producen importantes movimientos verticales y horizontales de las masas de aire (vientos) y hay relativa abundancia de agua. Es la zona más turbulenta de la atmósfera y en ella tienen lugar todos los fenómenos meteorológicos y climáticos: lluvias, vientos, cambios de temperatura, es además la capa de más interés para la ecología. Aquí se hace posible la vida, ya que se concentran la mayoría de los gases de la atmósfera proporcionando las condiciones necesarias para que pueda desarrollarse.

El ozono presente en capas más próximas a la superficie terrestre, como en la ya mencionada troposfera, es peligroso ya que es nocivo para los seres vivos pues

forma parte del denominado ***smog fotoquímico***. Se denomina smog fotoquímico a la contaminación del aire por ozono originado por reacciones fotoquímicas y otros compuestos. Como resultado se observa una atmósfera de un color marrón rojiza. El ozono como ya se mencionó anteriormente, es un compuesto oxidante y tóxico que puede provocar en el ser humano problemas respiratorios.

APENDICE C

```
**********************************************************************
PROGRAMACIÓN PARA CADA MODELO
**********************************************************************
t<-1:3640 ozono<- scan("respuesta.txt") o2 <-
read.table("ozono.txt",header=T) tem <-
read.table("temperatura.txt",header=T) rte <-
read.table("rango_temperatura.txt",header=T) h <-
read.table("humedad.txt",header=T) rh <-
read.table("rango_humedad.txt",header=T) v <-
read.table("velocidad.txt",header=T) rv <-
read.table("rango_velocidad.txt",header=T) wu <-
read.table("u.txt",header=T) wv <- read.table("v.txt",header=T)
**************MODELO NULO*****************************************
fgev(ozono, shape=0)
**********************************************************************
modelo autorregresivo de orden 0
**********************************************************************
datos <- data.frame(ozono,t,o2,tem,rte,h,rh,v,rv,wv,wu) fgev(ozono,
```

```
shape=0, nsloc= data.frame(model.matrix(~poly(t,7)+
(cos(2*pi*t/182.5))+poly(v[,1],1)+poly(tem[,1],1)+
poly(rte[,1],1)+poly(h[,1],1)+poly(rv[,1],1)+
poly(wu[,1],1)+poly(wv[,1],1)-1,data=datos)))
**********************************************************************
modelo autorregresivo de orden 1
**********************************************************************
datos <- data.frame(ozono,t,o2,tem,rte,h,rh,v,rv,wv,wu) fgev(ozono,
shape=0, nsloc= data.frame(model.matrix(~poly(t,7)+
(cos(2*pi*t/182.5))+poly(v[,1],1)+poly(tem[,1],1)+
poly(rte[,1],1)+poly(rv[,1],1)+poly(wu[,1],1)+
poly(wv[,1],1)+poly(o2[,1],1)+poly(tem[,2],1)+
poly(v[,2],1)+poly(h[,2],1)+poly(rv[,2],1)-1,data=datos)))
**********************************************************************
modelo autorregresivo de orden 2
**********************************************************************
datos <- data.frame(ozono,t,o2,tem,rte,h,rh,v,rv,wv,wu) fgev(ozono,
shape=0, nsloc= data.frame(model.matrix(~poly(t,7)+
(cos(2*pi*t/182.5))+poly(v[,1],1)+poly(tem[,1],1)+poly(rte[,1],1)+
poly(rv[,1],1)+poly(wu[,1],1)+poly(wv[,1],1)+poly(o2[,1],1)+
poly(tem[,2],1)+poly(v[,2],1)+poly(h[,2],1)+poly(rv[,2],1)+
poly(o2[,2],1)+poly(wu[,3],1)-1,data=datos)))
**********************************************************************
modelo autorregresivo de orden 3
**********************************************************************
datos <- data.frame(ozono,t,o2,tem,rte,h,rh,v,rv,wv,wu) fgev(ozono,
shape=0, nsloc= data.frame(model.matrix(~poly(t,7)+
(cos(2*pi*t/182.5))+poly(v[,1],1)+poly(tem[,1],1)+poly(rte[,1],1)+
poly(rv[,1],1)+poly(wu[,1],1)+poly(wv[,1],1)+poly(o2[,1],1)+
poly(tem[,2],1)+poly(v[,2],1)+poly(h[,2],1)+poly(rv[,2],1)+
```

```
poly(o2[,2],1)+poly(wu[,3],1)+poly(o2[,3],1)+
poly(v[,4],1)+poly(rv[,4],1)-1,data=datos)))
*********************************************************************
modelo autorregresivo de orden 4
*********************************************************************
datos <- data.frame(ozono,t,o2,tem,rte,h,rh,v,rv,wv,wu) fgev(ozono,
shape=0, nsloc= data.frame(model.matrix(~poly(t,7)+
(cos(2*pi*t/182.5))+poly(v[,1],1)+poly(tem[,1],1)+
poly(rte[,1],1)+poly(rv[,1],1)+poly(wu[,1],1)+
poly(wv[,1],1)+poly(o2[,1],1)+poly(tem[,2],1)+
poly(v[,2],1)+poly(h[,2],1)+poly(rv[,2],1)+
poly(o2[,2],1)+poly(wu[,3],1)+poly(o2[,4],1)+
poly(v[,5],1)+poly(rv[,5],1)+poly(wu[,5],1)-1,data=datos)))
*********************************************************************
modelo autorregresivo de orden 5
*********************************************************************
datos <- data.frame(ozono,t,o2,tem,rte,h,rh,v,rv,wv,wu) fgev(ozono,
shape=0, nsloc= data.frame(model.matrix(~poly(t,7)+
(cos(2*pi*t/182.5))+poly(v[,1],1)+poly(tem[,1],1)+poly(rte[,1],1)+
poly(rv[,1],1)+poly(wu[,1],1)+poly(wv[,1],1)+poly(o2[,1],1)+
poly(tem[,2],1)+poly(v[,2],1)+poly(h[,2],1)+poly(rv[,2],1)+
poly(o2[,2],1)+poly(wu[,3],1)-1,data=datos)))
*********************************************************************
modelo autorregresivo de orden 6
*********************************************************************
datos <- data.frame(ozono,t,o2,tem,rte,h,rh,v,rv,wv,wu) fgev(ozono,
shape=0, nsloc= data.frame(model.matrix(~poly(t,7)+
(cos(2*pi*t/182.5))+poly(v[,1],1)+poly(tem[,1],1)+
poly(rte[,1],1)+poly(rv[,1],1)+poly(wu[,1],1)+
poly(wv[,1],1)+poly(o2[,1],1)+poly(tem[,2],1)+
```

```
poly(v[,2],1)+poly(h[,2],1)+poly(rv[,2],1)+
poly(o2[,2],1)+poly(wu[,3],1)+poly(v[,7],1)+
poly(rv[,7],1)+poly(wu[,7],1)-1,data=datos)))
*********************************************************************
modelo autorregresivo de orden 7
*********************************************************************
datos <- data.frame(ozono,t,o2,tem,rte,h,rh,v,rv,wv,wu) fgev(ozono,
shape=0, nsloc= data.frame(model.matrix(~poly(t,7)+
(cos(2*pi*t/182.5))+poly(v[,1],1)+poly(tem[,1],1)+
poly(rte[,1],1)+poly(rv[,1],1)+poly(wu[,1],1)+
poly(wv[,1],1)+poly(o2[,1],1)+poly(tem[,2],1)+
poly(v[,2],1)+poly(h[,2],1)+poly(rv[,2],1)+
poly(o2[,2],1)+poly(wu[,3],1)+poly(v[,7],1)+
poly(o2[,7],1)-1,data=datos)))
```

APENDICE D

```
*****************************************************************
GRÁFICAS PARA EL MEJOR MODELO MODELO
*****************************************************************
t<-1:3640
ozono<- scan("respuesta.txt") o2 <- read.table("ozono.txt",header=T)
tem <- read.table("temperatura.txt",header=T) rte <-
read.table("rango_temperatura.txt",header=T) h <-
read.table("humedad.txt",header=T) rh <-
read.table("rango_humedad.txt",header=T) v <-
read.table("velocidad.txt",header=T) rv <-
read.table("rango_velocidad.txt",header=T) wu <-
read.table("u.txt",header=T) wv <- read.table("v.txt",header=T)

*****************************************************************
datos < data.frame(ozono,t,o2,tem,rte,h,rh,v,rv,wv,wu)
*****************************************************************
my.best<-fgev(ozono, shape=0, nsloc=
data.frame(model.matrix(~poly(t,7)+(cos(2*pi*t/182.5))+poly(o2[,1],2)+
```

```
poly(o2[,2],1)+poly(v[,1],1)*poly(tem[,1],1)+poly(rte[,1],1)+
poly(rv[,1],1)+poly(wu[,1],1)+poly(wv[,1],1)+poly(tem[,2],1)*poly(v[,2],1)+
poly(h[,2],1)+poly(rv[,2],1)+poly(wu[,3],1)-1,data=datos)))

********************LAS BASES DE TIEMPO CON GRADO 7***************
bases.t <- poly(datos$t, degree=7)

*********LOS ESTIMADORES PUNTUALES DEL POLINOMIO DE TIEMPO*********
point.est <- bases.t %*% my.best$estimate[2:8]

*******LOS VECTORES QUE CONTENDRAN LAS BANDAS DE CONFIANZA*********
lim.sup <- 1:3640 lim.inf <- 1:3640

****************************CARGA PAQUETE************************
library(mvtnorm)
*************************************************************************
********************LA SIMULACION DE MONTECARLO********************
for(i in 1:3640) { bases <- bases.t[i,]
aleatorios <- rmvnorm(n=100000, mean=my.best$estimate[2:8],
         sigma=my.best$var.cov[2:8,2:8])
simulaciones <- aleatorios %*% bases
lim.inf[i] <- quantile(simulaciones, probs=.025) lim.sup[i] <-
quantile(simulaciones, probs=.975)}

***********************LA GRAFICA DE TIEMPO************************
plot(datos$t, point.est, type="l", lwd=.8, ylim=c(-0.015, 0.03),
xlab="t i e m p o", ylab="f (t i e m p o)", main="GRAFICA DE
TIEMPO") lines(datos$t, lim.inf, lty=4) lines(datos$t, lim.sup,
lty=4)
```

```
******************LAS BASES DE OZONO CON GRADO 2*****************
bases.o<- poly(o2[,1],2)

**********LOS ESTIMADORES PUNTUALES DEL POLINOMIO DE OZONO*********
point.est.o <- bases.o %*% my.best$estimate[9:10]

*****LOS VECTORES QUE CONTENDRAN LAS BANDAS DE CONFIANZA*****
 lim.sup.o <- 1:3640 lim.inf.o <- 1:3640

**********************LA SIMULACION DE MONTECARLO*****************
for(i in 1:3640) {bases <- bases.o[i,] aleatorios <-
rmvnorm(n=100000, mean=my.best$estimate[9:10],
         sigma=my.best$var.cov[9:10,9:10])
simulaciones <- aleatorios %*% bases
lim.inf.o[i] <- quantile(simulaciones, probs=.025) lim.sup.o[i] <-
quantile(simulaciones, probs=.975)}

**********************LA GRAFICA DE OZONO*************************
plot(o2[,1][order(o2[,1])]) plot(point.est.o[order(o2[,1])])
plot(o2[,1][order(o2[,1])], point.est.o[order(o2[,1])], type="l",
main="GRAFICA DE OZONO", lwd=.9, ylim=c(-0.02, 0.13), xlab="o z o n
o", ylab="f (o z o n o)") rug(o2[,1]) lines(o2[,1][order(o2[,1])],
lim.inf.o[order(o2[,1])], lty=2) lines(o2[,1][order(o2[,1])],
lim.sup.o[order(o2[,1])], lty=2)

**************************DIAGNOSTICOS*****************************
El componente lineal ajustado es: base.design <- as.matrix(my.best$nsloc)

*******************************************************************
fitted.loc <- my.best$loc + base.design %*% my.best$estimate[2:25]
```

```
fitted.scale <- my.best$estimate[26]

**********DISTRIBUCION AJUSTADA**********
fitted.G <- pgev(ozono, loc= fitted.loc, scale=fitted.scale, shape=0)

**********LOS RESIDUALES ESTANDARIZADOS**********
r.t <- qnorm(fitted.G)

**********
#MI_VECTOR<- ts(r.t,start=c(1997,1),frequency=365) #arima(MI_VECTOR)
arima(r.t) RESULTADO R Call: arima(x = r.t) Coefficients:
      intercept
        -0.0007
s.e.     0.0169 sigma^2 estimated as 1.038:  log likelihood =
-5233.43,  aic = 10470.85

**********
*******GRAFICA LA DENSIDAD ESTIMADA DE LOS RESIDUALES**********
hist(r.t, prob=T, xlab="R E S I D U A L E S", ylab= "D E N S I D A
D", main="DENSIDAD ESTIMADA", sub="(a)") #lines(densidad.smooth)
lines(density(r.t,bw=0.24)) rug(r.t)

**********GRAFICA EL Q-Q PLOT**********
qqnorm(r.t, xlab="THEORETICAL QUANTILES", ylab="RESIDUAL QUANTILES",
       main="Q-Q", sub="(b)")
lines(seq(from=-3, to=3, length=10), seq(from=-3, to=3, length=10))

**********GRAFICA DE LA FUNCION DEAUTOCORRELACION**********
acf(r.t, ci=0.99,ylim=c(-0.02, 1), main="", xlab="lag", sub="(c)")
```

```
**********GRAFICA DE LA FUNCION DE AUTOCORRELACION PARCIAL**********
pacf(r.t, ci=0.99,ylim=c(-0.05, 1), main="",
xlab="Lag", ylab="PACF", sub="(d)")

**********************GRAFICA DE RESID VS TIEMPO******************
plot(1:3640, r.t,
xlab="t i e m p o", ylab="Residuals", sub="(e)")
```

BIBLIOGRAFÍA

[1] Anderson, David R.; Sweeney Dennis J.; Williams Thomas A. 2008. *estadística para Administración y Economía.* Cengage Learning.

[2] Benjamin, Michael A.; Rigby Robert A.; Mikis Stasinopoulos D. 2003. *Generalized Autorregressive Moving Average Models.* Journal of the American Statistical Association, **Vol. 98.** pp. 214-223.

[3] Bloomfield, P.; Royle, A.; Yang, Q. 1996. *Accounting for Meteorological Effects in Measuring Urban Ozone Levels and Trends.* Atmospheric Environment, **30** pp. 3067-3078.

[4] Boros, G.; Moll, V. 2004. *Irresistible Integrals, Symbolics, Analysis and Experiments in the Evaluation of Integrals.* Cambridge.

[5] Coles, S. G. 2001. *An Introduction to Statistical Modeling of Extreme Values.* London: Springer-Verlag.

[6] Cox, W. M. y Chu, S. H. 1992. *Meteorologically AdjustedOzone Trends in Urban Areas: A Probability Approach.* Atmospheric Environment, **27B** pp. 425-434.

[7] DRAPER N. R.; SMITH H. 1981. *Applied Regression Analysis.* Jhon Wiley & Sons.

[8] DUNN, P. K. Y SMYTH, G. K. 1996. *Randomized Quantile Residuals.* Journal of Computational and Graphical Statistics, **5** pp. 236-244.

[9] DUPUIS, D. J. Y TAWN, J. A. 2001. *Effects of Mis-Specification in Bivariate Extreme Value Problems.* Extremes, **4**, pp 315-330.

[10] HEREDIA, M. 1992. *¿Como se forma el ozono?.* Contactos, **5**, pp 101.

[11] HUANG, L.; SMITH, R. 1999. *Meteorologically-dependent Trends in Urban Ozone.* Environmetrics, **10** pp. 103-108.

[12] MONTGOMERY D. C.; PECK E. A. 1982. *Introducction to Linear Regression Analysis.* Jhon Wiley & Sons.

[13] NATIONAL RESEARCH COUNCIL 1991. *Rethinking the Ozone Problems in Urban and Regional Air Pollution.* Washington, DC: National Academic Press.

[14] NIU, X.-F. 1996. *Nonlinear Additive Models for Environmental Time Series, with Applications to Ground-Level Ozone Data Analysis.* Journal of the American Statistical Association, **91** pp. 1310-1321.

[15] PAGNOTTI, V. 1990. *Seasonal Ozone Levels and Control by Seasonal Metereology.* Journal of the air and Waste Management Association, **40**, pp. 206-210.

[16] R DEVELOPMENT CORE TEAM (2008). *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.

[17] RAFTERY A. E. 1995. *Bayesian Model Selection in Social Research.* Sociological Methodology, **25** pp. 111-163.

[18] SMITH, R. L. 1989. *Extreme Value Analysis of Environmental Time Series: An Application to Trend Detection in Ground-Level Ozone.* Statistical Science, **4**, pp. 367-393.

[19] STEPHENSON, A.; GILLELAND, E. (2006). *Software for the Analisys of Extreme Events: The Current State and Future Directions.* Extreme, **8** pp. 87-109.

[20] WHO - WORLD HEALTH ORGANIZATION. 1987. *Air Quality Guidelines for Europe.* WHO Regional Publications, European Series **23**, Copenhagen, Regional Office for Europe.

[21] ZEGER, S. L. Y QAQISH, B. 1988. *Markov Regression Models for Time Series: A Quasi-Likelihood Approach.* Biometrics, **44**, pp. 1019-1032.