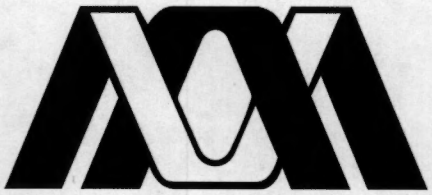
Análisis Bivariado de Extremos para Evaluar los niveles de Ozono Troposférico en la zona Metropolitana de Guadalajara.

Tesis que presenta:
Tania Moreno Zúñiga

Para la obtención del grado:
Maestro en Ciencias Matemáticas

18 febrero 2009



Casa abierta al tiempo

UNIVERSIDAD AUTÓNOMA

METROPOLITANA

IZTAPALAPA

Maestría en Ciencias Aplicadas e Industriales

P R E S E N T A

Mat. Tania Moreno Zúñiga

Asesor de Tesis: Dr. Gabriel Escarela Pérez

Sinodales: Dr. Alberto Castillo Morales

Dra. Blanca Rosa Pérez Salvador

Dr. Eduardo Gutiérrez Peña

Agradezco a mi madre porque sin el amor y apoyo incondicional no habría llegado hasta aquí, ha mi hermano que influyó en gran medida en mi educación, ha mi padre por su confianza.
A mi asesor Dr Gabriel Escarela por brindarme la oportunidad de trabajar con el, a mis sinodales Dr. Alberto Castillo, Dr. Eduardo Gutiérrez, Dr. Blanca Rosa Pérez por sus comentarios y sugerencias en la mejora de esta tesis.

Contenido

Capítulo 1

Introducción

El problema de los contaminantes emitidos en la mayor parte de las grandes ciudades, se encuentra estrechamente relacionado con su esquema de desarrollo urbano, tecnológico e industrial. La exposición a los contaminantes atmosféricos representa un riesgo para la salud de la población, lo que obliga a gobernantes y legisladores a implementar planes para el mejoramiento de la calidad del aire y crear políticas para reducir la emisión de dichos contaminantes.

Para revertir las tendencias de deterioro de la calidad de aire y así proteger la salud de la población que habita la zona metropolitana de Guadalajara, las autoridades responsables implementaron el "Programa para el mejoramiento de la calidad del aire en la zona metropolitana de Guadalajara 1997-2001". Sin embargo, a la fecha no existe un estudio que evalúe los beneficios reales de la implementación de dicho programa, con el cual se pueda diagnosticar la tendencia de los niveles de contaminación de la zona. Los mecanismos químicos que controlan la formación del ozono troposférico son complejos y las volátiles condiciones meteorológicas contribuyen, adicionalmente, a la dificultad de predecir periodos de ozono

alto con exactitud. Es bien sabido que la variación de los niveles de contaminantes corresponde a varias razones; entre las más importantes, se pueden mencionar los cambios estacionales de condiciones meteorológicas y el incremento de las diversas fuentes contaminantes. En particular, las altas temperaturas junto con bajas velocidades de viento están asociadas con observaciones altas de ozono (e.g. Pagnotti, 1990; Bloomfield, *et al.*, 1996; Huang y Smith, 1999).

La motivación principal de este estudio, es evaluar los beneficios reales de la implementación de programas en los últimos diez años para reducir las concentraciones de contaminantes en la zona urbana de Guadalajara. La ciudad de Guadalajara ha experimentado una expansión en la industria y en el comercio desde 1934. Esta expansión ha traído como consecuencia una contaminación atmosférica cuyas concentraciones alcanzaron con frecuencia niveles riesgosos para la salud a mediados de los noventas. En un inventario de la SEMARNAP, se advierte que las principales fuentes de contaminación atmosférica en dicha área son el transporte, el suelo, la industria y los servicios. En ese mismo estudio, se documentó que el sector transporte generó alrededor de 73.5 % del total de las emisiones; los suelos, 21.2 %; los servicios, 4.2 % y finalmente la industria, 1.1 % (ver e.g. Ramírez-Sánchez *et al.*, 2006).

En la actualidad, no se ha aplicado algún método estadístico bivariado para localizar la tendencia de los niveles de ozono; sin embargo, varios procedimientos estadísticos han sido propuestos para detectar la tendencia de los niveles de ozono en forma univariada en el tiempo, sin usar variables explicativas o tomando en consideración variables meteorológicas y de periodicidad.

Las grandes metrópolis cuentan en la actualidad con estaciones de monitoreo, las cuales tienden a estar dispersas en una región y ubicadas a varias altitudes sobre el nivel del mar. La Figura 1.1 muestra las ocho estaciones de la Red Automática de Monitoreo Atmosférico (RAMA) de la zona metropolitana de Guadalajara (Las Águilas, Atemajac, Centro, Loma Dorada, Miravalle, Oblatos, Tlaquepaque y Vallarta), las cuales operan y generan información cada 10 minutos, durante las 24 horas del día, los 365 días del año; registrando ozono O_3, humedad relativa `RH`, temperatura `TMP`, dirección del viento `WDR` y velocidad del viento `WSP`. Este sistema permite obtener información de forma oportuna y en un formato accesible para procesarse en programas de cómputo.

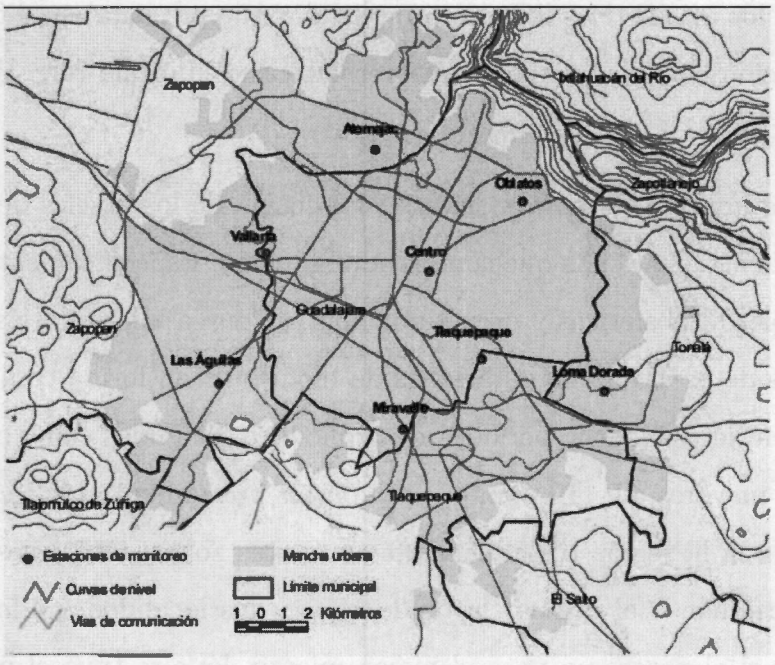


Fig. 1.1. Red Automática de Monitoreo Atmosférico (RAMA) de la zona metropolitana de Guadalajara.

Una forma efectiva de incrementar el conocimiento del problema para perfeccionar la capacidad de prever la concentración del ozono troposférico y mejorar estrategias de control de emisiones, se obtiene al aplicar una metodología que pueda tomar en consideración la relación que tiene el ozono con los procesos meteorológicos, la cual debe permitir formulaciones que puedan separar los efectos de las tendencias causadas por emisiones precursoras y los efectos debidos a la variabilidad meteorológica (ver e.g. National Research Council, 1991). En los Estados Unidos el National Research Council ha dividido esfuerzos para cuantificar el impacto meteorológico sobre el ozono, basándose en métodos de clasificación y métodos de regresión. Algunos de sus modelos tuvieron la ventaja de flexibilidad no paramétrica, pero son, en esencia, difíciles de interpretar. Otras especificaciones fueron modelos estadísticos lineales complicados que presentaban dificultad al capturar la relación entre las variables meteorológicas y el ozono (e.g. Bloomfield *et al.*, 1996).

Una hipótesis importante para estimar la tendencia de los niveles de ozono es que éstos varían entre las zonas que comprende la región. Específicamente, en presencia de variables atmosféricas y de periodicidad, se busca inferir varios aspectos sobre los niveles de ozono de las estaciones de monitoreo en forma conjunta. Por ejemplo, se desea determinar si periodos de ozono alto, definidos como concentraciones de ozono mayores a 0.110 ppm (1 ppm partes por millon= 1962 $\mu g/m^3$), se encuentran en toda la región o simplemente en ciertas zonas. Otra cuestión relacionada es determinar si al evaluar la tendencia en dos localidades diferentes, es posible verificar si el programa para el mejoramiento del aire tuvo el mismo impacto.

Han existido varios intentos para evaluar la tendencia. Bloomfield *et al.* (1996) crearon un modelo de regresión no lineal para los niveles de ozono en el área metropolitana de Chicago de 1981 a 1991, basándose en las condiciones meteorológicas, estacionalidad y tendencia anual; el modelo mostró que la adición de estas variables redujo el error predicho en aproximadamente la mitad y su error estándar alrededor del 30 %. Bloomfield *et al.* eligieron la relación de concentraciones de ozono con una combinación no lineal de variables meteorológicas, y especificaron el modelo semejante que producía errores casi normales. Cox & Chu (1992) también crearon un modelo predictivo para determinar la tendencia del ozono pero usando un modelo lineal generalizado, asumiendo la condición de una distribución Weibull para las concentraciones de ozono; los efectos meteorológicos están descritos en términos de combinaciones lineales de las variables meteorológicas con interacción entre los términos de temperatura y velocidad del viento. Aunque el ozono ambiental parece seguir una distribución Weibull, en dicho estudio no hay evidencia suficiente de que ésta sea la especificación correcta.

En particular, suponer una familia paramétrica con 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 linealidad en los efectos del tiempo (e.g. Bloomfield *et al.*, 1996), lo cual puede ser muy cuestionable pues -intuitivamente- se puede discernir que los niveles de ozono siguen un comportamiento curvilíneo.

El propósito del presente estudio es el de analizar los máximos locales semanales de ozono de dos estaciones de monitoreo ambiental ubicadas en el oriente y

poniente de la ciudad, en el período 1997-2006, usando un modelo con la capacidad de evaluar los efectos de la tendencia en presencia de variables periódicas y atmosféricas. La justificación de usar los máximos locales de dos estaciones, en vez de encontrar el máximo global en la zona, se basa en la hipótesis de que los niveles de ozono y las tendencias pueden variar dependiendo de la localidad; un estudio conjunto de las dos localidades provee un análisis más informativo. Como la principal meta de un estudio de datos bivariados de ozono es modelar sus extremos al estimar el comportamiento de la cola superior, una elección adecuada es usar una estructura paramétrica que esté compuesta por distribuciones de valor extremo que permitan localizar la tendencia mientras se mitigan los efectos meteorológicos y de periodicidad a través de formas funcionales convenientes (WHO, 1987).

La Figura 1.2 y Figura 1.3 muestran los máximos semanales locales de ozono para las estaciones Vallarta y Tlaquepaque respectivamente medidos en partes por millón. En las gráficas se sobrepone una línea horizontal que indica el límite de 0.110 ppm de ozono, el cual es el máximo diario permitido por la Organización Mundial de la Salud. Es posible observar en las gráficas que los máximos tienen una tendencia a la baja para las primeras 150 semanas y el comportamiento es más irregular para las semanas subsecuentes; lo cual coincide con la introducción del programa para el mejoramiento de la calidad del aire. Por lo tanto, se busca probar la hipótesis de que los niveles de ozono son satisfactorios y obedecen a una tendencia estable a la baja.

En este estudio, la variable respuesta es el máximo semanal de ozono en ambas estaciones de monitoreo, y las variables explicativas consisten en medidas apropia-

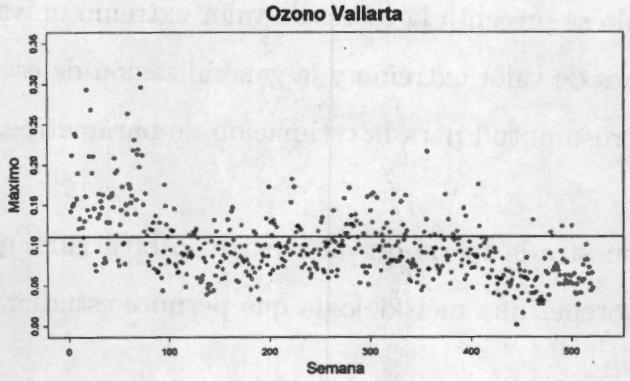


Fig. 1.2. Máximos semanales de concentraciones de ozono medidos en partes por millón del 1 de enero de 1997 al 31 de diciembre de 2006 en la estación Vallarta.

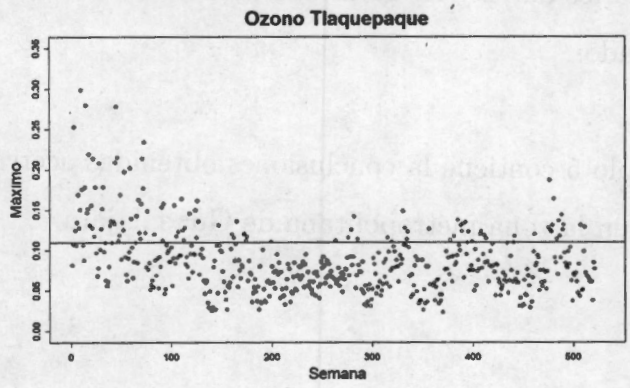


Fig. 1.3. Máximos semanales de concentraciones de ozono medidos en partes por millón del 1 de enero de 1997 al 31 de diciembre de 2006 en la estación Tlaquepaque.

das de velocidad del viento, dirección del viento, temperatura, humedad relativa, periodicidad y tendencia en el tiempo, las cuales se describirán en forma precisa en el capítulo 4. La construcción del modelo bivariado se hará a través de la técnica de la cópula. Parte del estudio comprende encontrar una forma paramétrica conveniente con la que se pueda evaluar la dependencia entre las respuestas de ambas estaciones de monitoreo.

En el siguiente capítulo se presenta la teoría de valor extremo univariado, algunos tipos de distribuciones de valor extremo y la generalización de éstas; así como el método de máxima verosimilitud para la estimación de parámetros.

En el capítulo 3 se muestra la teoría de valor extremo bivariado, que utiliza la teoría de cópulas para obtener una metodología que permite estudiar casos extremos bivariados.

El capítulo 4 presenta los datos y los resultados del análisis utilizando la teoría de valor extremo bivariado.

Finalmente, el capítulo 5 contiene la conclusiones obtenidas acerca del análisis de los niveles de ozono en la zona metropolitana de Guadalajara.

Capítulo 2

Teoría de Valor Extremo

En el análisis clásico de datos, los valores inusuales causados por eventos de baja probabilidad de ocurrencia, pero de alto impacto, son llamados *outliers* y, generalmente, son estudiados por separado. Esto es correcto si se buscan estimadores de los casos comunes, por lo que no debe importar si se quitan los valores extremos. Por otro lado, si se busca describir los eventos que no suceden comúnmente, eventos extremos, entonces es incorrecto. Una de las disciplinas estadísticas más importantes para la ciencia aplicada en los últimos 50 años ha sido la teoría de valor extremo.

Las técnicas de valor extremo son usadas en varias disciplinas, por ejemplo, para ajustar portafolios en la industria de seguros, para análisis de riesgo en el mercado financiero o para la predicción del tráfico en telecomunicaciones. En particular, el análisis de valor extremo requiere la estimación de la probabilidad de eventos que son más extremos que cualquier otro observado.

Los métodos estadísticos para evaluar eventos extremos necesitan poder modelar adecuadamente la cola de la distribución que se analiza. La Teoría de Valor

Extremo (**TVE**) es la que se encarga de este análisis. La **TVE** no solo genera modelos para la muestra que se está trabajando, también puede ser utilizada para extrapolar la probabilidad de un evento aún más extremo (ver e.g. Coles, 2001).

2.1. Formulación del Modelo

En esta sección se desarrollará el modelo que representa la teoría de valor extremo univariada, la cual permite estudiar la conducta estadística de

$$M_n = \text{máx}\{X_1, ..., X_n\},$$

donde $\{X_1, ..., X_n\}$, es una colección de variables aleatorias independientes (Nota: en caso contrario ver Teorema 2.3.1) que tienen una función de distribución común F. En la práctica, las X_i representan valores de un proceso de mediciones periódicas; por ejemplo: mediciones del nivel del mar, la temperatura diaria, contaminación (ozono o partículas). Así M_n representa el máximo del proceso sobre n unidades de tiempo observadas. Si n es el número de observaciones en un año, entonces M_n corresponde al máximo anual.

En teoría, la distribución de M_n puede ser derivada de la distribución de las n variables X_i, usando la suposición de independencia entre ellas:

$$\begin{aligned} \Pr\{M_n \leq z\} &= \Pr\{X_1 \leq z, ..., X_n \leq z\} \\ &= \Pr\{X_1 \leq z\} \times \cdots \times \Pr\{X_n \leq z\} \\ &= \{F(z)\}^n. \end{aligned} \tag{2.1}$$

Sin embargo, esto no es práctico ya que la función de distribución F es desconocida y pequeñas perturbaciones en el estimador de F pueden crear errores

sustanciales en F^n. Es decir, una posibilidad para obtener la distribución de M_n sería emplear técnicas de estadística estándar para estimar a F a partir de los datos observados y entonces sustituir esta estimación en la ecuación 2.1.

Otra alternativa, es aceptar que F es desconocida y buscar modelos de familias aproximadas para F^n, la cual puede ser estimada con base en los datos extremos. Esto es similar a la práctica usual de aproximar la distribución de la suma de una muestra suponiendo normalidad cuando se justifica con el teorema central del límite. Por tanto, es conveniente observar la conducta de F^n cuando $n \to \infty$.

En este caso, se puede notar que para cualquier z fijo tal que $F(z) < 1$ se tiene que $F^n(z) \to 0$ cuando $n \to \infty$. Esto significa que asintóticamente M_n crece indefinidamente. Para estudiar el comportamiento se utiliza una renormalización lineal de la variable M_n de la siguiente forma:

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

para sucesiones de constantes $a_n > 0$ y b_n. Elecciones apropiadas de a_n y b_n representarán los parámetros de localización y escala de M_n^*, evitando así las dificultades que surgiesen con la variable M_n (ver Coles, 2001).

2.1.1. Tipos de Extremales

El rango entero de posibles distribuciones límite para M_n^* está dado por el siguiente teorema.

Teorema 2.1.1 *Si existen sucesiones de constantes $a_n > 0$ y b_n tales que*

$$\Pr\{(M_n - b_n)/a_n \leq z\} \to G(z)$$

cuando $n \to \infty$, *donde* G *es una función de distribución no degenerada, entonces* G *pertenece a una de las siguientes familias:*

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

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

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

para parámetros $a > 0$, $b \in \mathbb{R}$ *y, en el caso de las familias II y III,* $\alpha > 0$.

En palabras, el Teorema 2.1.1 establece que el máximo de la muestra de una variable reescalada $(M_n - b_n)/a_n$ converge a una distribución de las familias nombradas I, II ó III. Estas tres clases de distribuciones son conocidas como las familias **Gumbel**, **Fréchet** y **Weibull** respectivamente. Cada familia tiene un parámetro de localización y uno de escala, b y a respectivamente; adicionalmente, las familias Fréchet y Weibull tienen un parámetro de forma α.

Lo importante de este resultado es que los tres tipos de distribución de valor extremo son los únicos posibles límites para la distribución de M_n^*, independientemente de la distribución F de la población.

2.2. Generalización de Distribuciones de Valor Extremo

Los tres tipos de límites que se presentan en el Teorema 2.1.1 tienen distintas conductas, las cuales corresponden a diferentes comportamientos de la cola de la función de distribución F. Esto puede ser hecho precisamente para considerar la conducta de la distribución límite G con $z_+ = \text{máx}\{0, z\}$, el punto más grande. Para la distribución Weibull z_+ es finito, para las distribuciones Fréchet y Gumbel $z_+ = \infty$ (ver Figura 2.1). Sin embargo, la densidad de G decae exponencialmente para la distribución Gumbel y polinomialmente para la distribución Fréchet, correspondiendo relativamente a tasas diferentes de decaimiento en la cola de F. En aplicaciones las tres diferentes familias dan representaciones distintas de la conducta de valor extremo; entonces se puede adoptar alguna de las tres familias y estimar el parámetro relevante para la distribución.

Un mejor análisis es ofrecer una reformulación del modelo en el Teorema 2.1.1. Las familias Gumbel, Fréchet y Weibull pueden ser representadas en una única familia de modelos, teniendo función de distribución de la forma:

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \tag{2.2}$$

donde $z \in \{z : 1+\xi(z-\mu)/\sigma > 0\}$, y los parámetros satisfacen $-\infty < \mu < \infty$, $\sigma > 0$ y $-\infty < \xi < \infty$. Esta es la familia de distribución de **valor extremo generalizada** (VEG). El modelo tiene tres parámetros: uno de localización, μ; uno de escala, σ; y uno de forma, ξ (ver Dupuis, 2005). Los tipos de clases II y III de distribuciones de valor extremo corresponden respectivamente a los casos $\xi > 0$ y $\xi < 0$ en esta parametrización; el subconjunto de familias VEG con $\xi = 0$ es

Tipo I (Gumbel). Sea $a = 1$, $b = 0$

$$g(z) = \exp[-z]\exp\{-\exp[-z]\}, \quad -\infty < z < \infty;$$

Tipo II (Frechet). Sea $a = 1$, $b = 0$, $\alpha = 1$

$$G(z) = \begin{cases} 0, & \text{si } z \leq 0, \\ z^{-2}\exp\{-z^{-1}\}, & \text{si } z > 0; \end{cases}$$

Tipo III (Weibull). Sea $a = 1$, $b = 0$, $\alpha = 1.5$

$$G(z) = \begin{cases} 1.5[-z]^{0.5}\exp\{-[-z]^{1.5}\}, & \text{si } z < 0, \\ 0, & \text{si } z \geq 0, \end{cases}$$

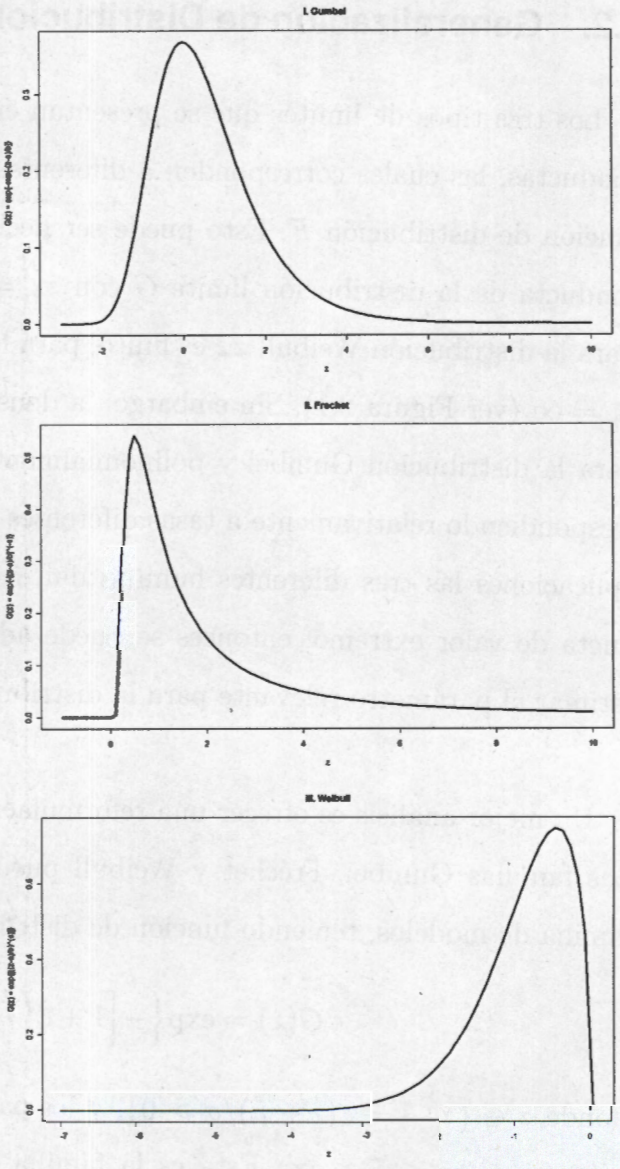


Fig. 2.1. Funciones de densidad I. Gumbel, II. Fréchet y III. Weibull

interpretado como el límite de (2.3) cuando $\xi \to 0$, lo cual nos lleva a la **familia Gumbel** con función de distribución:

$$G(z) = \exp\Big[-\exp\Big\{-\Big(\frac{z-\mu}{\sigma}\Big)\Big\}\Big], \quad -\infty < z < \infty \tag{2.3}$$

La unificación de las tres familias originales de distribución de valor extremo dentro de una sola familia de la forma (2.2) simplifica en gran medida la implementación estadística. A través de inferencia sobre ξ, los datos determinan el tipo más apropiado de conducta de cola, y no hay necesidad de hacer juicios subjetivos a priori acerca de cual familia individual de valor extremo adoptar. Además, la incertidumbre en la inferencia del valor de ξ mide la falta de certidumbre sobre cuál de los tres tipos es más apropiado para un conjunto de datos dado.

Por conveniencia, el Teorema 2.1.1 es modificado de la siguiente forma.

Teorema 2.2.1 *Si existen sucesiones de constantes* $\{a_n > 0\}$ *y* $\{b_n\}$ *tales que*

$$\Pr\{(M_n - b_n)/a_n \leq z\} \to G(z) \qquad \textit{cuando} \qquad n \to \infty \tag{2.4}$$

para una función de distribución no degenerada G*, entonces* G *es un miembro de la familia* **VEG**

$$G(z) = \exp\Big\{-\Big[1+\xi\Big(\frac{z-\mu}{\sigma}\Big)\Big]^{-1/\xi}\Big\} \tag{2.5}$$

definida sobre $\{z : 1+\xi(z-\mu)/\sigma > 0\}$*, donde* $-\infty < \mu < \infty$, $\sigma > 0$ *y* $-\infty < \xi < \infty$.

2.3. Distribución VEG

La familia VEG proporciona un modelo para la distribución del máximo, M_n; en la práctica, la elección del tamaño de muestra, es decir, el valor de n para cualquier

conjunto de datos particular, puede ser crítico. La elección del tamaño se realiza tratando de alcanzar un equilibrio entre el sesgo y la varianza.

Denotando cada máximo como $Z_1, \ldots, Z_m$ y con la suposición de que son variables independientes de una distribución VEG, entonces los parámetros pueden ser estimados.

Las Z_i pueden provenir de series de X_i independientes o dependientes; si son independientes es claro que las Z_i lo son también, pero si son dependientes la conclusión de que las Z_i tengan una distribución VEG aún es razonable, si se considera que la condición para la independencia se cumple si las observaciones están lo suficientemente distantes en el tiempo.

Definición 1: Una serie estacionaria $X_1, X_2, \ldots$ se dice que satisface la condición de $D(u_n)$ si, para toda $i_1 < \ldots < i_p < j_1 < \ldots < j_q$ con $j_1 - i_p > l$,

$$\begin{aligned} &|Pr\{X_{i1} \leq u_n, \ldots, X_{ip} \leq u_n, X_{j1} \leq u_n, \ldots, X_{jq} \leq u_n\} \\ &-Pr\{X_{i1} \leq u_n, \ldots, X_{ip} \leq u_n\} Pr\{X_{j1} \leq u_n, \ldots, X_{jq} \leq u_n\}| \leq \alpha(n, l_n) \end{aligned} \tag{2.6}$$

para alguna función $\alpha(n, l_n)$ tal que $\alpha(n, l_n) \longrightarrow 0$ donde l_n es una sucesión de forma que $l_n/n \longrightarrow 0$ cuando $n \longrightarrow \infty$.

Para sucesiones de variables independientes, la diferencia en las probabilidades expresada en (2.6) es exactamente cero para cualquier sucesión u_n. Ahora para una sucesión específica de umbrales u_n que se incrementan con n, la condición $D(u_n)$

asegura que para conjuntos de variables que están lo suficientemente separados en el tiempo, la diferencia de probabilidades expresada en (2.6) puede no ser cero, pero está lo suficientemente cerca de cero para no tener efecto sobre las leyes de los límites para extremos. Esto se formaliza en el siguiente teorema.

Teorema 2.3.1 *Sea* $X_1, X_2, \ldots$ *un proceso estacionario y sea* $M_n = \text{máx}\{X_1, X_2, \ldots, X_n\}$. *Si* $\{a_n > 0\}$ *y* $\{b_n\}$ *son sucesiones de constantes tales que*

$$\Pr\{(M_n - b_n)/a_n \leq z\} \longrightarrow G(z)$$
$$\Pr\{M_n < a_n z + b_n\} \longrightarrow G(z)$$

donde G *es una función de distribución no degenerada y la condición* $D(u_n)$ *se satisface con* $u_n = a_n z + b_n$ *para cualquier* z *real, entonces* G *es un miembro de la familia de distribuciones de valor extremo generalizada.*

2.3.1. Estimación de los Parámetros de la Distribución VEG por el Método de Máxima Verosimilitud

Se sabe que M_n^* se distribuye asintóticamente de acuerdo a una distribución VEG; sin embargo esta información no identifica completamente la función de distribución pues se desconoce los parámetros de localización, escala y forma, por lo que se deben estimar.

El estimador de un parámetro es una función de los datos que garantiza una aproximación satisfactoria del parámetro poblacional desconocido, siempre que cumplan ciertas propiedades tales como insesgamiento o máxima simetría; varianza

mínima o máxima concentración de los datos alrededor del parámetro estimado, consistencia y suficiencia.

Se han propuesto diversas técnicas para la estimación de parámetros en los modelos de valor extremo, como el método de momentos y el método de máxima verosimilitud.

En este trabajo se considera únicamente los estimadores obtenidos mediante el método de máxima verosimilitud.

El método de máxima verosimilitud permite, en el caso de un parámetro o un vector de parámetros poblacionales desconocidos, determinar el estimador o vector de estimadores que maximizan la función de probabilidad conjunta de una muestra de n variables aleatorias seleccionadas de la población de estudio.

Sea $f(x;\theta)$ la función de densidad o de probabilidad de una población en la cual se busca determinar $\theta \in \Theta$. Sea $x = (x_1, x_2, \ldots, x_n)'$ una muestra de variables aleatorias independientes, idénticamente distribuidas seleccionadas de dicha población; a la función de probabilidad conjunta $L(\theta; x)$ de las n variables aleatorias de la muestra, vista como función de θ, se le llama función de verosimilitud, es decir:

$$L(\theta; x) = f(x_1, x_2, \ldots, x_n; \theta).$$

Pero como las variables aleatorias son independientes se tiene que:

$$L(\theta; x) = f_1(x_1, \theta) f_2(x_2, \theta), \ldots, f_n(x_n, \theta)$$

Es decir: $L(\theta; x) = \prod_{i=1}^{n} f_i(x_i; \theta)$.

El estimador de máxima verosimilitud (EMV) del parámetro θ es $\widehat{\theta}$ siempre que $L(\widehat{\theta}; x)$ sea el valor máximo de la función de verosimilitud L sobre el espacio parametral Θ, es decir: $\widehat{\theta}$ es el EMV de θ si y solo si $L(\widehat{\theta}; x)$ es máximo. En la expresión $L(\theta; x) = \prod_{i=1}^{n} f_i(x_i; \theta)$ la función de verosimilitud varía con el parámetro θ y para el proceso de optimización se considera que las x_i son constantes luego de haber determinado la muestra.

Como la función logaritmo natural es siempre creciente, el EMV de $L(\theta; x)$ también optimiza a $\log(L(\theta; x))$ y es posible definir:

$$l(\theta; x) = \log(L(\theta; x)) = \log(\prod_{i=1}^{n} f_i(x_i, \theta)) = \Sigma_{i=1}^{n} \log f_i(x_i, \theta) \tag{2.7}$$

y maximizar la función log-verosimilitud $l(\theta, x)$. Entonces el EMV $\widehat{\theta}$ es tal que:

$$\log L(\widehat{\theta}, x) \geq \log L(\theta; x) \qquad \forall \theta$$

Bajo la suposición de que $Z_1, \ldots, Z_m$ son variables independientes que tienen la distribución VEG, la función de verosimilitud entonces está dada por:

$$L(\overrightarrow{\theta}; z) = \prod_{i=1}^{n} g(z_i; \overrightarrow{\theta}) \tag{2.8}$$

Donde $\overrightarrow{\theta} = (\mu, \sigma, \xi)$ y $g(z; \theta) = \frac{dG(z;\theta)}{dz}$.

$$g(z; \theta) = \exp\Big\{-\Big[1 + \xi\Big(\frac{z-\mu}{\sigma}\Big)\Big]^{-1/\xi}\Big\}\Big(1 + \xi\Big(\frac{z-\mu}{\sigma}\Big)\Big)^{-1/\xi-1}\Big(\frac{1}{\sigma}\Big) \tag{2.9}$$

Entonces la log-verosimilitud para la distribución VEG para parámetros cuando $\xi \neq 0$ es:

$$l(\mu,\sigma,\xi) = -m\log\sigma - (1+1/\xi)\sum_{\iota=1}^{m}\log\Big[1+\xi\Big(\frac{z_i-\mu}{\sigma}\Big)\Big] - \sum_{\iota=1}^{m}\Big[1+\xi\Big(\frac{z_i-\mu}{\sigma}\Big)\Big]^{-1/\xi} \tag{2.10}$$

y como consecuencia

$$1+\xi\Big(\frac{z_\iota-\mu}{\sigma}\Big) > 0, \iota = 1,\ldots,m. \tag{2.11}$$

Ahora, una combinación de parámetros donde la ecuación (2.11) es violada, corresponde a una configuración para la cual al menos uno de los datos observados cae más allá del punto final de la distribución, y entonces la verosimilitud es cero y la log-verosimilitud es igual a $-\infty$.

El caso $\xi = 0$ requiere un tratamiento separado, usando el límite de la distribución VEG que nos lleva al caso Gumbel. Utilizando log-verosimilitud:

$$l(\mu,\sigma) = -m\log\sigma - \sum_{\iota=1}^{m}\Big(\frac{z_i-\mu}{\sigma}\Big) - \sum_{\iota=1}^{m}\exp\Big\{-\Big(\frac{z_i-\mu}{\sigma}\Big)\Big\} \tag{2.12}$$

La maximización de las ecuaciones (2.10) y (2.12) con respecto al vector de parámetros (μ,σ,ξ) nos lleva a un estimador de máxima verosimilitud con respecto a toda la familia VEG. No hay solución analítica, pero para un conjunto de datos dado la maximización se sigue directamente usando algoritmos de optimización numérica estándar. Algunas precauciones son necesarias para asegurar que tales algoritmos no produzcan combinaciones de parámetros que violan la ecuación (2.11), y que las dificultades numéricas que aparecen de la evaluación de la ecuación (2.10) en la vecindad de $\xi = 0$ sean evitadas. Este último problema se resuelve usando la

ecuación (2.12) en lugar de la ecuación (2.10) para valores de ξ que caen dentro de una pequeña vecindad alrededor de cero (ver Coles, 2001).

Una dificultad potencial en el uso de métodos de verosimilitud para la distribución VEG, se refiere a la validez de las condiciones de regularidad que son requeridas para las propiedades asintóticas usuales asociadas con el estimador de máxima verosimilitud. Las condiciones no son satisfechas por el modelo VEG debido a que los puntos finales de la distribución son funciones del valor del parámetro: $\mu - \sigma/\xi$ es el mayor punto final de la distribución cuando $\xi < 0$, y es el menor punto final cuando $\xi > 0$. Esta violación de la condición de regularidad significa que los resultados asintóticos estándar no se pueden aplicar automáticamente. Smith (1985) estudia este problema a detalle y obtiene los siguientes resultados:

- Cuando $\xi > -0.5$ los estimadores de máxima verosimilitud son regulares, en el sentido de que tienen las usuales propiedades asintóticas de varianza mínima e insesgamiento.
- Cuando $-1 < \xi < -0.5$ los estimadores de máxima verosimilitud se pueden obtener, pero no tienen las propiedades asintóticas estándar.
- Cuando $\xi < -1$ no es posible obtener los estimadores de máxima verosimilitud.

El caso $\xi \leq -0.5$ corresponde a distribuciones con un pequeño salto en la cola superior, pero esta situación es raramente encontrada en aplicaciones de modelado de valores extremos.

Definición 2. La primera derivada de la función log-verosimilitud se denomina *función score de Fisher* y está dada por:

$$\mathrm{u}(\theta) = \frac{\partial \log L(\theta; x)}{\partial \theta} \tag{2.13}$$

El score es un vector de primeras derivadas parciales, una para cada elemento de θ.

Definición 3. El score es un vector de variables aleatorias con algunas propiedades estadísticas interesantes. En particular, el score evaluado en el valor verdadero de θ tiene *media* cero

$$\mathrm{E}[\mathrm{u}(\theta)] = 0$$

y su matriz de *varianza-covarianza* está dada por la matriz de información:

$$\mathrm{var}[\mathrm{u}(\theta)] = \mathrm{E}[\mathrm{u}(\theta)\mathrm{u}'(\theta)] = \mathrm{I}(\theta) \tag{2.14}$$

Bajo condiciones de regularidad, la matriz información puede ser obtenida como:

$$\mathrm{I}(\theta) = -\mathrm{E}\left[\frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta'}\right] \tag{2.15}$$

$\mathrm{I}(\theta)$ indica el grado de curvatura de la función de log-verosimilitud en el punto θ.

2.3.2. Newton-Raphson & Fisher Scoring

El cálculo del EMV a menudo requiere de procedimientos iterativos. Considerando el desarrollo de la función score evaluada en el EMV $\widehat{\theta}$ alrededor de un cierto

valor de prueba θ usando una aproximación de Taylor de primer orden, se tiene que:

$$\mathrm{u}(\widehat{\theta}) \approx \mathrm{u}(\theta_0) + \frac{\partial \mathrm{u}(\theta)}{\partial \theta}(\widehat{\theta} - \theta_0) \tag{2.16}$$

Sea H el Hessiano ó matriz de segundas derivadas de la función log-verosimilitud:

$$\mathrm{H}(\theta) = \frac{\partial^2 \log L}{\partial \theta \partial \theta'} = \frac{\partial \mathrm{u}(\theta)}{\partial \theta} \tag{2.17}$$

Igualando a cero la parte izquierda de la ecuación (2.16) y resolviendo para $\widehat{\theta}$ se obtiene la aproximación de primer orden:

$$\widehat{\theta} = \theta_0 - \mathrm{H}^{-1}(\theta_0)\mathrm{u}(\theta_0) \tag{2.18}$$

Este resultado da lugar a la llamada técnica de *Newton-Raphson*. Dado un valor de prueba, de la ecuación 2.18 se obtiene un estimador y se repite el proceso hasta que las diferencias entre las sucesivas estimaciones sean cercanas a cero o hasta que los elementos del vector de la primera derivada sean lo suficientemente cercanos a cero. Este procedimiento converge rápidamente si la log-verosimilitud tiene una conducta buena en una vecindad del máximo y si el valor de prueba con el que comienza el método es cercano al EMV.

Un procedimiento alternativo propuesto por Fisher es sustituir "Menos el Hessiano" por su valor esperado, que es la matriz de información. El procedimiento da como resultado:

$$\widehat{\theta} = \theta_0 + \mathrm{I}^{-1}(\theta_0)\mathrm{u}(\theta_0). \tag{2.19}$$

(2.19) es conocida como el *Fisher Scoring*.

2.3.3. Prueba de Hipótesis Sobre el Parámetro

Un vez que obtenemos un estimador del vector de parámetros, es posible realizar una prueba de hipótesis sobre él; a continuación se presentan tres tipos de pruebas de hipótesis.

Prueba de Wald

Bajo condiciones de regularidad, el estimador de máxima verosimilitud $\widehat{\theta}$ tiene, para muestras grandes, aproximadamente una distribución normal multivariada con media igual al verdadero valor del parámetro y matriz de varianza-covarianza dada como la inversa de la matriz de información, así que:

$$\widehat{\theta} \sim NMV(\theta, \mathrm{I}^{-1}(\theta)) \tag{2.20}$$

Las condiciones de regularidad son tales que: el verdadero valor del parámetro θ debe estar dentro del espacio de parámetros, la función log-verosimilitud debe ser de clase C^3 (tridiferenciable), y la tercera derivada acotada. Este resultado provee una base para la construcción de pruebas de hipótesis y regiones de confianza. Bajo la hipótesis:

$$H_0 : \theta = \theta_0$$

suponiendo que H_0 es cierta, se tiene que

$$W = (\widehat{\theta} - \theta_0)' \mathrm{var}^{-1}(\widehat{\theta})(\widehat{\theta} - \theta_0) \tag{2.21}$$

para muestras grandes se distribuye aproximadamente de acuerdo a la distribución Ji-cuadrada con p grados de libertad.

Prueba de Score

Bajo las condiciones de regularidad el score en sí mismo tiene una distribución normal asintótica con media 0 y matriz de varianza-covarianza igual a la matriz de información, de forma que:

$$\mathrm{u}(\theta) \sim N_p(0, \mathrm{I}(\theta)). \tag{2.22}$$

Este resultado brinda otra base para la construcción de pruebas de hipótesis y regiones de confianza. Por ejemplo, bajo:

$$H_0 : \theta = \theta_0$$

la forma cuadrática:

$$Q = \mathrm{u}(\theta_0)'\mathrm{I}^{-1}(\theta_0)\mathrm{u}(\theta_0) \tag{2.23}$$

tiene aproximadamente, para muestras grandes, una distribución Ji-cuadrada con p grados de libertad. Esta matriz de información puede ser evaluada en el valor θ_0 de la hipótesis o en el EMV $\widehat{\theta}$. Bajo H_0, ambas versiones de la prueba son válidas; ya que son asintóticamente equivalentes. Una ventaja en el uso de θ_0 es que puede evitarse el cálculo del EMV.

Prueba de Razón de Verosimilitud

El tercer tipo de prueba está basado en una comparación de máxima verosimilitud para modelos anidados. Suponiendo que se tienen 2 modelos definidos por $\omega_1 \subseteq \Theta$ y $\omega_2 \subseteq \Theta$ con $\omega_1 \subset \omega_2$, es decir, ω_1 es un subconjunto o puede ser considerado un caso especial de ω_2. En este caso, dada una parametrización adecuada,

se puede obtener el modelo más simple igualando a cero algunas componentes del vector de parámetros si se quiere probar la hipótesis de que en realidad son cero.

La idea básica de esta prueba es comparar la máxima verosimilitud en ω_1 y en ω_2. La máxima verosimilitud bajo el modelo mas pequeño ω_1 es:

$$\max_{\theta \in \omega_1} L(\theta; x) = L(\widehat{\theta}_{\omega_1}; x) \tag{2.24}$$

donde $\widehat{\theta}_{\omega_1}$ denota el EMV de θ bajo el modelo ω_1. La máxima verosimilitud bajo el modelo grande ω_2 tiene la misma forma:

$$\max_{\theta \in \omega_2} L(\theta; x) = L(\widehat{\theta}_{\omega_2}; x) \tag{2.25}$$

donde $\widehat{\theta}_{\omega_2}$ denota el EMV de θ bajo el modelo ω_2. La razón de estas 2 cantidades:

$$\lambda = \frac{L(\widehat{\theta}_{\omega_1}; x)}{L(\widehat{\theta}_{\omega_2}; x)}, \tag{2.26}$$

está acotada entre cero y uno; la región de rechazo es $\lambda < \lambda_0$. Valores cercanos a cero indican que la conjetura $\theta \in \omega_1$ no es aceptable ya que es poco probable tener datos muestrales como los obtenidos con este supuesto. Valores cercanos a uno indican que la conjetura $\theta \in \omega_1$ es aceptable, pues con este supuesto había mucha probabilidad de obtener la muestra que se observó, haciendo que los datos sean más probables. Bajo ciertas condiciones de regularidad, menos dos veces el logaritmo de la razón de verosimilitud tiene una distribución Ji-cuadrada (para muestras grandes) con grados de libertad igual a la diferencia del número de parámetros entre los dos modelos. Entonces,

$$-2\log\lambda = 2\log L(\widehat{\theta}_{\omega_2}; x) - 2L(\widehat{\theta}_{\omega_1}; x) \to \mathrm{W} \sim \chi^2_\nu \tag{2.27}$$

donde los grados de libertad son $\nu = dim(\omega_2) - dim(\omega_1)$, el número de parámetros en el modelo grande ω_2 menos el número de parámetros en el modelo pequeño ω_1.

Sujetos a las limitaciones discutidas en la sección 2.3.1, la distribución aproximada de $(\widehat{\mu}, \widehat{\sigma}, \widehat{\xi})$ es normal multivariada con medias (μ, σ, ξ) y matriz de varianza-covarianza igual a la inversa de la matriz de información evaluada en el estimador de máxima verosimilitud. Esta matriz puede ser evaluada analíticamente y es posible usar técnicas de diferenciación numérica para evaluar las segundas derivadas, y rutinas numéricas estándar para calcular la inversión. Los intervalos de confianza y otras formas de inferencia se siguen inmediatamente de la distribución normal aproximada del estimador.

Añadiendo más variables a un modelo la función de verosimilitud mejorará y si la muestra es grande será difícil distinguir mediante el contraste del cociente de verosimilitud entre una mejora "real" y una aportación trivial. El modelo perfecto no existe, puesto que todos constituyen simplificaciones de la realidad y siempre son preferibles modelos con menos variables, puesto que además de ser más sencillos, son más estables y menos sometidos a sesgo. Por ello se han propuesto otras medidas de contraste entre modelos que penalizan en alguna medida que éstos tengan muchos parámetros.

La devianza del modelo se calcula como $-2 \times \log L$ y es una medida del grado de diferencia entre las frecuencias observadas y predichas por el modelo de la variable dependiente, de forma que a mayor devianza, peor es el modelo. La devianza nos puede orientar durante la etapa de selección del modelo final. Idealmente, el

modelo final debería tener la menor devianza de los modelos analizados.

Otras medidas para comparación entre modelos son el *Akaike information criterion* (AIC) y el *Bayesian information criterion* (BIC), definidos por:

$$AIC = -2\log L + 2n_p$$
$$BIC = -2\log L + n_p \log n$$

donde n es el número total de observaciones de la muestra y n_p es el número total de parámetros estimados. En cualquier caso se busca el modelo que minimice alguna de estas medidas (ver e.g. Smith, 2003).

Si comparamos el AIC y el BIC vemos que la diferencia básica entre ambos criterios radica en que este último penaliza más los modelos con un número mayor de parámetros estimados, obteniéndose así modelos de orden inferior a los obtenidos a partir del AIC y corrigiendo, por tanto, la tendencia a la sobrestimación observada en este último.

Capítulo 3

Teoría de Valor Extremo Bivariado

Para analizar el impacto de los planes para el mejoramiento de la calidad del aire en la zona metropolitana de Guadalajara, se deben de tomar en cuenta tres aspectos importantes: el primero, es que este problema atañe al análisis de valor extremo debido a que se está hablando de altos niveles de ozono, peligrosos para la salud; el segundo, es hacer mejor uso de la información limitada disponible y, por último, realizar un análisis bivariado, basándose en la hipótesis de que los niveles de ozono y las tendencias pueden variar dependiendo de la localidad. Los métodos estadísticos univariados de valor extremo tradicionales, ignoran información importante sobre la relación entre las variables, ya sean los niveles de ozono en estaciones cercanas o las mediciones en una misma estación. A continuación se presenta la teoría de valor extremo bivariado, donde se hace referencia a la teoría de valor extremo univariado.

3.1. Existencia

Tomando la misma idea del caso univariado, supóngase que (X_{1i}, X_{2i}), para $i = 1, ..., n$, son parejas aleatorias independientes e idénticamente distribuidas cuya función de distribución conjunta es G. Sean $Y_1 = \text{máx}(X_{11}, ..., X_{1n})$ y $Y_2 = \text{máx}(X_{21}, ..., X_{2n})$.

La meta principal de este estudio es modelar la distribución conjunta de (Y_1, Y_2), denotada $F(y_1, y_2)$. De la teoría univariada de valor extremo, se sabe que Y_j tiene asintóticamente la distribución de valor extremo generalizada, para $j = 1, 2$. Bajo ciertas condiciones de regularidad (e.g. Resnick, 1987), la distribución conjunta de los máximos converge a una clase multivariada de distribuciones de valor extremo.

Una forma válida y conveniente de construir $F(y_1, y_2)$ es a través del modelo de cópula (ver e.g. Dupuis, 2005). Las cópulas son un instrumento adecuado para representar las relaciones de dependencia entre distintas variables aleatorias a través de la distribución de probabilidad conjunta. De hecho, la función de distribución conjunta de una serie de variables aleatorias puede expresarse como la función cópula aplicada sobre las distribuciones marginales consideradas individualmente (ver Yan, 2007).

Definición 4: Una cópula, C , es una función de distribución multivariante cuyas leyes marginales se distribuyen uniformemente entre [0,1]. En el caso bivariante, $C(u, v) = \Pr[U \leq u, V \leq v]$ es una función definida en $[0, 1]^2 \rightarrow [0, 1]$ que verifica las siguientes tres propiedades:

1. $C(u, v)$ es una función creciente para cada una de sus componentes.
2. $C(u, 1) = u$ y $C(1, v) = v$.
3. $\forall a_1 \leq a_2$ y $\forall b_1 \leq b_2$, $C(a_1, b_1) + C(a_2, b_2) - C(a_1, b_2) - C(a_2, b_1) \geq 0$.

Teorema 3.1.1 *(*Teorema de Sklar*) Sea una función de distribución bidimensional cuyas marginales son F_X y F_Y. Entonces existe una cópula C tal que $\forall (x, y) \in [-\infty, \infty]^2$, satisface:*

$$F(x, y) = C(F_X(x), F_Y(y)) \tag{3.1}$$

Si las distribuciones marginales son continuas, la cópula es única.

Por tanto, a partir de las cópulas, es posible crear distribuciones bivariantes con distribuciones marginales definidas. De esta forma, si C es una cópula y F_X y F_Y son dos distribuciones marginales, $C(F_X(x), F_Y(y))$ es una distribución bivariante; donde cada una de las marginales tiene asintóticamente una distribución de valor extremo generalizada (2.2).

Las cópulas Arquimedianas son una familia de cópulas cuyos elementos se generan a partir de una función φ denominada generador de la cópula. Dicha familia contiene las siguientes cópulas: **Cópula de Frank, Cópula de Clayton y Cópula de Gumbel.**

La cópula de Frank es una cópula Arquimediana simétrica, considera la misma dependencia entre todas las observaciones. La cópula de Clayton tiene tendencia a correlacionar los pequeños valores y no los grandes, exhibe mayor dependencia en la cola negativa que en la positiva. La cópula Gumbel (cópula Gumbel-Hougaard)

muestra una asimetría adecuada de los datos, exhibe mayor dependencia en la cola positiva que en la negativa. Por ello es importante mencionar que la familia Gumbel es la única familia arquimediana de cópulas de valor extremo.

3.2. Cópula Gumbel. Cópula Positiva Estable.

En este estudio se empleará la cópula *positiva estable*, debido a que toma en cuenta una asimetría adecuada de los datos respecto a valores extremos y los valores "normales". Dicha cópula está dada por:

$$C_\theta(v_1, v_2) = \exp\left\{-\left[(-\log v_1)^{1/\theta} + (-\log v_2)^{1/\theta}\right]^\theta\right\}, \qquad \theta \in (0,1), \qquad (3.2)$$

y es útil para modelar dependencias positivas. Cuando $\theta \to 0$ se obtiene la cópula superior de Fréchet, mientras que valores de θ cercanos a 1 proveen estructuras de dependencia cercanas a la independencia, i.e. $\lim_{\theta \to 1} C_\theta(u_1, u_2) = u_1\, u_2$.

Entonces para medir la concordancia de u_1 y u_2 (Y_1 y Y_2 respectivamente) es posible usar la τ de Kendall cuyo valor es $\tau_\theta = 1 - \theta$ cuando se usa la cópula positiva estable. Dicha cópula exhibe dependencia en la cola superior, por lo que una forma de cuantificar a la dependencia entre eventos extremos es a través del *coeficiente de dependencia de la cola superior* dado por (ver Joe, 1996):

$$\lambda_u = \lim_{u \to 1^-} \Pr\{Y_2 > F_2^{-1}(u) \mid Y_1 > F_1^{-1}(u)\} = 2 - 2^\theta, \qquad \theta \in (0,1).$$

3.3. Funciones Marginales

Las marginales Y_1 y Y_2 tienen asintóticamente la distribución de valor extremo generalizada; entonces se toman de la familia generalizada de distribuciones de valor extremo:

$$F_j(z_j) = \exp\Big\{-\Big[1+\xi\Big(\frac{z_j-\mu_j}{\sigma_j}\Big)\Big]_+^{-1/\xi_j}\Big\}, \qquad z_j > \mu_j \tag{3.3}$$

La clase de distribuciones dada por la ecuación anterior, como se vio en el capítulo anterior, contiene varias distribuciones importantes útiles para ajustar máximos; la Gumbel, la cual se obtiene cuando $\xi \longrightarrow 0$, la Fréchet, la cual se obtiene cuando $\xi > 0$, y la Weibull, la cual se obtiene cuando $\xi < 0$.

Para tomar en cuenta a las variables explicativas en ambas marginales, uno puede especificar al parámetro de localización de cada marginal de la siguiente forma:

$$\mu_j = \beta_j^T \mathbf{x}_{jt}, \qquad j = 1, 2, \tag{3.4}$$

donde $\mathbf{x}_{jt}$ es un vector de variables explicativas del componente j, el cual incluye a la ordenada y es observado en el tiempo t, y β_j es el vector de coeficientes de regresión correspondientes.

3.4. Función de Verosimilitud

Para obtener los estimadores de los parámetros en el caso bivariado se aplica la teoría vista en la sección 2.4. En este caso, la *función de verosimilitud* está definida como:

$$L = \prod_{i=1}^{n} f(y_{1i}, y_{2i}), \tag{3.5}$$

donde f es la función de densidad correspondiente a F, la cual corresponde a la función cópula cuyas marginales pertenecen a la familia generalizada de valor extremo y cumplen con las propiedades vistas en el capítulo anterior.

Capítulo 4

Desarrollo

La respuesta de interés en el presente estudio, es la pareja cuyas entradas son los máximos semanales de ozono registrados por las dos estaciones de monitoreo ubicadas en el oriente de la ciudad (Vallarta), y en el poniente (Tlaquepaque) de 1997 a 2006.

La justificación de tomar máximos niveles de ozono se debe a que, de acuerdo a la NORMA Oficial Mexicana de la Salud Ambiental (ver e.g. NORMA Oficial Mexicana, 1993), la concentración de ozono no debe rebasar el límite máximo normado de 0.11 ppm aplicable en todo el territorio mexicano; un análisis realizado a los datos diarios de concentraciones de ozono reveló que los niveles de ozono aumentan conforme transcurre la semana hábil, es decir, el lunes comienza con niveles moderados de contaminación aumentando hasta el sábado, mientras que el domingo baja; en otras palabras, se limpia la ciudad. La primera semana del conjunto de datos comprende del lunes 6 al domingo 12 de enero de 1997 y la última semana del lunes 25 al domingo 31 de diciembre de 2006, teniendo un total de 521 máximos semanales de niveles de ozono.

Los mecanismos químicos que controlan la formación del ozono troposférico son complejos y las volátiles condiciones meteorológicas contribuyen adicionalmente a la dificultad de predecir periodos de ozono alto con exactitud. Por esta razón el objetivo principal de este trabajo es evaluar las tendencias de los niveles máximos de ozono en presencia de las variables meteorológicas y de periodicidad.

4.1. Variables Explicativas

Para capturar la dependencia remanente, la cual se debe a la no estacionalidad y a las variables atmosféricas, se incluye la información de las siguientes variables en el componente lineal de los modelos de localización especificados en la ecuación (3.4) para la modelación de los niveles máximos de ozono:

Notación	Nombre
`tiempo`	número de semana
`maxTemp`	temperatura máxima
`rangoTemp`	rango de la temperatura
`minHum`	humedad mínima
`rangoHum`	rango de la humedad
`minVel`	mínimo de velocidad
`rangoVel`	rango de velocidad
`anual`	periodicidad anual
`semestral`	periodicidad semestral
`vientouSabado`	vectores de viento registrados en sábado
`vientovSabado`	

A continuación se explica brevemente en qué consiste cada una de ellas:

- `tiempo`. Es el número de semana que va de 1 a 521 semanas registradas.

- `maxTemp`. Es el promedio de los niveles máximos diarios registrados entre las 12 y las 17 horas; en este intervalo se presentan los máximos niveles de temperatura los cuales propician la presencia de niveles altos de contaminación.
- `rangoTemp`. Es la diferencia entre la temperatura más alta y la más baja de la semana.
- `minHum` Es el promedio de los mínimos niveles de humedad diarios entre las 12 y las 17 horas.
- `rangoHum`. Es la diferencia entre el nivel más alto de humedad y el mínimo nivel de la semana.
- `minVel`, es el promedio de los mínimos diarios de velocidad del viento entre las 12 y las 17 horas.
- `rangoVel`. Es la diferencia entre la velocidad semanal máxima del viento y la mínima.
- `anual`. La periodicidad anual se incluye con la función:

$$\cos\left(\frac{2\pi * \texttt{tiempo}}{52}\right) + \sin\left(\frac{2\pi * \texttt{tiempo}}{52}\right)$$

  donde 52 corresponde al número de semanas en un año.
- `semestral`. La periodicidad semestral se incluye de la misma forma con la función:

$$\cos\left(\frac{2\pi * \texttt{tiempo}}{26}\right) + \sin\left(\frac{2\pi * \texttt{tiempo}}{26}\right)$$

  donde 26 corresponde al número de semanas en un semestre.
- `vientouSabado` y `vientovSabado` son vectores de viento registrados en sábado, que es el día más contaminado, utilizando la siguiente transformación (ver e.g. Huang and Smith, 1999)

$$WIND.U = WSPD \times \sin(2\pi WDIR/360)$$

$$WIND.V = WSPD \times \cos(2\pi WDIR/360)$$

donde $WSPD$ es la velocidad del viento y $WDIR$ es la dirección del viento para cada una de las semanas.

Para la aplicación se utilizó el paquete **evd** del lenguaje R (ver Stephenson, 2006) para minimizar $-2\times\log L$, la *devianza*. Una característica importante del modelado bivariado es que el proceso para encontrar un modelo parsimonioso puede seguir ideas del cociente de verosimilitud, análogo a su uso en el análisis de los modelos lineales generalizados.

4.2. Polinomios Ortogonales

Debido a que la suposición lineal de la no estacionalidad es cuestionable, se incluyó a la variable `tiempo` en términos de bases ortogonales de una regresión polinomial; la misma idea fue implementada para las variables atmosféricas. A continuación se resumirán algunos aspectos importantes de la teoría de polinomios ortogonales (ver Draper & Smith, 1981).

Los polinomios ortogonales son usados para ajustar un modelo polinomial de cualquier orden de una variable. Si se tienen n observaciones (X_i, Y_i) con $i = 1, 2, \ldots, n$ donde X es una variable explicativa y Y es la variable respuesta, el modelo ajustado tiene la forma:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_p X^p + \varepsilon \tag{4.1}$$

Que se puede escribir en forma matricial como:

$$Y = X\beta + \varepsilon \tag{4.2}$$

En general, las columnas de la matriz X pueden no ser ortogonales, es decir, sus vectores columnas pueden ser linealmente dependientes. Para solventar esto se propone construir un polinomio de la forma:

$$\begin{aligned}
\psi_0(X_i) &= 1 \\
\psi_1(X_i) &= X_i + a_{10} \\
\psi_2(X_i) &= X_i^2 + a_{21}X_i + a_{20} \\
&\vdots \\
\psi_r(X_i) &= X_i^r + a_{rr-1}X_i^{r-1} + \ldots + a_{r0} \\
&\vdots
\end{aligned} \tag{4.3}$$

con la propiedad de que sean ortogonales entre ellos, es decir, que sean linealmente independientes y cumplan:

$$\sum_{i=1}^{n} \psi_j(X_i)\psi_l(X_i) = 0 \qquad (j \neq l) \tag{4.4}$$

para todo j, $l < n - 1$. Entonces, reescribiendo el modelo se tiene:

$$Y = \alpha_0\psi_0(X) + \alpha_1\psi_1(X) + \ldots + \alpha_p\psi_p(X) + \varepsilon \tag{4.5}$$

así que :

$$X'X = \begin{pmatrix} A_{00} & & & \\ & A_{11} & 0 & \\ & & A_{22} & \\ 0 & & & \ddots \\ & & & & A_{pp} \end{pmatrix}$$

donde $A_{jj} = \sum_{i=1}^{n}\{\psi_j(X_i)\}^2$.

4.3. Método de Eliminación Recursiva

Cada variable atmosférica junto con el `tiempo` fueron incluidas en la forma de un polinomio ortogonal de grado ocho, posteriormente se procedió a usar el algoritmo de eliminación recursiva (*backward elimination*) para encontrar el modelo más parsimonioso. Dicho algoritmo se basó como primer acercamiento en el criterio BIC (Bayesian Information Criterion), el cual consiste en escoger el modelo para el cual $\text{BIC} = -2\log L + n_p \log n$ alcanza el mínimo. Aquí n_p representa el número de parámetros en el modelo. La Tabla 4.1 muestra el proceso de elección del mejor modelo de forma que $\text{BIC}_{n_k} \leq \text{BIC}_{n_p}$ cuando $n_k < n_p$; así sucesivamente hasta encontrar el modelo con todas las variables explicativas significativas y que minimice el BIC (programa en R disponible en el Apéndice).

Tabla 4.1. Método backward elimination. Criterio del BIC.

Modelo	Devianza	No Parámetros	BIC	Dependencia
1	-4816.198	158	-3827.78949	0.687
2	-4792.908	138	-3929.61450	0.697
3	-4783.543	122	-4020.34150	0.703
4	-4779.935	108	-4104.31400	0.705
5	-4770.997	98	-4157.93350	0.702
6	-4762.568	91	-4193.29475	0.698
7	-4754.307	82	-4241.33550	0.706
8	-4736.509	72	-4286.09500	0.721
9	-4727.3	63	-4333.18775	0.716
10	-4723.814	59	-4354.72475	0.720
11	-4723.361	57	-4366.78325	0.720
12	-4717.98	55	-4373.91375	0.720
13	-4715.613	54	-4377.80250	0.721

El Modelo 13 contiene las variables explicativas que fueron significativas y, como se puede observar, el BIC disminuye conforme se eliminan las variables no significativas.

4.3.1. Resultados

Los mejores modelos obtenidos para cada marginal incluyeron a las siguientes variables explicativas:

Marginal 1 (Vallarta) : $\texttt{tiempo}^5, \texttt{maxTemp}^2, \texttt{rangoTemp}, \texttt{anual}, \texttt{semestral}$.

Marginal 2 (Tlaquepaque) : $\texttt{tiempo}^8, \texttt{maxTemp}^5, \texttt{minVel}^2, \texttt{rangoVel}^7, \texttt{minHum}^5, \texttt{rangoHum}^5, \texttt{vientouSabado}^4, \texttt{anual}, \texttt{semestral}$.

La Figura 4.1 y Figura 4.2 muestran la función de `tiempo` ajustada con las bases ortogonales de los polinomios obtenidos en el mejor modelo en presencia de las demás variables atmosféricas y de periodicidad. Las líneas punteadas corresponden a bandas de confianza de 95 % calculadas con la aproximación $\hat{\boldsymbol{\beta}}_j \sim \text{NMV}(\boldsymbol{\beta}_j, \mathbf{V}(\boldsymbol{\beta}_j))$, donde NMV denota la función de distribución normal multivariada, $\hat{\boldsymbol{\beta}}_j$ es el estimador de máxima verosimilitud de $\boldsymbol{\beta}_j$ y $\mathbf{V}(\boldsymbol{\beta}_j)$ es la matriz de varianza-covarianza correspondiente.

Aunque las curvas difieren significativamente, es posible observar que en las primeras 200 semanas hay una tendencia a la baja y que este comportamiento se repite entre las semanas 350 y 450. Como se puede observar en las gráficas, ambas estaciones presentan un tendencia a la baja en las primeras 150 semanas, este hecho puede ligarse a la implementación del plan para el mejoramiento del aire, el

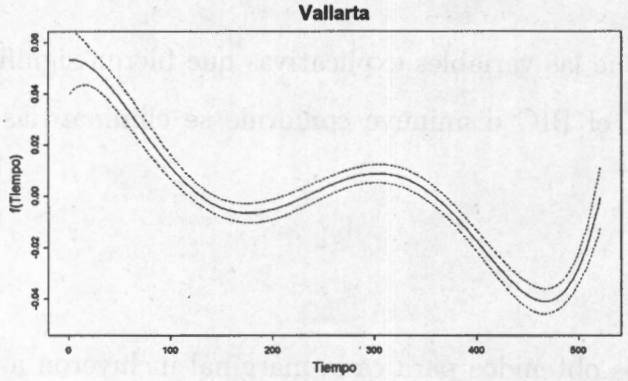


Fig. 4.1. Polinomio en la variable tiempo obtenido en el mejor modelo para Vallarta.

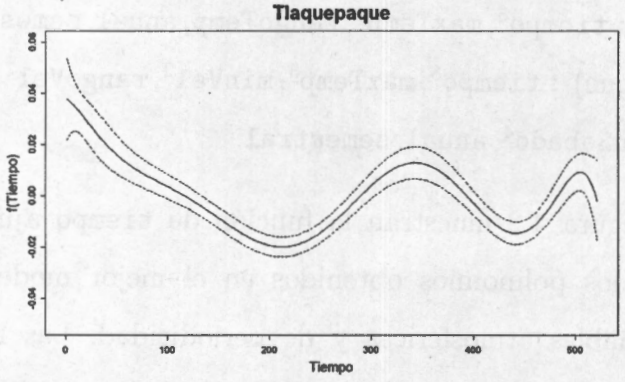


Fig. 4.2. Polinomio en la variable tiempo obtenido en el mejor modelo para Tlaquepaque.

cual se aplicó de 1997 a 2000, es decir, en las primeras 156 semanas; posteriormente se tiene un comportamiento distinto en cada estación. La tendencia a la baja es consistente en ambas estaciones, pero es claro que ambas continuan teniendo la presencia de niveles altos de ozono. Además la parte derecha de la gráfica en la estación Tlaquepaque tiene un comportamiento a la baja mientras que para la estación Vallarta se observa un incremento.

La Tabla 4.2 muestra los estimadores de la ordenada, σ_1, σ_2, γ_2, y los estimadores puntuales de los coeficientes lineales en el mejor modelo, cuando hay mayor grado, se presenta gráficamente. La Figura 4.3 muestra el polinomio máximo de temperatura para Vallarta; la Figura 4.4 y Figura 4.5 muestra los polinomios de máximo de temperatura, mínimo de humedad, mínimo de velocidad, rango de humedad, rango de velocidad y viento Sábado, las variables atmosféricas en el mejor modelo para Tlaquepaque.

Tabla 4.2. Estimadores y errores estándares de los coeficientes lineales y de los parámetros

Vallarta			Tlaquepaque		
Parámetro	Estimador	Error Estándar	Parámetro	Estimador	Error Estándar
Ordenada	0.0830993	0.0011322	Ordenada	0.0722477	0.0009875
RangoTemp	0.1423565	0.0316228			
$\cos\left(\frac{\pi \times \text{tiempo}}{26}\right)$	0.0103863	0.0021461			
$\sin\left(\frac{\pi \times \text{tiempo}}{26}\right)$	0.0069984	0.0017792	$\sin\left(\frac{\pi \times \text{tiempo}}{26}\right)$	0.0070859	0.0022168
$\sin\left(\frac{\pi \times \text{tiempo}}{13}\right)$	−0.0074167	0.0018167	$\sin\left(\frac{\pi \times \text{tiempo}}{13}\right)$	−0.0050758	0.0015313
σ_1	0.0246562	0.0008404	σ_2	0.0194219	0.00073
			γ_2	0.1266817	0.0341919

En ambas estaciones, las altas temperaturas tienden a aumentar la posibilidad de que se incremente la severidad de los niveles de ozono, un hecho bien conocido en las ciencias atmosféricas. Las variables humedad y viento influyen de forma importante en la estación Tlaquepaque, a diferencia de Vallarta. Es posible notar que las variables presentadas gráficamente muestran una influencia significativa en los niveles de ozono troposférico.

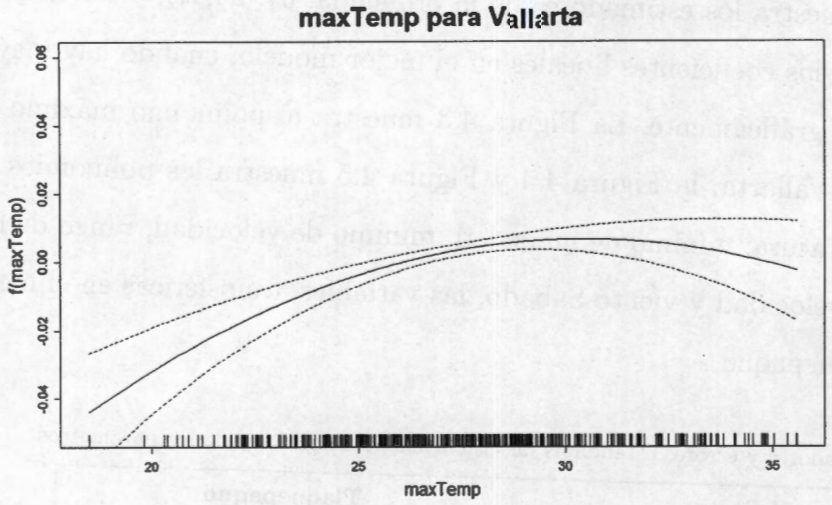


Fig. 4.3. Polinomios con bandas de confianza de 95 % para las variables atmosféricas en el mejor modelo para Tlaquepaque.

Para analizar el ajuste del modelo se grafican los residuales de la siguiente forma: $\frac{m+1-j}{m+1}$ contra $\exp(-U_{(j)})$, para $j = 1, \ldots, m$, donde m es el número de observaciones y $U_{(j)}$ son los valores ordenados de $U = -\log S(Z)$, con $S(Z)$ la función de superviviencia de Z. En este caso Z es la variable aleatoria marginal para cada estación obtenida en el mejor modelo (ver e.g. Collet, 1996).

La densidad de la variable aleatoria $U = -\log S(Z)$ está dada por:

$$f_U(u) = f_Z\{\mathfrak{z}(u)\}\frac{\mathrm{d}\mathfrak{z}(u)}{\mathrm{d}u} \; con \; \mathfrak{z}(u) = S^{-1}(e^{-u}) \tag{4.6}$$

Por otro lado, se sabe que $S(Z) = \Pr(Z \geq \mathfrak{z}) = 1 - F_Z(\mathfrak{z}) = e^{-u}$ y derivando ambas partes de la ecuación $1 - F_Z(\mathfrak{z}) = e^{-u}$ con respecto a u, se tiene que:

$$\begin{aligned} -f_Z(\mathfrak{z})\tfrac{\mathrm{d}\mathfrak{z}(u)}{\mathrm{d}u} &= -e^{-u} \\ \tfrac{\mathrm{d}\mathfrak{z}}{\mathrm{d}u} &= \tfrac{e^{-u}}{f_Z(\mathfrak{z})} \end{aligned} \tag{4.7}$$

y al sustituir este término en (4.6) se llega a que

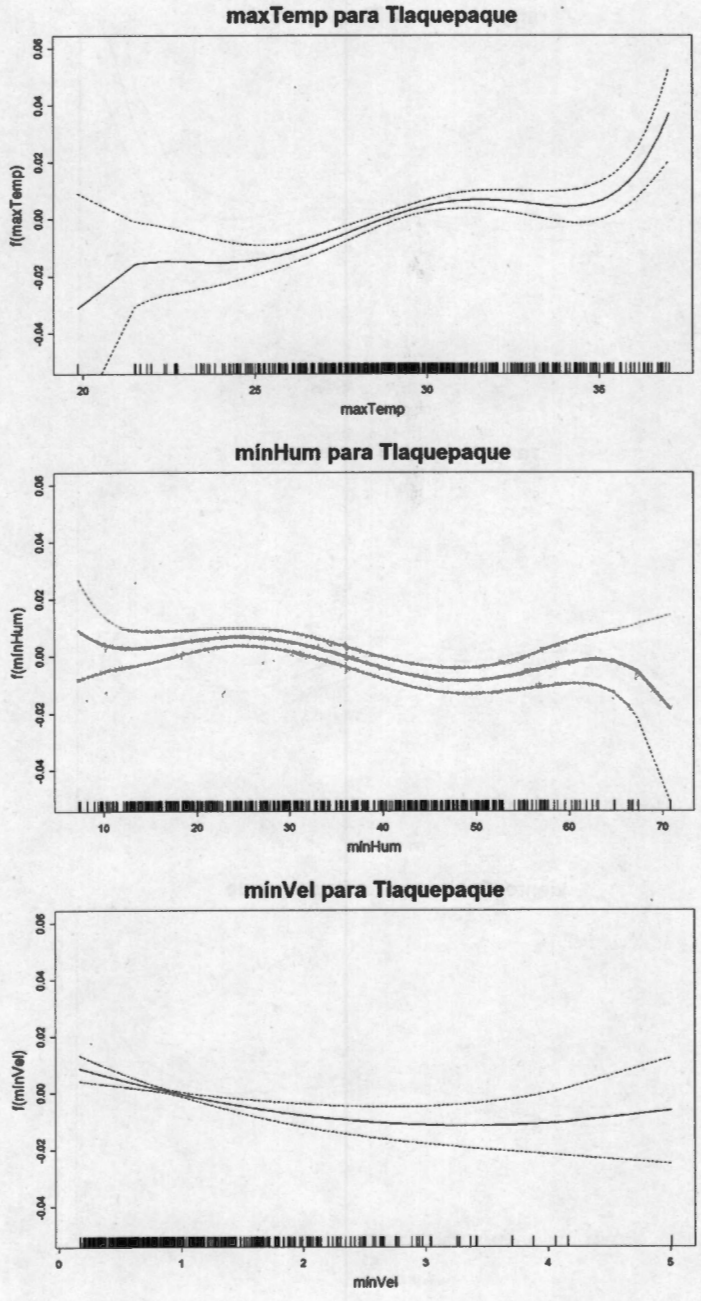


Fig. 4.4. Polinomios con bandas de confianza de 95 % para las variables atmosféricas en el mejor modelo para Tlaquepaque.

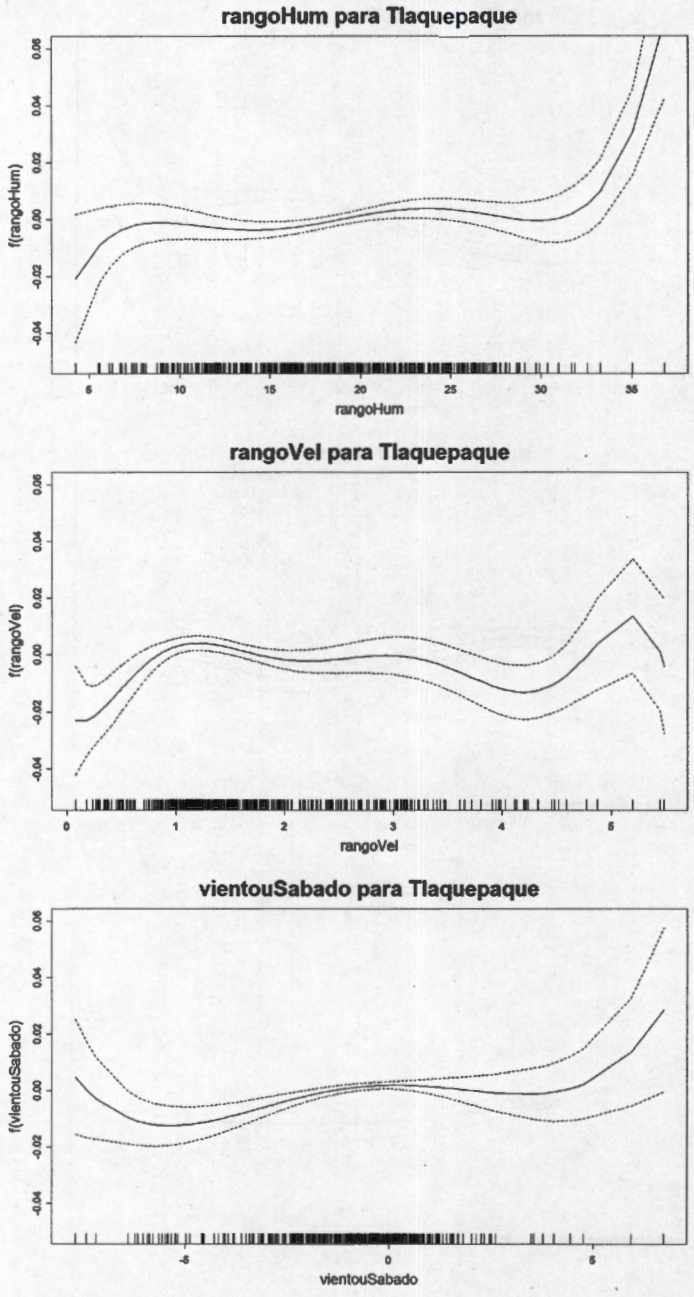


Fig. 4.5. Polinomios con bandas de confianza de 95 % para las variables atmosféricas en el mejor modelo para Tlaquepaque.

$$f_U(u) = f_Z(\jmath(u))\frac{e^{-u}}{f_Z(\jmath(u))} = e^{-u} \tag{4.8}$$

la cual es la función de densidad de probabilidad de una variable aleatoria exponencial con media uno. Entonces, si el modelo ajustado es satisfactorio, los estimadores de la función de superviviencia para el i-ésimo individuo al tiempo t_i, tiempo de supervivencia del individuo, podría ser cercano al verdadero valor $S_i(t_i)$. Esto sugiere que si el modelo correcto ha sido ajustado, los valores $\widehat{S}_i(t_i)$ tendrían propiedades similares a $S_i(t_i)$. Por lo tanto, el logaritmo negativo de las funciones de superviviencia estimadas, $-\log \widehat{S}_i(t_i)$ $i = 1, 2, \ldots, n$, tienen aproximadamente una distribución exponencial con parámetro uno. Entonces de acuerdo a la ecuación (4.8), es posible concluir que en ambos casos u se distribuye exponencial con parámetro igual a uno. En la Figura 4.6 y Figura 4.7 se presentan las gráficas de los residuales para cada una de las marginales, en ambos casos puede observarse que el modelo obtenido, tiene un buen ajuste.

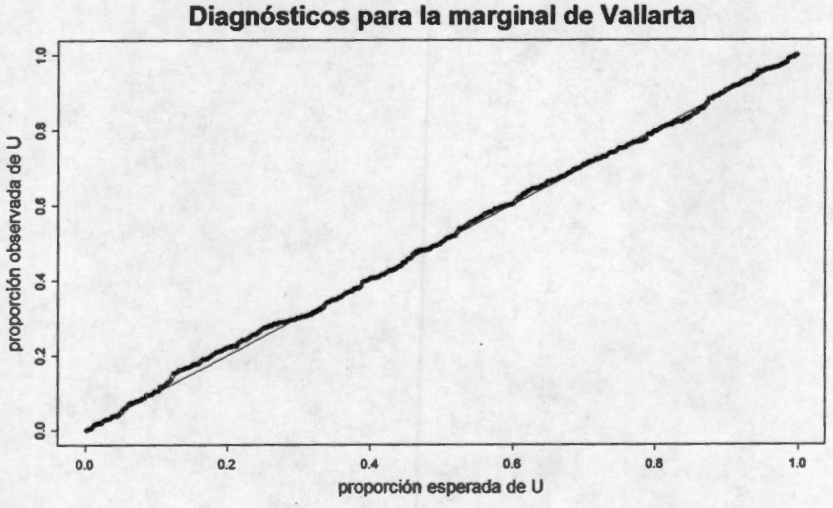


Fig. 4.6. Gráficas de los residuales para Vallarta

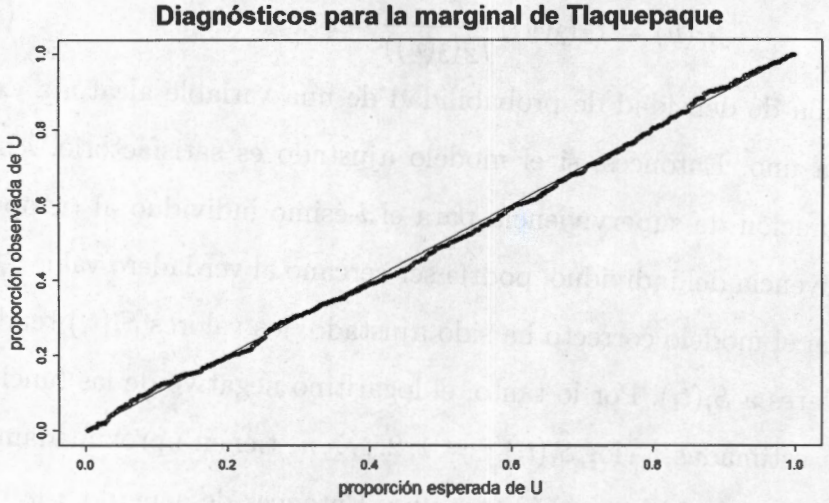


Fig. 4.7. Gráficas de los Residuales para Tlaquepaque

Capítulo 5

Conclusiones

Para el tratamiento del análisis bivariado es posible utilizar la teoría de cópulas, ya que a través de la distribución de probabilidad conjuntas es posible representar las relaciones de dependencia entre distintas variables aleatorias. Es más, la función de distribución conjunta de una serie de variables aleatorias puede expresarse como la función cópula aplicada sobre las distribuciones marginales consideradas individualmente. La cópula Gumbel es útil para evaluar la dependencia de datos extremos debido a que muestra una asimetría adecuada de los datos, tanto en los valores grandes como en los pequeños.

De la teoría de valor extremo se tiene que cada marginal tiene asintóticamente la distribución de valor extremo generalizada y es posible agregar en el parámetro de localización μ, las variables explicativas; a través de los polinomios ortogonales es posible encontrar el grado en que cada una influye en la variable dependiente, debido a que no siempre existe influencia solo lineal.

El estudio presentado refuerza así la conclusión de que la implementación de un programa global para reducir los niveles de contaminación en el área metropolitana de Guadalajara no se ve reflejada en una clara y estable mejora de la calidad del aire a largo plazo. Las marginales de cada estación son esencialmente distintas, lo que muestra que los proyectos para el mejoramiento del aire tuvieron un impacto diferente en cada región, por lo que se sugiere la ejecución de proyectos especializados para cada población, en lugar de un plan global.

La afirmación "Altas temperaturas junto con bajas velocidades de viento, están asociadas a observaciones de máximos de ozono", es confirmada con la presencia en ambas marginales de la variable maxTemp; en el caso de la estación de monitoreo Tlaquepaque hay presencia de efectos de viento contrario a la estación Vallarta. El efecto del rango de temperatura emula a la turbulencia vertical; este efecto solo es significativo para Vallarta.

Las gráficas de los polinomios en el tiempo para el mejor modelo muestran claramente los efectos del programa llamado: "Programa para el mejoramiento de la calidad del aire en la zona metropolitana de Guadalajara 1997-2000"; con una disminución constante de la concentración de ozono en las primeras 150 semanas para Vallarta y las primeras 200 semanas para Tlaquepaque. En las gráficas de las variables restantes de grado mayor a uno, es posible observar la presencia de la variable máximo de Temperatura para Vallarta; y las variables máximo de Temperatura, mínimo de Humedad, rango de Humedad, mínimo de Velocidad, rango de Velocidad y viento en Sábado para Tlaquepaque.

El parámetro de forma ξ fue estadísticamente significativo para Vallarta pero no para Tlaquepaque. El parámetro de dependencia estimado $\hat{\theta}$ es 0.721 (con error estándar 0.028), lo cual indica una concordancia moderada entre los máximos de ambas estaciones de monitoreo.

Las gráficas de los residuales muestra que el modelo aplicado tiene un buen ajuste, ya que los residuales tienen aproximadamente una distribución exponencial con parámetro uno.

La interacción de las variables temperatura y velocidad del viento vistas por Cox & Chu (1992) se asumió en ambas marginales, pero al aplicar el método de eliminación recursiva no se encontró una influencia significativa para alguna estación.

En el arbitraje de la Memoria del XXII Foro Nacional de Estadística se comenta: "la gráfica presentada es muy probablemente similar a la que se obtendría con el solo ajuste del polinomio en el tiempo"; pero los efectos de periodicidad resultaron ser significativos y - de hecho - absorbieron información a la que los polinomios de `tiempo` hubieran sido sensitivos; esto es, en ausencia de la periodicidad se obtendría un polinomio de mayor orden en el mejor modelo. Si solo se toma la variable `tiempo`, el polinomio ajustado de tiempo es de grado 22 para Vallarta y 23 para Tlaquepaque (ver Figura 5.1 y Figura 5.2).

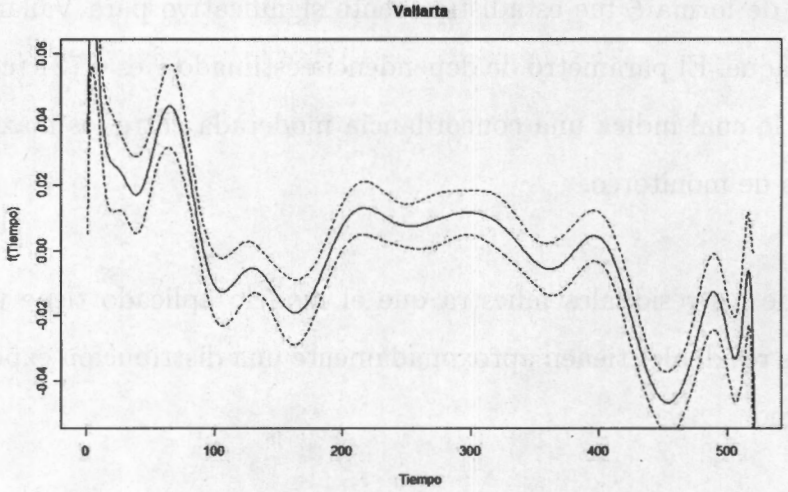


Fig. 5.1. Polinomio en el tiempo correspondiente a Vallarta.

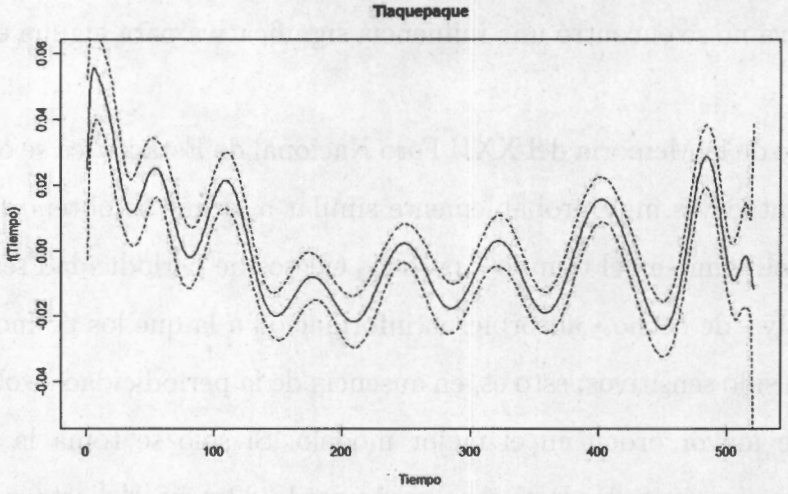


Fig. 5.2. Polinomio en el tiempo correspondiente a Tlaquepaque.

Apéndice

R es un paquete estadístico que ofrece una extensa colección de herramientas que permiten la manipulación de datos, el análisis estadístico de estos y la producción de gráficas. Además, R ofrece la posibilidad de ser utilizado como un lenguaje efectivo de programación. R debe su nombre al humor inspirado del nombre del lenguaje de programación S creado en los laboratorios Bell por John Chambers y colegas. Por lo antes mencionado, R contiene la sintaxis del lenguaje S y esto los hace muy parecidos. Sin embargo, algunas funciones varían, muchas veces porque las versiones de R tratan de simplificar las cosas para el usuario.

La versión más actualizada se puede bajar gratis del sitio oficial de internet de R titulado CRAN (*The Coimprehensive R Archive Network*), en http://www.r-project.org/. El paquete esta disponible para su instalación en Linux, MacOS X y Windows.

```
#Programa R

Cargar los paquetes: EVD y MVTNORM en el software R

#Datos: cada archivo contiene los registros de ambas
#estaciones de monitoreo; Tlaquepaque (primera columna)
#y Vallarta (segunda columna).

ozono<- read.table(file="ozono.txt", header = T)
maxtemp<- read.table(file="max_temp.txt", header = T)
rangotemp<- read.table(file="rango_temp.txt", header = T)
minvel<- read.table(file="min_vel.txt", header = T)
rangovel<- read.table(file="rango_vel.txt", header = T)
minhum<- read.table(file="min_hum.txt", header = T)
rangohum<- read.table(file="rango_hum.txt", header = T)
windusabado <- read.table(file="windu_sabado.txt", header = T)
windvsabado <- read.table(file="windv_sabado.txt", header = T)
tiempo <-1:521

#Comienzo del método BACKWARD ELIMINATION

# Se definen los polinomios de grado 8 en cada marginal
#agregando las variables explicativas, se comprueba
#significancia para verificar el grado adecuado del
# polinomio.
```

```
pol.tla<- data.frame(model.matrix(~poly(tiempo,8)+
                                  poly(maxtemp[,1],8)+
                                  poly(rangotemp[,1],8)+
                                  poly(minvel[,1],8)+
                                  poly(rangovel[,1],8)+
                                  poly(minhum[,1],8)+
                                  poly(rangohum[,1],8)+
                                  poly(windusabado[,1],8)+
                                  poly(windvsabado[,1],8)+
            I(cos(2*pi*tiempo/52))+I(sin(2*pi*tiempo/52))+
            I(cos(2*pi*tiempo/26))+I(sin(2*pi*tiempo/26))-1))

pol.val<- data.frame(model.matrix(~poly(tiempo,8)+
                                  poly(maxtemp[,2],8)+
                                  poly(rangotemp[,2],8)+
                                  poly(minvel[,2],8)+
                                  poly(rangovel[,2],8)+
                                  poly(minhum[,2],8)+
                                  poly(rangohum[,2],8)+
                                  poly(windusabado[,2],8)+
                                  poly(windvsabado[,2],8)+
            I(cos(2*pi*tiempo/52))+I(sin(2*pi*tiempo/52))+
            I(cos(2*pi*tiempo/26))+I(sin(2*pi*tiempo/26))-1))
```

```
# Se incluyen los polinomios en el parámetro de localización.

modelo <- fbvevd(ozono,"log", nsloc1 = pol.tla.1,
                 nsloc2 = pol.val.1)

#El valor de la devianza en el primer modelo es:
deviance(modelo)
-4816.198

#Verificando la significancia de cada variable, se excluyen
#las no significativas; hasta #tener todas las variables
#significativas y devianza mínima del modelo.

#En el último paso del método se obtuvo la siguiente combinación

pol.tla<- data.frame(model.matrix(~poly(tiempo,8)+
                                  poly(maxtemp[,1],5)+
                                  poly(minvel[,1],2)+
                                  poly(rangovel[,1],7)+
                                  poly(minhum[,1],5)+
                                  poly(rangohum[,1],5)+
                                  poly(windusabado[,1],4)+
          I(sin(2*pi*tiempo/52))+I(sin(2*pi*tiempo/26))-1))
```

```
pol.val<- data.frame(model.matrix(~poly(tiempo,5)+
                                  poly(maxtemp[,2],2)+
                                  poly(rangotemp[,2],1)+
                                 I(cos(2*pi*tiempo/52))+
        I(sin(2*pi*tiempo/52))+I(sin(2*pi*tiempo/26))-1))

# Se incluyen los polinomios en el parámetro de localización.

modelo <- fbvevd(ozono,"log", shape2=0, nsloc1 = pol.tla.1,
                 nsloc2 = pol.val.1)

#El valor de la devianza en la última iteración del método es:
deviance(modelo)
-4715.613

#Gráficas con bandas de confianza para Vallarta de cada
#variable explicativa no lineal.

#Tiempo

#Las bases de tiempo con grado 5:
bases.t   <- poly(tiempo, degree=5)
#Los estimadores puntuales del polinomio de tiempo
point.est <- bases.t %*%  modelo$estimate[43:47]
```

```
#Los vectores que contendrán las bandas de confianza:
lim.sup <- 1:521
lim.inf   <- 1:521

#La simulacion de Montecarlo:
for(i in 1:521){
bases      <- bases.t[i,]
aleatorios <- rmvnorm(n=100000, mean= modelo$estimate[43:47],
                      sigma=modelo$var.cov[43:47,43:47])
simulaciones <- aleatorios %*% bases
lim.inf[i]   <- quantile(simulaciones, probs=.025)
lim.sup[i] <- quantile(simulaciones, probs=.975)
}

#La grafica de tiempo:
plot(tiempo, point.est, type="l", lwd=1.6, ylim=c(-0.05, 0.06),
     xlab="Tiempo",ylab="f(Tiempo)", plot.window(xlim=c(0,530),
     ylim=c(-0.02,0.04)))
lines(tiempo, lim.inf, lty=2)
lines(tiempo, lim.sup, lty=2)
title("Vallarta")

# Max Temperatura

#Las bases de max_temp con grado 2:
```

```
bases.maxtemp     <- poly(maxtemp[,2], degree=2)
#Los estimadores puntuales del polinomio de max_temp
point.est.maxtemp <- bases.maxtemp %*% modelo$estimate[48:49]

#Los vectores que contendran las bandas de confianza:
lim.sup.mt <- 1:521
lim.inf.mt   <- 1:521

#La simulacion de Montecarlo:
for(i in 1:521){
bases        <- bases.maxtemp[i,]
aleatorios <- rmvnorm(n=100000, mean=modelo$estimate[48:49],
                      sigma=modelo$var.cov[48:49,48:49])
simulaciones <- aleatorios %*% bases
lim.inf.mt[i]   <- quantile(simulaciones, probs=.025)
lim.sup.mt[i] <- quantile(simulaciones, probs=.975)
}

#La grafica de max_temp:
plot(maxtemp[order(maxtemp[,2]),2], point.est.maxtemp[order(
     maxtemp[,2])], type="l",lwd=1.6, ylim=c(-0.05, 0.06),
     xlab="maxTemp", ylab="f(maxTemp)")
rug(maxtemp[,2])
lines(maxtemp[order(maxtemp[,2]),2],
      lim.inf.mt[order(maxtemp[,2])], lty=2)
```

```
lines(maxtemp[order(maxtemp[,2]),2],
      lim.sup.mt[order(maxtemp[,2])], lty=2)
title("maxTemp para Vallarta")

#Gráficas con bandas de confianza para Tlaquepaque de cada
#variable explicativa no #lineal.

#Tiempo

#Las bases de tiempo con grado 8:
bases.t   <- poly(tiempo, degree=8)
#Los estimadores puntuales del polinomio de tiempo
point.est <- bases.t %*% modelo$estimate[2:9]

#Los vectores que contendran las bandas de confianza:
lim.sup <- 1:521
lim.inf   <- 1:521

#La simulacion de Montecarlo:
for(i in 1:521){
bases <- bases.t[i,]
aleatorios <- rmvnorm(n=100000, mean=modelo$estimate[2:9],
                      sigma=modelo$var.cov[2:9,2:9])
simulaciones <- aleatorios %*% bases
lim.inf[i]   <- quantile(simulaciones, probs=.025)
```

```
lim.sup[i] <- quantile(simulaciones, probs=.975)
}

#La grafica de tiempo:
plot(tiempo, point.est, type="l", lwd=1.6, ylim=c(-0.05, 0.06),
     xlab="Tiempo",ylab="f(Tiempo)", plot.window(xlim=c(0,530),
     ylim=c(-0.02,0.04)))
lines(tiempo, lim.inf, lty=2)
lines(tiempo, lim.sup, lty=2)
title("Tlaquepaque")

#Max Temperatura

#Las bases de max_temp con grado 5:
bases.maxtemp     <- poly(maxtemp[,1], degree=5)
#Los estimadores puntuales del polinomio de max_temp
point.est.maxtemp <- bases.maxtemp %*% modelo$estimate[10:14]

#Los vectores que contendran las bandas de confianza:
lim.sup.mt <- 1:521
lim.inf.mt   <- 1:521

#La simulacion de Montecarlo:
for(i in 1:521){
bases <- bases.maxtemp[i,]
```

```
aleatorios <- rmvnorm(n=100000, mean=modelo$estimate[10:14],
                      sigma=modelo$var.cov[10:14,10:14])
simulaciones <- aleatorios %*% bases
lim.inf.mt[i]   <- quantile(simulaciones, probs=.025)
lim.sup.mt[i] <- quantile(simulaciones, probs=.975)
}

#La grafica de max_temp:
plot(maxtemp[order(maxtemp[,1]),1],
     point.est.maxtemp[order(maxtemp[,1])], type="l",
     lwd=1.6, ylim=c(-0.05, 0.06),
     xlab="maxTemp", ylab="f(maxTemp)")
rug(maxtemp[,1])
lines(maxtemp[order(maxtemp[,1]),1],
     lim.inf.mt[order(maxtemp[,1])], lty=2)
lines(maxtemp[order(maxtemp[,1]),1],
     lim.sup.mt[order(maxtemp[,1])], lty=2)
title("maxTemp para Tlaquepaque")

# Mínimo Velocidad

#Las bases de min_vel con grado 2:
bases.minvel   <- poly(minvel[,1], degree=2)
#Los estimadores puntuales del polinomio de min_vel
point.est.minvel <- bases.minvel %*% modelo$estimate[15:16]
```

```
#Los vectores que contendran las bandas de confianza:
lim.sup.mt <- 1:521
lim.inf.mt   <- 1:521

#La simulacion de Montecarlo:
for(i in 1:521){
bases <- bases.minvel[i,]
aleatorios <- rmvnorm(n=100000, mean=modelo$estimate[15:16],
                      sigma=modelo$var.cov[15:16,15:16])
simulaciones <- aleatorios %*% bases
lim.inf.mt[i]   <- quantile(simulaciones, probs=.025)
lim.sup.mt[i] <- quantile(simulaciones, probs=.975)
}

#La grafica de min_vel:
plot(minvel[order(minvel[,1]),1],
      point.est.minvel[order(minvel[,1])], type="l",
      lwd=1.6, ylim=c(-0.05, 0.06),
      xlab="mínVel", ylab="f(mínVel)")
rug(minvel[,1])
lines(minvel[order(minvel[,1]),1],
      lim.inf.mt[order(minvel[,1])], lty=2)
lines(minvel[order(minvel[,1]),1],
      lim.sup.mt[order(minvel[,1])], lty=2)
```

```
title("mínVel para Tlaquepaque")

# Rango Velocidad

#Las bases de rango_vel con grado 3:
bases.rangovel     <- poly(rangovel[,1], degree=7)
#Los estimadores puntuales del polinomio de rango_vel
point.est.rangovel <- bases.rangovel %*% modelo$estimate[17:23]

#Los vectores que contendran las bandas de confianza:
lim.sup.rv <- 1:521
lim.inf.rv   <- 1:521

#La simulacion de Montecarlo:
for(i in 1:521){
bases       <- bases.rangovel[i,]
aleatorios <- rmvnorm(n=100000, mean=modelo$estimate[17:23],
                      sigma=modelo$var.cov[17:23,17:23])
simulaciones <- aleatorios %*% bases
lim.inf.rv[i]   <- quantile(simulaciones, probs=.025)
lim.sup.rv[i] <- quantile(simulaciones, probs=.975)
}

#La grafica de rango_vel:
plot(rangovel[order(rangovel[,1]),1],
```

```
      point.est.rangovel[order(rangovel[,1])], type="l",
      lwd=1.6, ylim=c(-0.05, 0.06),
      xlab="rangoVel", ylab="f(rangoVel)")
rug(rangovel[,1])
lines(rangovel[order(rangovel[,1]),1],
      lim.inf.rv[order(rangovel[,1])], lty=2)
lines(rangovel[order(rangovel[,1]),1],
      lim.sup.rv[order(rangovel[,1])], lty=2)
title("rangoVel para Tlaquepaque")

# Mínimo Humedad

#Las bases de min_hum con grado 5:
bases.minhum     <- poly(minhum[,1], degree=5)
#Los estimadores puntuales del polinomio de min_hum
point.est.minhum <- bases.minhum %*% modelo$estimate[24:28]

#Los vectores que contendran las bandas de confianza:
lim.sup.mh <- 1:521
lim.inf.mh   <- 1:521

#La simulacion de Montecarlo:
for(i in 1:521){
bases <- bases.minhum[i,]
aleatorios <- rmvnorm(n=100000, mean=modelo$estimate[24:28],
```

```
                    sigma=modelo$var.cov[24:28,24:28])
simulaciones <- aleatorios %*% bases
lim.inf.mh[i]   <- quantile(simulaciones, probs=.025)
lim.sup.mh[i] <- quantile(simulaciones, probs=.975)
}

#La grafica de min_hum:
plot(minhum[order(minhum[,1]),1],
     point.est.minhum[order(minhum[,1])], type="l",
     lwd=1.6, ylim=c(-0.05, 0.06),
     xlab="mínHum", ylab="f(mínHum)")
rug(minhum[,1])
lines(minhum[order(minhum[,1]),1],
     lim.inf.mh[order(minhum[,1])], lty=2)
lines(minhum[order(minhum[,1]),1],
     lim.sup.mh[order(minhum[,1])], lty=2)
title("mínHum para Tlaquepaque")

#Rango Humedad

#Las bases de rango_hum con grado 5:
bases.rangohum <- poly(rangohum[,1], degree=5)
#Los estimadores puntuales del polinomio de rango_hum
point.est.rangohum <- bases.rangohum %*% modelo$estimate[29:33]
```

```
#Los vectores que contendran las bandas de confianza:
lim.sup.rh <- 1:521
lim.inf.rh   <- 1:521

#La simulacion de Montecarlo:
for(i in 1:521){
bases <- bases.rangohum[i,]
aleatorios <- rmvnorm(n=100000, mean=modelo$estimate[29:33],
                      sigma=modelo$var.cov[29:33,29:33])
simulaciones <- aleatorios %*% bases
lim.inf.rh[i]   <- quantile(simulaciones, probs=.025)
lim.sup.rh[i] <- quantile(simulaciones, probs=.975)
}

#La grafica de rango_hum:
plot(rangohum[order(rangohum[,1]),1],
      point.est.rangohum[order(rangohum[,1])], type="l",
      lwd=1.6, ylim=c(-0.05, 0.06),
      xlab="rangoHum", ylab="f(rangoHum)")
rug(rangohum[,1])
lines(rangohum[order(rangohum[,1]),1],
      lim.inf.rh[order(rangohum[,1])], lty=2)
lines(rangohum[order(rangohum[,1]),1],
      lim.sup.rh[order(rangohum[,1])], lty=2)
title("rangoHum para Tlaquepaque")
```

```
# windusabado[,1]

#Las bases de windusabado con grado 4:
bases.windusabado      <- poly(windusabado[,1], degree=4)
#Los estimadores puntuales del polinomio de windusabado
point.est.windusabado <- bases.windusabado %*% modelo$
                         estimate[34:37]

#Los vectores que contendran las bandas de confianza:
lim.sup.wus <- 1:521
lim.inf.wus   <- 1:521

#La simulacion de Montecarlo:
for(i in 1:521){
bases <- bases.windusabado[i,]
aleatorios <- rmvnorm(n=100000, mean=modelo$estimate[34:37],
                      sigma=modelo$var.cov[34:37,34:37])
simulaciones <- aleatorios %*% bases
lim.inf.wus[i]   <- quantile(simulaciones, probs=.025)
lim.sup.wus[i] <- quantile(simulaciones, probs=.975)
}

#La grafica de windusabado:
plot(windusabado[order(windusabado[,1]),1],
```

```
      point.est.windusabado[order(windusabado[,1])], type="l",
      lwd=1.6, ylim=c(-0.05, 0.06),
      xlab="vientouSabado", ylab="f(vientouSabado)")
rug(windusabado[,1])
lines(windusabado[order(windusabado[,1]),1],
      lim.inf.wus[order(windusabado[,1])], lty=2)
lines(windusabado[order(windusabado[,1]),1],
      lim.sup.wus[order(windusabado[,1])], lty=2)
title("vientouSabado para Tlaquepaque")

#Aproximación USANDO solo TIEMPO, así como sugiere el arbitraje
#de la Memoria #del XXII Foro Nacional de Estadística

polinomio1 <- data.frame(model.matrix(~poly(tiempo,23)-1))
#polinomio1--> Tlaquepaque

polinomio2 <- data.frame(model.matrix(~poly(tiempo,22)-1))
#polinomio2--> Vallarta

modelo.tiempo <- fbvevd(ozono, "log", nsloc1 = polinomio1,
                        nsloc2 = polinomio2)
#El valor de la devianza en el primer modelo es:
deviance(modelo)
-4622.237
```

```
#Gráficas con bandas de confianza para Tlaquepaque de la
#variable tiempo.

#Las bases de tiempo con grado 23:
bases.t.t <- poly(tiempo, degree=23)
#Los estimadores puntuales del polinomio de tiempo
point.est.t <- bases.t.t %*% modelo.tiempo$estimate[2:24]

#Los vectores que contendran las bandas de confianza:
lim.sup.t <- 1:521
lim.inf.t <- 1:521

#La simulacion de Montecarlo:
for(i in 1:521){
bases <- bases.t.t[i,]
aleatorios <- rmvnorm(n=100000, mean=modelo.tiempo$
        estimate[2:24],
        sigma=modelo.tiempo$var.cov[2:24,2:24])
simulaciones <- aleatorios %*% bases
lim.inf.t[i] <- quantile(simulaciones, probs=.025)
lim.sup.t[i] <- quantile(simulaciones, probs=.975)
}

#La grafica de tiempo:
plot(tiempo, point.est.t, type="l", lwd=1.6, ylim=c(-0.05, 0.06),
```

```
     xlab="Tiempo", ylab="f(Tiempo)")
lines(tiempo, lim.inf.t, lty=2)
lines(tiempo, lim.sup.t, lty=2)
title("Tlaquepaque")

#Gráficas con bandas de confianza para Vallarta de la
#variable tiempo.

#Las bases de tiempo con grado 22:
bases.t.v <- poly(tiempo, degree=22)
#Los estimadores puntuales del polinomio de tiempo
point.est.v <- bases.t.v %*% modelo.tiempo$estimate[28:49]

#Los vectores que contendran las bandas de confianza:
lim.sup.v <- 1:521
lim.inf.v <- 1:521

#La simulacion de Montecarlo:
for(i in 1:521){
bases <- bases.t.v[i,]
aleatorios <- rmvnorm(n=100000, mean=modelo.tiempo$
              estimate[28:49],
         sigma=modelo.tiempo$var.cov[28:49,28:49])
simulaciones <- aleatorios %*% bases
lim.inf.v[i] <- quantile(simulaciones, probs=.025)
```

```
lim.sup.v[i] <- quantile(simulaciones, probs=.975)
}

#La grafica de tiempo:
plot(tiempo, point.est.v, type="l", lwd=1.6, ylim=c(-0.05, 0.06),
     xlab="Tiempo", ylab="f(Tiempo)")
lines(tiempo, lim.inf.v, lty=2)
lines(tiempo, lim.sup.v, lty=2)
title("Vallarta")

#Gráfica de los residuales para Tlaquepaque.

beta1 <- modelo$estimate[1:39]

x1 <- model.matrix(~poly(tiempo,8)+poly(maxtemp[,1],5)+
                    poly(minvel[,1],2)+poly(rangovel[,1],7)+
                    poly(minhum[,1],5)+poly(rangohum[,1],5)+
                    poly(windusabado[,1],4)+
              I(sin(2*pi*tiempo/52))+I(sin(2*pi*tiempo/26)))

c.lineal <- c(x1 %*% beta1)

ozono1 <- ozono$TLA

S.t.1 <- pgev(ozono1, loc=c.lineal, scale=modelo$estimate[40],
```

```
            shape=modelo$estimate[41], lower.tail = F)

u.t <- -1*log(na.omit(S.t.1))

u.ord <- sort(u.t)

cuantil.obs <- exp(-1*u.ord)

cuantil.esp <- (518+1-(1:518))/(518+1)

plot(cuantil.esp, cuantil.obs, xlab="proporción esperada de U",
       ylab="proporción observada de U",
       main="Diagnósticos para la marginal de Tlaquepaque")
lines(seq(from=0, to=1, length=10), seq(from=0, to=1, length=10))

#Gráfica de los residuales para Vallarta.

beta2 <- modelo$estimate[42:53]

x2 <- model.matrix(~poly(tiempo,5)+poly(maxtemp[,2],2)+
                    poly(rangotemp[,2],1)+
                   I(cos(2*pi*tiempo/52))+I(sin(2*pi*tiempo/52))+
                   I(sin(2*pi*tiempo/26)))

c.lineal <- c(x2 %*% beta2)
```

```
ozono2 <- ozono$VAL

S.t.2 <- (pgev(ozono2, loc=c.lineal, scale=modelo$estimate[54],
                    lower.tail = F))

u.t <- -1*log(na.omit(S.t.2))

u.ord <- sort(u.t)

cuantil.obs <- exp(-1*u.ord)

cuantil.esp <- (521+1-(1:521))/(521+1)

plot(cuantil.esp, cuantil.obs, xlab="proporción esperada de U",
       ylab="proporción observada de U",
      main="Diagnósticos para la marginal de Vallarta")
lines(seq(from=0, to=1, length=10), seq(from=0, to=1, length=10))
```

Referencias

1. Bloomfield, P.; Royle, A.; Yang, Q. (1996). Accounting for meteorological effects in measuring urban ozone levels and trends. *Atmospheric Environment*, 30: 3067-3078.
2. Chihara T. S. (1978), An introduction to orthogonal polynomials, Gordon and Breach, Nueva York.
3. Coles S. (2001). An Introduction to Statistical Modeling of Extreme Values, Great Britain: Springer-Verlang.
4. Collet D. (1996). Modelling Survival Data in Medical Research, Great Britain: Chapman-Hall.
5. Cox W. M. and Chu S-H (1992). Meteorologically Adjusted Ozone trends in Urban Areas: A Probabilistic Approach *Atmospheric Environment*, Vol. 27B, No. 4, 425-434.
6. Draper N.R.; Smith H. (1981).Applied Regression Analysis, John Wiley & Sons, Inc.
7. Dupuis, D.J. (2005). Ozone concentrations: A robust analysis of mutivariate extremes. *Technometrics*, 47: 191-201.
8. Huang, L.; Smith, R. (1999). Meteorologically-dependent trends in urban ozone. *Environmetrics*, 10:103-108.
9. Joe H.; Hu T. (1996). Multivariate distributions from mixtures of mas-infinitely divisible distributions. *Journal of multivariate analysis* 57, 240-265. Article No.0032.
10. Joe H. (1997). Multivariate Models and Dependence Concepts. New York. Chapman & Hall.
11. National Research Council (1991). Rethinking the Ozone Problems in Urban and Regional Air Pollution. Washington, DC: National Academic Press.

12. 12-23-94 NORMA Oficial Mexicana NOM-020-SSA1-1993, Salud Ambiental. Criterio para evaluar la calidad del aire ambiente con respecto al ozono (O_3). Valor normado para la concentración de ozono (O_3) en el aire ambiente como medida de protección a la salud de la población.
13. Pagnotti, V. (1990). Seasonal ozone levels and control by seasonal metereology. *Journal of the Air and Waste Management Association*, 40: 206-210.
14. Ramírez-Sánchez H.U.; Andrade García, M.D.; González Castañeda, M.E.;Celis- De la Rosa, A.J. (2006). Contaminantes atmosféricos y su correlación con infecciones agudas de las vías respiratorias en niños de Guadalajara, Jalisco. *Salud Pública de México*, 48: 385-394.
15. Resnick, S. I. (1987). Extreme Values, Regular Variation, and Point Processes. Springer Verlang, New York.
16. Smith, R. L. (1985). Maximum Likeliohood Estimation in a Class of Nonregular Cases. *Biometrika*, 72: 67-90.
17. Smith, R. L.; Wigley, T. L.; Santer, B. D. (2003), A Bivariate Time Series Approach to Anthropogenic Trend Detection in Hemispheric Mean Temperatures, *Journal of Climate*, vol. 16, Issue 8, pp.1228-1240
18. Stephenson A., A User's Guide to the evd Package (Version 2.2), Department of Statistic and Applied Probability, National University of Singapore, March 2006.
19. WHO - World Health Organization. (1987). Air Quality Guidelines for Europe. WHO regional publications, European series 23, Copenhagen, Regional Office for Europe.
20. Yan J. (2007). Enjoy the Joy of Copulas: With a Package copula, University of Connecticut, October 2007, Volume 21, Issue 4.
21. http://www.r-project.org/