



COLEGIO DE POSTGRADUADOS

INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN
EN CIENCIAS AGRÍCOLAS

CAMPUS MONTECILLO

POSTGRADO DE SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA
ESTADÍSTICA

**Estimación de Tendencias en Niveles Máximos de
Contaminación usando Regresión por Cuantiles
ajustando el efecto por variables meteorológicas**

Hortensia Josefina Reyes Cervantes

T E S I S

PRESENTADA COMO REQUISITO PARCIAL PARA
OBTENER EL GRADO DE:

DOCTORA EN CIENCIAS

MONTECILLO, TEXCOCO, EDO. DE MÉXICO
2008

La presente tesis titulada: **Estimación de Tendencias en Niveles Máximos de Contaminación usando Regresión por Cuantiles ajustando el efecto por variables meteorológicas**, realizada por la alumna: **Hortensia Josefina Reyes Cervantes**, bajo la dirección del Consejo Particular indicado ha sido aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:

DOCTORA EN CIENCIAS

**SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA
ESTADÍSTICA**

CONSEJO PARTICULAR

CONSEJERO _____
Dr. Humberto Vaquera Huerta

ASESOR _____
Dr. José A. Villaseñor Alva

ASESOR _____
Dr. Félix V. González Cossio

ASESOR _____
Dr. Filemón Ramírez Pérez

ASESOR _____
Dr. Barry C. Arnold

Montecillo, Texcoco, Estado de México, Abril de 2008

Estimación de Tendencias en Niveles Máximos de Contaminación usando Regresión por Cuantiles ajustando el efecto por variables meteorológicas

Hortensia Josefina Reyes Cervantes, Dr.

Colegio de Postgraduados, 2008

En este trabajo se proponen dos metodologías estadísticas de análisis basada en la teoría de valores extremos para modelar tendencias de contaminación, aplicándose en el caso particular de la contaminación ambiental de la Ciudad de México. En la primera metodología, suponiendo que los datos de medición de ozono se pueden modelar de acuerdo a un modelo paramétrico, pues existen las condiciones de regularidad para la función GVE, se pueden aplicar resultados de la teoría asintótica para investigar si hay condiciones para que los parámetros estimados tengan distribución normal y se pueda estimar la función cuantil de la distribución GVE. Para la función cuantil a diferentes niveles, en este caso para $p = 0.05$, se obtiene la tendencia para las 20 estaciones meteorológicas de la Zona Metropolitana de la Ciudad de México durante el periodo de 1986 a 2005. El análisis de la información indica una disminución en los niveles muy altos de ozono para los últimos años.

En la segunda metodología, se involucran las covariables ambientales al parámetro de escala de la distribución GVE, para investigar tendencias en las concentraciones del ozono. Los modelos propuestos tienen la ventaja de que se pueden hacer ajustes de ozono como los modelos de regresión lineal a partir de la convergencia de la Función Cuantil a una normal estandarizada. Por lo cual se pueden obtener pruebas de hipótesis e intervalos de confianza de los parámetros de interés.

Palabras clave: Teoría de valores extremos, ozono troposférico, excedencias, ozono.

Estimation of Trends in High levels of pollution using the quantiles regression adjusting the effects by meteorological variables

Hortensia Josefina Reyes Cervantes, Dr.

Colegio de Postgraduados, 2008

In this work, two methodologies of statistical analysis based on the extreme value theory to model pollution trends are proposed and applied to environmental pollution in Mexico City. In the first methodology, assuming that the ozone measurement data can be modeled according to a parametric model, since the conditions for regularity of the GEV function are met, we can apply the results of the asymptotic theory to investigate if there are the conditions for the estimated parameters to follow a normal distribution, and the quantile function of the GEV distribution can be obtained. For the quantile function at different levels, in this case for estimated, we estimated the trend for the 20 meteorological stations of the Metropolitan Zone of Mexico City during the period from 1986 to 2005. The results of indicates a down ward in the very high ozone levels in the later years.

In the second methodology, we incorporate environmental variables in the scale parameter of the GEV distribution, for investigate trend in the ozone measurements. The proposed models have the advantage of adjustment the trends of ozone as the linear regression quantile function. Also it was obtained confidence intervals and hypotheses test of parameters.

Key words: Theory of extreme values, tropospheric ozone, exceedances, ozone.

AGRADECIMIENTOS

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico brindado durante la realización de mis estudios.

A la Benemérita Universidad Autónoma de Puebla (BUAP), a la Facultad de Ciencias Físico-Matemáticas y, en particular, a la Academia de Matemáticas de la FCFM por todo el apoyo recibido durante la realización de mis estudios.

Al Colegio de Postgraduados, por haberme brindado la oportunidad de seguir aprendiendo en sus aulas.

Al Doctor Humberto Vaquera, le agradezco la oportunidad que me dio, así como sus enseñanzas, su apoyo y su paciencia en toda mi estancia en el C. P. Su optimismo siempre me ayudó para seguir adelante.

Dr. Sergio Pérez Elizalde, por sus observaciones y consejos.

A los integrantes de mi Consejo Particular:

Al Dr. José Villaseñor Alva, por ser un excelente profesor.

A los Doctores Dr. Félix V. González Cossio, Dr. Filemón Ramírez Pérez y Barry C. Arnold por sus comentarios y por haber revisado el escrito.

A mis profesores, compañeros de clases y al personal administrativo del C.P., que siempre me trataron bien y me hicieron muy agradable la estancia en esta honorable institución pública.

La vida nos presenta a miles de personas y cada una de ellas viene a cumplir un papel en nuestra vida diaria.

DEDICATORIAS

A Dios por permitirme estar aquí.

A mis padres (†) que aunque no tuvieron la fortuna de estudiar me dieron esa oportunidad, que ahora la veo como la mejor herencia para un hijo.

A Manuel, Atziri y Cintli, por su amor, apoyo y comprensión. Son la luz que me ayuda en todos los momentos de mi vida.

A mis hermanitos: Fernando, Margarita, Gabriel, Jorge, Alejandra, Blanca y Juan.

Hace mucho tiempo los aztecas dejaron como testimonio la siguiente inscripción, que hoy me parece muy cercana a nuestra realidad...

O aprendemos a vernos a nosotros mismos tal como somos y cambiamos de actitud frente al planeta y frente a nosotros mismos, o el planeta se encargará de acabar con nosotros (El quinto sol).

Índice

Prólogo	13
1. Introducción	15
1.1. Antecedentes	16
1.2. Objetivo General	19
1.3. Estructura de la tesis	20
2. Distribución de Valores Extremos	21
2.1. Función de Valores Extremos	23
2.1.1. La α -parametrización de la GVE	23
2.1.2. Media y varianza de una variable aleatoria con distribución VE	24
2.2. Estimación paramétrica de la Distribución Generalizada de Valores Extremos	25
2.3. Estimadores de Máxima Verosimilitud	27
2.4. Estimadores de Momentos para los parámetros μ y σ	28
2.4.1. Procedimiento para obtener los estimadores de momentos de la DGVE	29
2.5. Estimadores de orden para el parámetro de forma	31
2.6. Intervalos de confianza	33

2.7. Intervalo de Confianza para la media de una distribución normal cuando σ es desconocida	34
2.8. Normalidad Asintótica de los Estimadores Máximo Verosímiles	35
2.9. Pruebas de Hipótesis	36
2.10. Pruebas de Razón de verosimilitud generalizada.	38
2.11. Diagnósticos de Modelos	39
2.12. Método de los umbrales	40
3. Modelos Estadísticos para investigar tendencias en ozono	42
3.1. El problema de la Contaminación por Ozono	44
3.1.1. Modelos tipo Regresión	47
3.1.2. Modelos tipo Valor Extremo	48
3.1.3. Modelos Espacio tiempo	51
4. Los Modelos Propuestos	53
4.1. La convergencia de la función cuantil	53
4.2. Evaluación Numérica de la función cuantil	56
4.3. Modelo de Regresión basado en la función cuantil	62
4.4. Modelo de Regresión incorporando covariables	63
5. Un ejemplo de aplicación	65
5.1. Características de la Ciudad de México.	66
5.2. Construcción de bloques.	68
5.3. Procedimiento para la primera propuesta	69
5.4. Procedimiento para la segunda propuesta	70

5.5. Resultados	73
5.5.1. Resultados de la primera propuesta	73
5.5.2. Resultados de la segunda propuesta	87
5.6. Discusión	90
5.6.1. Discusión de la primera propuesta	90
5.6.2. Discusión de la segunda propuesta	91
6. Conclusiones	92
6.1. Conclusiones	92
6.2. Recomendaciones e investigaciones futuras	93
Referencias	93

Índice de tablas

3.1. Variables Meteorológicas más usuales.	45
4.1. Simulaciones de GVE con $p = 0.05$	57
5.1. Estaciones Meteorológicas (Información del SIMAT)	67
5.2. Estimadores de Máxima verosimilitud anual en la estación Plateros con $p=0.05$	69
5.3. Información mensual de las variables ambientales en la estación Plateros (1990).	71
5.4. Modelo con Ozono y Velocidad de viento (wsp)	89
5.5. Modelo con Ozono y tiempo (t)	89
5.6. Modelo con Ozono y covariable: temperatura	90
5.7. Modelo con Ozono y covariable: tiempo	90

Índice de figuras

2.1. Gráfica de las funciones de densidad GVE.	26
3.1. Modelo gráfico de relaciones causales entre algunas variables meteorológicas(Torres,1997)	46
4.1. Distribución asintótica del estimador de la función cuantil para la Gumbel 1.	57
4.2. Distribución asintótica del estimador cuantil para la Gumbel 2.	58
4.3. Distribución asintótica del estimador de la función cuantil para la Weibul 1.	59
4.4. Distribución asintótica del estimador de la función cuantil para la Weibul 2.	59
4.5. Distribución asintótica del estimador de la función cuantil para la Fréchet 1.	60
4.6. Distribución asintótica del estimador de la función cuantil para la Fréchet 2.	60
4.7. Distribución asintótica del estimador de la función cuantil para la Fréchet 3.	61
5.1. Ubicación de las estaciones de la ZMCM (Fuente: SIMAT, RAMA)	66
5.2. Estación Merced, niveles de ozono en 1986 y 2005.	68
5.3. Comportamiento mensual de los máximos de la velocidad del viento, en la Estación Plateros.	71

5.4. Comportamiento mensual de los máximos de la humedad relativa, en la Estación Plateros.	72
5.5. Comportamiento mensual de los máximos de la temperatura, en la Estación Plateros.	72
5.6. Estimaciones usando la Regresión Cuantil en $p = 0.05$, en las Estaciones Meteorológicas para el Valle de México 1.	74
5.7. Estimaciones usando la Regresión Cuantil en $p = 0.05$, en las Estaciones Meteorológicas para el Valle de México 2.	75
5.8. Estimaciones usando la Regresión Cuantil en $p = 0.05$, en las Estaciones Meteorológicas para el Valle de México 3.	76
5.9. Estimaciones usando la Regresión Cuantil en $p = 0.05$, en las Estaciones Meteorológicas para el Valle de México 4.	77
5.10. Estimaciones usando la Regresión Cuantil en $p = 0.05$, en las Estaciones Meteorológicas para el Valle de México 5.	78
5.11. Estimaciones usando la Regresión Cuantil en $p = 0.05$, en las Estaciones Meteorológicas para el Valle de México 6.	79
5.12. Estimaciones usando la Regresión Cuantil en $p = 0.05$, en las Estaciones Meteorológicas para el Valle de México 7.	80
5.13. Comparaciones entre los resultados teóricos y empíricos en las Estaciones Pedregal, Plateros y Sur, en $p=0.05$	81
5.14. Comparaciones entre los resultados teóricos y empíricos en las Estaciones Tacuba, Tláhuac y Taxqueña, al $p=0.05$	82
5.15. Comparaciones entre los resultados teóricos y empíricos en las Estaciones Tlanepantla, Tlalpan y UAM Iztapalapa, al $p=0.05$	83
5.16. Comparaciones entre los resultados teóricos y empíricos en las Estaciones Xalostoc, Azcapozalco y Benito Juárez, al $p=0.05$	84
5.17. Comparaciones entre los resultados teóricos y empíricos en las Estaciones, Cerro de la Estrella, Chapingo, Merced, al $p=0.05$	85
5.18. Comparaciones entre los resultados teóricos y empíricos en las Estaciones Cuajimalpa, Ecatepec y Hangares, al $p=0.05$	86

5.19. Comparaciones entre los resultados teóricos y empíricos en las Estaciones Lagunilla y San Agustín, al $p=0.05$	87
5.20. Análisis descriptivo de los supuestos distribucionales bajo GVE.	88

Prólogo

Como disciplina, la estadística ha sido influenciada por otras ciencias con las cuales interactúa. Este fenómeno ocurre muy frecuentemente en las ciencias ambientales pues allá se ligan problemas de la atmósfera, ecología, geología, toxicología, ciencias biomédicas y económicas con el objetivo de mejorar la salud pública. Esta interacción genera una gran cantidad de datos, cuyo análisis en general, motiva nuevos desafíos que se convierten en nuevos desarrollos estadísticos (Olkin *et al.*, 1990). Estos avances dan nuevos parámetros a los investigadores para re-evaluar sus diseños y análisis globales, enriqueciendo así el panorama estadístico. Piegorsch *et al.* muestra diversos estudios realizados modelando la contaminación ambiental y el análisis de mortalidad, el espacio-tiempo de la lluvia ácida, el análisis de tendencias y evaluaciones ecológicas. También se han hecho investigaciones sobre exploraciones de dosis mínimas con efectos ambientales sobre poblaciones de los animales. Lo anterior es una ilustración de la importancia de los avances de la estadística ambiental moderna. Lo cierto es que muchos problemas importantes en estas áreas permanecen sin resolver.

El ozono es un constituyente importante de la atmósfera y éste varía con la altura. El estudio de la concentración de ozono cerca de la superficie terrestre (tropósfera) es de gran interés por los efectos adversos que produce en los seres vivos. La contaminación del aire en zonas urbanas existe porque en el ambiente hay grandes cantidades de óxidos de nitrógeno (NOx) y compuestos orgánicos volátiles (VOC's), emitidos por vehículos de automotores y fuentes estacionarias que dan lugar a su formación. La Revista Electrónica Tierramérica reporta que, en opinión de Mario Molina, que la producción de ozono en el ambiente a partir de los anteriores precursores es una función compleja de muchos factores que incluyen la intensidad de la luz solar, el mezclado atmosférico, las concentraciones y las relaciones entre los precursores y la reactividad de los precursores orgánicos (Molina, 2006). Los compuestos orgánicos volátiles (VOC's) son sustancias que contienen carbono y se encuentran en todos los elementos vivos. Los VOC's se convierten en gases o vapores y también son liberados por la quema de combustibles como gasolina, madera, carbón o gas natural.

La exposición al ozono causa irritación de las mucosas, enfermedades respiratorias y reducción de la resistencia a las infecciones. Se considera que los primeros efectos sobre el hombre aparecen cuando su concentración alcanza a alrededor de 0.100

ppm. Para proteger la salud pública de la contaminación del ozono se han establecido estándares que especifican distintos límites, desde 0.060 ppm para 1 hora de exposición (Organización Mundial para la Salud) hasta 0.120 ppm y a no ser excedido por más de un día por año (U.S. Environmental Protection Agency, EPA). En México, la norma de calidad del aire estipula que no debe rebasarse la concentración de 0.11 ppm como promedio máximo en una hora, una vez en un periodo de 3 años.

Respecto a la vegetación, las plantas son más sensibles aún y pueden observarse daños en algunas especies; estas se manifiestan en un deterioro visible en la vegetación ya que se reduce el tamaño en las semillas y la producción de las cosechas. La acción sobre la salud humana y la vegetación continúa siendo objeto de una intensa investigación. Además de estos efectos ambientales, la Revista Electrónica Milagro ha consultado a diversos investigadores que establecen que el ozono también actúa como un gas invernadero y que absorbe radiación infrarroja emitida por la Tierra. La concentración de ozono troposférico podría duplicar la temperatura para incrementarse alrededor de 1 °C. Este efecto radiante depende de la masa total de ozono en toda la tropósfera.

Existen diversos indicadores de exposición al ozono, algunos son más robustos que aquellos asociados con National Ambient Air Quality Standard (NAAQS) y pueden ser usados para monitorear las tendencias de ozono. La efectividad de un programa ambiental reduce las concentraciones de ozono en un área particular que es rastreada en un periodo de años junto con un conteo de los efectos en las fluctuaciones meteorológicas. Las condiciones de altas concentraciones de ozono inducen altas temperaturas, bajos vientos, radiación solar intensa y ausencia de precipitación.

Investigadores mexicanos, entre ellos Mario Molina (Molina, 2006), están investigando las relaciones entre el ozono, cambio climático y calidad del aire, para crear una delicada ecuación que sea la base de nuevos compuestos químicos que no afecten al medio ambiente y sirvan de energía. Él comenta que en México se analiza el problema de calidad del aire que beneficiará al cambio climático y que mejorando la eficiencia de transporte, se disminuye las emisiones de dióxido de carbono, por lo cual, recomienda cambiar la fórmula de la gasolina, modernizar el cambio del parque vehicular y hacer reprogramación del transporte público, entre otras medidas.

Capítulo 1

Introducción

Es conocido que las grandes ciudades sufren problemas de contaminación ambiental. Algunas de ellas, como Los Angeles, Beijing, Tokyo, Cairo, Moscú y la Ciudad de México, entre otras, monitorean las concentraciones de ozono en el aire (tropósfera) para documentar la severidad del problema del ozono y medir el progreso en cuanto a la reducción de las concentraciones de ozono emitidas por los habitantes en sus actividades, debido al cumplimiento de nuevas normas ambientales. La Organización Mundial para la Salud (WHO) y el Programa Ambiental de las Naciones Unidas (UNEP) en E.U., han establecido en NAAQS (National Ambient Air Quality Standard) algunos controles de salud que intentan dar un margen de seguridad y bienestar a los seres humanos. Una disminución en las concentraciones de ozono debe sugerir que los controles sobre las emisiones están teniendo efecto, mientras que la ausencia de tales tendencias sugiere la necesidad de adicionar más controles. Sin embargo, en algunos sitios, los análisis de datos de monitoreo ambiental del ozono, indican que el número de días donde las concentraciones de ozono exceden el nivel estándar de concentración promedio por hora, 120 partes por billón (ppb), varía ampliamente de un año al siguiente. En el reporte de Seinfeld (1991) se menciona que las medidas estadísticas principales de la EPA para tendencias de concentraciones de ozono son el inmediato anterior al máximo diario por hora para un año dado, y éste varía considerablemente de un año a otro. Una porción importante de esta variabilidad es atribuible no a cambios anuales en las emisiones de los precursores sino a fluctuaciones en el ambiente. A su vez reportan los mismos autores, que la variabilidad meteorológica provoca dificultad para comprender la identificación de la tendencia en las concentraciones de ozono ocasionando cambios en las cantidades y los tipos de emisiones precursoras en sus distribuciones geográficas.

Para monitorear las concentraciones de ozono urbana y medir tendencia es común usar dos tipos de mediciones: indicaciones de concentraciones (tales como el inmediato anterior al máximo diario por hora en concentraciones de ozono) y los indicadores de los umbrales (tal como el número de días sobre el cual hay un máximo diario por hora que exceda 120 ppb).

Wolff *et al.* (2001) en una investigación que analiza la calidad del aire urbano, reporta que detectar la tendencia del ozono y atribuirla a los controles de emisión es difícil por varias razones:

- ◇ Los cambios en las concentraciones de ozono por implementaciones en políticas de control de emisiones, son fácilmente disimulados por las condiciones meteorológicas. Estas últimas tienen conductas cuasi-periódicas sobre rangos de escalas de tiempo, por lo que hay razón para sospechar que existen ciclos naturales en las concentraciones de ozono que ocurren en escalas similares.
- ◇ El alcance temporal y espacial de las bases de datos de ozono, sus precursores y los lugares en que se evalúan las variables meteorológicas limitan las magnitudes de cambios que uno puede suponer detectar.
- ◇ El impacto de un programa de control de emisiones será más fácilmente distinguible en las emisiones precursoras que en el ozono. Sin embargo, los datos confiables sobre precursores de ozono son mucho más limitados que los datos de ozono con lo que contamos actualmente.

1.1. Antecedentes

En Seinfeld (1991), se discute toda una problemática que existe en la contaminación ambiental de ozono, basándose en los estudios hechos por diferentes expertos en áreas relacionadas y los análisis estadísticos que hasta la fecha se habían realizado. El mismo autor, Molina y L. (2004) en sus investigaciones identifican las tendencias de emisiones precursoras, por sus mediciones en:

- Emisiones de óxidos de nitrógeno y los VOC's acompañado del análisis del ozono.
- Las emisiones de VOC's antropogénicas y biogénicas,
- Tendencias normales de ozono en variaciones meteorológicas.

Cuando se involucra al medio ambiente es importante usar indicadores más robustos para las tendencias del ozono. Algunos investigadores como Chock (1989) han probado que el número de excedencias de una región, estadísticamente, puede ser reemplazado por el rango percentil (como el 95 percentil). Esto reduce la influencia aleatoria del estado del tiempo y ésta información podría apoyar el establecimiento de estrategias de reducción de emisiones. Los indicadores robustos que potencialmente han sido propuestos son: el máximo, el inmediato anterior al máximo diario, el 99 percentil, el 95 percentil, el 90 percentil, la media estacional y la mediana. Los

”indicadores de los umbrales” fueron el número de días o de horas en que se excede un umbral de concentración. Se encontró que los indicadores más robustos fueron aquellos que dependían menos de valores extremos. Posteriormente se compararon ambos indicadores, encontrándose que los indicadores de los umbrales fueron mucho más variables que los indicadores de concentraciones.

Seinfeld (1991) menciona que para medir tendencias en los métodos estadísticos que ajustan el ozono, las variables más usadas son: máximos diarios, promedio de máximos diarios de todas las estaciones de monitoreo en una red, el promedio anual de mediciones diarias, el promedio en la red de máximos diarios por hora de concentraciones de ozono y el valor más grande del 75 percentil de 3 meses con máximos diarios por hora en la red de monitores. La forma más simple para involucrar el efecto meteorológico puede ser determinando por las condiciones meteorológicas asociadas con aquellos días donde existen altas concentraciones de ozono, con lo que se pueden predecir altos, moderados y bajos días de ozono basándose en observaciones meteorológicas. Existen diversos factores importantes que impiden realizar buenas predicciones confiables sobre el ozono, uno de ellos es la variable de transporte que no siempre está relacionada al estado del tiempo, ya que las cantidades y tipos de precursores del ozono que son transportados dentro de un área impactada con contaminación ambiental, frecuentemente provoca los peores episodios de contaminación de ozono.

En el artículo de Wolff *et al.* (2001) se menciona que The North American Research Strategy for Tropospheric Ozone (NARSTO), ha hecho diversos reportes de ozono, NO_x y HCs (hidrocarburos no quemados), para investigar sistemáticamente las tendencias en Canadá y los Estados Unidos, evaluando la efectividad de sus políticas ambientales. En ellos se usan diversas técnicas estadísticas, algunas incluyendo variables meteorológicas con diversas escalas de medición y aplicándolas en diversos lugares.

En México hay intentos en muy diversos campos científicos e instituciones tratando de evaluar la contaminación por ozono (Bravo (1992); I.N.E.G.I. (1999); Molina y L. (2004)). Tratan de identificar las zonas con mayor riesgo para los habitantes de la Zona Metropolitana de la Ciudad de México (ZMCM) y se hacen comparaciones de tendencias entre algunas ciudades.

El trabajo de González (2000) muestra el uso de un método para detectar tendencias de los días en que el valor promedio de máximos diarios de ozono excedió las 0.24 ppm usando un Proceso Poisson no Homogéneo con parámetro de intensidad dado por una forma lineal en el tiempo con la información del Sistema de Monitoreo Atmosférico en los años de 1986 a 1999. Se hizo un análisis de la tendencia del ozono en las zonas representativas y se encontró que, en la zona suroeste su tendencia es creciente.

La investigación de Jaimes (2004), utiliza inicialmente la prueba estadística de Kolmogorov Smirnov para verificar normalidad en las observaciones. Se usaron pruebas no paramétricas como la Mann-Kendall para investigar tendencias en una serie de tiempo, se hicieron evaluaciones con diferentes indicadores (promedio de los 30 valores máximos en un año, el inmediato anterior al máximo diario y el cuarto valor más alto en tres años de datos). Se hacen diferentes cortes por periodos de tiempo como comparaciones entre días laborales y fines de semana, entre meses importantes del año contra meses no importantes, encontrando que en general la tendencia es significativamente decreciente en el ozono y que las concentraciones muy grandes han disminuido, pero no las concentraciones intermedias, pues son las más difíciles de abatir, lo que implica que las estrategias de control no han sido tan eficientes como se esperaba.

Otro trabajo con ese mismo propósito, que usa datos de sólo un sitio de monitoreo de la Ciudad de México con peores episodios de ozono lo hace Garfias *et al.* (2005), analiza tres modelos para pronosticar concentraciones de ozono sobre la base de 19 datos del valor medio semestral. El mejor ajuste encontrado en este trabajo es el modelo logarítmico para la información de 1996 a 2004. Finalmente, este trabajo sugiere reforzar los controles políticos y cambiar el tipo de gasolina en sus compuestos químicos para poder recuperar la calidad del aire de la Ciudad de México.

Desde hace años se han utilizado diversos enfoques estadísticos con combinaciones entre ellos para manejar la información en forma conveniente a sus intereses de investigación, haciendo variaciones en escalas de medición, usando o definiendo métricas para evaluar y modelar tendencias de concentraciones de ozono relacionadas con algún nivel de interés. En estos trabajos es común contar con mucha información de diversas estaciones de monitoreo que recopilan información de concentraciones de ozono por diversos intervalos de tiempo, así como las mediciones de variables ambientales. Esta información ayuda y complica a los investigadores en sus diseños de hipótesis, para establecer modelos empíricos que expliquen relaciones del comportamiento de la formación de ozono extremo. Posteriormente con esta información se pueden establecer modelos estadísticos para investigar la tendencia de las concentraciones de ozono extremo.

Uno de los trabajos tomados a manera de ejemplo es la investigación de Kuchenhoff y Thamerus (1996) que incluyó promedios diarios de concentraciones de ozono y de nitrógeno en la Ciudad de Munich. Se usaron varios modelos como: el modelo generalizado de valores extremos para máximos mensuales, la distribución Pareto Generalizada para las excedencias y el modelo de regresión logística para la respuesta binaria de las excedencias y como variables predictorias fueron las últimas concentraciones de ozono y las variables meteorológicas. Los resultados del modelo fueron difíciles de interpretar debido a la cantidad de variables.

Como segundo ejemplo está la investigación de Gao *et al.* (1996). Debido a que evaluar las tendencias del ozono es complicado por la variabilidad meteorológica, las concentraciones de ozono de Chicago fueron modeladas como un efecto anual más un modelo semi-paramétrico en las variables ambientales. La estimación de tendencias a largo plazo son importantes para evaluar la efectividad de las políticas ambientales. Tomando en cuenta el tamaño de la región hubo variabilidad en los niveles de ozono, por lo que se observó la relación espacial del ozono con la fuentes locales y la transportación del ozono. La modelación de datos ambientales fue compleja debido a la no linealidad y las interacciones de las variables. El modelo dio buenas predicciones de concentraciones de ozono usando la información meteorológica y del tiempo.

A continuación se propone un modelo sencillo basado sobre la Distribución Generalizada de Valores Extremos que utiliza la convergencia de la función cuantil a una distribución Normal, para una probabilidad de excedencia fija y alta. En el modelo se puede introducir sólo a la variable tiempo y/o alguna combinación de variables importantes para el investigador. La sencillez del procedimiento permite que cualquier individuo que tenga conocimientos mínimos del Modelo de Regresión Lineal Simple pueda construir intervalos de confianza, pruebas de hipótesis de los parámetros de interés y poder hacer investigaciones sobre tendencias de la variable dependiente. Finalmente, los resultados son muy congruentes con otros modelos reportados.

1.2. Objetivo General

En el presente trabajo se realiza modelación estadística en la Distribución Generalizada de Valores Extremos, con el fin de estimar tendencias en los niveles máximos de un contaminante.

Objetivos específicos:

- Desarrollar una metodología para el análisis de tendencia usando la Función Cuantil de un modelo Generalizado de valores Extremos.
- Justificar matemáticamente el uso de la Función Cuantil para poder realizar análisis de tendencia.
- Estudiar la propiedad de convergencia del estimador de la Función Cuantil usando simulación.
- Incorporar en el modelo Generalizado de Valores Extremos el parámetro de escala con algunas covariables meteorológicas, para estimar tendencias.
- Generar programas en R para realizar las estimaciones de los parámetros.
- Analizar un caso concreto de datos de ozono para investigar tendencias usando la metodología propuesta.

1.3. Estructura de la tesis

Esta tesis se organiza de la siguiente manera: en el capítulo 1, se presenta una revisión general de las medidas y problemas que se involucran en los temas de Contaminación Ambiental, así como su interés en la modelación estadística de concentraciones de ozono urbano, para evaluar y calificar las medidas de los planes de contingencia ambiental, encaminados a mejorar la calidad del aire de una ciudad. Finalmente, se presentan los objetivos y las metas del trabajo de investigación.

En el capítulo 2, se presentan los conceptos básicos acerca de la Teoría de Valores Extremos, Estimación paramétrica de la distribución Generalizada de Valores Extremos, algunas propiedades de los estimadores de máxima verosimilitud, los estimadores de momentos, algunas características del método de los umbrales, los intervalos de confianza y las pruebas de hipótesis.

En el capítulo 3, se hace revisión de la literatura sobre los modelos estadísticos que investigan tendencias de ozono, de la importancia del problema de contaminación por ozono y de los tres tipos de modelos estadísticos que generalmente analizan tendencias de ozono. En las últimas dos secciones 3.1.2 y 3.1.3, se presentan las propuestas de investigación para analizar tendencias: la primera, usa a la Función Cuantil para investigar tendencias y la segunda, el parámetro de escala se pone en combinación lineal de las variables meteorológicas para incorporar el efecto de covariables ambientales al modelo propuesto y poder así analizar tendencias.

En el capítulo 4, se propone una metodología para analizar problemas de valores extremos; es la contribución que se presenta en este trabajo y se basa en el resultado 4.1 que se refiere a la convergencia asintótica de la función cuantil a una distribución Normal estándar. También se presenta la demostración de la convergencia de la función cuantil y algunas simulaciones de la convergencia a la distribución normal.

En el capítulo 5, la metodología se aplica a un caso real, la información generada por las mediciones de contaminación de la Ciudad de México. La primera propuesta se trabaja con las 20 estaciones de monitoreo de la ZMCM. La segunda propuesta involucra sólo a la estación Plateros de la ZMCM, debido a que esta estación es una de las dos estaciones con peores emisiones de ozono diario que presentan información de 1986 a 2005. De donde resulta interesante poder aplicar el procedimiento usando la información de las variables ambientales y las observaciones de ozono para poder realizar pronósticos de tendencia de niveles altos de ozono.

Por último, en el capítulo 6, se presenta una discusión de los resultados obtenidos en este trabajo, las conclusiones y posibles investigaciones futuras.

En este trabajo se realizaron algunos programas usando el software R y algunas de las librerías de R llamadas *ismev* y *evir*.

Capítulo 2

Distribución de Valores Extremos

Introducción

La teoría de valores extremos es una de las disciplinas de la estadística que se ha desarrollado rápidamente en pocos años. Algunos investigadores como Galambos (1978), Davison y Smith. (1990) y Johnson *et al.* (1995), hacen un recuento histórico de los avances, fundamentos y aplicaciones de la Teoría de Valores Extremos. Con esta teoría se modelan eventos de poca frecuencia y probabilidades casi siempre con impactos desastrosos. La estadística de valores extremos trata con estimaciones de cantidades que están relacionadas con la cola de la distribución con la que se modelan los eventos. Esta teoría se comenzó a aplicar en el área de la hidrología, específicamente en el análisis de inundaciones, pues es importante tener estimaciones de los promedios de niveles de agua que exceden un umbral dado en un periodo de 100 años (Reiss y Thomas, 2001). Otras aplicaciones interesantes son en el análisis de ráfagas de viento, contaminación de aire, corrosión de materiales, desastres de huracanes, altas concentraciones de ozono, tamaños fuera de lo común en reclamaciones e ingresos, estimaciones en la estructura de diseños ambientales y en la construcción de puentes y presas.

En el artículo de Chavez y Embrechts (2004) se presentan teorías y aplicaciones de cómo una aseguradora debe compensar la pérdida debida al pago de reclamaciones de políticas de tenencia por medio de un precio justo. Un investigador en seguros debe estar interesado en estimar el precio neto, que es la media de todos los reclamos de un individuo, o más general, estimar un portafolio de riesgos. El número total de reclamos dependerá del tamaño de reclamos (valor de la póliza) y la frecuencia de los reclamos dentro de un periodo dado.

En el trabajo de Reiss y Thomas (2001) ejemplifica con un problema de ingeniería, cuyo interés es analizar la corrosión en láminas. Las corrosiones en forma de hoyos que afectan negativamente la estructura metálica de los tanques, tubos, cascos de barcos y de buques. En este caso, el análisis de valores extremos se aplica debido a que los hoyos profundos son considerados de interés primario porque pueden causar daños mayores.

Reiss y Thomas y otros autores como Finkenstadt y Rootzen (2004) mencionan más ejemplos en los que se está aplicando la Teoría de Valores extremos a las áreas de ciencia de alimentos, a las predicciones de tráfico en telecomunicaciones, a la termodinámica de terremotos y en las estimaciones de cambios climatológicos. Piegorsch *et al.* (1998), Lou *et al.* (2001), Porter *et al.* (2001) y Wolff *et al.* (2001) hacen una revisión de la literatura estadística sobre las estimaciones meteorológicas de ozono troposférico.

De los trabajos anteriores se puede resumir que, variadas metodologías se manejan con datos de diversos tipos, con variadas mediciones y con características socio-políticas propias de los países o lugares donde se efectúa la investigación, pues depende de la conciencia colectiva de cada ciudad ante la salud humana.

Lou *et al.* (2001) clasifica los diversos estudios estadísticos que modelan las concentraciones de máximos de ozono diario en cuatro grandes áreas de investigación: modelos basados en la regresión, enfoques de valores extremos, redes neuronales y modelos de espacio tiempo.

Concluyendo: existen trabajos con distintas metodologías que dan contribuciones al desarrollo y análisis de la tendencia de los niveles altos de ozono troposférico.

En este capítulo se hace una breve revisión de la literatura sobre la teoría de los valores extremos, su distribución generalizada, la estimación paramétrica, y las propiedades de los estimadores máximo verosímiles y de momentos. También se trata de la estimación por intervalo y pruebas de hipótesis.

2.1. Función de Valores Extremos

Diversos problemas en la vida diaria están relacionados con la teoría de valores extremos en donde la solución matemática para evitar riesgos de catástrofes, está en términos de la observación mayor o la observación menor obtenida en la recopilación de la información (Galambos, 1978). Por ejemplo, en Alemania se hacen investigaciones industriales sobre la construcción de trenes de alta velocidad, con estructura formada de materiales ligeros que soporten sistemas dinámicos de vientos como son las tormentas tropicales fuertes que ocurren en las trayectorias cercanas a las costas donde se ubican las estaciones del tren (Payer, 2006). A continuación se da una pequeña revisión sobre la teoría de valores extremos y la estructura estadística en que se basa.

Sea X_1, \dots, X_n una muestra de variables aleatorias independientes con idéntica función de distribución F , con límite superior $x_F = \sup\{x \in \mathbb{R} : F(x) < 1\}$. La función de distribución del máximo $M_n = \max\{X_1, \dots, X_n\}$ es

$$F_{M_n}(x) = P(M_n \leq x) = P(X_1 \leq x, \dots, X_n \leq x) = [F(x)]^n = F^n(x).$$

En la práctica F suele ser desconocida y por lo tanto también F_{M_n} es desconocida. Para encontrar un resultado más general es natural considerar el comportamiento asintótico de M_n . Así cuando $n \rightarrow \infty$, F^n es una distribución degenerada que converge al límite superior x_F . Por lo cual el crecimiento de F^n tiene que ser ajustado para evitar degeneración. En analogía al teorema del límite central hay que seleccionar una transformación lineal adecuada, esto se verá en la sección 2.2.

Por otro lado, la estadística extrema mínima es obtenida del caso anterior escribiendo $\min X_i = -\max(-X_i)$ con $i \leq n$, de donde

$$\begin{aligned} F_{\min X_i}(x) &= P(\min X_i \leq x) = 1 - P(X_1 > x, \dots, X_n > x) = \\ &= 1 - [P(X_i > x)]^n = 1 - [1 - F(x)]^n. \end{aligned}$$

Por lo cual, con $i \leq n$,

$$F_{\min X_i}(x) = 1 - [1 - F(x)]^n.$$

2.1.1. La α -parametrización de la GVE

Reiss y Thomas (2001) dan una lista de tres submodelos diferentes para la distribución de valores extremos (VE) estandarizados cuando se varía el parámetro de forma α . A continuación se definen las distribuciones Gumbel, Fréchet y Weibul.

La función de distribución tipo Gumbel, $\alpha = 0$

$$(VE = 0) : G_0(x) = \exp(-e^{-x}), \quad -\infty < x < \infty. \quad (2.1)$$

La distribución Fréchet, $\alpha > 0$

$$(VE = 1) : G_{1,\alpha}(x) = \begin{cases} \exp(-x^{-\alpha}), & x \geq 0; \\ 0, & x < 0. \end{cases} \quad (2.2)$$

La distribución Weibul, $\alpha < 0$

$$(VE = 2) : G_{2,\alpha}(x) = \begin{cases} \exp(-(-x)^{-\alpha}), & x \leq 0; \\ 1, & x > 0. \end{cases} \quad (2.3)$$

Si X es una variable aleatoria con función de distribución F entonces $\mu + \sigma X$ tiene una función de distribución $F_{\mu,\sigma}(x) = F(\frac{x-\mu}{\sigma})$ donde μ y σ son los parámetros de localización y de dispersión. Con lo anterior a los modelos de valores extremos se les adiciona los parámetros de localización y de escala, de esta manera.

$$G_{0,\mu,\sigma}(x) = \exp(-e^{-\frac{x-\mu}{\sigma}}) \quad -\infty < x < \infty. \quad (2.4)$$

$$G_{1,\alpha,\mu,\sigma}(x) = \begin{cases} \exp(-(\frac{x-\mu}{\sigma})^{-\alpha}), & x \geq 0; \\ 0, & x < 0. \end{cases} \quad (2.5)$$

$$G_{2,\alpha,\mu,\sigma}(x) = \begin{cases} \exp(-(-\frac{x-\mu}{\sigma})^{-\alpha}), & x \leq 0; \\ 1, & x > 0. \end{cases} \quad (2.6)$$

2.1.2. Media y varianza de una variable aleatoria con distribución VE

Los momentos j -ésimos de una variable aleatoria Weibull, se obtienen de la expresión de la función Generadora de Momentos

$$m_{j,G_{2,\alpha}} = E(X^j) = (-1)^j \Gamma(1 - \frac{j}{\alpha}).$$

En particular la esperanza es $E(X) = -\Gamma(1 - \frac{1}{\alpha})$.

Los momentos j -ésimos de una variable aleatoria Fréchet, se obtienen de la expresión

$$m_{j,G_{1,\alpha}} = E(X^j) = \Gamma(1 - \frac{j}{\alpha}) \text{ si } \alpha > j.$$

En particular se tiene la $E(X) = \Gamma(1 - \frac{1}{\alpha})$ si $\alpha > 1$.

En el caso de la varianza de las distribuciones Fréchet y Weibull estas coinciden y está dada por $Var(X) = \Gamma(1 - \frac{2}{\alpha}) - \Gamma^2(1 - \frac{1}{\alpha})$ con $\frac{1}{\alpha} < \frac{1}{2}$.

En el caso de la distribución Gumbel, la media y la varianza son

$$E(X) = \mu + \gamma\sigma, \text{ y } Var(X) = \frac{\sigma^2\pi^2}{6},$$

con $\gamma = 0.57722..$ que es conocida como la constante de Euler. Para mayor información sobre estos temas ver Reiss y Thomas (2001).

2.2. Estimación paramétrica de la Distribución Generalizada de Valores Extremos

La distribución Generalizada de Valores Extremos (GVE) unifica dentro de una única representación paramétrica a las tres diferentes familias de distribución VE (Reiss y Thomas (2001), Coles (2004)).

La expresión de la distribución GVE estandarizada incorpora el parámetro de localización μ , el parámetro de escala σ , y el parámetro de ε , la expresión esta dada por:

$$G_{\mu,\sigma,\varepsilon}(z) = \exp\{-[1 + \varepsilon(\frac{z-\mu}{\sigma})]^{-\frac{1}{\varepsilon}}\} \text{ si } 1 + \varepsilon(\frac{z-\mu}{\sigma}) > 0 \quad (2.7)$$

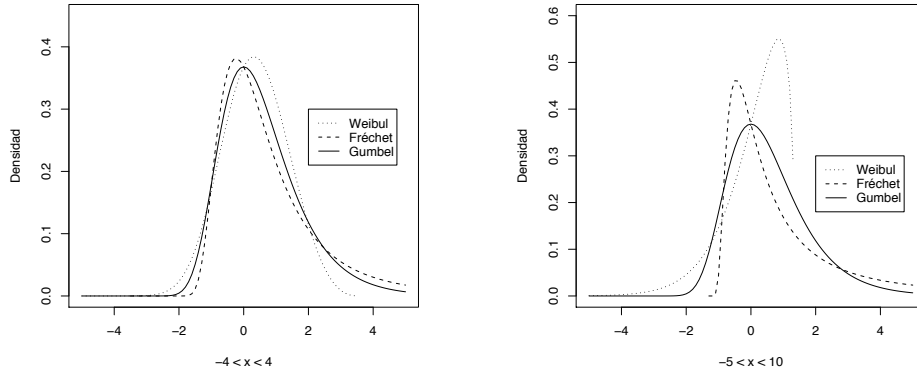
$$G_{\mu,\sigma,0}(z) = \exp(-\exp^{-\frac{z-\mu}{\sigma}}) \text{ en el caso de } \varepsilon = 0 \quad (2.8)$$

con $\theta = (\mu, \sigma, \varepsilon) \in \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}$.

La función de densidad de probabilidad de una GVE es

$$g_{\mu,\sigma,\varepsilon}(z) = \begin{cases} \exp\{-[1 + \varepsilon(\frac{z-\mu}{\sigma})]^{-\frac{1}{\varepsilon}}\} \frac{1}{\sigma} [1 + \varepsilon(\frac{z-\mu}{\sigma})]^{-\frac{1}{\varepsilon}-1} & \text{si } 1 + \varepsilon(\frac{z-\mu}{\sigma}) > 0; \\ \exp(-\exp^{-\frac{z-\mu}{\sigma}}) \exp^{-\frac{z-\mu}{\sigma}} \frac{1}{\sigma} & \text{en el caso de } \varepsilon = 0. \end{cases} \quad (2.9)$$

En la Figura 2.1 se presentan algunas gráficas de las densidades de valores extremos (usando el paquete evir de McNeil and Stephenson (2004)).



(a) Funciones de densidad: Gumbel ($\mu = 0, \sigma = 1, \varepsilon = 0n$), Fréchet $GEV(\mu = 0, \sigma = 1, \varepsilon = 7/25)$ y Weibul $GEV(\mu = 0, \sigma = 1, \varepsilon = -7/25)$.

(b) Funciones de densidad: Gumbel ($\mu = 0, \sigma = 1, \varepsilon = 0n$), Fréchet ($\mu = 0, \sigma = 1, \varepsilon = 3/4$) y Weibul ($\mu = 0, \sigma = 1, \varepsilon = -3/4$).

Figura 2.1: Gráfica de las funciones de densidad GVE.

Los autores Reiss y Thomas (2001), Finkenstadt y Rootzen (2004) y Payer (2006) establecen que los valores del parámetro de forma determinan si la distribución tiene un límite superior. Las distribuciones asociadas con $\varepsilon > 0$ son llamadas de tipo Fréchet, no tienen límite superior y tienen la cola pesada como las distribuciones Pareto, Cauchy, t-student y las mixtas. Cuando se tiene $\varepsilon = 0$, se interpreta como $\varepsilon \rightarrow 0$, la cola es moderadamente pesada y la distribución resultante es de tipo Gumbel. Finalmente, cuando $\varepsilon < 0$ se tiene una distribución de cola ligera con límite superior $x_F = \mu - \frac{\sigma}{\varepsilon}$ y corresponde a la distribución Weibull.

De acuerdo a resultados de Fisher y Tippet (1928) y Gnedenko (1943), se establece que si existen constantes $c_n > 0, d_n \in \mathbb{R}$ tales que para $M_n = \max\{X_1, \dots, X_n\}$, cuando $n \rightarrow \infty$,

$$P\left(\frac{M_n - d_n}{c_n} \leq x\right) \rightarrow G(x)$$

para alguna función de distribución no degenerada G , entonces G pertenece a la familia GVE. Ésta familia tiene las siguientes propiedades para cada miembro de G (Pickands (1975);Payer (2006)),

1. $G(ax + b) = G^*(x)$ con $a > 0$ y $b \in \mathbb{R}$, G y $G^*(x)$ pertenecen a la misma familia.
2. Si G pertenece a un miembro de la familia GVE, entonces para una n existe

$a_n > 0$ y b_n tal que se cumple $G^n(a_n x + b_n) = G(x)$.

Coles y Dixon (1999) dicen que una forma natural de realizar modelación estadística es tener observaciones x_1, x_2, \dots , de F para formar bloques de máximos, estos son tomados convenientemente por año y se estima la familia GVE por bloques máximos. Suponiendo que el modelo de máximos bloques es correcto para la información z_1, z_2, \dots, z_k . La estimación es juzgada en términos de las estimaciones aproximadas de los cuantiles extremos, estos cuantiles son usualmente requeridos para el análisis de valores extremos.

Dado un valor $0 < \alpha = 1 - p < 1$, se define el α -cuantil de la función de distribución $F_X(x; \theta)$ como el valor q_α tal que $\alpha = P(X \leq q_\alpha) = F_X(q_\alpha; \theta) \Leftrightarrow F_X^{-1}(\alpha; \theta) = q_\alpha$. A q_α se le llama Función Cuantil de $F_X(x; \theta)$. Para valores muy grandes de q_α , se considera que hay un $(1-\alpha)\%$ de observaciones más altas.

En el caso de la distribución GVE (2.7 y 2.8), la función Cuantil para un umbral alto en las observaciones, con α fijo (Coles, 2004).

$$\alpha = P(Z \leq q_\alpha) = G_Z(q_\alpha) \Leftrightarrow$$

$$G_Z^{-1}(\alpha; \theta) = Z_\alpha(p) = q_\alpha$$

con

$$Z_\alpha(p) = \begin{cases} \mu - \frac{\sigma}{\varepsilon}(1 - y_p^{-\varepsilon}), & \varepsilon \neq 0; \\ \mu - \sigma \log(y_p), & \varepsilon = 0. \end{cases} \quad (2.10)$$

con $y_p = -\log(1 - p)$ con $0 \leq p \leq 1$.

A z_p se le conoce como el nivel de retorno asociado al periodo de retorno $\frac{1}{p}$. Se espera que el nivel z_p sea excedido en promedio por algún máximo anual con probabilidad p . En la gráfica del nivel de retorno, z_p es graficada contra y_p sobre una escala logarítmica, o z_p es graficada sobre $\log(y_p)$. La gráfica tiene forma lineal cuando $\varepsilon = 0$, si $\varepsilon < 0$, la gráfica resultante tiene forma cóncava con límites asintóticos cuando $p \rightarrow 0$ en $\mu - \frac{\sigma}{\varepsilon}$ y si $\varepsilon > 0$, la gráfica tiene forma cóncava y no tiene fronteras. Esta gráfica es simple de interpretar por la reducción de escala en la cola de la distribución pues clarifica la extrapolación. Esta gráfica se usa para validar y analizar los modelos.

2.3. Estimadores de Máxima Verosimilitud

La Función de verosimilitud de m variables aleatorias Z_1, \dots, Z_m se define como la densidad conjunta de m variables, es decir, $f(z_1, \dots, z_m; \mu, \sigma, \varepsilon)$ es una función de

μ, σ, ε . La función de verosimilitud esta definida como

$$L(\mu, \sigma, \varepsilon; z_1, \dots, z_m) = \prod_i^m f(z_i; \mu, \sigma, \varepsilon)$$

La factorización en la anterior expresión se debe a la independencia de las observaciones. Conveniente tomar el logaritmo de la verosimilitud,

$$l(\mu, \sigma, \varepsilon) = \log L(\mu, \sigma, \varepsilon) = \sum_i^m \log f(z_i; \mu, \sigma, \varepsilon). \quad (2.11)$$

Smith en 1985, prueba que los estimadores de máxima verosimilitud existen siempre que $\varepsilon > -1$ y son regulares siempre que $\varepsilon > -0.5$ pues tienen propiedades asintóticas usuales, cuando $-1 < \varepsilon < -0.5$ los estimadores son fácilmente encontrados aunque no tienen propiedades asintóticas y finalmente, si $\varepsilon < -1$ los estimadores de máxima verosimilitud no existen.

$$l(\mu, \sigma, \varepsilon) = -m \log \sigma - (1 + 1/\varepsilon) \sum_i^m \log \left[1 + \varepsilon \left(\frac{z_i - \mu}{\sigma} \right) \right] - \sum_i^m \left[1 + \varepsilon \left(\frac{z_i - \mu}{\sigma} \right) \right]^{-1/\varepsilon} \quad (2.12)$$

dado que $1 + \varepsilon \left(\frac{z_i - \mu}{\sigma} \right) > 0$, para $i = 1, \dots, m$.

Cuando $\varepsilon = 0$ se usa el límite Gumbel de la distribución GVE. Esto lleva a el logaritmo de la verosimilitud

$$l(\mu, \sigma) = -m \log \sigma - \sum_i^m \left(\frac{z_i - \mu}{\sigma} \right) - \sum_i^m \exp \left\{ -\frac{z_i - \mu}{\sigma} \right\} \quad (2.13)$$

La maximización de las ecuaciones (2.12) y (2.13) con respecto al vector paramétrico $(\mu, \sigma, \varepsilon)$ lleva a los estimadores máximo verosímiles con respecto a la familia GVE. La solución no es analítica, aunque para algunos conjuntos de datos la maximización es directa usando algoritmos numéricos (Casella y Berger (2002); Coles (2004)).

2.4. Estimadores de Momentos para los parámetros μ y σ

En esta sección se encuentran los estimadores de momentos de la DGVE para los parámetros de localización y de escala. Se inicia obteniendo los estimadores de la distribución Gumbel que no tiene parámetro de forma y posteriormente se encuentran los casos de las distribuciones Weibul y Fréchet, suponiendo que tienen parámetro de forma conocido.

Un procedimiento para obtener estos estimadores consiste en usar los resultados de las distribuciones de valores extremos dados por ejemplo en el libro de Reiss (Reiss y Thomas, 2001).

2.4.1. Procedimiento para obtener los estimadores de momentos de la DGVE

1. La función de distribución Gumbel tiene dos parámetros (sección 2.1), donde la

$E(X) = \mu + \sigma\gamma$ y $E(X^2) = \frac{\sigma^2\pi^2}{6} + (\mu + \sigma\gamma)^2$ con $\gamma = 0.57722..$
Igualando los momentos muestrales con los poblacionales, se tiene

$$\begin{aligned}\bar{X} &= \mu + \sigma\gamma \quad y \\ \frac{\sum_i X_i^2}{n} &= \frac{\sigma^2\pi^2}{6} + (\mu + \sigma\gamma)^2.\end{aligned}$$

Resolviendo el sistema de ecuaciones, se tiene que

$$\begin{aligned}\tilde{\mu}_{mo} &= \bar{X} - \sqrt{\frac{\sum_i 6(X_i^2 - n\bar{X}^2)}{n\pi^2}} \gamma \quad y \\ \tilde{\sigma}_{mo} &= \sqrt{\frac{\sum_i 6(X_i^2 - n\bar{X}^2)}{n\pi^2}} \gamma.\end{aligned}$$

2. La función de la distribución Weibull tiene dos parámetros (sección 2.1), y la $E(X) = -\Gamma(1 - \frac{1}{\alpha})$ y $V(X) = \Gamma(1 - \frac{2}{\alpha}) - \Gamma^2(1 - \frac{1}{\alpha})$ si $\frac{1}{\alpha} < \frac{1}{2}$. Por lo cual, el segundo momento poblacional, es $E(X^2) = \Gamma(1 - \frac{2}{\alpha})$. Igualando los momentos muestrales con los poblacionales, se tiene

$$\begin{aligned}\bar{X} &= -\Gamma(1 - \frac{1}{\alpha}) \quad y \\ \frac{\sum_i X_i^2}{n} &= \Gamma(1 - \frac{2}{\alpha}).\end{aligned}$$

Sea Z una variable aleatoria que se distribuye como GEV en la forma (2.3) con $\alpha < 0$, y que cumple la siguiente relación entre variables aleatorias

$$\begin{aligned}-X &= 1 + \frac{\varepsilon}{\sigma}(Z - \mu) \quad y \quad \varepsilon = \frac{1}{\alpha} \Leftrightarrow \\ Z &= (-X - 1)\frac{\sigma}{\varepsilon} + \mu \quad y \quad \varepsilon = \frac{1}{\alpha}\end{aligned}$$

De donde

$$E(Z) = \frac{\sigma}{\varepsilon}(-1 + \Gamma(1 - \frac{1}{\alpha})) + \mu \quad y \quad V(Z) = \frac{\sigma^2}{\varepsilon^2}[\Gamma(1 - \frac{2}{\alpha}) - \Gamma^2(1 - \frac{1}{\alpha})],$$

Por lo cual, finalmente se tiene

$$E(Z^2) = \frac{\sigma^2}{\varepsilon^2}[\Gamma(1 - \frac{2}{\alpha}) - \Gamma^2(1 - \frac{1}{\alpha})] + [\frac{\sigma}{\varepsilon}(-1 + \Gamma(1 - \frac{1}{\alpha})) + \mu]^2.$$

Para encontrar los estimadores por el método de momentos se igualan los momentos muestrales y poblacionales.

$$\bar{X} = \frac{\sigma}{\varepsilon}(-1 + \Gamma(1 - \frac{1}{\alpha})) + \mu$$

$$\frac{\sum_i X_i^2}{n} = \frac{\sigma^2}{\varepsilon^2}[\Gamma(1 - \frac{2}{\alpha}) - \Gamma^2(1 - \frac{1}{\alpha})] + [-1 + \frac{\sigma}{\varepsilon}(\Gamma(1 - \frac{1}{\alpha})) + \mu]^2.$$

De donde la solución es:

$$\tilde{\sigma}_{mo} = \sqrt[n]{\frac{\sum_i X_i^2}{n} - \bar{X}^2} \frac{|\varepsilon|}{\sqrt{\Gamma(1 - \frac{2}{\alpha}) - \Gamma^2(1 - \frac{1}{\alpha})}} \quad y$$

$$\tilde{\mu}_{mo} = \bar{X} - \sqrt{\frac{\sum_i X_i^2}{n} - \bar{X}^2} \frac{(-1 + \Gamma(1 - \frac{1}{\alpha}))}{\sqrt{\Gamma(1 - \frac{2}{\alpha}) - \Gamma^2(1 - \frac{1}{\alpha})}}.$$

Para la distribución Weibul $\varepsilon > 0$, se tienen los estimadores de momentos

$$\tilde{\sigma}_{mo} = \sqrt[n]{\frac{\sum_i X_i^2}{n} - \bar{X}^2} \frac{(-\varepsilon)}{\sqrt{\Gamma(1 - \frac{2}{\alpha}) - \Gamma^2(1 - \frac{1}{\alpha})}}$$

$$\tilde{\mu}_{mo} = \bar{X} - \sqrt{\frac{\sum_i X_i^2}{n} - \bar{X}^2} \frac{(-1 + \Gamma(1 - \frac{1}{\alpha}))}{\sqrt{\Gamma(1 - \frac{2}{\alpha}) - \Gamma^2(1 - \frac{1}{\alpha})}}.$$

3. Para el caso de la función de distribución Fréchet $\varepsilon > 0$, con $\alpha > 0$
 Se tiene que $E(X) = \Gamma(1 - \frac{1}{\alpha})$ con $\alpha > 1$ y $V(X) = \Gamma(1 - \frac{2}{\alpha}) - \Gamma^2(1 - \frac{1}{\alpha})$ si $\frac{1}{\alpha} < \frac{1}{2}$.
 Haciendo lo mismo que el caso anterior, se llega a obtener los mismos estimadores de momentos de la variable Z .

$$\tilde{\sigma}_{mo} = \sqrt[n]{\frac{\sum_i X_i^2}{n} - \bar{X}^2} \frac{\varepsilon}{\sqrt{\Gamma(1 - \frac{2}{\alpha}) - \Gamma^2(1 - \frac{1}{\alpha})}}$$

$$\tilde{\mu}_{mo} = \bar{X} - \sqrt{\frac{\sum_i X_i^2}{n} - \bar{X}^2} \frac{(-1 + \Gamma(1 - \frac{1}{\alpha}))}{\sqrt{\Gamma(1 - \frac{2}{\alpha}) - \Gamma^2(1 - \frac{1}{\alpha})}}.$$

2.5. Estimadores de orden para el parámetro de forma

En las distribuciones VEG el parámetro de forma es de gran interés debido a que determina el comportamiento extremo de la distribución. En la práctica se cuenta con una muestra finita y en algunos casos, no se cuenta con la observación más grande para inferir. Por lo cual se asume la distribución asintótica de la k -ésima observación (k es grande pero no tan grande como el tamaño de n) para poder estimar al parámetro. La selección de k no es una tarea sencilla pues muchos de los métodos propuestos son muy controvertidos por los investigadores.

Supongamos que tenemos una muestra de variables aleatorias independientes con X_1, \dots, X_n donde $(X_{1:n} \geq X_{2:n} \geq \dots \geq X_{n:n})$ son los estadísticos de orden en forma descendente) de una distribución F desconocida y continua.

En seguida se presentan algunos estimadores del parámetro de forma más comunes.

El **Estimador de Pickands** (1975) es el primero que sugiere un estimador para el parámetro de $\varepsilon \in \mathbb{R}$ de una DGVE, éste estimador ya se ha manejado en distribuciones de valores extremos por Rosen y Weissman en 1996 (Rosen y Weissman, 1996).

$$\hat{\varepsilon} = \frac{1}{\ln 2} \ln \left(\frac{X_{(k/4):n} - X_{(k/2):n}}{X_{(k/2):n} - X_{k:n}} \right). \quad (2.14)$$

Este estimador se basa en el método de estimación percentil para las diferencias entre los estadísticos de orden superior. Bajo ciertas condiciones de regularidad, tiene propiedades de consistencia débil y fuerte, así como de normalidad asintótica. La consistencia depende únicamente del comportamiento de k , mientras que la normalidad depende de condiciones de segundo orden sobre la definición de la distribución F . Una característica particular del estimador de Pickands, es el hecho de que la observación más grande no es explícitamente usada en la estimación. Este hecho puede justificarse a que la observación más grande puede adicionar mucha incertidumbre (Panaretos y Tsourti, 2003).

El **Estimador de Hill** (1975) es sólo para la distribución Fréchet $\varepsilon > 0$. El estimador está dado por

$$\widehat{\varepsilon}_H = \frac{1}{k} \sum_i^k \ln X_{1:n} - \ln X_{k+1:n}.$$

Este estimador se basa en la noción de los estimadores de máxima verosimilitud condicionales. Los investigadores han estudiado este estimador y le dan cualidades de tener consistencia débil y fuerte, así como tener normalidad asintótica si las muestras son aleatorias e independientes (Panaretos y Tsourti, 2003).

El **Estimador de Momentos** es una adaptación del estimador de Hill, para obtener consistencia para todos los valores de $\varepsilon \in \mathbb{R}$, ha sido propuesto por Dekkers y Haan (1989). El estimador está dado por

$$\widehat{\varepsilon}_M = M_1 + 1 - \frac{1}{2} \left(1 - \frac{M_1^2}{M_2}\right)^{-1} \text{ donde } M_j = \frac{1}{k} \sum_i^k (\ln X_{i:n} - \ln X_{k+1:n})^j$$

con $j = 1, 2$. Dekkers y Haan (1989), han probado que estos estimadores poseen consistencia débil y fuerte, tanto como normalidad asintótica (Panaretos y Tsourti, 2003).

El **Estimador de de Radio de Momentos**. Tomando sólo el caso cuando $\varepsilon > 0$, la principal desventaja del estimador de Hill es que puede ser sesgado, dependiendo de la conducta de segundo orden de la función de distribución F . Danielsson *et al.* (1996), basándose sobre cuestiones asintóticas de la expansión de segundo orden de la distribución consiguen obtener un estimador insesgado del estimador del Hill, que se define como

$$\widehat{\varepsilon}_{RM} = \frac{1}{2} \frac{M_2}{M_1}.$$

Este estimador tiene un error cuadrático asintótico menor que el estimador de Hill (cuando se evalúa sobre el mismo umbral, es decir la misma k) (Panaretos y Tsourti, 2003).

El **Estimador de de Radio de Pen's y W**. Un estimador relacionado con el estimador de momentos $\widehat{\gamma}_M$ es el estimador de Peng's sugerido por Deheuvels *et al.* (1997),

$$\widehat{\varepsilon}_L = \frac{M_2}{2M_1} + 1 - \frac{1}{2} \left[1 - \frac{M_1^2}{M_2}\right]^{-1}.$$

Este estimador fue construido para reducir el sesgo del estimador de momentos.

Otro estimador sugerido por los mismos autores es el **Estimador W**.

$$\widehat{\varepsilon}_W = 1 - \frac{1}{2} \left[1 - \frac{L_1^2}{L_2} \right]^{-1}.$$

donde $L_j = \frac{1}{k} \sum_i^k (X_{i:n} - X_{(k+1):n})^j$ $j = 1, 2$

El estimador $\widehat{\varepsilon}_L$ es consistente para algún $\gamma \in \mathbb{R}$ (bajo condiciones normales), mientras el estimador $\widehat{\varepsilon}_W$ es consistente únicamente para $\gamma < 1/2$. Más aún, bajo condiciones apropiadas sobre F y $k(n)$, $\widehat{\varepsilon}_L$ es asintóticamente normal, y para $\widehat{\varepsilon}_W$ la normalidad se tiene para $\gamma < 1/4$

2.6. Intervalos de confianza

La inferencia realizada por medio de un estimador puntual varía de una muestra a otra, por lo cual es necesario usar otro mecanismo que permita estimar los valores posibles del estimador puntual con un cierto grado de confiabilidad. En aras de la precisión y claridad en esta sección se presentan las definiciones de algunos conceptos importantes.

Definición 2.1 *Un intervalo aleatorio es un intervalo en el cual al menos uno de sus extremos es aleatorio.*

Definición 2.2 *Un estimador por intervalo es una variable aleatoria cuyos extremos son estadísticas. Una realización de un estimador por intervalo, es un intervalo de confianza, cuyos coeficientes corresponden a la probabilidad de que el estimador por intervalo contenga al parámetro desconocido.*

Definición 2.3 *Sea $Q = q(\underline{x}; \theta)$ una variable aleatoria, se dice que Q se llama Pivote si:*

1. *La distribución de Q no depende de θ y está completamente especificada.*
2. *La función $q(\underline{x})$ es invertible como función de θ . Es decir, para valores q_1 y q_2 tales que $\alpha = P(q_1 < q(\underline{x}; \theta) < q_2)$ para $\alpha \in [0, 1]$ existen las estadísticas $t_1(\underline{x})$ y $t_2(\underline{x})$ tales que:
 $\alpha = P(t_1(\underline{x}) < \theta < t_2(\underline{x}))$ $(t_1(\underline{x}), t_2(\underline{x}))$ es un estimador para el intervalo que contiene a θ con probabilidad α .*

2.7. Intervalo de Confianza para la media de una distribución normal cuando σ es desconocida

Sea x_1, \dots, x_n una muestra aleatoria con distribución normal con parámetros μ y σ^2 desconocida, entonces un intervalo al $(1 - \alpha)100\%$ de confianza para el parámetro μ esta dado por

$$\bar{x} - t_{\alpha/2} \frac{s_{n-1}}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2} \frac{s_{n-1}}{\sqrt{n}}$$

en donde $t_{\alpha/2}$ es el valor de la distribución t -student con $v = n - 1$ grados de libertad.

La inferencia en una estimación es establecer que conjunto $\theta \in C$ y $C = C(\underline{x}) \subset \Theta$ es un conjunto determinado por el valor de los datos $\underline{X} = \underline{x}$ que es observado. Si θ es un número real el conjunto C es un intervalo.

Para un valor de $0 < \alpha < 1$, es posible encontrar los límites. El intervalo de confianza al $(1 - \alpha)100\%$ de confianza para θ_0 al intervalo en el se encuentra el parámetro θ corresponde a una realización de la muestra

Coles (2004) establece que para estimar la media de una población con realizaciones independientes x_1, \dots, x_n con σ^2 desconocida. Dice, que la media muestral es un estimador insesgado para la μ con $var(\bar{x}) = \sigma^2/n$. Un intervalo de confianza asintótico puede ser obtenido, por el teorema central límite

$$\bar{x} - \mu \simeq N(0, \sigma^2/n)$$

$\bar{x} - \mu$ es un pivote. Realizando operaciones básicas se llega

$$Pr\{\bar{x} - z_{\alpha/2}\sigma/\sqrt{n} \leq \bar{x} - \mu \leq z_{\alpha/2}\sigma/\sqrt{n}\} = 1 - \alpha$$

donde $z_{\alpha/2}$ es el cuantil $(1 - \alpha/2)$ de la distribución normal estándar. Esta expresión es equivalente a

$$Pr\{\bar{x} - z_{\alpha/2} \sigma/\sqrt{n} \leq \mu \leq \bar{x} + z_{\alpha/2} \sigma/\sqrt{n}\} = 1 - \alpha \quad (2.15)$$

Por lo que

$$(\bar{x} - z_{\alpha/2} \sigma/\sqrt{n}, \bar{x} + z_{\alpha/2} \sigma/\sqrt{n})$$

es un intervalo asintótico para μ a la confianza $(1 - \alpha)$. Al tomar una muestra suficientemente grande esta aproximación no se afecta si se reemplaza σ desconocido por el estimador muestral s , llevando al intervalo de confianza

$$(\bar{x} - z_{\alpha/2} s/\sqrt{n}, \bar{x} + z_{\alpha/2} s/\sqrt{n})$$

2.8. Normalidad Asintótica de los Estimadores Máximo Verosímiles

Los estimadores de máxima verosimilitud tienen beneficios en la estimación paramétrica asintótica ya que tienen aproximaciones aplicables para manejos de distribuciones muestrales. Sea x_1, \dots, x_n realizaciones independientes de una variable aleatoria X que tiene distribución verdadera F , donde la familia $\mathcal{F} = \{f(\underline{x}, \theta) : \theta \in \Theta\}$ denota la familia de densidades conjuntas para un conjunto de observaciones está indexada por un parámetro d -dimensional θ y la distribución verdadera F tiene $\theta = \theta_0$. El estimador máximo verosímil de θ_0 es $\hat{\theta}$. Lo anterior es la base estadística para los siguientes teoremas (Coles (2004), Casella y Berger (2002)):

Teorema 2.1 Sean x_1, \dots, x_n realizaciones independientes de una distribución contenida en una familia paramétrica \mathcal{F} , sea $l(\cdot)$, la función de logaritmo de verosimilitud y $\hat{\theta}_0$, el estimador máximo verosímil de un modelo paramétrico d -dimensional θ_0 . Entonces suponiendo condiciones de regularidad, para n grande

$$\hat{\theta}_0 \xrightarrow{n \rightarrow \infty} N_d(\theta_0, I_E \theta_0^{-1}) \tag{2.16}$$

donde

$$I_E(\theta) = \begin{bmatrix} E\left\{-\frac{\partial^2}{\partial \theta_1 \partial \theta_1} l(\theta)\right\} & \dots & \dots & E\left\{-\frac{\partial^2}{\partial \theta_1 \partial \theta_d} l(\theta)\right\} \\ \dots & \dots & E\left\{-\frac{\partial^2}{\partial \theta_i \partial \theta_j} l(\theta)\right\} & \dots \\ \dots & E\left\{-\frac{\partial^2}{\partial \theta_j \partial \theta_i} l(\theta)\right\} & \dots & \dots \\ E\left\{-\frac{\partial^2}{\partial \theta_d \partial \theta_1} l(\theta)\right\} & \dots & \dots & E\left\{-\frac{\partial^2}{\partial \theta_d \partial \theta_d} l(\theta)\right\} \end{bmatrix}$$

$I_E(\theta)$, es la matriz de información esperada de Fisher .

Se puede utilizar el teorema anterior para obtener intervalos de confianza para un θ_i de $\theta_0 = (\theta_1, \dots, \theta_d)$. Tomando el inverso del elemento ij -ésimo de la matriz inversa $I_E(\theta_0)^{-1}$, por ejemplo el elemento $\kappa_{(i,j)}$, de las propiedades de una distribución normal multivariada, se tendría que para n grande (Anderson, 1958),

$$\hat{\theta}_i \xrightarrow{n \rightarrow \infty} N(\theta_i, \kappa_{(i,i)})$$

Si $\kappa_{(i,i)}$ fuera conocido, un intervalo de confianza asintótico $(1 - \alpha)$ para $\kappa_{(i,i)}$ sería

$$\hat{\theta}_i \pm z_{\alpha/2} \sqrt{\kappa_{(i,i)}} \tag{2.17}$$

donde $z_{\alpha/2}$ es el cuantil $(1 - \alpha/2)$ de una distribución normal estándar.

Generalmente θ_0 es desconocido es usual aproximar I_E por la matriz de información observada, definida por

$$I_{Obs}(\theta) = \begin{bmatrix} -\frac{\partial^2}{\partial\theta_1^2} l(\theta) & \dots & \dots & -\frac{\partial^2}{\partial\theta_1\partial a} l(\theta) \\ \dots & \dots & -\frac{\partial^2}{\partial\theta_i\partial_j} l(\theta) & \dots \\ \dots & -\frac{\partial^2}{\partial\theta_j\partial_i} l(\theta) & \dots & \dots \\ -\frac{\partial^2}{\partial\theta_a\partial_1} l(\theta) & \dots & \dots & -\frac{\partial^2}{\partial\theta_a^2} l(\theta) \end{bmatrix}$$

y evaluar en $\theta = \hat{\theta}$. Como en el anterior caso, denotando los términos inversos de la matriz por $\hat{\kappa}_{(i,j)}$, se sigue que un intervalo de confianza para θ_i es

$$\hat{\theta}_i \pm z_{\alpha/2} \sqrt{\hat{\kappa}_{(i,i)}}$$

A pesar de que los intervalos tienen buenas aproximaciones, las mejores son obtenidas en (2.17). El siguiente resultado utiliza las propiedades de los estimadores máximo verosímiles para obtener alguna función del estimador, simplemente substituyendo.

Teorema 2.2 *Sea $\phi = g(\theta)$ una función escalar y si $\hat{\theta}_0$ es un estimador máximo verosímil de θ_0 . Entonces el estimador máximo verosímil de ϕ_0 está dado por $\hat{\phi}_0 = g(\hat{\theta}_0)$*

A continuación el siguiente resultado es consecuencia del método delta, que puede ser usado para obtener intervalos de confianza para alguna componente individual de θ_0 .

Teorema 2.3 *Sea $\hat{\theta}_0$ un estimador máximo verosímil del parámetro d -dimensional θ_0 para muestras grandes, con matriz de varianzas-covarianzas aproximadas V_θ . Entonces si $\phi = g(\theta)$ es una función escalar, el estimador máximo verosímil de $\phi_0 = g(\theta_0)$ satisface*

$$\hat{\phi}_0 \xrightarrow{n \rightarrow \infty} N(g(\theta_0), (\nabla\phi)^T \Sigma (\nabla\phi))$$

con $\nabla\phi = \left[\frac{\partial\phi}{\partial\theta_1}, \dots, \frac{\partial\phi}{\partial\theta_a} \right]^T$ evaluado en $\hat{\theta}_0$

2.9. Pruebas de Hipótesis

Sea $\mathbf{X} = (X_1, \dots, X_n)$ una muestra aleatoria con distribución $F(x; \theta)$ con $\theta \in \Omega \subset \mathbb{R}^k$ con $k \geq 1$. Se asume que la forma funcional de $F(x; \theta)$ se conoce aunque no para el parámetro θ .

Definición 2.4 Una hipótesis estadística es una aseveración acerca de la distribución de una o más poblaciones. Si una hipótesis estadística tiene completamente especificada la distribución, es llamada una hipótesis simple; de otra forma se tiene una hipótesis compuesta (Mood et al., 1974).

Si se plantea la siguiente aseveración

$$H_0 : \theta \in \omega \quad \text{vs} \quad H_1 : \theta \in \Omega - \omega.$$

El espacio paramétrico Ω se particiona en dos regiones: la región ω que corresponde al parámetro de la hipótesis nula y la región $\Omega - \omega$ de la hipótesis alternativa.

Tomando en cuenta la muestra de \mathbf{X} , se desea conocer cuando la decisión de no rechazar o rechazar H_0 es adecuada.

Si χ es el conjunto de todos los valores posibles del vector de observaciones \mathbf{X} , entonces en una prueba de hipótesis se forma una partición en dos conjuntos que cumplen

$$\chi = \chi_A \cup \chi_R \text{ y } \chi_A \cap \chi_R = \emptyset.$$

Dado \mathbf{x} una realización de \mathbf{X} , se tienen las siguientes reglas de decisiones:

si $\mathbf{x} \in \chi_A$, no se rechaza H_0 o

si $\mathbf{x} \in \chi_R$, rechazar H_0 en favor de la hipótesis alternativa.

A χ_R se le conoce como región de rechazo o región crítica y a χ_A se le conoce como región de no rechazo. Generalmente, χ_R define una prueba $\chi_R = \{\mathbf{x} \in \chi : t(\mathbf{x}) > k\}$, con $t(\mathbf{x})$ estadística de prueba, donde a cada prueba se le asocia una función indicadora $\phi_{\chi_R} : \chi \rightarrow \{0, 1\}$

$$\phi_{\chi_R}(\mathbf{x}) = \begin{cases} 1, & \text{si } \mathbf{x} \in \chi_R ; \\ 0, & \text{si } \mathbf{x} \in \chi_A. \end{cases}$$

Cuando se toma una decisión siempre existe la posibilidad de equivocarse, por lo cual es posible cometer dos tipos de error:

Error tipo I: rechazar H_0 cuando H_0 es verdadera.

Error tipo II: no rechazar H_0 cuando H_0 es falsa.

Sus probabilidades son:

$$P(\text{Error tipo I usando } \chi_R) = \int_{\chi_R} f_{\mathbf{x}}(\mathbf{x}; \theta) d\mathbf{x}, \theta \in \omega$$

$$P(\text{Error tipo II usando } \chi_R) = \int_{\chi_A} f_{\mathbf{x}}(\mathbf{x}; \theta) d\mathbf{x}, \theta \in \Omega - \omega$$

Sería deseable escoger una prueba que minimice la probabilidad de cometer ambos tipos de errores, pero lo único que se puede hacer es fijar la probabilidad de cometer el error tipo I y minimizar el error tipo II (Villaseñor, 2004).

Definición 2.5 Se le llama función de potencia de una prueba de hipótesis sobre χ_R a la función de θ definida por

$$\beta_{\chi_R}(\theta) = P(\text{rechazar } H_0 \text{ usando } \chi_R)$$

(Casella y Berger, 2002).

Definición 2.6 Una prueba se llama de tamaño α , si $\alpha \in (0,1)$ y $\sup_{\theta \in \omega}(\theta) \leq \alpha$ y cuando se cumple la igualdad, entonces α se llama nivel de significancia

(Casella y Berger, 2002).

A continuación se da una descripción muy rápida de un método para desarrollar pruebas de hipótesis.

2.10. Pruebas de Razón de verosimilitud generalizada.

Sea X_1, \dots, X_n una muestra aleatoria de densidades $f(x; \theta)$, en donde $\theta \in \Omega$ y $L_X(\theta)$ es la función de verosimilitud de la muestra.

Definición 2.7 La estadística de razón de verosimilitudes para probar $H_0 : \theta \in \omega$ vs $H_1 : \theta \in \Omega - \omega$ esta dada por

$$\lambda(x) = \frac{\sup_{\theta \in \omega} L_X(\theta)}{\sup_{\theta \in \Omega} L_X(\theta)}$$

La prueba de razón de verosimilitud tiene una región de rechazo de la siguiente forma $\{\mathbf{x} : \lambda(\mathbf{x}) \leq c\}$ donde $c \in [0, 1]$. Es decir se rechaza H_0 para valores pequeños de $\lambda(x)$, esta estadística no depende de algún parámetro. En algunas ocasiones no se puede encontrar la distribución de $\lambda(x)$, por lo cual hay que recurrir a otros teoremas.

Una prueba estadística muy utilizada cuando al menos una de las hipótesis es compuesta, es la prueba de Razón de Verosimilitud Generalizada.

Definición 2.8 Para probar $H_0 : \theta \in \omega$ vs $H_1 : \theta \in \Omega - \omega$

$$\lambda(x) = \frac{\sup_{\theta \in \omega} L_X(\theta)}{\sup_{\theta \in \Omega} L_X(\theta)}$$

Bajo condiciones de regularidad sobre $f(x; \theta)$, la variable aleatoria $-2\log\lambda(X)$ es asintóticamente distribuida como una χ_v^2 , donde $v = a$ la diferencia entre el número de parámetros independientes en Ω y el número en ω .

Se rechaza H_0 si $-2\log\lambda(X) > \chi_{v,1-\alpha}^2$, donde $\chi_{v,1-\alpha}^2$ es el cuantil $1-\alpha$ de la distribución χ^2 con v grados de libertad (Mood *et al.* (1974); Rohatgi. (1976)).

2.11. Diagnósticos de Modelos

La razón de estimar un modelo estadístico de un conjunto de datos es hacer conclusiones acerca de algunos aspectos de la población de la cual los datos fueron seleccionados. Coles (2004) opina que normalmente es una opción buena juzgar la precisión del modelo en términos de sus concordancias con los datos que fueron usados para estimarlo.

Sea un conjunto de realizaciones independientes $x_{(1)}, \dots, x_{(n)}$ de una población con función de distribución F . Un estimador F , digamos \hat{F} obtenido por máximo verosimilitud, se desea evaluar que tan plausible es que las observaciones x_i sean de una muestra aleatoria de \hat{F} .

Definición 2.9 *Dada una muestra ordenada de observaciones independientes $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ de una población con función de distribución F , la función de distribución empírica es definida por*

$$\tilde{F}(x) = \frac{i}{n+1} \quad \text{para } x_{(i)} \leq x < x_{(i+1)}.$$

\tilde{F} es un estimador de la distribución de probabilidad F , debería de estar en concordancia con el modelo candidato \hat{F} . Varias procedimientos de bondad de ajuste son basadas comparando a \tilde{F} y \hat{F} . Sólo dos técnicas gráficas son usadas comúnmente.

Definición 2.10 *Dada una muestra ordenada de observaciones independientes $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ de una población con función de distribución estimada \hat{F} , una gráfica de Probabilidad consiste de los puntos*

$$\left\{ \left(\hat{F}(x_{(i)}), \frac{i}{n+1} \right) : i = 1, \dots, n \right\}$$

Cuando \hat{F} es un modelo razonable para la distribución de la población los puntos de

la gráfica de Probabilidad están cercanos a la diagonal, puntos alejados de la diagonal dan evidencia de la falla en \widehat{F} como un modelo para los datos.

Definición 2.11 *Dada una muestra ordenada de observaciones independientes $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ de una población con función de distribución estimada \widehat{F} , la gráfica Cuantil-Cuantil consiste de los puntos*

$$\{(\widehat{F}^{-1}(\frac{i}{n+1}), x_{(i)}) : i = 1, \dots, n\}$$

El nombre de este gráfico se debe a que cada pareja de $x_{(i)}$ y $\widehat{F}^{-1}(\frac{i}{n+1})$ da estimaciones del cuantil $\frac{i}{n+1}$ de la distribución F . Si \widehat{F} es un estimador razonable de F entonces la gráfica Cuantil-Cuantil consiste de los puntos cercanos a la diagonal.

La gráfica de Probabilidad y la gráfica Cuantil-Cuantil contienen la misma información expresada sobre diferentes escalas. Coles (2001) dice que se gana percepción con estas presentaciones porque en un gráfico puede parecer razonable una estimación mientras en el otro puede parecer una estimación pobre.

2.12. Método de los umbrales

Sea Y_1, \dots, Y_n una sucesión de n variables aleatorias que tienen una función de distribución común F . Es posible considerar como evento extremo al evento $\{Y_i > u\}$, donde u es un umbral alto fijo. La forma estocástica de este evento está dado por:

$$P[Y > u + x / Y > u] = \frac{1 - F(u + x)}{1 - F(u)} > 0. \quad (2.18)$$

Si la función de distribución de F fuera conocida, la distribución de excedencias del umbral puede también ser conocida (2.18). Usando el resultado de Pickands (1975) se puede aproximar a un valor alto del umbral, significando que para u suficientemente grande, la distribución de $\{Y > u\}$ dado que $\{Y_i > u\}$ es aproximadamente

$$G(x; \tilde{\sigma}, \varepsilon) = 1 - [1 + \frac{\varepsilon x}{\tilde{\sigma}}]^{-1/\varepsilon} \quad (2.19)$$

definida sobre el conjunto $\{x : x > 0 \text{ y } 1 + \frac{\varepsilon x}{\tilde{\sigma}} > 0\}$ donde $\tilde{\sigma} = \sigma + \varepsilon(u - \mu)$. De donde, G pertenece a la familia de la distribución Pareto Generalizada (DPG). Si $\varepsilon = 0$, se interpreta que $\varepsilon \rightarrow 0$ en la ecuación (2.19) que corresponde a la distribución exponencial con parámetro $1/\tilde{\sigma}$, común en el método de picos sobre umbrales. Usando técnicas sobre máximo-verosimilitud, se obtienen resultados similares a la distribución

de GVE (Davison y Smith. (1990);Bellanger y Tomassone (2004)).

En la DPG, el caso $\varepsilon > 0$ es de cola pesada, en la cual $1 - G(y)$ decae en la misma proporción que $x^{-\frac{1}{\varepsilon}}$ para x grande. Esto hace recordar a la distribución Pareto $G(x) = 1 - cx^{-\alpha}$, con $\varepsilon = \frac{1}{\alpha}$. Para $\varepsilon = 0$, se puede tomar el límite cuando $\varepsilon \rightarrow 0$ obteniendo

$$G(x; \sigma, 0) = 1 - \exp\left(-\frac{x}{\sigma}\right),$$

la distribución exponencial con media σ . Para $\varepsilon < 0$, la distribución tiene límite superior en $-\frac{\sigma}{\varepsilon}$. Algunos resultados elementales acerca de la DPG (Finkenstadt y Rootzen, 2004), son

$$E(X) = \frac{\sigma}{1 - \varepsilon}, \quad \varepsilon < 1$$

$$var(X) = \frac{\sigma^2}{(1 - \varepsilon)^2(1 - 2\varepsilon)}, \quad \varepsilon < \frac{1}{2}$$

$$E(X - x/X > x > 0) = \frac{\sigma + \varepsilon x}{1 - \varepsilon}, \quad \varepsilon < 1$$

Capítulo 3

Modelos Estadísticos para investigar tendencias en ozono

Introducción

En las últimas décadas se ha propuesto una variedad de métodos estadísticos para ajustes meteorológicos, para planes de pronósticos y para hacer estimaciones de tendencias del ozono en el tiempo. En los ajustes de tendencia se investiga la variación sistemática de las concentraciones de ozono en el tiempo en relación a un conjunto de covariables. Por ejemplo, el ozono troposférico se puede modelar estadísticamente por la asociación entre las concentraciones de ozono y las variables meteorológicas. Estos métodos se clasifican en tres enfoques estadísticos: Análisis de regresión, Valores Extremos y métodos de Espacio Tiempo. Los modelos de regresión y los valores extremos son generalmente métodos dirigidos para hacer pronósticos y tendencias sin comprometerse en aclarar los mecanismos fundamentales que los originan. Los modelos de espacio tiempo no se han usado mucho porque tienen la desventaja de incrementar la complejidad de la modelación y la toma de datos.

Los objetivos principales para ajustes meteorológicos de superficie de ozono incluyen:

- (a) La obtención de pronósticos de calidad de aire.
- (b) La investigación y estimación de tendencias del ozono en el tiempo.
- (c) Un creciente interés por comprender científicamente los mecanismos fundamentales.

Para el caso (a), los pronósticos de eventos de ozono son mensajes para la salud pública que pueden enfocarse hacia aquellas observaciones que exceden un nivel establecido

por algún organismo de salud y asociarse a predicciones meteorológicas sencillas. Para el caso (b), la valoración de tendencias puede involucrar diarios de superficie de ozono y sus relaciones con alguna covariable meteorológica máxima de importancia. Las estimaciones de tendencia están muy comprometidas a evaluar los efectos de cambios en emisiones. En cuanto al caso (c), este requiere mucho más trabajo para evaluar los modelos estadísticos que reconocen fundamentos químicos y físicos y que analizan las relaciones complejas del ozono troposférico.

En este capítulo se hace un recuento histórico de los problemas de la contaminación por ozono en las ciudades y zonas urbanas, de las variables que comúnmente se utilizan en la modelación estadística del ozono, así como la inclusión de covariables ambientales para explicar los pronósticos de ajustes meteorológicos.

3.1. El problema de la Contaminación por Ozono

Las mediciones de ozono se recolectan de resúmenes de cada cinco minutos a resúmenes diarios, aunque la mayoría de la información se maneja por máximos diarios por hora. Los máximos por hora de concentraciones de ozono promedio están relacionadas con NAAQS, por los efectos sobre la salud debido a las excedencias de los niveles fijados. Aunque estos efectos dañinos no han sido suficientemente investigados hasta el momento. Los resúmenes diarios son también la base para las más recientes evaluaciones sobre tendencias; se cree que serán apropiadas en objetivos de escalas de tiempo meteorológicas que impactan al ozono. Actualmente, se están tomando escalas finas en horarios que son relevantes para procesos de modelación, predicciones a corto tiempo y evaluaciones de modelos fotoquímicos.

Las mediciones de ozono se manejan con una variedad de escalas espaciales que ayudan en los análisis de mediciones de ozono. Algunos estudios usan datos de una sola localización con al menos una red de monitoreo, mientras otros consideran una red regional grande de estaciones de monitoreo. Los datos de la red regional requieren tomar decisiones sobre cómo dirigir los campos espaciales de ozono para la colección de la información, así como elegir la dependencia o independencia de los monitores. El diseño de una red de estaciones de monitoreo de ozono es muy importante porque se deben encontrar máximos de un campo aleatorio de monitoreos de ozono y tomar en cuenta la información sobre las dinámicas regionales de la meteorología para analizar tendencias del ozono. Los resúmenes pueden aparecer de dos formas: un promedio sobre la red de un sitio específico de diarios máximos por hora con máximos de red en máximos diarios por hora y resúmenes de red derivados de análisis en técnicas de reducción de dimensión multivariadas.

Los periodos con altas concentraciones de ozono son observados con lentos movimientos, sistemas ambientales de presión alta que resultan de tener un sol brillante, temperatura alta y aire estancado (Seinfeld, 1991). Estos sistemas típicamente están en un estrato troposférico estable con mezclas conductivas por medio de la inversión de temperatura que ayudan a contener los contaminantes precursores (NO_x y VOCs) en la tropósfera. Para su estudio es necesario incluir mediciones de meteorología que evalúen las características espacio temporal de los vientos y sus direcciones, la temperatura, la presión atmosférica, el tipo de nubes que se encuentran y si los días están soleados entre otras condiciones.

La selección de variables relevantes depende de la propuesta de metodología de análisis, de la región geográfica, la dinámica física y química de la formación de contaminantes fotoquímicos, así como los patrones de emisión de ozono y la disponibilidad de los datos. Unos trabajos en esta dirección de evaluar tendencias cortas tomando en cuenta la variedad de condiciones atmosféricas se encuentran en Torres (1997), Gao *et al.* (1996), Porter *et al.* (2001), Lou *et al.* (2001) y otros más. En la Tabla 3.1, se presentan algunas variables de interés en los estudios de ajustes meteorológicos revisados.

Tabla 3.1: Variables Meteorológicas más usuales.

Temperatura (tmp)	Maxtmp ¹	Maxtmp-Mintmp ²	
Dirección de viento (dv) NS ³	dv NE ⁴	dv NO ⁵	dv EO ⁶
Velocidad del viento(wsp)	Maxwsp ⁷	Prom. de wsp ⁸	Maxwsp-Minwsp ⁹
Presión			
Radiación Solar			
Nubosidad			
Humedad Relativa			
Altura Geopotencial			

Un camino interesante para involucrar los efectos meteorológicos sobre concentraciones de ozono es correlacionar la calidad de aire con uno o más aspectos de la meteorología y usar la asociación resultante para estimar al ozono. Las reacciones con organismos compuestos orgánicos son importantes en la formación de ozono. Seinfeld (1991) observa que periodos con altas concentraciones de ozono son observados con movimientos lentos, presiones ambientales altas ocasionados de sol ardiente con altas temperaturas y aires estancado.

Trabajos como el de Lou *et al.* (2001) han concluido que la temperatura es la única variable usada para ajustar la tendencia del ozono. Aunque se observó que una dependencia lineal de altas concentraciones sobre la temperatura podrían dificultar el análisis porque la temperatura es frecuentemente relacionada a otros factores que influyen en la formación del ozono, tales como tener cielos claros, vientos ligeros, mezcla del aire estancado a una cierta altura y la ocurrencia de posibles tormentas. Algo que sí ha quedado claro es que las temperaturas altas son necesarias pero no son condición suficiente para tener la ocurrencia de altas concentraciones de ozono. También se ha llegado a observar que las variables más incluidas en los modelos son: temperatura, velocidad de viento y dirección de viento.

En los aeropuertos frecuentemente se manejan superficies meteorológicas de alta calidad del aire y máximos de observaciones ambientales. En las mediciones de los monitoreos meteorológicos frecuentemente se realizan análisis de datos regionales, es-

¹Temperatura máxima

²Rango de temperatura

³Dirección de viento norte-sur

⁴Dirección de viento norte-este

⁵Dirección de viento norte-oeste

⁶Dirección de viento este-oeste

⁷Velocidad de viento máxima

⁸Velocidad de viento promedio

⁹Rango de la Velocidad de viento

tos son resultados de mediciones de una gran variedad de fuentes de información que incluyen datos de radar, de satélites y de globos aerostáticos.

En la figura 3.1, se presenta un esquema de algunas relaciones meteorológicas que se plantean por los expertos en la meteorología para explicar y poder simular la dinámica de formación de ozono por diversos agentes (Torres,1997).

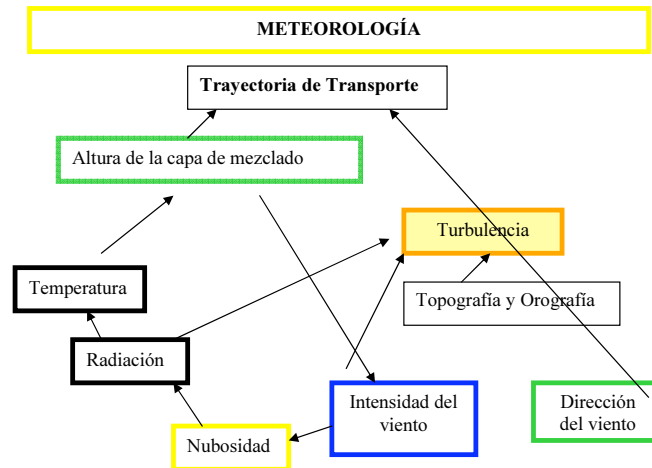


Figura 3.1: Modelo gráfico de relaciones causales entre algunas variables meteorológicas(Torres,1997)

Se ha observado que en las zonas urbanas, la radiación solar dispara una serie de reacciones químicas y fotoquímicas debido en gran medida a las concentraciones iniciales de los precursores. Estas reacciones no ocurren en un volumen constante de aire, sino que están bajo ciertas dinámicas de mezclado, dilución en movimiento horizontal, crecimiento y deformación vertical. Esto es un resultado de la combinación de las condiciones meteorológicas, las orográficas locales y las sinópticas, las cuales pueden llevar a eventos de ozono extremo.

Lou *et al.* (2001) dividen a los ajustes meteorológicos en tres amplias ramas con una variedad considerable de cada uno de ellos. Estos son los modelos basados en Regresión, modelos con el enfoque de Valores Extremos y los modelos de Espacio Tiempo. Una característica entre estos enfoques es la reducción de la variabilidad no explicada en el ozono por medio del ajuste meteorológico.

3.1.1. Modelos tipo Regresión

La mayoría de los enfoques de ajustes meteorológicos de ozono están basados en algún sentido en la regresión con una amplia variedad de grados de dificultad: como tener muchas variables dependientes, que las relaciones no lineales entre las variables, entre otras. Estos enfoques modelan la conducta promedio y su estructura puede considerarse en los modelos lineales de regresión, en árboles de regresión y la regresión no lineal. Dentro de cada categoría hay procedimientos para la introducción de variables meteorológicas y la incorporación de la dependencia espacial y temporal.

La mayoría de todas las metodologías empleadas en la literatura usan Regresión Lineal. Todos los modelos de regresión lineal están abiertos a la crítica de que los procesos químicos y físicos sean lineales y aditivos. Esta suposición afecta la manera de cómo se deben comprender y procesar las relaciones e interacciones de los datos.

Lou *et al.* (2001) argumentan que el ozono y las mediciones meteorológicas son típicamente multivariadas y las propuestas que se hagan en el análisis de Regresión son verdaderos retos para los científicos. Diversas metodologías de reducción se usan para realizar uno o más resúmenes usando la Descomposición en Valores Singulares o análisis de Covarianza Canónico (Reynolds, 1998), análisis de Factores (Spichtinger *et al.*, 1996) entre otros. Estos enfoques manejan grandes cantidades de información, las cuales son covariables multicolinales y reducen la complejidad de las interacciones que son consideradas .

Bloomfield *et al.* (1996) argumentan que los modelos Lineales complican la captura de las relaciones complejas entre las variables meteorológicas y el ozono, por lo cual desarrollan un modelo paramétrico no lineal para datos de 45 estaciones de monitoreo en Chicago de 1981-1991. Parametizan la tendencia y las relaciones entre los datos de ozono actual y retrasado, con tres variables ambientales. El modelo es no paramétrico, utiliza series de Fourier, también se ayuda de componentes principales y de gráficas multivariadas exploratorias para reducir dimensión.

El trabajo de Baur *et al.* (2004) se encuentra en el área de modelos de regresión Cuantil Condicional , que busca interpretar las relaciones no lineales entre concentraciones máximas de ozono por hora con la información meteorológica y la información persistente en las zonas urbanas que están sujetas a contaminación fotoquímica, tal como es el área metropolitana de Atenas. El enfoque extiende la estimación de mínimos cuadrados basada sobre la estimación de la función media condicional en los métodos de regresión paramétricos y los no paramétricos. Además este método da información acerca de la linealidad o no linealidad de la variable dependiente y las variables

explicatorias sin conocer el tipo de no linealidad. Se crearon algunos métodos computacionales avanzados basados en mínimos cuadrados para hacer la estimación de funciones medias condicionales, el procedimiento se puede considerar semiparamétrico, los términos del error se asumen que son independientes del valor de las covariables y no hay suposiciones acerca de la estructura de la varianza. Las variables que se midieron son: velocidad del viento nocturno, velocidad del viento matinal, velocidad del viento tarde, radiación solar, humedad relativa, máximo de temperatura en un valor fijo, cambio de días previos en la temperatura superior en el valor fijo, rango en la superficie de temperatura y superficie máxima de temperatura. Hubo 4 sitios de captura dispuestos conforme a los vientos principales de la cuenca; 3 suburbano y 1 urbano. Los resultados mostraron que las concentraciones de ozono alto son más persistentes que las concentraciones de ozono bajo y que la temperatura máxima está influida por las altas concentraciones de ozono que por bajas concentraciones de ozono. En general los efectos de las variables meteorológicas con los máximos diarios de concentraciones de ozono fueron muy complejos y las interpretaciones de los resultados sugieren que los efectos físicos o químicos causan las variaciones de ozono.

3.1.2. Modelos tipo Valor Extremo

La teoría de valores extremos ha sido desarrollada rápidamente en los años recientes y es ampliamente usada en varias disciplinas. Muchas trabajos han sido publicados, los más recientes son de Reiss y Thomas (2001) y Coles (2004).

Bellanger y Tomassone (2004) clasifican los diferentes enfoques para modelar valores extremos en una población dependiendo de la estructura y complejidad de los datos. Estos son :

- Los modelos Clásicos de Valores Extremos (estudios anuales o estudios sobre la k -ésima observación más grande): Si hay una secuencia grande de datos, este método consiste en el tratamiento de máximos consecutivos de igual longitud (por ejemplo años) de series independientes e idénticamente distribuidas, y que pertenezcan a la familia de la distribución GVE.
- Modelos de los umbrales: El método de picos sobre umbrales consiste en seleccionar un alto umbral y modelar el proceso de excedencias, condicionando sobre la selección de un nivel. La familia paramétrica natural de distribuciones de tales excedencias es la familia Pareto Generalizada (Pickands, 1975).
- Enfoque de procesos puntuales de extremos: Este técnica está inmersa dentro de la teoría de procesos límites que ocurren en la teoría de valores extremos y es una formulación natural de la no estacionalidad y la dependencia.

Al realizar estimaciones con modelos de análisis de regresión frecuentemente se producen malas predicciones de Valores Extremos (Seinfeld,1991). Un enfoque alternativo, que se usa comúnmente, es modelar excedencias de umbrales usando la Teoría de Valores extremos.

Vaquera (1997) investigó si existía alguna tendencia en los niveles muy altos de ozono usando el método de los umbrales. Supuso que el número de observaciones sobre un nivel alto de umbral sigue un Proceso Poisson no Homogéneo (PPNH) (depende del tiempo) y que el número de excedencias sobre tal umbral sigue una distribución Pareto Generalizada (DPG). Tomo tres umbrales con sus respectivos número de observaciones y de magnitudes de excesos, para investigar tres metodologías estadísticas para probar tendencias en niveles altos. Se asumió que el parámetro de forma es constante y que el parámetro de escala considera una liga logarítmica $\log\sigma_i = \beta_0 + \beta_1 t_i$, donde β_1 es el parámetro relacionado con la tendencia y t_i el tiempo (los años). Los resultados mostraron que el enfoque de Proceso Poisson no Homogéneo es superior analizando magnitudes de las excedencias sobre un alto umbral comparando con DPG y el enfoque no paramétrico.

Los investigadores Cox y Chu (1996) aplicaron el modelo de riesgo Weibul para valores de ozono, se estimaron promedios de máximos diarios de ozono para 43 áreas metropolitanas con información de 1981-1993. Se consideraron alrededor de 100 variables meteorológicas y encontraron importantes el máximo de superficie de temperatura, la velocidad de viento y la humedad relativa. Además, usaron mezclas de altura con cielos opacos, velocidad de viento y tomaron interacción con la temperatura. El parámetro de escala dependió de las variables meteorológicas M_1, \dots, M_n usadas de la forma

$$\sigma = \exp\left\{\sum_j^k \beta_j M_j + \zeta T\right\}$$

donde T son los años. La ventaja de este enfoque es que no está limitado a un umbral particular. Los parámetros fueron estimados usando máxima verosimilitud y se supuso independencia entre los días. Los resultados indicaron efectos anuales significantes en temperatura, velocidad de viento, interacción entre la velocidad de viento y la temperatura, así como en el número de excedencias. Sus predicciones meteorológicas fueron significantes sobre la mayoría de las áreas metropolitanas (Wolff *et al.*, 2001).

Gao *et al.* (1996) usaron los datos de máximos diarios de ozono de la Ciudad de Chicago de 1981 a 1991, propusieron un modelo semiparamétrico en los errores, con relaciones no lineales entre el ozono y la meteorología para estimar la tendencia a largo plazo. Bloomfield *et al.* (1996) usaron la misma información de Chicago e investigaron

la tendencia para dos estaciones de monitoreo a largo plazo ajustando los efectos por variables de ambientales en las concentraciones por día, por hora, en los niveles de la estación global, el efecto del día y el efecto de la hora, entre otras (cerca de doce variables). También en el estudio de Bloomfield *et al.* (1996), se estudio la relación espacial del ozono para comprender las fuentes locales y el fenómeno del transporte del ozono.

Los resultados mostraron que los niveles de ozono decrecieron una vez que los efectos de ozono se ajustaron. Los modelos que usaron estas técnicas mostraron predicciones muy cercanas a los niveles de ozono reales, aunque también hubo aspectos para ser investigados más adelante.

En el reporte técnico de Porter *et al.* (2001), se revisan los programas regulatorios de las evaluaciones que ayudan a la calidad del aire con respecto al ozono pues es considerado de interés para la comunidad científica y para ejecutar planes de acción. La detección y señalamiento de tendencias son difíciles por los cambios de concentraciones de contaminantes y pueden ser pequeños en comparación con la variaciones naturales debidas al clima y al tiempo. Las técnicas de estimación de tendencias usadas en la contaminación de aire se discuten con respecto a sus fortalezas y limitaciones en percepción y atribución a cambios de ozono debidas a controles políticos de emisión en el norte de América de la NARSTO. Se encuentra que hay diferencias considerables en reportes de tendencias basados sobre métodos estadísticos con bases de datos similares. Las diferencias surgen de las variedades de técnicas usadas en las series de tiempo que, incluyen la mitigación de efectos meteorológicos y la variedad de métricas usadas para rastrear cambios (Porter,2001).

En las investigaciones de Sánchez (2001), Sánchez *et al.* (2001b) y Sánchez *et al.* (2001a) el problema de la estimación paramétrica con el modelo Pareto Generalizado (PG) a partir de un umbral, se usa el modelo de regresión donde los dos parámetros de la PG dependen de covariables, planteando una metodología alternativa en el análisis de tendencias y comparaciones de poblaciones de esa misma familia. En el trabajo se propone un enfoque Bayesiano y un método no paramétrico para estimar el parámetro relacionado con la tendencia, se compara su desempeño en términos de potencia usando un experimento Monte Carlo. Las bases de datos pertenece a la Ciudad de Guadalajara. Se concluye que suponer el modelo de Regresión Pareto Generalizado es apropiado para los datos y que hay evidencia para suponer que las magnitudes de las concentraciones de ozono que exceden un umbral establecido de 110 ppb presentan tendencia negativa.

Continuando con estos estudios los investigadores Villaseñor, Vaquera y Reiss (Reiss y Thomas, 2001), toman datos de una estación de monitoreo muy contaminada de la Ciudad de México, para investigar sobre el modelo distribucional del cluster máximo de excedencias de veinte segmentos mensuales sobre un umbral alto de la distribución PG, se investiga la tendencia de los parámetros en el tiempo y se toma una prueba no paramétrica para analizar tres tipos de umbrales: bajo, medio y alto. Se concluye que cambios en el tiempo afecta a la distribución, que hay evidencia que la tendencia disminuye en los niveles altos para el conjunto de datos y que este procedimiento

puede ser usado en series de tiempo no estacionales.

Un estudio propuesto en el enfoque de procesos puntuales, es el de Bellanger y Tomassone (2004) con datos de París de 1988 a 2001, en siete estaciones de monitoreo. Este estudio es una continuación de dos estudios previos del 2000 y 2001. El propósito de esta investigación es conocer si las tendencias observadas en el tiempo de las excedencias de ozono que exceden un umbral, son reales o son resultados de cambios meteorológicos que afectan las condiciones en las que se genera el ozono. Se supone un Proceso Poisson no Homogéneo, con parámetros que dependen de las covariables meteorológicas, la tendencia temporal y de siete sitios de monitoreo. El enfoque considera un modelo de regresión logística para las excedencias de los umbrales y se desea describir el crecimiento regulatorio para detectar tendencias. Las variables manejadas son: máximo de temperatura (T_{max}), rango de temperatura (T_{rango}), temperatura mínima (T_{min}), viento y el año; se tomaron 3 altitudes diferentes en alguno de los 7 sitios de monitoreo y tipos de zona (urbano y suburbano). Se concluye que el análisis simultáneo del número de excedencias y el tamaño de las excedencias son buenos predictores para las variables meteorológicas clásicas, aunque no se debe de descuidar las mediciones en los sitios y debe tomarse precaución. Los resultados confirman la complejidad de las relaciones entre el ozono y las variables meteorológicas.

El trabajo de Tobías y Scotto (2005) maneja datos de ozono de la Ciudad de Barcelona que es una de las ciudades más contaminadas de Europa en donde no ha habido estudios sobre altas concentraciones de ozono. Esta investigación toma máximos diarios de ozono urbano, con información de 1991 a 1996 para dos estaciones de monitoreo de la ciudad que reflejan características distintas, como lo es la zona industrial y la urbana. La investigación se basó en el método de Picos sobre Umbrales (POT), se analizó las excedencias sobre tres niveles críticos de World Health Organisation air quality guidelines Organisation (1987), las cuales son $120\mu gm^{-3}$ basados en 8 datos de una hora que protegen a la salud humana, de $180\mu gm^{-3}$ para un promedio de una hora que da información de salud pública y $360\mu gm^{-3}$ para dar mensajes de advertencia al público. El objetivo de la investigación consistió en usar una versión paramétrica del método de umbrales sobre picos y estimar periodos de retorno para tres diferentes años (3, 10 y 40 años). El trabajo concluye que los valores del umbral para proteger a la salud humana son excedidos muchas veces, a pesar de que son zonas diferentes y no se reporto ninguna mensaje de salud local. El estudio se llevó a cabo usando una serie de tiempo grande y no se manejaron variables ambientales.

3.1.3. Modelos Espacio tiempo

Como se mencionaba al principio de este capítulo, la modelación de datos de ozono en una red requiere un balance entre la interpretación, la simplicidad del enfoque y

la incorporación de información de respuestas de escala regionales para incrementar la potencia estadística.

Schabenberger y Gotway (2005) dicen que los métodos de estudio de la estadística para datos espaciales presentan desafíos que son únicos dentro de la ciencia estadística. Como pocas áreas, la estadística espacial muestra y brinda filosofías, metodologías y técnicas que son típicamente enseñadas separadamente de los cursos curriculares.

En el trabajo de Huerta y Sansó (2005) se propone un enfoque para modelar extremos en el espacio-tiempo suponiendo que los datos provienen de una distribución GVE con parámetros de escala y de forma, que definen una estructura de espacio y tiempo. La componente temporal está definida por medio de un modelo lineal dinámico o un estado espacial que representa a los estimadores de la tendencia o estacionalidad de los datos en el tiempo. El elemento espacial está impuesto por la matriz de evolución del modelo lineal dinámico donde se adopta un proceso en forma de convolución. En el trabajo se muestra como producir estimadores espaciales y temporales usando simulación de Cadenas de Markov Monte Carlo. Se ejemplifica con los concentraciones de ozono de máximos diarios de las estaciones de la RAMA de la Ciudad de México en el período de 1990 a 2002 y los datos de lluvia ácida extrema del caribe Venezolano. Se concluye que el enfoque puede ser extendido a la variabilidad espacial usando procesos de convolución y se propone un enfoque dinámico que da flexibilidad a la captura de tendencias cortas y largas, así como a la dependencia de la variación en el tiempo.

Capítulo 4

Los Modelos Propuestos

Introducción

Diversas metodologías se están desarrollando en la práctica por los investigadores de diferentes disciplinas de un tiempo a la fecha sobre la problemática ambiental. Estos modelos se interesan en comprender e investigar las relaciones causales que ocasionan el aumento de los máximos de concentraciones de ozono. También hay análisis e investigación, evaluación y creación de nuevas tecnologías alternativas relacionadas con algunos precursores que ocasionan el aumento del ozono (García-Colín y Varela (1996); García-Colín y Varela (2001) ; García-Colín y Varela (2003)). Se ha observado que el problema es muy complejo y dinámico, que se necesita más información que no ha sido recopilada anteriormente, así como recolección de los datos tomando en cuenta las interacciones de las variables importantes. Además de mayor colaboración de investigadores de distintas áreas del conocimiento para poder establecer modelos más finos en el área micro ambiental.

En este capítulo se presenta una metodología sencilla y fácil de aplicar para analizar problemas de valores extremos que se basa en un resultado acerca del comportamiento distribucional del estimador de la función cuantil. Se realizan simulaciones de variables generalizadas de valores extremos. Se presentan dos posibles maneras de analizar tendencias en los niveles alcanzados de las concentraciones del ozono. Un modelo solo depende del tiempo y en el otro modelo, se pueden involucrar a las variables ambientales de humedad relativa, velocidad del viento, temperatura y al tiempo.

4.1. La convergencia de la función cuantil

Resultado 4.1 *Sea Y_1, \dots, Y_n observaciones independientes idénticamente distribuidas con función de distribución VEG. Para $y_p = -\log(1 - p) = \log(G_\theta(Z(p)))$ and*

$\alpha = 1 - p$ fijos, con $0 \leq p \leq 1$, sea $\widehat{Z}(p)$ el estimador en 2.10 de la función Cuantil $Z(p)$, donde μ es conocida y $\widehat{\sigma}, \widehat{\varepsilon}$ son los estimadores de máxima verosimilitud. Entonces

$$\begin{aligned} \sqrt{n} \left[\widehat{Z}(p) - \left(\mu - \sigma \log(y_p) + \frac{(\log(y_p))^2}{2} \right) (\text{cov}(\widehat{\sigma}, \widehat{\varepsilon}) + \sigma \varepsilon) \right] &\xrightarrow{n \rightarrow \infty} \\ N\left(0, \left[\log(y_p) - \varepsilon \frac{(\log(y_p))^2}{2} \right]^2 \text{var}(\widehat{\sigma}) + \left[\frac{(\log(y_p))^2 \sigma}{2} \right]^2 \text{var}(\widehat{\varepsilon}) \right. \\ &\left. - 2 \left[\log(y_p) - \varepsilon \frac{(\log(y_p))^2}{2} \right] \frac{(\log(y_p))^2 \sigma}{2} \text{cov}(\widehat{\sigma}, \widehat{\varepsilon}) \right). \end{aligned}$$

Demostración (Usando el método Delta)

Supongamos que Y_1, \dots, Y_n tienen una distribución (2.7) tomando su Función Cuantil para un umbral alto y fijo en las observaciones con

$$y_p = -\log(1 - p) = \log(G_\theta(Z(p)))$$

and $\alpha = 1 - p$ fijos, con $0 \leq p \leq 1$.

Para encontrar la distribución de probabilidad de $\widehat{Z}(p)$ cuando se tiene un tamaño de muestra grande, se analiza la ecuación de la función Cuantil (2.10), entonces

$$Z(p) = \mu - \frac{\sigma}{\varepsilon} (1 - y_p^{-\varepsilon}) \quad \varepsilon \neq 0. \quad (4.1)$$

Tomando $y_p^{-\varepsilon}$ de la expresión (2.7), se puede escribir como

$$y_p^{-\varepsilon} = \exp(-\varepsilon \log(y_p)).$$

Por otro lado, al tomar la aproximación en serie de Taylor de \exp^{-x} (Casella y Berger, 2002), con $x = \varepsilon$,

$$y_p^{-\varepsilon} = \exp(-\varepsilon \log(y_p)) = \sum_{i=0}^{\infty} \frac{(-\varepsilon \log(y_p))^i}{i!} = 1 - \varepsilon \log(y_p) + \frac{(\varepsilon \log(y_p))^2}{2} - \dots$$

Sustituyendo lo anterior en (2.7) se tiene

$$Z(p) = \mu - \frac{\sigma}{\varepsilon} \left(\varepsilon \log(y_p) - \frac{(\varepsilon \log(y_p))^2}{2} + \dots \right).$$

La anterior expresión es equivalente a tomar sólo los dos primeros términos.

$$Z(p) \approx \mu - \sigma \left[\log(y_p) - \frac{\varepsilon (\log(y_p))^2}{2} \right] \quad \varepsilon \neq 0. \quad (4.2)$$

Para encontrar la distribución asintótica de $Z(p)$, se observa que esta función es continua y diferenciable en su dominio de definición, por lo que se puede aplicar el método delta para encontrar la media y varianza de la distribución asintótica resultante.

$$\widehat{Z}(p) \approx \widehat{\mu} - \widehat{\sigma} \left[\log(y_p) - \frac{\widehat{\varepsilon}}{2} (\log(y_p))^2 \right].$$

Para calcular su varianza.

$$\begin{aligned} \text{var}(Z(p)) &\approx \left(\frac{\partial}{\partial \sigma} Z(p) \right)^2 \text{var}(\sigma) + \left(\frac{\partial}{\partial \varepsilon} Z(p) \right)^2 \text{var}(\varepsilon) \\ &\quad + 2 \frac{\partial}{\partial \sigma} Z(p) \frac{\partial}{\partial \varepsilon} Z(p) \text{cov}(\sigma, \varepsilon). \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \varepsilon} Z(p) &= \frac{(\log(y_p))^2 \sigma}{2}, \\ \frac{\partial}{\partial \sigma} Z(p) &= - \left[\log(y_p) - \frac{(\log(y_p))^2 \varepsilon}{2} \right]. \end{aligned}$$

Así substituyendo

$$\begin{aligned} \text{var}(\widehat{Z}(p)) &\approx \left[\log(y_p) - \frac{(\log(y_p))^2 \varepsilon}{2} \right]^2 \text{var}_\sigma(\widehat{\sigma}) + \left[\frac{(\log(y_p))^2 \sigma}{2} \right]^2 \text{var}_\varepsilon(\widehat{\varepsilon}) \\ &\quad - 2 \left[\log(y_p) - \frac{(\log(y_p))^2 \varepsilon}{2} \right] \frac{(\log(y_p))^2 \sigma}{2} \text{cov}(\widehat{\sigma}, \widehat{\varepsilon}). \end{aligned}$$

Si $\widehat{\sigma}$ y $\widehat{\varepsilon}$ son variables dependientes, al tomar la esperanza de la expresión (4.2). Por propiedades de los estimadores máximos verosímiles y propiedades de esperanzas.

$$\begin{aligned} E(\widehat{Z}(p)) &\approx \mu - E \left[\widehat{\sigma} \left(\log(y_p) - \frac{\widehat{\varepsilon} (\log(y_p))^2}{2} \right) \right] = \mu - \left[\log(y_p) E(\widehat{\sigma}) - \frac{E(\widehat{\sigma} \widehat{\varepsilon} (\log(y_p))^2)}{2} \right] \\ &= \mu - \left[\log(y_p) E(\widehat{\sigma}) - \frac{(\log(y_p))^2}{2} (\text{cov}(\widehat{\sigma}, \widehat{\varepsilon}) + E(\widehat{\sigma}) E(\widehat{\varepsilon})) \right]. \end{aligned}$$

$$\therefore E(\widehat{Z}(p)) \approx \mu - \sigma \log(y_p) + \frac{(\log(y_p))^2}{2} (\text{cov}(\widehat{\sigma}, \widehat{\varepsilon}) + \sigma \varepsilon).$$

Por hipótesis, el vector de los parámetros converge asintóticamente a una distribución normal bivariada, así

$$\begin{pmatrix} \hat{\sigma} \\ \hat{\varepsilon} \end{pmatrix} \rightsquigarrow N_2 \left[\begin{matrix} E(\hat{\sigma}) = \sigma \\ E(\hat{\varepsilon}) = \varepsilon \end{matrix}, \begin{pmatrix} var(\hat{\sigma}) & cov(\hat{\sigma}, \hat{\varepsilon}) \\ cov(\hat{\sigma}, \hat{\varepsilon}) & var(\hat{\varepsilon}) \end{pmatrix} \right]$$

Aplicando el método Delta y centrado para tener una normal estandarizada, se tiene

$$\begin{aligned} N(0, [log(y_p) - \varepsilon \frac{(log(y_p))^2}{2}]^2 var(\hat{\sigma}) + [\frac{(log(y_p))^2 \sigma}{2}]^2 var(\hat{\varepsilon}) \\ - 2[log(y_p) - \varepsilon \frac{(log(y_p))^2}{2}] \frac{(log(y_p))^2 \sigma}{2} cov(\hat{\sigma}, \hat{\varepsilon})). \end{aligned}$$

4.2. Evaluación Numérica de la función cuantil

En esta sección se realizan simulaciones numéricas de las distribuciones de valores extremos para verificar la convergencia del estimador de la Función Cuantil a una normal. Se usa el método de la función inversa para realizar simulaciones Monte Carlo (Vaquera, 2004).

Como la distribución GVE tiene tres parámetros, para los parámetros de localización y de escala se usan los estimadores de momentos y para el parámetro de forma se maneja el estimador de Pickands (ecuación 2.14). Por lo cual, los estimadores que se usan en esta sección serán una mezcla de estimadores: de momentos y de Pickands. Tomando los valores fijos de los parámetros como en la tabla 4.1, se hicieron simulaciones para distintas distribuciones GVE.

A continuación se presentan algunas simulaciones realizadas en el paquete R. Las Figuras 4.1 y 4.2 son tipo Gumbel. Las Figuras 4.3 y 4.4 son tipo Weibul y las Figuras 4.5, Figura 4.6 y Figura 4.7 son tipo Fréchet.

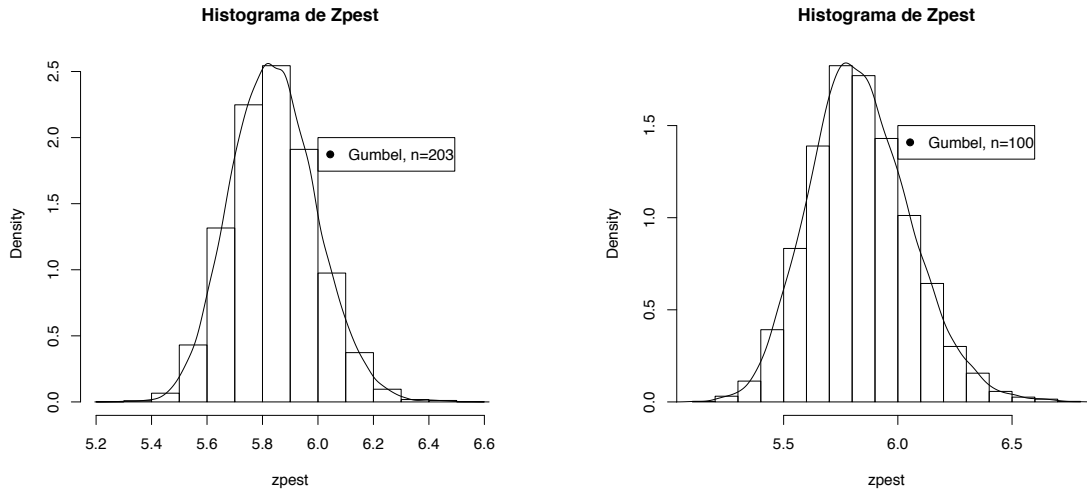
Usando los estimadores de momentos para los dos parámetros de la distribución Gumbel, se calcula la distribución asintótica del estimador de la función Cuantil para diferentes tamaños de muestra.

Usando los estimadores de momentos para dos parámetros de la distribución Weibul $\varepsilon > 0$ y el estimador de Pickands para el parámetro de forma, se calcula la distribución asintótica del estimador de la función Cuantil para diferentes tamaños de muestra.

$$\tilde{\sigma}_{mo} = \sqrt{\frac{\sum_i X_i^2}{n} - \bar{X}^2} \frac{(-\varepsilon)}{\sqrt{\Gamma(1 - \frac{2}{\alpha}) - \Gamma^2(1 - \frac{1}{\alpha})}}$$

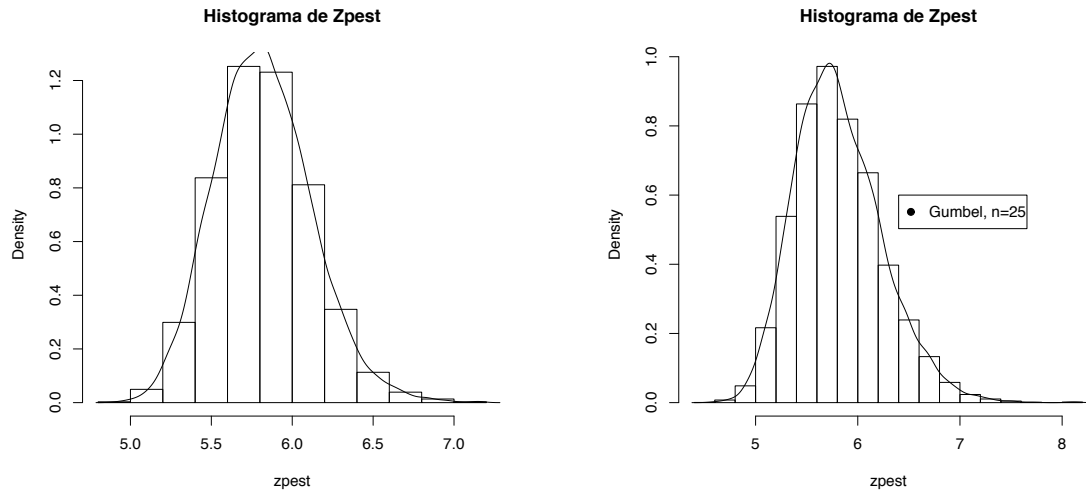
Tabla 4.1: Simulaciones de GVE con $p = 0.05$

Parámetros	Gumbel	Weibul	Fréchet
μ	3.87	0.198	0.198
σ	0.57	1.87	1.87
ε		-1.3	3.5
Tamaños de n y k			
T	203	50(48)	100(95)
a	100	50(47)	100(90)
m	50	40(39)	50(46)
a	25	40(38)	50 (45)
\tilde{n}			40 (38)
o			40 (37)

(a) Distribución asintótica del estimador de la función cuantil para la Gumbel tomando $n=203$ (b) Distribución asintótica del estimador de la función cuantil para la Gumbel tomando $n=100$.**Figura 4.1:** Distribución asintótica del estimador de la función cuantil para la Gumbel 1.

$$\tilde{\mu}_{mo} = \bar{X} - \sqrt{\frac{\sum_i X_i^2}{n} - \bar{X}^2} \frac{(-1 + \Gamma(1 - \frac{1}{\alpha}))}{\sqrt{\Gamma(1 - \frac{2}{\alpha}) - \Gamma^2(1 - \frac{1}{\alpha})}}$$

$$\hat{\varepsilon} = \frac{1}{\ln 2} \ln\left(\frac{X_{(k/4):n} - X_{(k/2):n}}{X_{(k/2):n} - X_{k:n}}\right).$$



(a) Distribución asintótica del estimador de la función cuantil para la Gumbel tomando $n=50$

(b) Distribución asintótica del estimador de la función cuantil para la Gumbel tomando $n=25$.

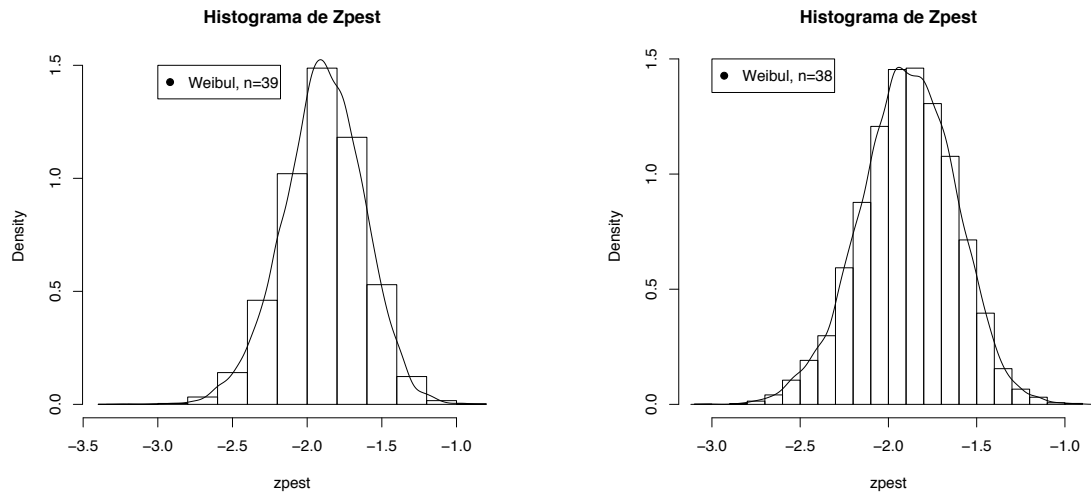
Figura 4.2: Distribución asintótica del estimador cuantil para la Gumbel 2.

Usando los estimadores de momentos para dos parámetros de la distribución Fréchet $\varepsilon < 0$ y el estimador de Pickands para el parámetro de forma, se calcula la distribución asintótica del estimador de la función Cuantil para diferentes tamaños de muestra.

La distribución Gumbel converge sin muchos problemas para tamaños de muestra pequeños a una distribución normal. Situación que no ocurre para las distribuciones Weibul y Fréchet pues son distribuciones de cola pesada, su convergencia a una normal se obtiene para valores grandes de n . Además, en el caso de la distribución Fréchet los resultados no son tan inmediatos debido al comportamiento asintótico del estimador de la función cuantil que no es simétrico. En general, se observó que las simulaciones de la distribución GVE, la convergencia del estimador cuantil a una distribución normal depende de:

1. El tamaño de la muestra, del valor de k en (ecuación 2.14) con $k < n$.
2. El número de repeticiones en la simulación y del valor de $p = 0.05$, que en este trabajo se mantiene fijo.
3. En el estimador del Pickands,

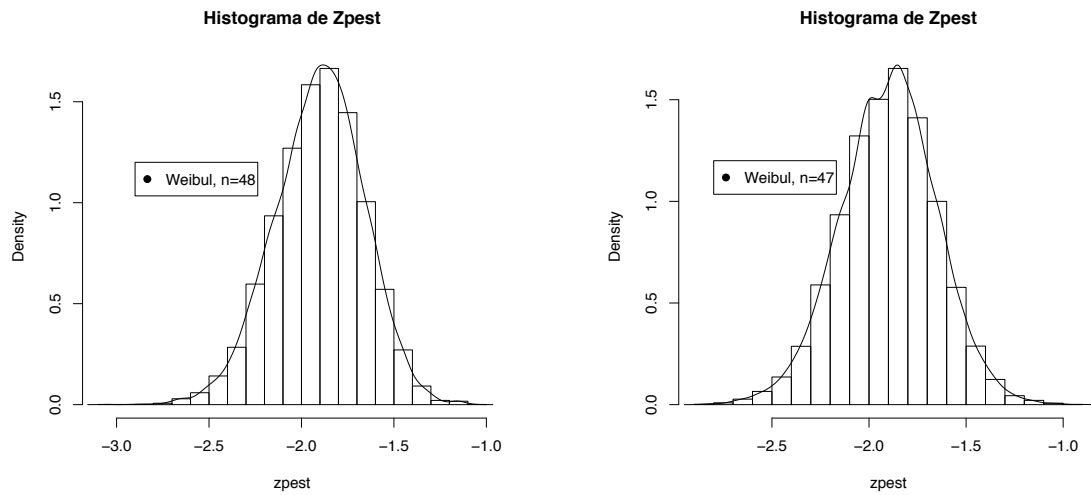
$$\hat{\varepsilon} = \frac{1}{\ln 2} \ln \left(\frac{X_{(k/4):n} - X_{(k/2):n}}{X_{(k/2):n} - X_{k:n}} \right),$$



(a) Distribución asintótica del estimador de la función cuantil para la Weibul tomando $n=40$ y $k=39$

(b) Distribución asintótica del estimador de la función cuantil para la Weibul tomando $n=40$ y $n=38$.

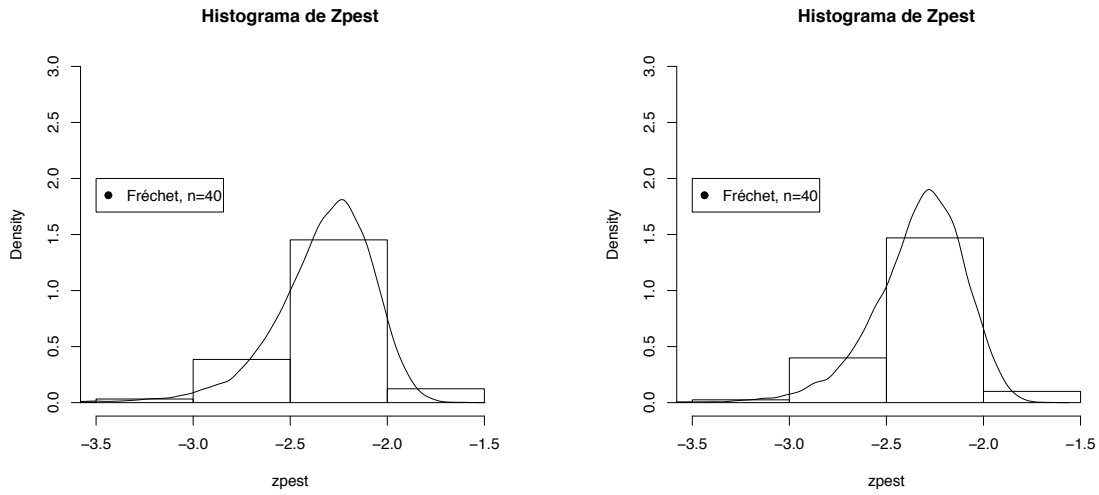
Figura 4.3: Distribución asintótica del estimador de la función cuantil para la Weibul 1.



(a) Distribución asintótica del estimador de la función cuantil tomando $n=50$ y $k=48$

(b) Distribución asintótica del estimador de la función cuantil tomando la Weibul con $n=50$ y $n=47$.

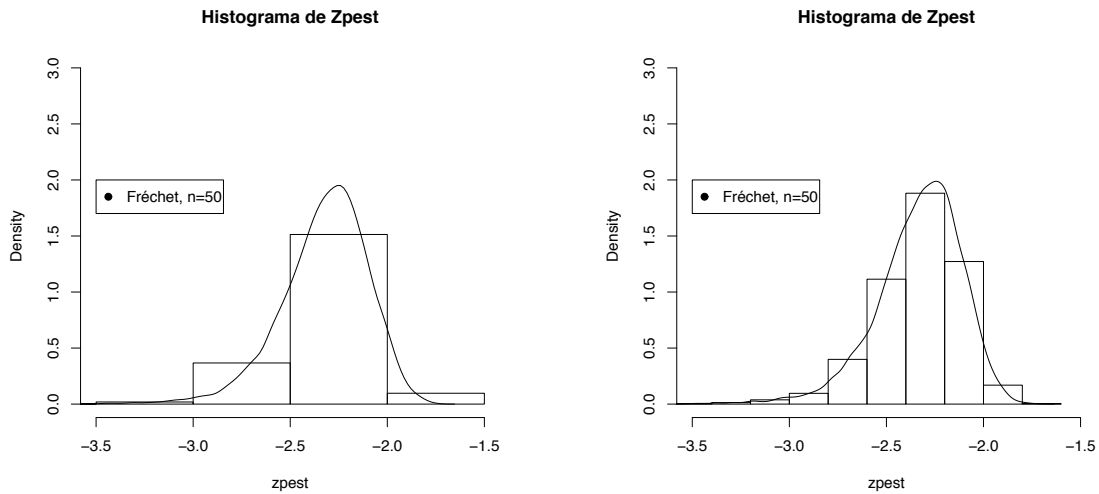
Figura 4.4: Distribución asintótica del estimador de la función cuantil para la Weibul 2.



(a) Distribución asintótica del estimador de la función cuantil para la Fréchet tomando $n=40$ y $k=37$

(b) Distribución asintótica del estimador de la función cuantil para la Fréchet tomando $n=40$ y $n=38$.

Figura 4.5: Distribución asintótica del estimador de la función cuantil para la Fréchet 1.



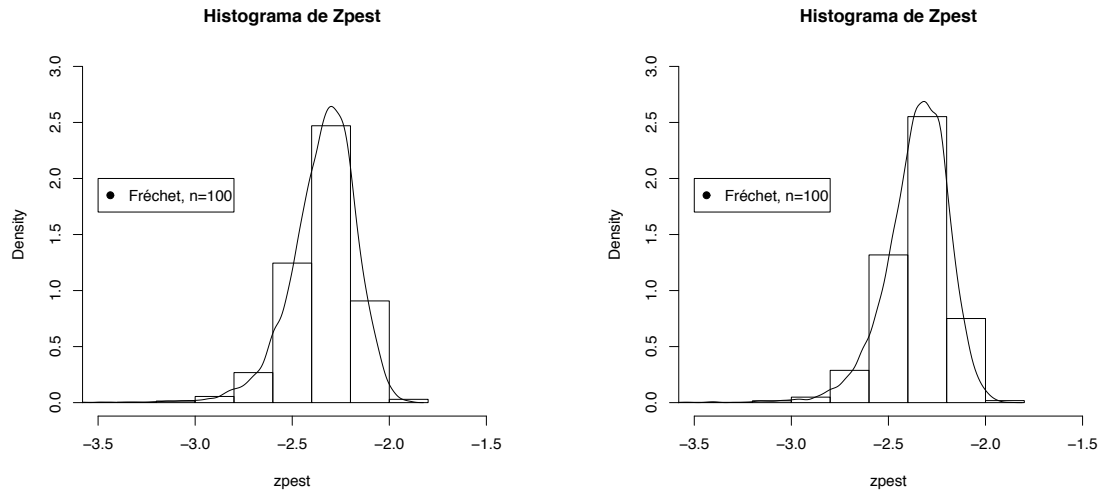
(a) Distribución asintótica del estimador de la función cuantil para la Fréchet tomando $n=50$ y $k=45$

(b) Distribución asintótica del estimador de la función cuantil para la Fréchet tomando $n=50$ y $n=46$.

Figura 4.6: Distribución asintótica del estimador de la función cuantil para la Fréchet 2.

el logaritmo natural

$$\frac{X_{(k/4):n} - X_{(k/2):n}}{X_{(k/2):n} - X_{k:n}} > 0 \Leftrightarrow$$



(a) Distribución asintótica del estimador de la función cuantil para la Fréchet tomando $n=100$ y $k=90$

(b) Distribución asintótica del estimador de la función cuantil para la Fréchet tomando $n=100$ y $n=95$.

Figura 4.7: Distribución asintótica del estimador de la función cuantil para la Fréchet 3.

$$X_{(k/4):n} - X_{(k/2):n} > 0 \quad y \quad X_{(k/2):n} - X_{k:n} \neq 0$$

para que exista el estimador de Pickands y poder realizar las simulaciones.

La función cuantil depende del nivel que se fije de α y este debe ser muy pequeño para que esté sobre los valores de la cola superior.

Revisando en la literatura sobre la elección de la k estadística extrema a usar para poder ajustar la cola derecha de una distribución de Valores Extremos, no se ha encontrado una solución teórica. El trabajo más cercano a esta problemática la da Velasco (2000), cuando trabaja con las distribuciones exponenciales utilizando la distribución de los excesos, la función de la varianza de los excesos y propone un algoritmo que selecciona la k , que maximice la varianza estimada.

Este es uno de varias situaciones en la literatura en donde todavía hay que realizar más investigaciones para poder precisar los resultados. Davison y Smith. (1990), para el caso de DPG, observaron que la selección del valor del umbral en la investigación, es un punto muy importante porque implica el balance entre el sesgo y la varianza de los estimadores. Existen dos métodos para poder elegirlos: Un método exploratorio basado sobre la media de la DPG y la gráfica de los excesos medios, también conocida como gráfica de la vida media residual o bien analizar la estabilidad de los parámetros,

llamado propiedad de invarianza del umbral; es decir, si la función de distribución de $\{Y > u\}$ dado que $\{Y_i > u\}$ es DPG entonces para algún umbral $v \geq u$, la función de distribución de $\{Y > v\}$ dado que $\{Y_i > v\}$ es también DPG (Davison y Smith, (1990); Coles (2004)).

4.3. Modelo de Regresión basado en la función cuantil

Sean x_1, \dots, x_n realizaciones independientes de una variable aleatoria X que tiene densidad de probabilidad $GVE(x; \mu, \sigma, \varepsilon)$, siendo ésta información para un periodo de tiempo (por ejemplo, datos de los máximos de cada cuatro días para un año específico).

La función de verosimilitud está dada por

$$L(X; \mu, \sigma, \varepsilon) = \prod_{i=1}^n g(x_i; \mu, \sigma, \varepsilon) = \prod_{i=1}^n \exp[-(1 + \frac{\varepsilon(x_i - \mu)}{\sigma})^{-\frac{1}{\varepsilon}}] \frac{1}{\sigma} (1 + \frac{\varepsilon(x_i - \mu)}{\sigma})^{-\frac{1}{\varepsilon}-1}.$$

Otra expresión de ella más compacta es

$$L(X; \mu, \sigma, \varepsilon) = \exp[-\sum_{i=1}^n [1 + \frac{\varepsilon(x_i - \mu)}{\sigma}]^{-\frac{1}{\varepsilon}}] \frac{1}{\sigma^n} \left[\prod_{i=1}^n (1 + \frac{\varepsilon(x_i - \mu)}{\sigma}) \right]^{-\frac{1}{\varepsilon}-1}. \quad (4.3)$$

Se supone que existen condiciones de regularidad en la función, para obtener los estimadores máximo verosímiles de la función de log-verosimilitud $-\log L(X; \mu, \sigma, \varepsilon)$. Como puede suceder que el problema no tenga una solución cerrada, se necesita un algoritmo numérico para obtenerlos. Usando la biblioteca de funciones ismev de software libre R, se logran encontrar los estimadores de los tres parámetros. Como la función cuantil, $\widehat{Z}(p)$ es una función de estos estimadores, se usa el principio de invarianza de los estimadores máximo verosímiles (Casella y Berger, 2002). Por lo cual, se tiene el estimador de $\widehat{Z}(p)$ de (2.10) para el periodo de análisis fijando un nivel alto de p .

Para cada periodo de tiempo se obtiene un $\widehat{Z}(p)$. Todos los periodos de interés a analizar (todos los años) conforman el conjunto de todos los periodos a analizar, se tiene $\widehat{Z}_1(p), \widehat{Z}_2(p), \dots, \widehat{Z}_n(p)$ como una muestra de observaciones normales (**resultado 4.1**) con un p alto fijo. Suponiendo condiciones de independencia y de normalidad en las observaciones. También que el estimador de la varianza de los errores se distribuyan como una variable aleatoria normal con media cero, varianza común y conocida, se

puede ajustar un modelo de regresión con variable dependiente a $\widehat{Z}(p)$ y como variable explicativa al tiempo, para analizar su comportamiento.

Con esta información se gráfica la función cuantil $\widehat{Z}(p)$ en el tiempo y ajusta un modelo de Regresión para pronosticar el comportamiento en las observaciones futuras.

4.4. Modelo de Regresión incorporando covariables

Sean y_1, \dots, y_n realizaciones independientes de una variable aleatoria Y que tiene densidad de probabilidad $GVE(y; \mu, \sigma, \varepsilon)$, donde $\sigma = \sigma(X^t \underline{\beta}) = \exp^{X^t \underline{\beta}}$ ésta información son máximos de un periodo de la variable de interés para un tiempo fijo (por ejemplo, el de un mes en particular). Cox-Lewis (Cox y Chu, 1996) proponen a X^t como el vector de covariables y $\underline{\beta}$ es el vector de parámetros desconocidos; esta idea la dieron inicialmente Davison y Smith en 1990 para un modelo de Regresión PG.

$$\sigma = \sigma(\underline{x}^t \underline{\beta})$$

Como $\sigma > 0$, es natural suponer que

$$\sigma(\underline{x}^t \underline{\beta}) = \exp(\underline{x}^t \underline{\beta}) = \exp[\beta_0 + \sum_{j=1}^k \beta_j x_{ij}]$$

En esta expresión se introducen los máximos mensuales de alguna o algunas de las covariables, $j = 1, \dots, k$, para el periodo de tiempo que se maneje.

La función de verosimilitud en esta propuesta es

$$L(Y; \mu, \underline{\beta}, \varepsilon) = \prod_{i=1}^n g(y_i; \mu, \underline{\beta}, \varepsilon),$$

donde n denotan los años que se tiene información.

En forma equivalente se tiene

$$= \exp\left[-\sum_{i=1}^n \left[1 + \frac{\varepsilon(y_i - \mu)}{\exp[\beta_0 + \sum_{j=1}^k \beta_j x_{ij}]}\right]^{-\frac{1}{\varepsilon}}\right] \exp[\beta_0 + \sum_{j=1}^k \beta_j x_{ij}]^{-n}$$

$$\left[\prod_{i=1}^n \left(1 + \frac{\varepsilon(y_i - \mu)}{\exp[\beta_0 + \sum_{j=1}^k \beta_j x_{ij}]}\right) \right]^{-\frac{1}{\varepsilon} - 1}. \quad (4.4)$$

Al igual que el caso anterior, para obtener los estimadores máximo verosímiles de la función de verosimilitud $-\log L(X; \mu, \underline{\beta}, \varepsilon)$ se necesita un algoritmo numérico. Usando el módulo ismev de software libre R (Coles and Stephenson, 2006), se obtienen los estimadores de los parámetros $\mu, \underline{\beta}$ y ε .

Para el caso particular de tener sólo la covariable tiempo y sólo una base de datos de una estación meteorológica, se tendría que el parámetro β_1 es el parámetro que se está asociando a la tendencia y que t_i representa al i-ésimo periodo de tiempo. Es decir,

$$-\log L(X; \mu, \sigma = \exp(\beta_0 + \beta_1 t_i), \varepsilon).$$

Como β_1 es el parámetro asociado a la tendencia, si $\beta_1 = 0$ se tiene que no existe tendencia, si $\beta_1 > 0$ implica que hay tendencia creciente y por último, si $\beta_1 < 0$ habrá tendencia decreciente. Para analizar tendencia en la variable de interés hay que probar la hipótesis

$$H_0 : \beta_1 = 0 \quad vs \quad H_a : \beta_1 \neq 0.$$

Cuando se tiene una muestra de n observaciones y k covariables meteorológicas se tendría que probar que alguna de las covariables ($j=1, \dots, k$) está contribuyendo en la tendencia de los máximos de concentraciones de ozono. Se tiene que probar la hipótesis

$$H_0 : \beta_j = 0 \quad vs \quad H_a : \beta_j \neq 0.$$

con todos los coeficientes de las k variables. La prueba consiste en rechazar H_0 si $|\frac{\widehat{\beta}_j}{\sqrt{v_{jj}}}| > z_{\alpha/2}$ donde α es el nivel de confianza y v_{jj} es el j-ésimo término de la matriz diagonal $[E(-\frac{\partial^2 \log L}{\partial \beta_i \partial \beta_j} |_{\beta_j = \widehat{\beta}_j})]^{-1}$

Otra forma estadísticamente más poderosa es usar la prueba de Razón de Verosimilitudes Generalizada,

$$\Lambda = \frac{L(\widehat{\mu}, \beta_0 = 0, \beta_j = 0, \widehat{\varepsilon})}{L(\widehat{\mu}, \widehat{\beta}_0, \widehat{\beta}_j, \widehat{\varepsilon})}$$

Por la teoría asintótica (Rohatgi. (1976); Lehmann (1999)), se rechazará H_0 si $-2 \log \Lambda > \chi_{4-3=1}^2$, al nivel de significancia α .

En el caso de que se rechace la hipótesis nula, habrá evidencia a un nivel α de significancia, que existe tendencia en las observaciones de Y. Si todas las covariables resultaran significativas que fueran cero, entonces se tendría que las covariables no influyen en la tendencia de Y. El modelo más recomendable sería aquel que solo usa a la covariable tiempo.

Capítulo 5

Un ejemplo de aplicación

Introducción

La Ciudad de México es considerada una Megaciudad, pues en ella habitan más de 20 millones de personas; ésta Megaciudad tiene graves problemas de Contaminación Atmosférica y quizá, los más críticos en el mundo actualmente.

En el año 2000, se registraron en el mundo más de 411 ciudades con un millón de habitantes cada una. Actualmente este crecimiento es cada vez más explosivo, se presenta en mayor cantidad en las naciones que están en vías de desarrollo, principalmente en Asia (Molina y Molina (2002);Ambiente (2002);New (2006)).

Los sitios presentan ciertas características muy peculiares, hay una creciente demanda de infraestructura y de servicios básicos, un uso creciente de energía, demanda sin límites de fuentes no renovables, con necesidades de más vías de comunicación, necesidad de espacios físicos para vivir sin importar la destrucción de las áreas verdes y acabar con los pulmones de la ciudad que limpian el aire contaminado. También hay graves desequilibrios en las áreas económicas y sociales del país que obligan a las individuos a emigrar a las grandes ciudades en busca de mejores oportunidades económicas.

En este capítulo se trabaja con los procedimientos descritos en el capítulo anterior usando dos bases de datos: una contiene a todas las estaciones de monitoreo y la otra sólo se realiza para una estación de monitoreo.

5.1. Características de la Ciudad de México.

La Zona Metropolitana de la Ciudad de México está situada en una cuenca que restringe la mayor parte del tiempo la libre circulación de vientos y no hay una buena ventilación. Su crecimiento en los pasados 15 años, tanto de vehículos automotores, zonas habitacionales y en menor grado las industrias, han contribuido a un gran reactor químico virtual, en el que todas las condiciones de alimentación de reactivos están presentes para que los fenómenos físicos naturales locales funcionen como agitadores de esa mezcla química activa que, aunado a la radiación solar, es un activador energético.

La dirección dominante de los vientos es del noreste al noroeste. Muchos de sus

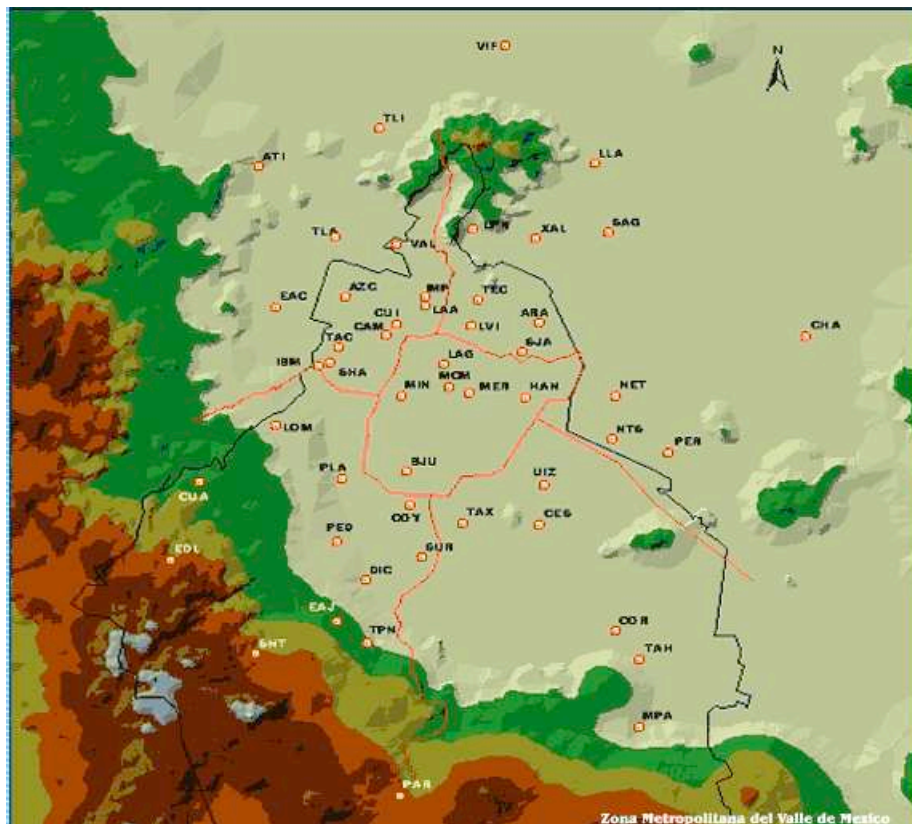


Figura 5.1: Ubicación de las estaciones de la ZMCM (Fuente: SIMAT, RAMA)

habitantes están en pobreza extrema, cubriendo aproximadamente 1500 kilómetros cuadrados y contando aproximadamente con 3.5 millones de vehículos. Operan más de 30,000 industrias, 12,000 empresas de servicios y alrededor de 3 millones de vehículos automotores. De acuerdo con el inventario de emisiones de la Zona Metropolitana de la Ciudad de México (ZMCM), se emiten alrededor de 4 millones de toneladas de contaminantes atmosféricos anualmente (I.N.E.G.I. (1999); Ambiental (2002); Correa (2004)).

En este capítulo se aplica la metodología desarrollada en el capítulo anterior para el contaminante más dañino, el ozono. La base de datos de la Ciudad de México tiene aspectos muy interesantes de investigación, como son los mecanismos de formación, de transporte y de distribución del ozono y de las partículas secundarias en las atmósferas urbanas.

Enseguida se presenta la ubicación geográfica de las Estaciones de Monitoreo del SIMAT (Figura 5.1); éstas se fueron creando como una necesidad de medir y evaluar la los altos niveles de contaminación Ambiental que han existido en la Ciudad de México.

Con el fin de obtener un registro continuo y extenso de la contaminación del aire en la Ciudad de México, se estableció un sistema denominado Red Automático de Monitoreo Atmosférico (RAMA), el cual se puso en operación a finales de 1985, contando en ese tiempo con 25 estaciones automáticas, distribuidas en diferentes puntos de la ciudad. En el 2004 contaba con 51 estaciones de monitoreo, integradas en 3 subsistemas: la Red Automática, la Red Manual de Monitoreo Atmosférico (REDMA), Red de Parámetros Meteorológicos (REDMET) y la Red de Depósito Atmosférico (REDDA). Las variables meteorológicas que se miden son: la temperatura (temp) en grados centígrados ($^{\circ}\text{C}$), la humedad relativa (hr) en porcentaje (%), la dirección del viento (wdp) en grados azimut, y la velocidad del viento (wsp) en metros por segundo (m/s) (Ambiental (2002)).

A continuación, en la Tabla 5.1, se presenta el nombre de cada una de los sitios de monitoreo, sus abreviaturas, el año en que iniciaron a transmitir las mediciones de ozono y la zona asignada por su ubicación geográfica. Actualmente también se miden otras variables ambientales y químicas de interés para las investigaciones ambientales.

Tabla 5.1: Estaciones Meteorológicas (Información del SIMAT)

Red Meteorológica de la ZMCM					
Estaciones	Año	Zona	Estaciones	Año	Zona
Pedregal (PED)	1986	Suroeste	Lagunilla (LAG)	1991	Centro
Merced (MER)	1986	Centro	ENEP Acatlán (EAC)	1991	Noroeste
Tlanepantla (TLA)	1986	Noroeste	Benito Juárez (BJU)	1992	Centro
Xalostoc (XAL)	1986	Noreste	Tacuba (TAC)	1993	Noroeste
Cerro Estrella (CES)	1986	Sureste	Taxqueña (TAX)	1993	Sureste
Plateros (PLA)	1986	Suroeste	Cuajimalpa (CUA)	1994	Suroeste
Hangares (HAN)	1986	Centro	Tlalpan (TPN)	1994	Suroeste
San Agustín (SAG)	1987	Noreste	Tláhuac (TAH)	1994	Sureste
Azcapozalco (AZC)	1988	Noroeste	Chapingo (CHA)	1995	Noreste
UAM Iztapalapa (UIZ)	1989	Sureste	Santa Ursula (SUR)	2000	Suroeste

Los investigadores han observado que las zonas geográficas de la Ciudad de México reflejan las características del tipo de contaminación que se produce principalmente

en la zona estudiada. En la zona Noroeste abunda el ozono y el bióxido de carbono, en la zona sureste predominan las partículas suspendidas y en la zona suroeste pulula el ozono (Correa, 2004). Por ejemplo, en la Estación Merced, que no es una estación con peores emisiones de ozono de la Ciudad, se violan los niveles de la norma oficial en general. A manera de ejemplo se presenta la gráfica de las emisiones de ozono en la Figura 5.2 Observando la base de datos se tiene que la Zona Suroeste (PED, PLA,

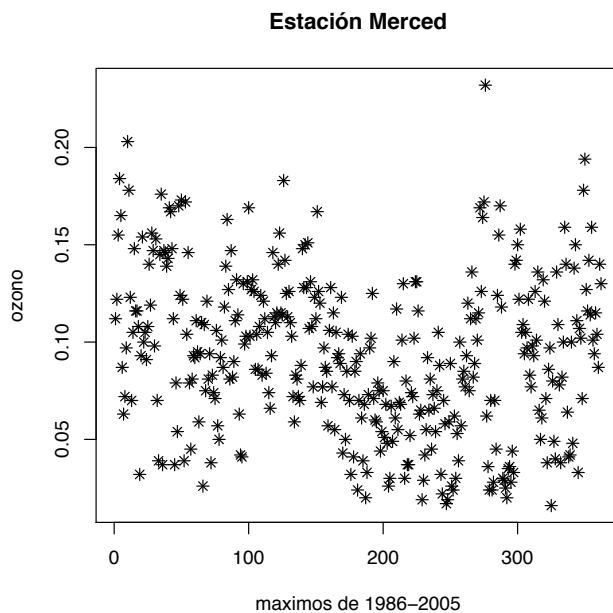


Figura 5.2: Estación Merced, niveles de ozono en 1986 y 2005.

TLA, CUA, SUR), tiene las peores emisiones en concentraciones de ozono

5.2. Construcción de bloques.

El modelo de distribución GVE tiene como hipótesis el tener una muestra aleatoria e independencia en los datos, por lo que la información dada por el SIMAT no cumple con este requisito ya que los datos están bajo las mismas condiciones ambientales y son consecutivas. Por lo tanto, se necesita hacer bloques con la información para remediarlo (Coles, 2004), mediante el siguiente procedimiento: La información dada por el SIMAT, de máximos por día en un año, se toma haciendo grupos de cada 4 días, de este subconjunto de máximos se elige al máximo del subconjunto y los demás se desechan. Este proceso se repite para cada subgrupo y la unión de todos los máximos por subconjunto forma una muestra aleatoria independiente de una distribución

$GVE(z, \mu, \sigma, \varepsilon)$ (Villaseñor (1996); Coles (2004)).

5.3. Procedimiento para la primera propuesta

Se supone que el modelo de máximos de bloques es correcto para la información z_1, z_2, \dots, z_k y que existen condiciones de regularidad en la distribución GVE. Para un año fijo, con la información anterior se forman bloques, se estiman los parámetros de $GVE(\hat{\mu}, \hat{\sigma}, \hat{\varepsilon})$ por medio de los estimadores de Máxima Verosimilitud que se pueden obtener usando el módulo ismev del paquete R (R Development Core Team (2007); McNeil and Stephenson (2004)) y así se estima la función Cuantil de ese año. A manera de ejemplo se presenta la tabla 5.2, con los valores de los estimadores obtenidos para la Estación Plateros de 1986 a 2005.

Tabla 5.2: Estimadores de Máxima verosimilitud anual en la estación Plateros con $p=0.05$.

año	$\hat{\varepsilon}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{Z}(p)$	año	$\hat{\varepsilon}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{Z}(p)$
1986	-0.123	0.037	0.1342	0.132	1996	-0.426	0.051	0.181	0.179
1987	-0.600	0.058	0.193	0.190	1997	-0.373	0.054	0.155	0.152
1988	-0.109	0.047	0.198	0.195	1998	-0.393	0.050	0.160	0.158
1989	-0.311	0.047	0.187	0.185	1999	-0.278	0.047	0.171	0.169
1990	-0.369	0.058	0.201	0.198	2000	-0.293	0.037	0.167	0.165
1991	-0.254	0.046	0.239	0.237	2001	-0.315	0.041	0.146	0.144
1992	-0.052	0.064	0.206	0.203	2002	-0.314	0.045	0.130	0.127
1993	-0.355	0.056	0.198	0.195	2003	-0.286	0.030	0.135	0.134
1994	-0.344	0.038	0.205	0.203	2004	-0.366	0.033	0.118	0.116
1995	-0.243	0.060	0.149	0.146	2005	-0.376	0.034	0.120	0.118

Este procedimiento se repite para todos los años en donde haya información de la estación de monitoreo de interés. Para las estaciones más antiguas con 20 años como PED, MER, TLA, XAL, CES, PLA y HAN, habrá 20 estimaciones de la forma $\hat{Z}(p)$, es decir, se tiene una muestra aleatoria $\hat{Z}_{1986}, \dots, \hat{Z}_{2005}$ con p fijo y alto. Además, esta es una muestra de observaciones normales (**Resultado 4.1**), por lo cual, se puede ajustar una Regresión Lineal simple, para investigar la tendencia de las concentraciones de ozono en el tiempo, para un p fijo y alto. Suponiendo que se cumplen los supuestos del Modelo de Regresión lineal, que el ozono es una variable aleatoria cuya distribución de probabilidad depende del tiempo, por lo cual dependencia lineal entre ellas, que las varianzas son constantes comunes entre los errores del modelo, que los valores de ozono son estadísticamente independientes y que la distribución asintótica de Y , para

cualquier valor del tiempo, es Normal.

Al realizar el ajuste de los datos, el mejor ajuste se obtuvo tomando en cuenta el coeficiente de correlación usando el modelo cuadrático, es decir, $Z(p) = \beta_0 + \beta_1 t + \beta_2 t^2 + \epsilon_i$ con $i = 1, \dots, n$.

Esta información permite hacer pruebas de hipótesis o intervalos de confianza para los parámetros y además realizar pronósticos acerca del valor esperado del ozono en el tiempo.

El mismo algoritmo anterior se repite para las 13 estaciones meteorológicas restantes de la ZMCM (SAG, AZC, UIZ, LAG, EAC, BJU, TAC, TAX, CUA, TPN, TAH, CHA, SUR).

5.4. Procedimiento para la segunda propuesta

Suponiendo que existen condiciones de regularidad en la distribución GVE. Para un año fijo, se forman bloques por mes. En este procedimiento se van a involucrar los máximos por mes de las variables ambientales de interés a partir de 1990 a 2006, porque la información esta mas cercana a la realidad ambiental. La relación que se establece para cada observación de ozono mensual máxima está en combinación lineal de máximos mensuales de covariables.

El paso siguiente es estimar los parámetros de la distribución $GVE(\hat{\mu}, \hat{\sigma} = \rho^{X^t \hat{\beta}}, \hat{\epsilon})$ por año.

Usando el módulo ismev de R, se calculan los estimadores de máxima Verosimilitud. Con esta información se pueden hacer pruebas de hipótesis o intervalos de confianza sobre la tendencia de la serie en las concentraciones de máximos de ozono $(\beta_0 + \beta_1 t)$, o investigar si hay efecto de alguna covariable, como la velocidad de viento $\beta_0 + \beta_1 wsp$.

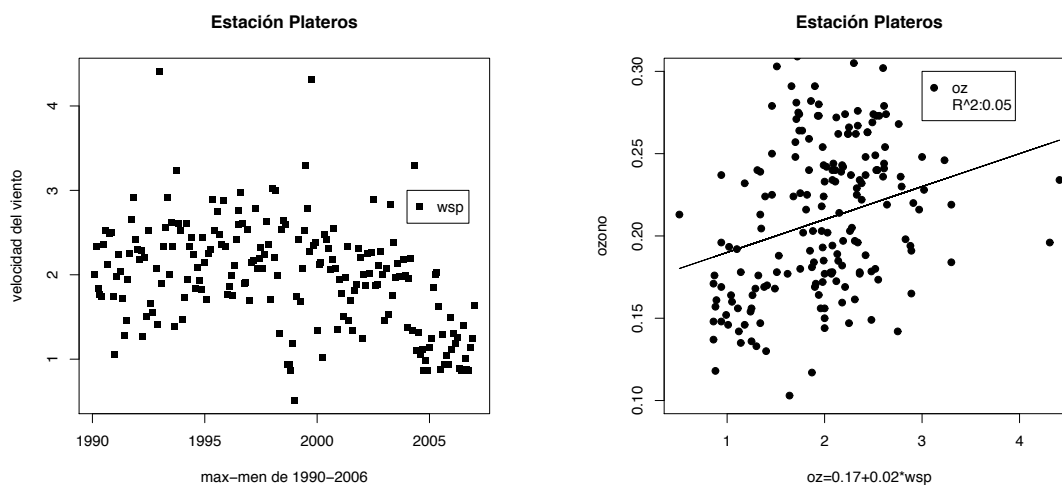
Se presenta la tabla 5.3 con las variables ambientales de la Estación Plateros para el año 1990. Los valores más grandes de humedad relativa se encuentran en los meses de julio, septiembre y junio, en cuanto a la variable temperatura los valores más grandes estan en mayo, marzo y junio, en cuanto a los valores mas grandes de la variable velocidad de viento, se encuentran en julio, octubre y septiembre.

A continuación se presenta la gráfica de los datos de ozono conjuntamente con cada variable ambiental de 1990 a 2006, Figura 5.3, Figura 5.4 y Figura 5.5.

La velocidad del viento estuvo en aumento en los años de 1990 a 1999, alrededor del año 2000 disminuyo en promedio. La regresión lineal entre el ozono y la velocidad del

Tabla 5.3: Información mensual de las variables ambientales en la estación Plateros (1990).

año 1990	tiempo	humedad relativa	temperatura	velocidad del viento
enero	1	77.8	23.3	2
febrero	2	71.6	25.3	2.34
marzo	3	76	27.6	1.84
abril	4	82.5	26.5	1.77
mayo	5	83.4	30.2	1.74
junio	6	95.7	27.5	2.36
julio	7	99.2	23.6	2.52
agosto	8	87.3	24.8	2.12
septiembre	9	96.2	24.1	2.49
octubre	10	82	25.3	2.5
noviembre	11	80.9	24.4	1.74
diciembre	12	99.1	25	1.05



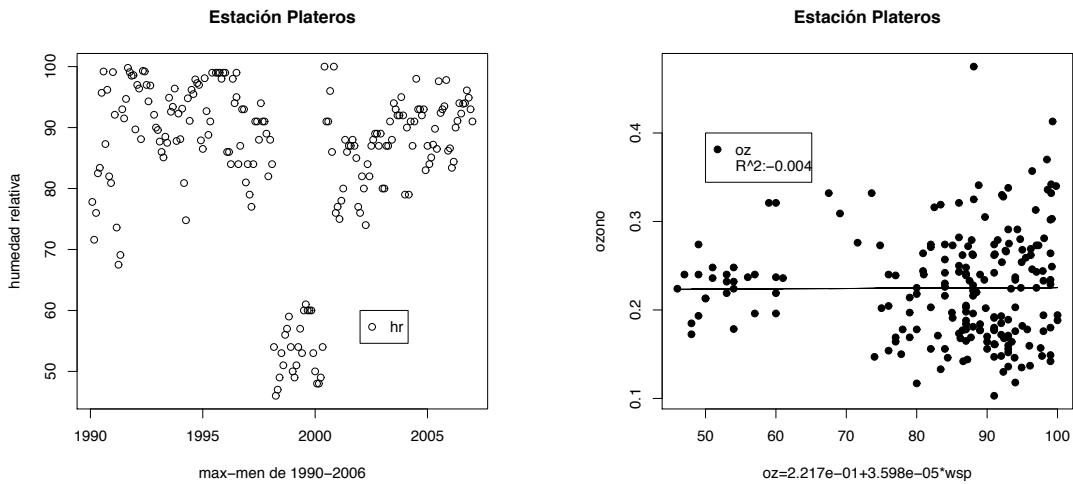
(a) Información mensual de la velocidad del viento.

(b) Línea estimada mensual del ozono con velocidad del viento .

Figura 5.3: Comportamiento mensual de los máximos de la velocidad del viento, en la Estación Plateros.

viento no tiene un buen ajuste (Figura 5.3).

La humedad relativa estuvo en aumento en los años de 1990 a 1997 con una variabilidad de 20 unidades, alrededor del año 2000 disminuye notablemente, teniendo un comportamiento no tan disperso, pero en el 2002 se vuelve a tener un comportamiento casi parecido al inicio de los 90. La regresión lineal entre el ozono y la humedad

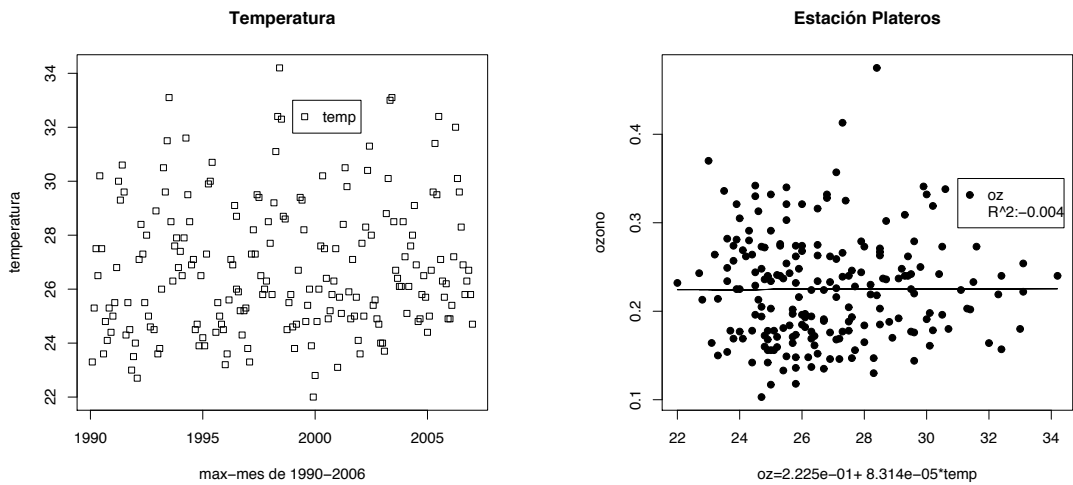


(a) Información mensual de la humedad relativa.

(b) Línea estimada del ozono con humedad relativa. .

Figura 5.4: Comportamiento mensual de los máximos de la humedad relativa, en la Estación Plateros.

relativa no tiene un buen ajuste (Figura 5.4).



(a) Información mensual de la temperatura.

(b) Línea estimada del ozono con temperatura. .

Figura 5.5: Comportamiento mensual de los máximos de la temperatura, en la Estación Plateros.

La temperatura estuvo en aumento de 1990 a 1999 con una variabilidad de 9 unidades, alrededor del año 2000 disminuye notablemente, teniendo un comportamiento

disperso, en el 2000 se disminuye notablemente, pero después de esta fecha la temperatura vuelve a ir aumentando en el tiempo. La regresión lineal entre el ozono y la temperatura no tiene un buen ajuste (Figura 5.5).

5.5. Resultados

En todas las estaciones meteorológicas de la ZMCM, los niveles más grandes de concentraciones de ozono se encuentran en el año 1991. La mayoría de las concentraciones de ozono en la Ciudad de México siempre están arriba de la norma oficial decretada por la ONU (0.11 ppm). De las estimaciones realizadas se observa que la tendencia del ozono va a la baja hasta 2005. Estos resultados coinciden con los trabajos de Jaimes (2004), Correa (2004) y Garfias *et al.* (2005), en donde se menciona que la tendencia de las concentraciones de ozono van a la baja.

5.5.1. Resultados de la primera propuesta

A continuación se muestran las figuras Figura 5.6, Figura 5.7, Figura 5.8, Figura 5.9, Figura 5.10, Figura 5.11 y Figura 5.12, de las gráficas obtenidas por el procedimiento.

Las estimaciones reflejan la tendencia del comportamiento de los máximos de las concentraciones de ozono y éstas son por abajo de los datos reales. Solo en la Estación Plateros casi coinciden.

Las estimaciones reflejan la tendencia del comportamiento de los máximos de las concentraciones de ozono y son por abajo de los datos reales.

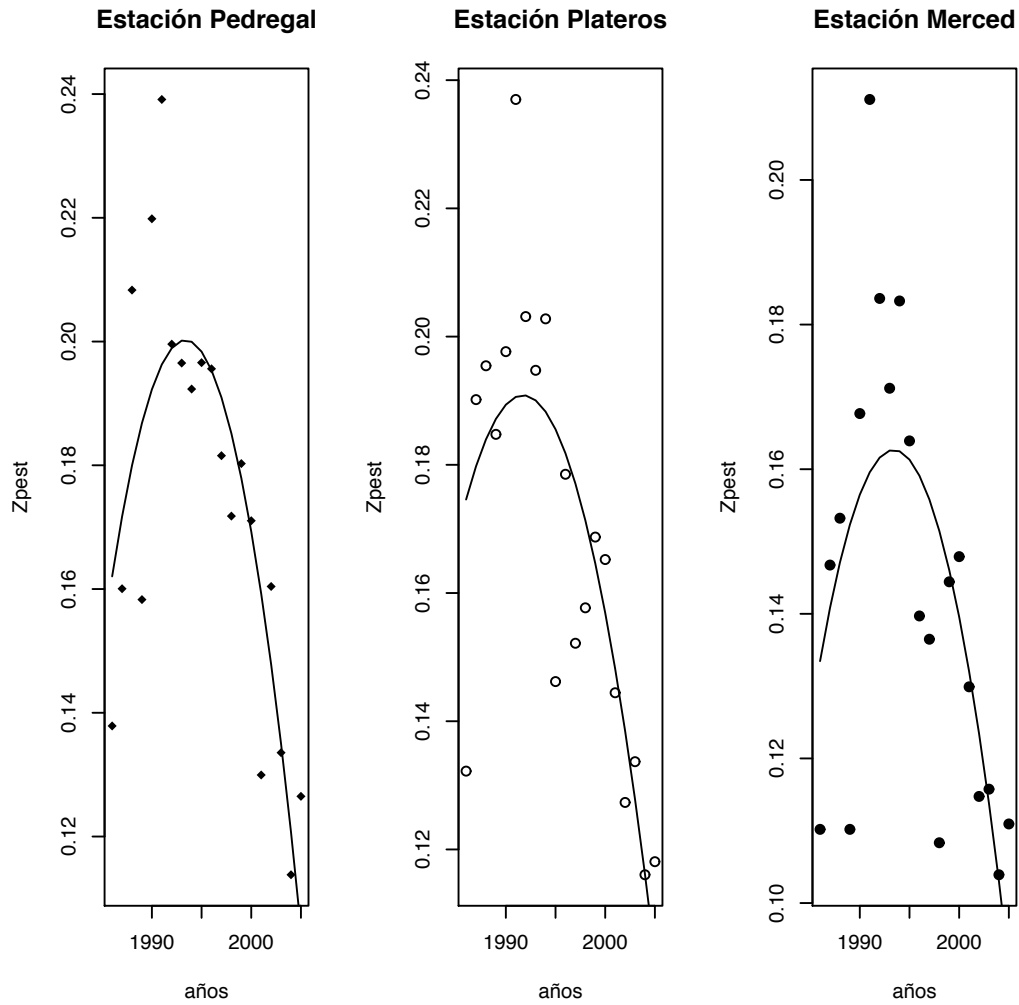


Figura 5.6: Estimaciones usando la Regresión Cuantil en $p = 0.05$, en las Estaciones Meteorológicas para el Valle de México 1.

Las estimaciones reflejan la tendencia del comportamiento de los máximos de las concentraciones de ozono y están por abajo de los datos reales.

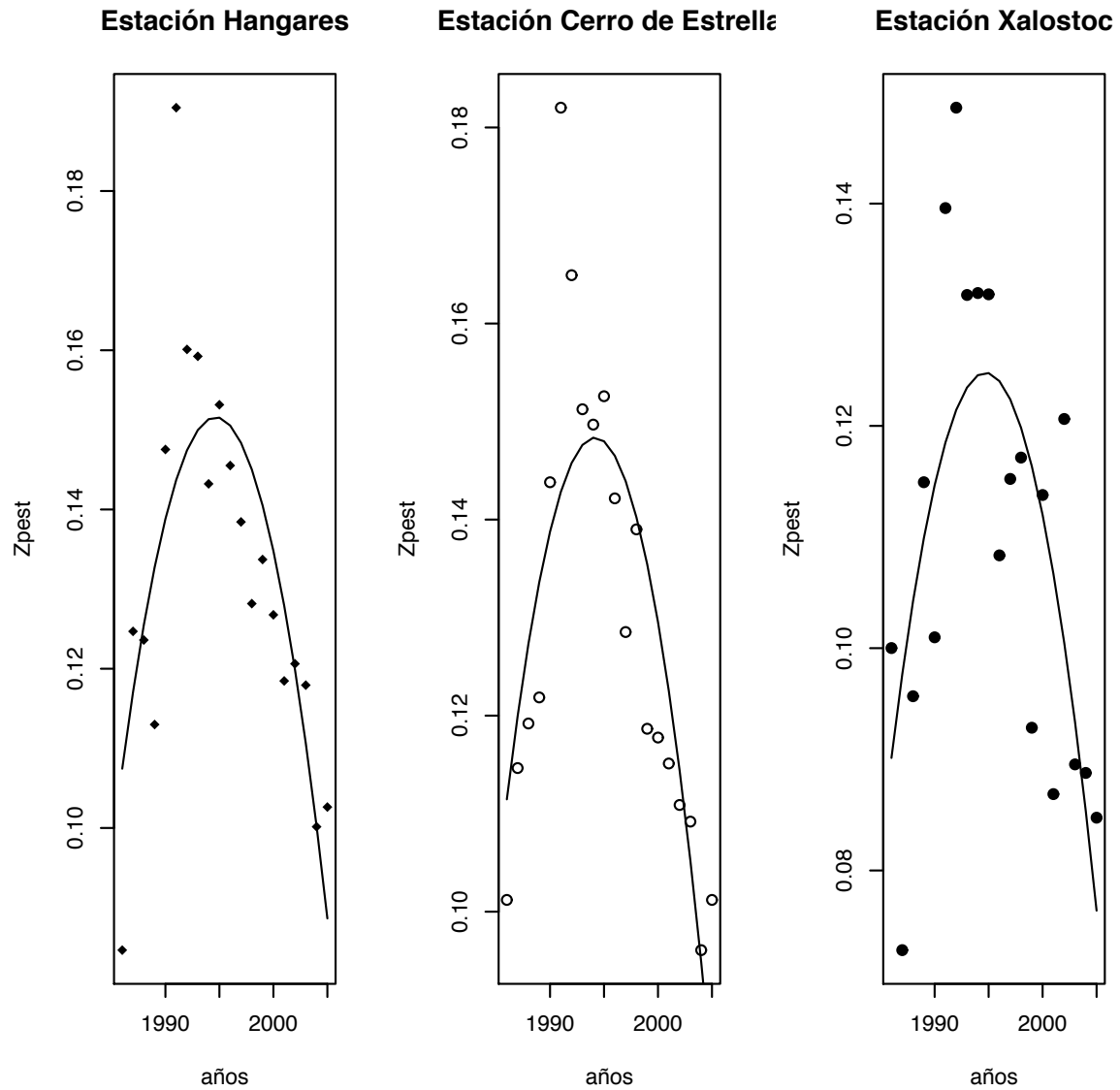


Figura 5.7: Estimaciones usando la Regresión Cuantil en $p = 0.05$, en las Estaciones Meteorológicas para el Valle de México 2.

Las estimaciones reflejan la tendencia del comportamiento de los máximos de las concentraciones de ozono por abajo de los datos reales.

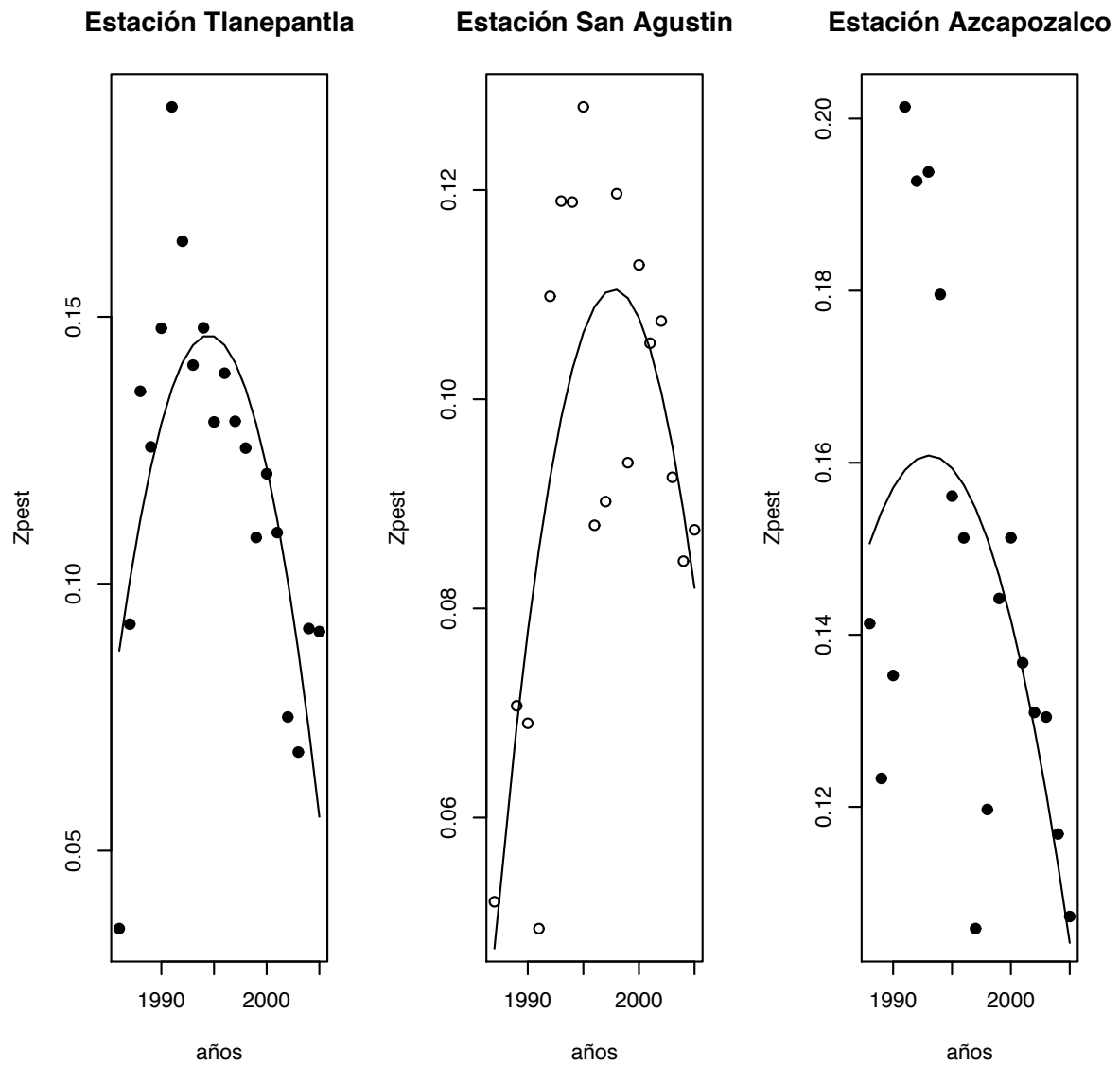


Figura 5.8: Estimaciones usando la Regresión Cuantil en $p = 0.05$, en las Estaciones Meteorológicas para el Valle de México 3.

Las estimaciones reflejan la tendencia del comportamiento de los máximos de las concentraciones de ozono por abajo de los datos reales.

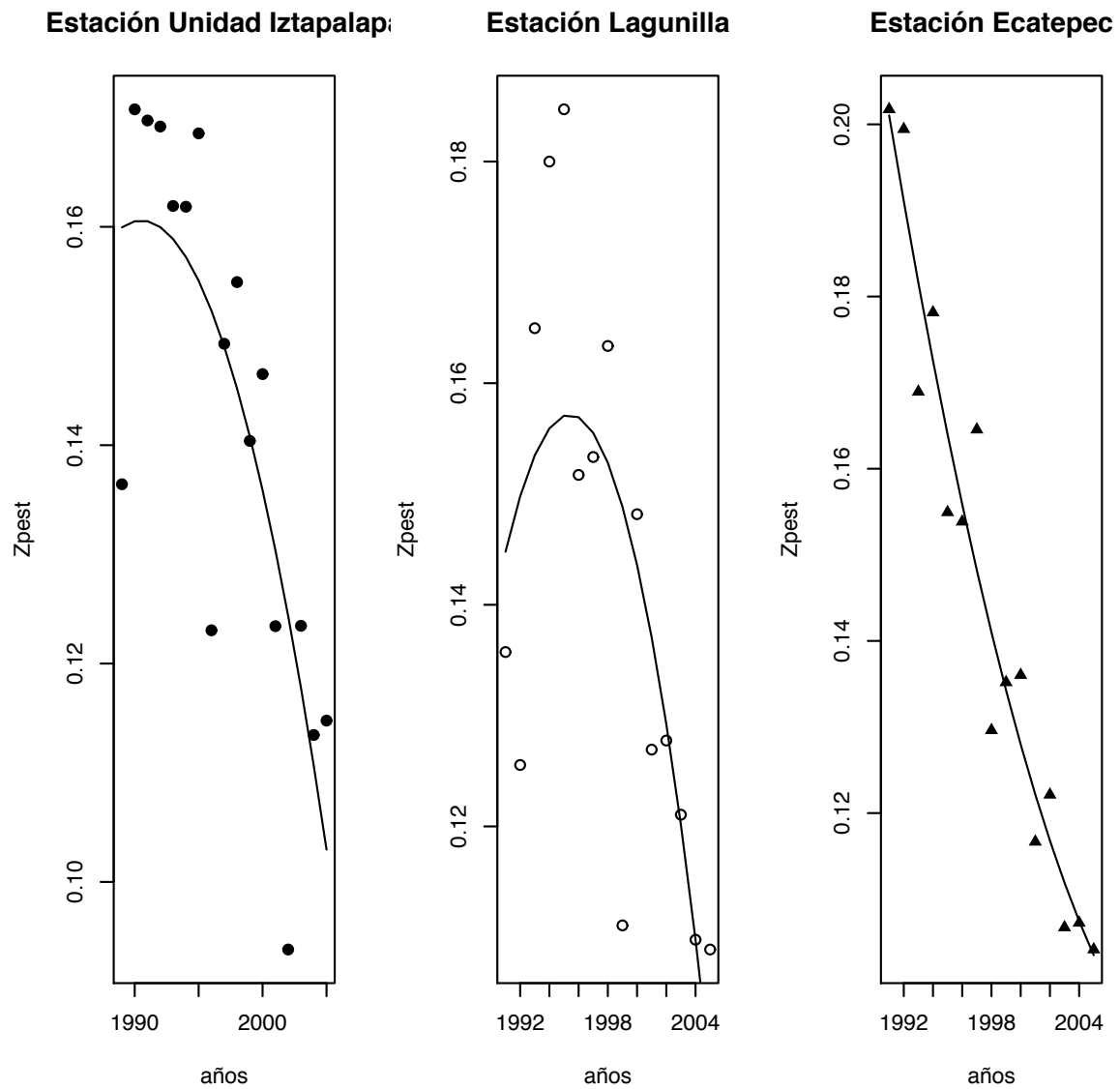


Figura 5.9: Estimaciones usando la Regresión Cuantil en $p = 0.05$, en las Estaciones Meteorológicas para el Valle de México 4.

Las estimaciones reflejan la tendencia del comportamiento de los máximos de las concentraciones de ozono por abajo de los datos reales.

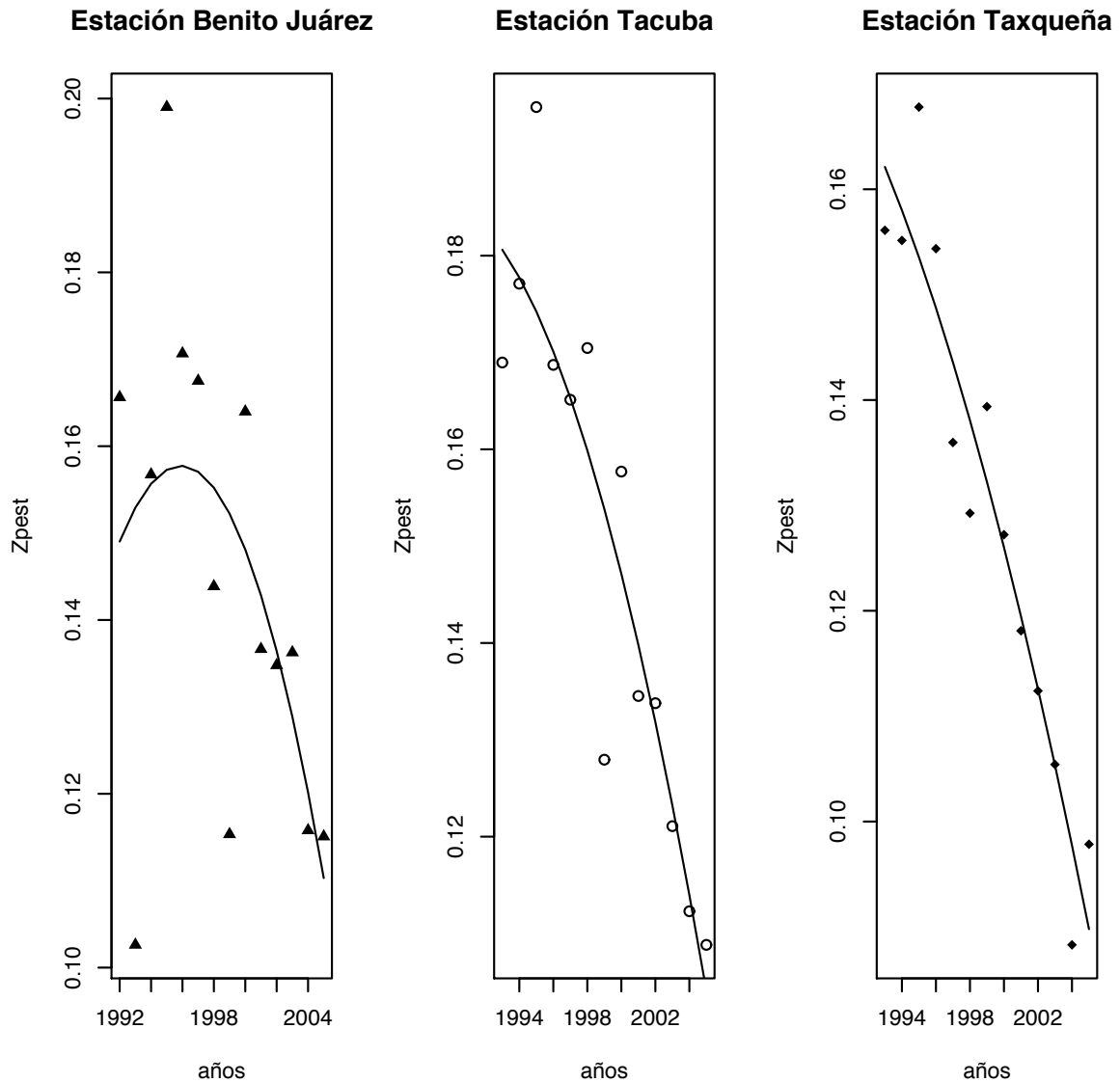


Figura 5.10: Estimaciones usando la Regresión Cuantil en $p = 0.05$, en las Estaciones Meteorológicas para el Valle de México 5.

Las estimaciones reflejan la tendencia del comportamiento de los máximos de las concentraciones de ozono por abajo de los datos reales. En la Estación Sur gráficamente parece que si hay un buen ajuste.

Realizando otros gráficos, se presentan las comparaciones de las distribuciones teóricas y empíricas para las 20 estaciones, el comportamiento en general es de dos formas: 18 de las gráficas de las distribuciones teóricas están subestimadas, con respecto a la distribución empírica, sólo en la estación Plateros su ajuste es perfecto, ya que se

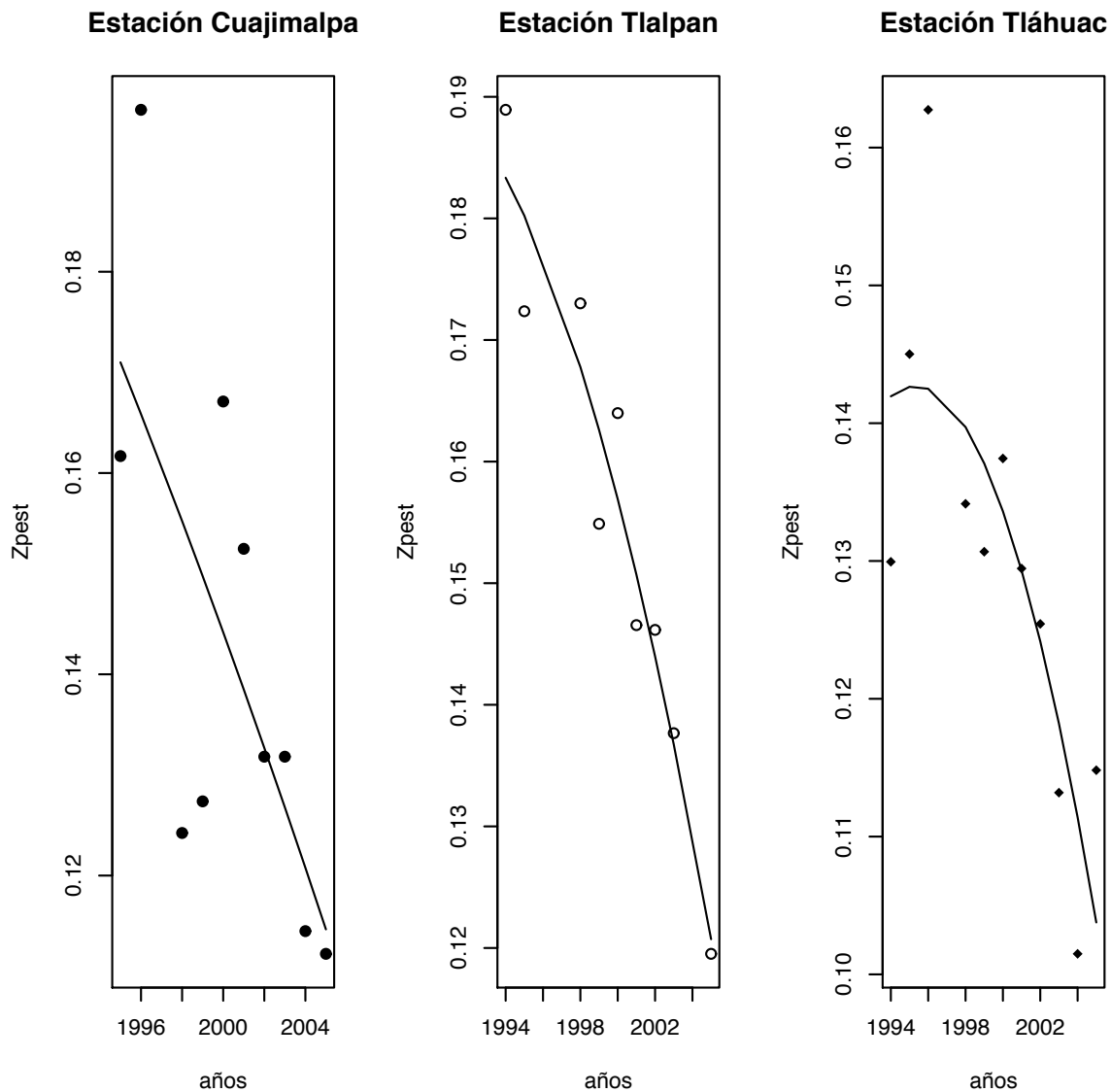


Figura 5.11: Estimaciones usando la Regresión Cuantil en $p = 0.05$, en las Estaciones Meteorológicas para el Valle de México 6.

empalman casi ambas gráficas y en la Estación Chapingo el ajuste fue malo porque no siguió la tendencia de la información. Todo lo dicho anteriormente se puede observar en la Figura 5.13, Figura 5.14, Figura 5.15, Figura 5.16, Figura 5.17, Figura 5.18 y Figura 5.19. Este resultado puede que se deba a que falta información en el modelo para poder ajustar mejor. De aquí, se presenta una segunda propuesta de modelación para tratar de realizar un mejor ajuste.

En estas gráficas se presentan los datos teóricos del modelo propuesto y la información real por cada estación. Siendo notable la concordancia de esta información en la

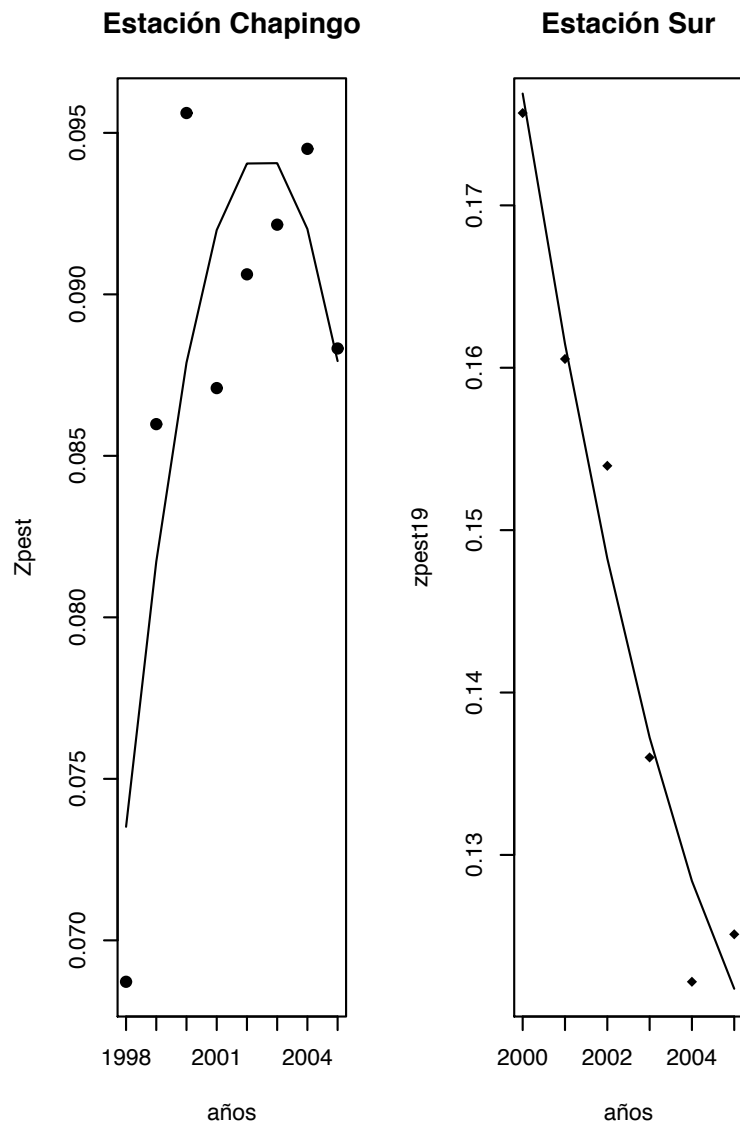


Figura 5.12: Estimaciones usando la Regresión Cuantil en $p = 0.05$, en las Estaciones Meteorológicas para el Valle de México 7.

Estación Plateros.

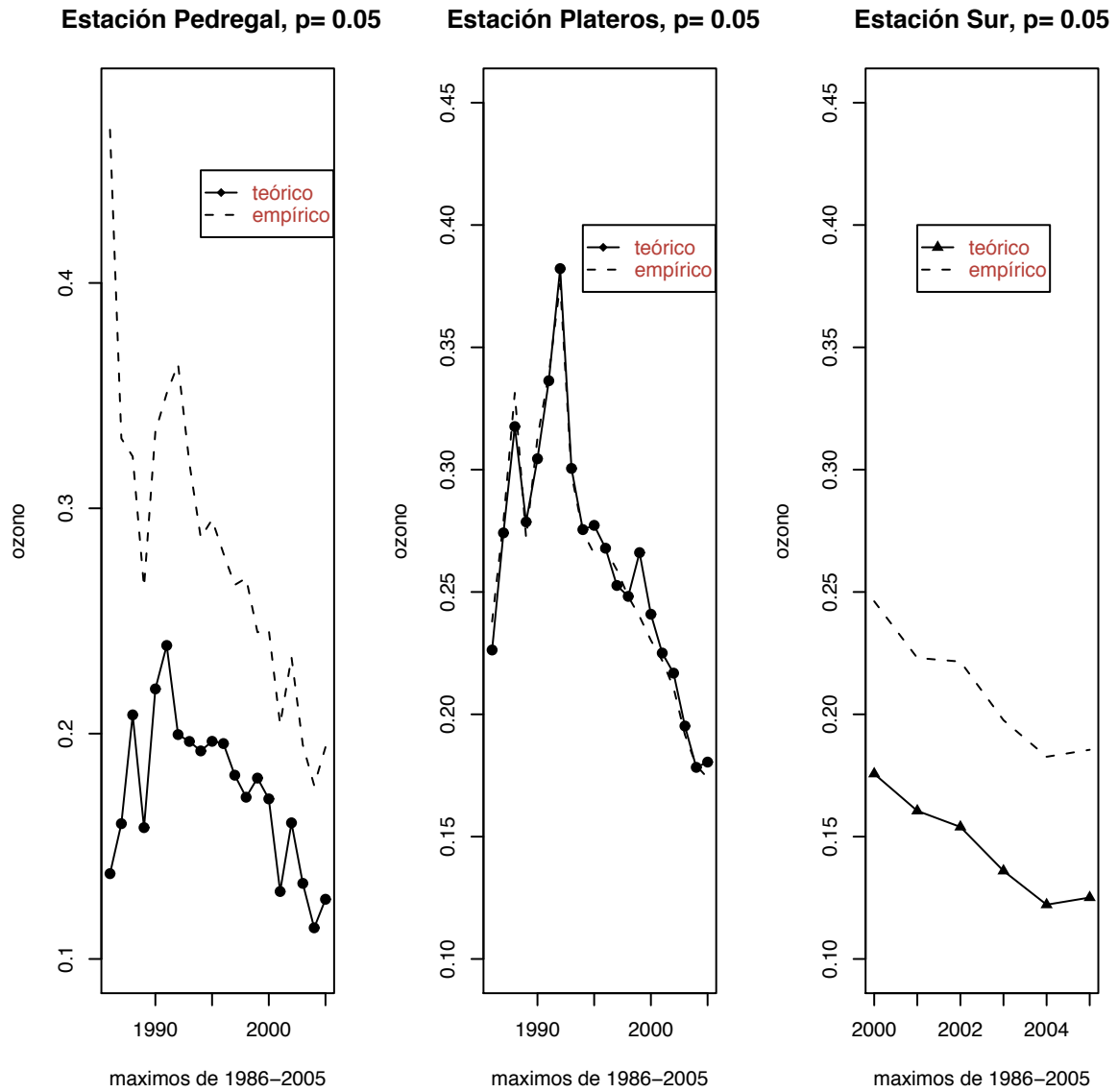


Figura 5.13: Comparaciones entre los resultados teóricos y empíricos en las Estaciones Pedregal, Plateros y Sur, en $p=0.05$.

En estas gráficas se presentan los datos teóricos del modelo propuesto y la información real por cada estación.

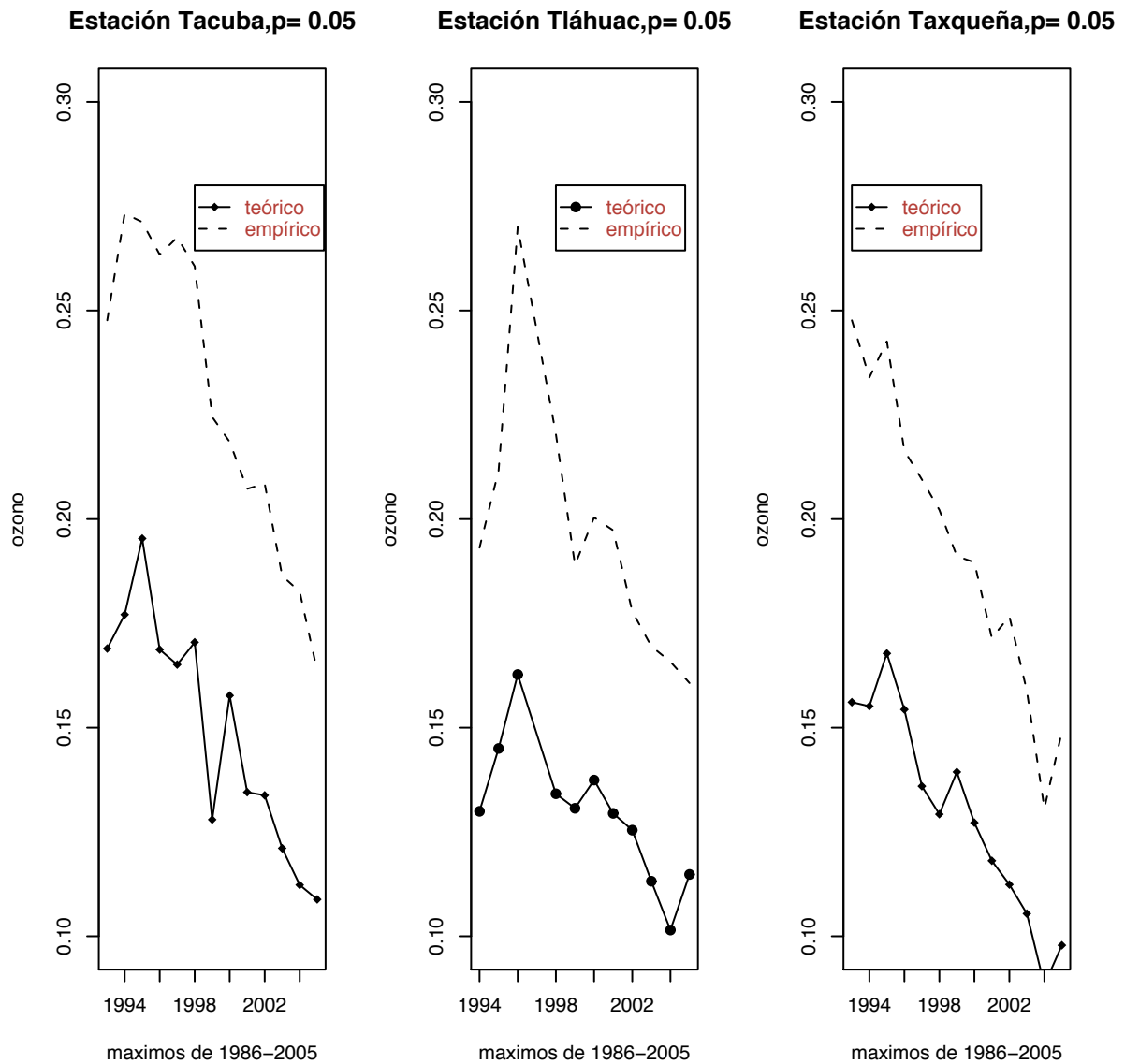


Figura 5.14: Comparaciones entre los resultados teóricos y empíricos en las Estaciones Tacuba, Tláhuac y Taxqueña, al $p=0.05$.

En estas gráficas se presentan los datos teóricos del modelo propuesto y la información real por cada estación.

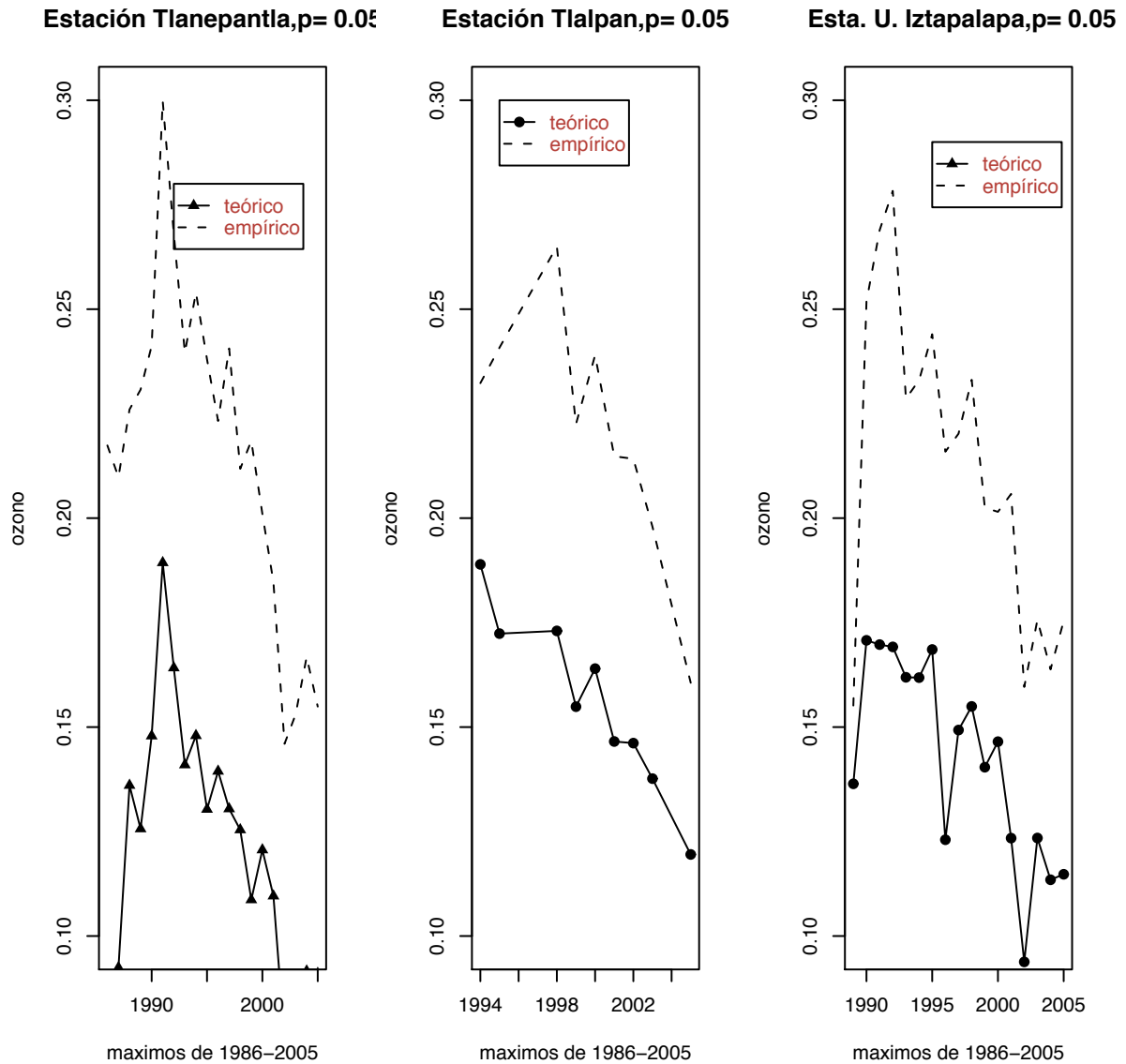


Figura 5.15: Comparaciones entre los resultados teóricos y empíricos en las Estaciones Tlanepantla, Tlalpan y UAM Iztapalapa, al $p=0.05$.

En estas gráficas se presentan los datos teóricos del modelo propuesto y la información real por cada estación.

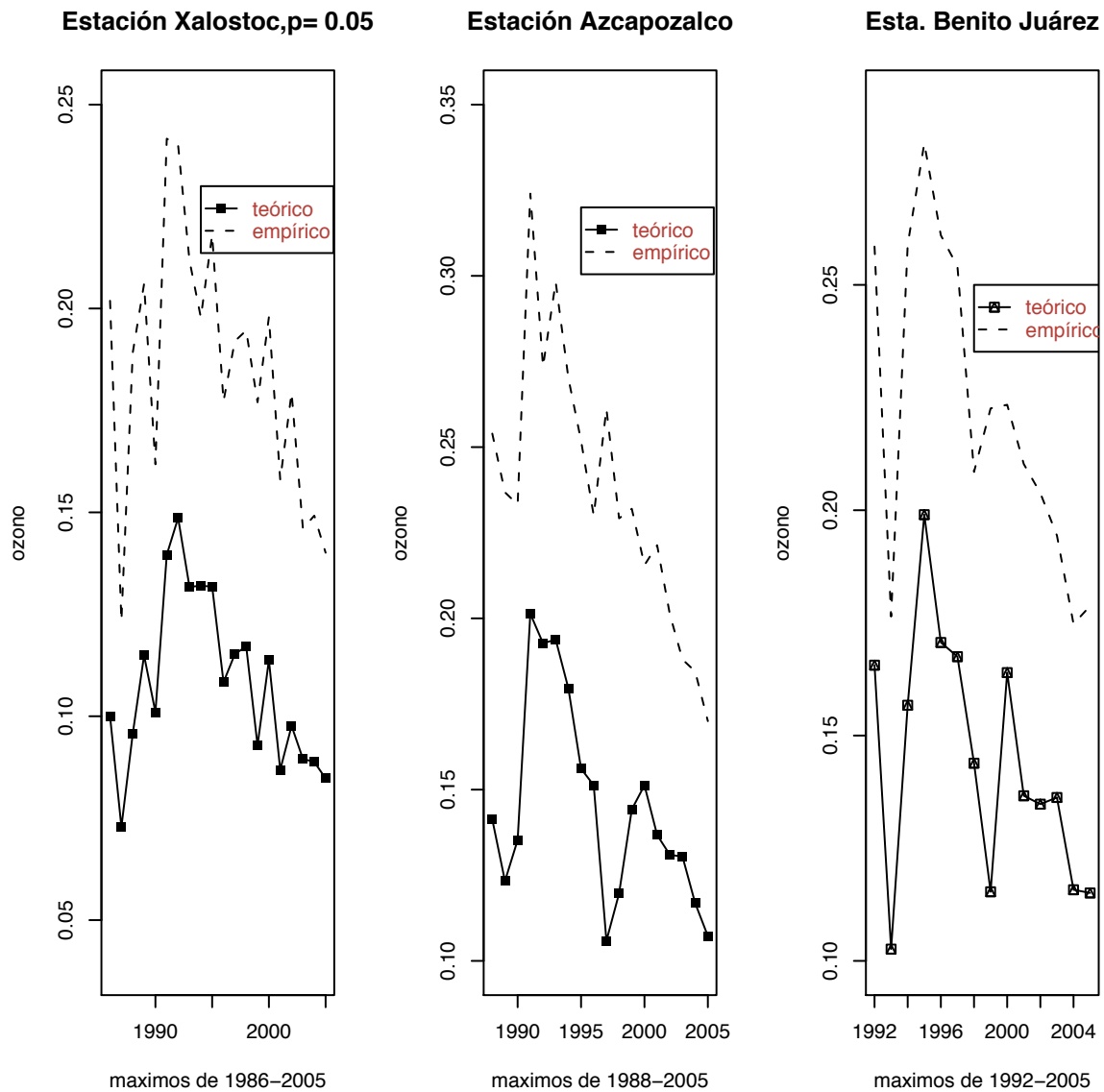


Figura 5.16: Comparaciones entre los resultados teóricos y empíricos en las Estaciones Xalostoc, Azcapozalco y Benito Juárez, al $p=0.05$.

En estas gráficas se presentan los datos teóricos del modelo propuesto y la información real por cada estación. En la Estación Chapingo el modelo propuesto no sigue la tendencia de los datos (quizá se deba que la estación ha tenido problemas con el mal funcionamiento de los aparatos).

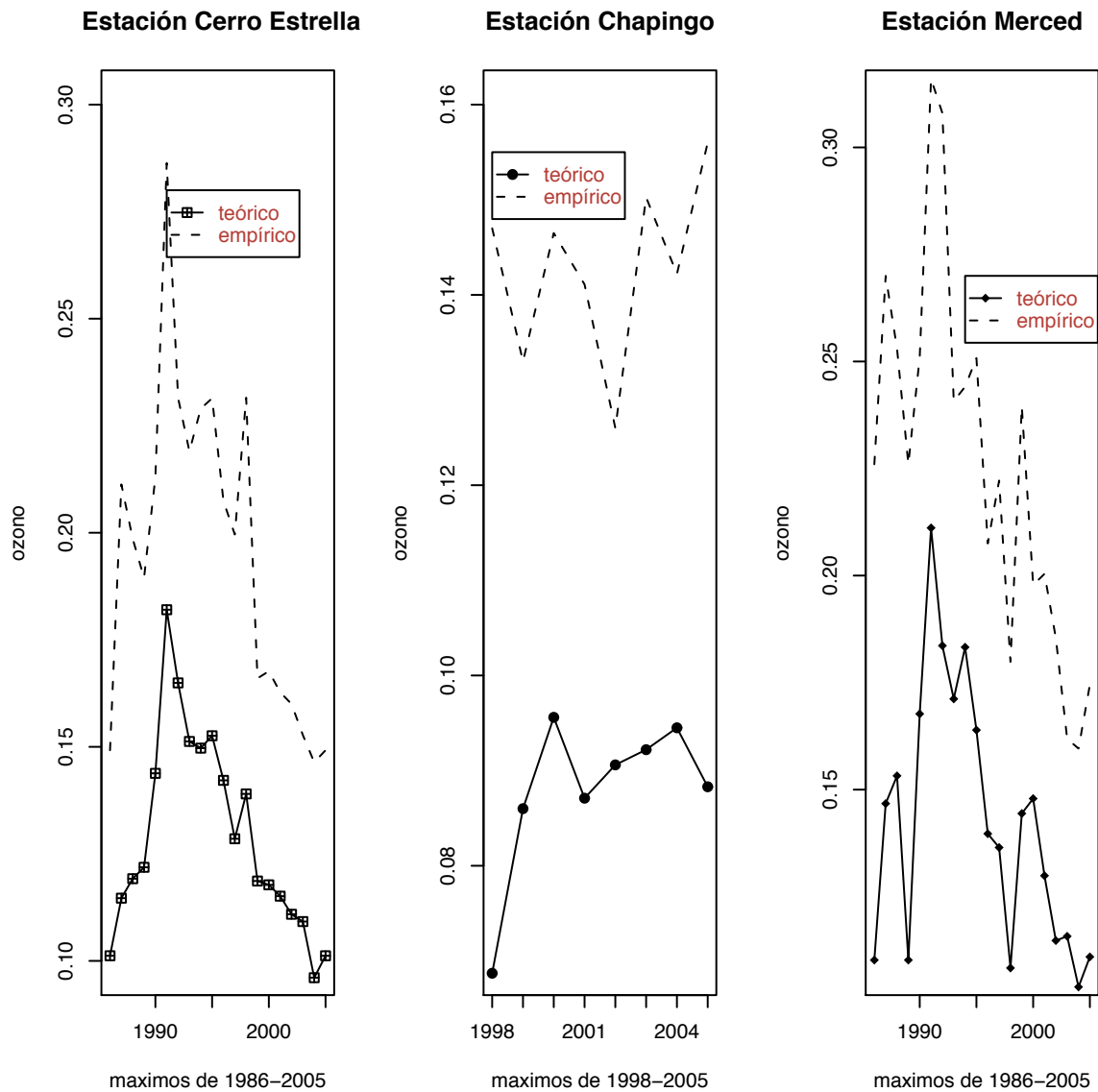


Figura 5.17: Comparaciones entre los resultados teóricos y empíricos en las Estaciones, Cerro de la Estrella, Chapingo, Merced, al $p=0.05$.

En estas gráficas se presentan los datos teóricos del modelo propuesto y la información real por cada estación.

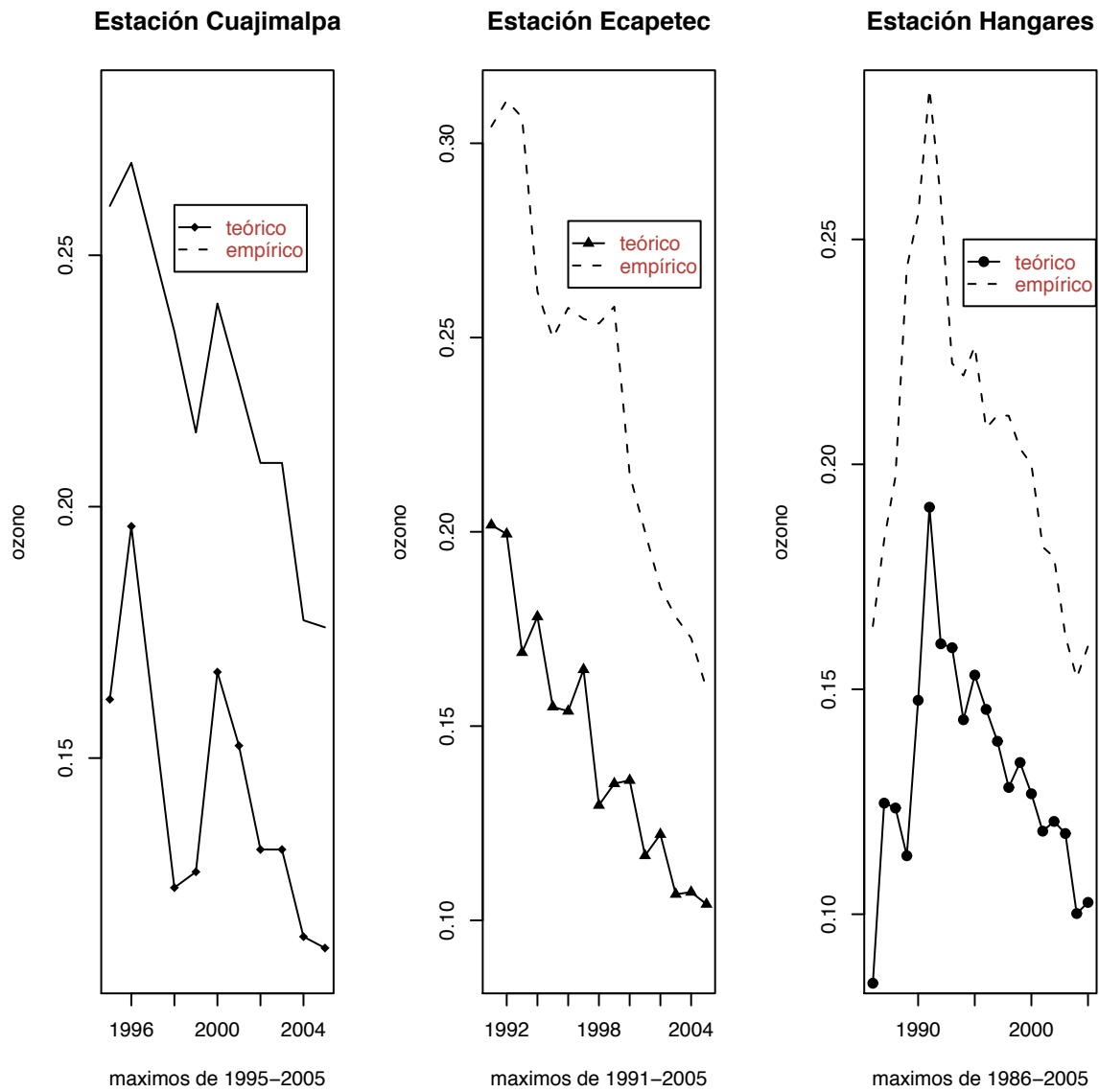


Figura 5.18: Comparaciones entre los resultados teóricos y empíricos en las Estaciones Cuajimalpa, Ecatepec y Hangares, al $p=0.05$.

En estas gráficas se presentan los datos teóricos del modelo propuesto y la información real por cada estación.

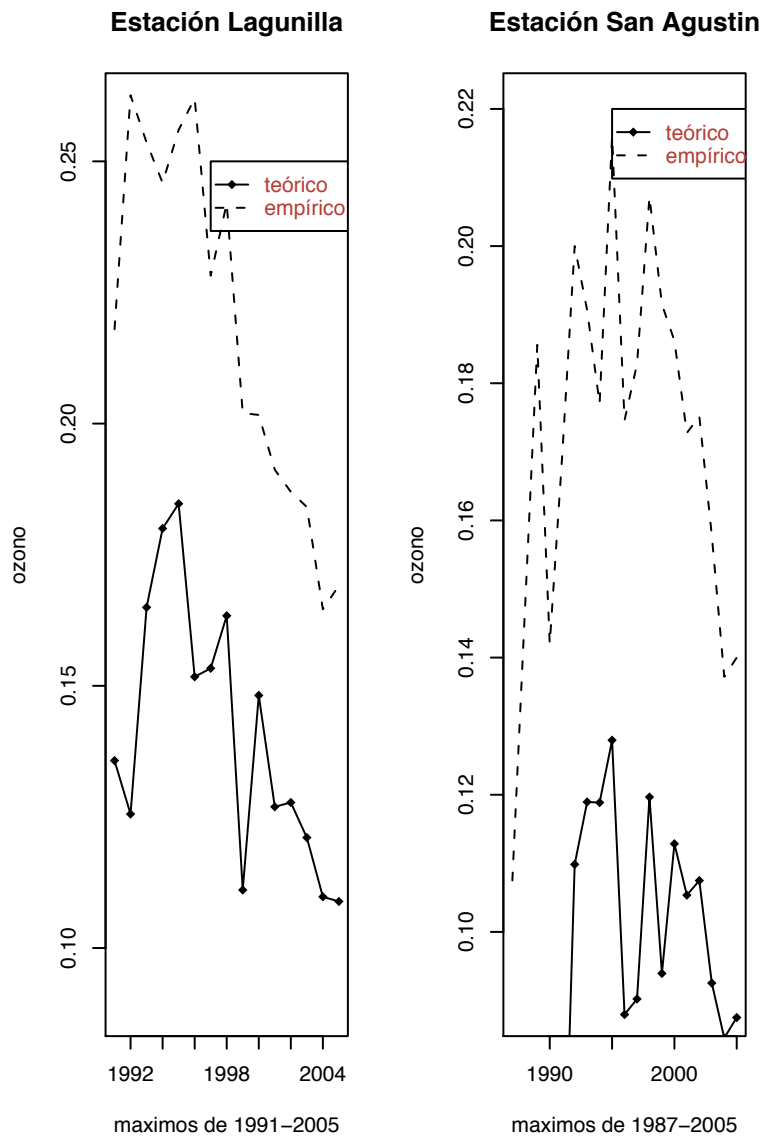


Figura 5.19: Comparaciones entre los resultados teóricos y empíricos en las Estaciones Lagunilla y San Agustín, al $p=0.05$.

5.5.2. Resultados de la segunda propuesta

Para esta propuesta, se maneja una salida del módulo `isnev` (**gev.fit**), donde se utilizan aproximaciones numéricas para calcular los estimadores máximo verosímil ($\hat{\mu}, \hat{\sigma}, \hat{\varepsilon}$) y la matriz de covarianzas del modelo (V),

$$\text{var}(\hat{z}_p) \approx \nabla z_p^T V \nabla z_p \quad y$$

$$\nabla z_p^T = \left[\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \varepsilon} \right]$$

evaluado en $(\hat{\mu}, \hat{\sigma}, \hat{\varepsilon})$ (Coles, 2001). Por lo que se pueden construir intervalos de confianza o hacer pruebas de hipótesis para poder incorporar alguna covariable en el modelo. Posteriormente, con la salida **gev.fit**, se pueden hacer gráficos de diagnósticos para verificar si el modelo propuesto estima bien (**gev.diag**).

A continuación en la Figura 5.20 se puede observar diagnósticos del comportamiento del modelo supuesto. La gráfica de la Probabilidad o de Cuantil-Cuantil aplicada a la serie de tiempo de la estación Plateros, indican que las observaciones son adecuadas al modelo de distribución GVE: cada conjunto de puntos graficados esta cercano a una línea recta. En la gráfica de Nivel de Retorno, los estimadores están cercanos a cero, la curva estimada es cercana a una línea recta excepto el último dato que todavía está dentro de la banda de confianza. Finalmente, la función de densidad estimada parece consistente con el histograma de los datos. Por lo cual, las cuatro gráficas de diagnostico llevan a respaldar que el modelo estimado es de tipo distribución GVE.

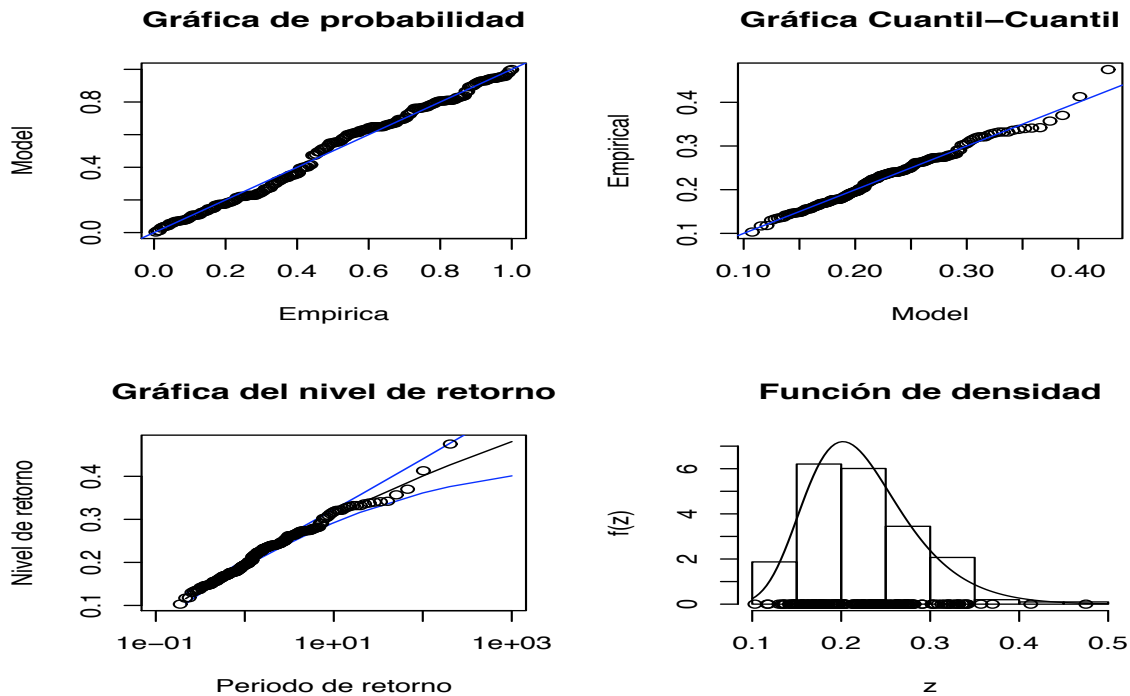


Figura 5.20: Análisis descriptivo de los supuestos distribucionales bajo GVE.

En seguida, se presentan algunos resultados (Tablas 5.4, 5.5, 5.6 y 5.7) sobre las asociaciones de las covariables en el modelo.

En la Tabla 5.4, se muestra al modelo

$$GVE(\sigma = \exp\{\beta_0 + \beta_1 \text{wsp}\}).$$

Note que se relaciona al parámetro de escala (σ) de la distribución GVE con la covariable velocidad del viento (wsp). Usando la salida del módulo ismev se obtienen los estimadores máximo verosímil, los cuales se reportan en el siguiente cuadro.

Tabla 5.4: Modelo con Ozono y Velocidad de viento (wsp)

Parámetros	Esti-M.V.	des.estan.	Intervalo de Confianza al 95 %
$\hat{\beta}_0$	-0.89	0.238	(-1.85,0.063)
$\hat{\beta}_1$	-0.95	0.101	(-1.57,-0.326)

Como ejemplo 2, se presenta el modelo estadístico siguiente

$$GVE(\sigma = \exp\{\beta_0 + \beta_1 t\}).$$

Note que se relaciona al parámetro de escala (σ) de la distribución GVE con la covariable tiempo (t). Usando la salida del módulo ismev se obtienen los estimadores máximo verosímil, los cuales están en la tabla siguiente.

Tabla 5.5: Modelo con Ozono y tiempo (t)

Parámetros	Esti-M.V.	des.estan.	Intervalo de Confianza al 95 %
$\hat{\beta}_0$	-2.017	0.055	(-2.48,-1.553)
$\hat{\beta}_1$	-0.0008	0.000001	(-1.01,-0.005)

Como ejemplo 3, se presenta el modelo estadístico siguiente

$$GVE(\sigma = \exp\{\beta_0 + \beta_1 \text{temp}\}).$$

Note que se relaciona al parámetro de escala (σ) de la distribución GVE con las covariable temperatura (temp). Usando la salida del módulo ismev se obtienen los estimadores máximo verosímil, los cuales están en la tabla siguiente.

Tabla 5.6: Modelo con Ozono y covariable: temperatura

Parámetros	Esti-M.V.	des.estan.	Intervalo de Confianza al 95 %
$\hat{\beta}_0$	-0.502	0.056	(-0.96,-0.04)
$\hat{\beta}_1$	-0.057	0.000001	(-0.06,-0.05)

Ahora como ejemplo 4, se presenta el modelo estadístico siguiente

$$GVE(\sigma = \exp\{\beta_0 + \beta_1 t\}).$$

Note que se relaciona al parámetro de escala (σ) de la distribución GVE con la covariable tiempo (t). Usando la salida del módulo ismev se obtienen los estimadores máximo verosímil, los cuales están en la tabla siguiente.

Tabla 5.7: Modelo con Ozono y covariable: tiempo

Parámetros	Esti-M.V.	des.estan.	Intervalo de Confianza al 95 %
$\hat{\beta}_0$	-1.68	0.19	(-2.53,-0.82)
$\hat{\beta}_1$	-0.008	0.000001	(-0.010,-0.004)

5.6. Discusión

5.6.1. Discusión de la primera propuesta

Se observaron en los años 1991 y 1992, los niveles más altos de concentraciones de ozono en todas las estaciones meteorológicas. En 1993, se empezó a bajar los niveles de concentraciones de ozono en la Ciudad de México debido a que hubo cambios de políticas en el tiempo, se cambiaron los convertidores catalíticos de dos a tres vías para mejorar la combustión de los motores, se hicieron cambios químicos en la fórmula de la gasolina para no emitir plomo y mejorar la combustión en la gasolina. Además, la refinera 18 de marzo cerró sus instalaciones para trasladarse a otro lugar y hubo restricciones para poder circular en los autos y transportes que atraviesan la Ciudad de México (el hoy no circula). Finalmente se establecieron instituciones y lugares físicos que vigilan que no se sobrepasen ciertas normas establecidas de contaminación.

Para 1995 en adelante los niveles de ozono vuelven a aumentar, aunque no tanto como en el periodo de 1991-1992. Hay justificaciones para esta situación que esta en relación de ser una Megaciudad, pues en la Ciudad de México sigue creciendo la población,

con demanda de transporte, de construcción de más casas habitacionales, más centros comerciales, más industrias y siguen eliminando áreas verdes para construir más vías de comunicación.

Los resultados obtenidos en este trabajo son: los niveles más altos de ozono en la zona metropolitana de la Ciudad de México se tienen en la estación Plateros, Pedregal y Merced. Las zonas más contaminadas están en la zona Suroeste, Noroeste y Centro. Este resultado es debido a la dirección de los vientos y a la importancia de las zonas culturales y económicas que representan. Estos resultados están de acuerdo con otros trabajos como de González (2000), Jaimes et al. (2004), Correa (2004) y Garfias et al. (2005).

La falta de ajuste del procedimiento quizá se deba a no haber incluido variables ambientales en el modelo, debido a que la ZMCM es un lugar muy afectado por ellas.

5.6.2. Discusión de la segunda propuesta

Esta estación es la estación más contaminada desde 1986 a la fecha, se ha visto en la serie de tiempo, que su contaminación ha disminuído a la fecha. En el procedimiento, la Estación Plateros la velocidad del viento, el tiempo y temperatura son variables importantes que están contribuyendo a los altos niveles de concentración de ozono en forma mensual. En este lugar viene a acumularse todo lo que está en el aire de la ciudad de México. Hay poca velocidad de viento y condiciones propicias para que se realicen las reacciones químicas que generan al ozono en el tiempo. Las variables ambientales tienen relaciones muy complejas a nivel micro-ambiental, que se están investigando, tienen que ver con modelos dinámicos de capas de vientos, humedad relativa, temperatura, dirección de viento, velocidad del viento y muchas variables físico-químicas. En esta segunda propuesta el parámetro de escala es una función que depende del efecto de las covariables en forma mensual. En los modelos encontrados los coeficientes de los parámetros de las variables ambientales tienen pendientes negativas.

Capítulo 6

Conclusiones

6.1. Conclusiones

Se desarrollo un modelo paramétrico sencillo con fin de estimar tendencias en series de tiempo bajo el supuesto de tener observaciones con distribución Generalizada de Valores Extremos.

También se propone un procedimiento para estimación de tendencia en los niveles de ozono usando covariables mensuales.

Los modelos propuestos tienen la ventaja de que se pueden hacer ajustes de ozono como los modelos de regresión lineal a partir de la convergencia de la Función Cuantil a una normal estandarizada. Por lo cual se pueden obtener pruebas de hipótesis e intervalos de confianza de los parámetros de interés.

Se obtiene la distribución de la función cuantil estimada.

Aplicando los métodos estadísticos propuesto en este trabajo a la base de datos de ozono urbano en el valle de México, se concluye que las niveles altos de ozono van a la baja en todas las estaciones meteorológicas, aunque no se ha alcanzado el nivel de ozono impuesto por la Organización Mundial para la Salud. La tendencia del comportamiento del ozono estimado no es estrictamente decreciente en todo el periodo analizado, pues las primeras estaciones que iniciaron en 1986 presentan en general tres cambios de pendiente hasta 2005. Los cambios en las estaciones que comenzaron después de 1987, son más notables en las estimaciones de niveles de ozono, presentan cambios de pendiente en 1990-1991, 1995-1996 y 2000.

En general, se puede observar que la mayoría de las gráficas se pueden dividir en dos partes, antes de año 1996 con tendencia creciente y después de 1996 con tendencia decreciente. Las estimaciones de ozono que presentan clara tendencia en disminución año con año son: Cuajimalpa, Sur, Tlalpán y Taxqueña. Se detectaron cambios en las concentraciones de ozono en el tiempo y ajustando con las variables meteorológicas (temperatura y velocidad de viento).

6.2. Recomendaciones e investigaciones futuras

En la literatura revisada muchos investigadores coinciden, en que no hay un método más apropiado para todos las propuestas de interés y todas los escenarios meteorológicos (Lou:2001; Wolff:2001). En general se sugiere realizar un análisis regional en escala, revisar la dependencia no lineal de ozono sobre la meteorología y hacer análisis de valor extremo para investigar tendencias. También se observa en la revisión bibliográfica, que los investigadores actualmente utilizan escalas más finas en las observaciones relacionadas con el ozono y variables ambientales para tratar de encontrar relaciones causales para posteriormente probar sus hipótesis con los modelos estadísticos.

- Investigar sobre la pérdida de información de un periodo corto de tiempo que algunas veces es muy común y frecuente en un día, cuando los aparatos se descomponen o se tienen que calibrar para coleccionar la información, suponiendo observaciones de una distribución GVE.
- Investigar y verificar sobre los supuestos de independencia del modelo GVE.
- Investigar en valores grandes de niveles de ozono urbano como es el comportamiento de las tendencias de las concentraciones de ozono.
- Investigar sobre los supuestos de independencia en el Modelo de Regresión Lineal de la Función Cuantil.
- Analizar e introducir un número mayor de variables meteorológicas en el modelo GVE para tener un modelo que ajuste mejor la tendencia de las concentraciones de ozono; como dirección del viento, radiación solar, nubosidad y distintas mediciones de la altura vertical.

◇

Referencias

- (2006). Proyecto milagro. Contaminación del Aire.
- Ambiental, C. (2002). *GEM, Programa para Mejorar la Calidad del Valle de México 2002-2010*. GDF, México.
- Ambiente, S. M. (2002). *Biblioteca Virtual*. www.sma.df.gob.mx/simat, México.
- Anderson, T. (1958). *An Introduction to Multivariate Statistical Analysis*. John Wiley, New York.
- Baur, D., Saisana, M. y Schulze, N. (2004). Modelling the effects of meteorological variables on ozone concentration. *Atmospheric Environment*, 38, 4689–4699.
- Bellanger, L. y Tomassone, R. (2004). Trend in high tropospheric ozone levels. Application to Paris monitorings sites. *Statistics*, 38(3), 217–241.
- Bloomfield, P., Royle, J., Steinberg, L. y Yang, Q. (1996). Accounting for meteorological effect in measuring urban ozone levels and trends. *Atmospheric Environment*, 30, 17, 3067–3077.
- Bravo, H. (1992). La contaminación atmosférica por ozono en la zona metropolitana de la Ciudad de México: evolución histórica y perspectivas. *IX Comisión Nacional de los Derechos Humanos*.
- Casella, G. y Berger, R. (2002). *Statistical Inference*. DUXBURY, Thomson Learning, segunda edición.
- Chavez, V. y Embrechts, P. (2004). Advances Extremal Models for Operational Risk. *Department of Mathematics, Zurich Switzerland*.
- Chock, D. (1989). *J. Air Pollut. Control Assoc.*, 39, 1063–1072.
- Coles, S. (2004). *An introduction to Statistical Modeling of Extreme values*. Springer Verlag, UK.
- Coles, S. y Dixon, M. (1999). Likelihood-Based Inference for Extreme Value Models. *Kluwer Academic Publishers*, 5, 21.
- Coles and Stephenson (2006). *ismev Package: An Introduction to Statistical Modeling of Extreme Values*. R Foundation for Statistical Computing, London.

- Correa, A. (2004). *Contaminantes Atmosféricos en la Zona Metropolitana de la Ciudad de México*. IPN-UAM, México.
- Cox, W. y Chu, S. (1996). Assessment of interannual ozone variation in urban areas from a climatological perspective. *Atmospheric Environment*, 30, 2615–2625.
- Danielsson, J., Jansen, D. y Vries, C. (1996). The Method of Moment Ratio Estimator for the Tail Shape Distribution. *Communication in Statistics (A) Theory and Methods*, 25(4), 711–720.
- Davison, C. y Smith., R. (1990). Models for Exceedances over High Thresholds. *J.R. Statist Soc. B*, 52(3), 393–442.
- Deheuvels, P., L. Haan, L. P. y Pereira, T. (1997). Comparison of Extreme Value Index Estimators. *NEPTUNE T400:EUR-09*.
- Dekkers, A. y Haan, L. (1989). On the Estimation of the Extreme-Value Index and Large Quantile Estimation. *The Annals of Statistics*, 17(4), 1795–1832.
- Finkenstadt, B. y Rootzen, H. (2004). *Extreme Values in Finance, Telecommunications, and the Environment*. Chapman & Hall/CRC, Sweden.
- Galambos, J. (1978). *The Asymptotic Theory of Extreme Order Statistics*. Wiley, New York.
- Gao, F., Sacks, J. y Welch, W. (1996). Predicting Urban Ozone Levels and Trends With Semiparametric Modeling. *American Statistics Association and the International Biometric Society, Journal of Agricultural, Biological and Environmental Statistics*, 1(4), 404–425.
- García-Colín, L. y Varela, J. (1996). *Contaminación Atmosférica*. El Colegio Nacional, México.
- García-Colín, L. y Varela, J. (2001). *Contaminación Atmosférica III*. El Colegio Nacional, México.
- García-Colín, L. y Varela, J. (2003). *Contaminación Atmosférica IV*. El Colegio Nacional, México.
- Garfias, M., Audry, J. y Garfias, F. (2005). Ozone Trend Analysis at Pedregal Station in the Metropolitan Area of Mexico City. *J. Mex. Chem. Soc.*, 49(4), 322–323.
- González, E. (2000). Estimación de tendencias a largo plazo de los niveles extremos de ozono en el Valle de México usando el proceso Poisson no homogéneo. *Tesis de Licenciatura, Chapingo*.
- Huerta, G. y Sansó, B. (2005). Time-Varying Models for Extreme Values. *Environmental and Ecological Statistics*.
- I.N.E.G.I. (1999). *Estadísticas del Medio Ambiente*. Semarnap, México.
- Jaimes, M. (2004). *Análisis de la tendencia y Comparación de la Contaminación Atmosférica por Ozono para la Ciudad de México y los Angeles, California, EUA*. Proyecto Fin de Carrera, Tesis de Maestría en Ingeniería Ambiental-UNAM.

- Johnson, N., Kotz, S. y Balakrishnan, N. (1995). *Continuous Univariate Distributions, Vol.2*. Wiley, New York.
- Kuchenhoff, H. y Thamerus, M. (1996). Extreme value analysis of Munich air pollution data.
- Lehmann, E. (1999). *Elements of Large Sample Theory*. Springer-Verlag.
- Lou, M., Reynolds, J., Cox, L., Guttorp, P. y Sampson, P. (2001). A review of statistical methods for the meteorological adjustment of tropospheric ozone. *Atmospheric environment*, 35, 617–630.
- McNeil and Stephenson (2004). *evir Package: Extreme Values in R*. R Foundation for Statistical Computing, London.
- Molina, M. (2006). Revista tierraamérica. Cambio Climático.
- Molina, M. y L., M. (2004). The impacts of megacities on air pollution, environmental aspects of urbanization. Goteborg Sweden.
- Molina, M. y Molina, L. (2002). *Air Quality in the Mexico Megacity: An Integrated Assessment*. Kluwer Academia Publishers, U.S.A.
- Mood, A., Graybill, A. y Boes, D. (1974). *Introduction to the Theory of Statistics*. McGraw Hill.
- Olkin, I., Sacks, J., Blumstein, A., Eddy, A., Eddy, W., Jurs, P., Kruskal, W., Kurtz, T., McDonald, G., Peierls, R., Shaman, P. y Spurgeon, W. (1990). Cross-disciplinary research in the statistical sciences. *Statist. Sci.*, 5, 121–146.
- Organisation, W. H. (1987). Air Quality Guidelines for Europe. *WHO regional publications, European series 23, Copenhagen, Regional Office for Europe*.
- Panaretos, J. y Tsourti, Z. (2003). Extreme Values Index Estimators and Smoothing alternatives: A critical review. *Stochastic Musings Perspectives from the Pioneers of the late 20th century*.
- Payer, T. (2006). Modeling Extreme Wind Speeds.
- Pickands, J. (1975). Statistical Inference using Extreme Order Statistics. *The Annals of Statistics*, 3(1), 119–131.
- Piegorsch, W., Smith, E., E. y Smith, R. (1998). Statistical Advances in Environmental Science. *Statistical Science*, 13(2), 186–208.
- Porter, S., Trivikrama, S., Zurbenko, I., Dunker, A. y Wolff, G. (2001). Ozone Air Quality over North América: Part II- A Review of Reported Trends. *J. Air & Waste Manage Assoc.*, 51, 283–306.
- R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

- Reiss, R. y Thomas, M. (2001). *Statistical Analysis of extreme values*. Birkhauser Verlag, Germany.
- Reynolds, H. (1998). Meteorological adjustment of Western Washington and Northwest Oregon surface ozone observations with investigation of trends. *National Research Center for Statistics and the Environment Technical Report 15*.
- Rohatgi., V. (1976). *An Introduction to Probability Theory and Mathematical Statistics*. John Wiley & Sons, U.S.A.
- Rosen, O. y Weissman, I. (1996). Comparison of estimation methods in extreme value theory. *Commun. Statist.*, 25(4), 759–773.
- Schabenberger, O. y Gotway, C. (2005). *Statistical Methods for Spatial Data Analysis*. Texts in Statistical Science Chapman & Hall/Crc.
- Seinfeld, J. (1991). *Comittee on tropospheric ozone formation and measurement; Board on Environmente Studies and Toxicology; Board on Atmospheric Sciences and Climate; Commission on Geosciencies, Environment and Resources; National Research Council, Rethinking the on ozone problem in urban and regional air Pollution*. National Academic Press, Washington.
- Sánchez, R. (2001). Análisis de tendencia en exedencias sobre un umbral alto, con aplicación en ozono urbano.
- Sánchez, R., Cantú, M. y Vaquera, H. (2001a). Uso del modelo de regresión Pareto generalizado para analizar tendencias temporal en las concentraciones de ozono en la Zona Metropolitana de Guadalajara. *Cuadernos de Investigación de la U.A. de C.*
- Sánchez, R., Vaquera, H., Villaseñor, J. y Barry, A. (2001b). Analysis of Trend in Exceedances over a hight Threshold with applications to Urban Ozone Levels.
- Spichtinger, N., Winterhalter, M. y Fabian, P. (1996). Ozone and Grosswetterlagen: analysis for the Munich metropolitan area. *Environmental Science and Pollution Research*, 3, 145–152.
- Tobías, A. y Scotto, M. (2005). Prediction of Extreme Ozono levels in Barcelona, Spain. *Environmental Monitoring and Assessment*, 100, 23–32.
- Torres, A. (1997). Construcción y uso de Trayectorias de retroseso de parcelas de aire para el estudio de la Contaminación Atmosférica Fotoquímica en la ZMCM. *Tesis de Maestría en Ingeniería Ambiental, UNAM*.
- Vaquera, H. (1997). On the Statistical of Trend in Troposheric Ozone Levels.
- Vaquera, H. (2004). Notas del curso de simulación. C.P-ISEI.
- Velasco, C. (2000). Determinación de las k Estadísticas Extremas para la estimación de la cola de una distribución.
- Villaseñor, J. (1996). A model for cluster maxima of exceedances over a threshold for ozone data. *Revista de Estadística, INEGI*, 8(10), 7–16.

Villaseñor, J. (2004). Notas del curso de estadística. C.P-ISEI.

Wolff, G., Dunker, A., Trivikrama, S., Porter, P. y Zurbenko, I. (2001). Ozone Air Quality over North América: Part I- A Review of Reported Trends. *J. Air & Waste Manage, Assoc.*, 51, 273-282.