COLEGIO DE POSTGRADUADOS

INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN
EN CIENCIAS AGRÍCOLAS

CAMPUS MONTECILLO
POSTGRADO EN SOCIOECONOMÍA, ESTADÍSTICA E
INFORMÁTICA, ESTADÍSTICA

Modelación de valores máximos de ozono

María Guzmán Martínez

T E S I S

PRESENTADA COMO REQUISITO PARCIAL PARA
OBTENER EL GRADO DE:

MAESTRA EN CIENCIAS

MONTECILLO,TEXCOCO, EDO. DE MÉXICO
2011

La presente tesis titulada: **Modelación de valores máximos de ozono**, realizada por la alumna: **María Guzmán Martínez**, bajo la dirección del Consejo Particular indicado, ha sido aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:

MAESTRO EN CIENCIAS
SOCIOECONOMÍA-ESTADÍSTICA E INFORMÁTICA-ESTADÍSTICA

CONSEJO PARTICULAR

CONSEJERO **DR. José A. Villaseñor Alva**

ASESOR **DR. Javier Suarez Espinoza**

ASESOR **DR. Enrique Raúl Villa Diharce**

Montecillo, Texcoco, Estado de México, Enero de 2011

Modelación de valores máximos de ozono

Resumen

El cambio climático se ha convertido en un tema *IMPORTANTE* durante los últimos años. Hay una gran preocupación en muchos países por disminuir las constantes emisiones de contaminantes a la atmósfera de la Tierra. Algunas Ciudades más desarrolladas y por lo tanto más contaminadas, tienen su propio sistema de monitoreo para evaluar la calidad del aire. Tal es el caso de la zona metropolitana de Guadalajara, México. Tiene su red automática de monitoreo atmosférico, que consiste en ocho estaciones de monitoreo.

Este trabajo de tesis, está interesado en las lecturas máximos diarias del ozono procedentes de la estación de monitoreo Centro, ubicada en la zona centro de Guadalajara. La base de datos que se considera contiene las observaciones correspondientes al período $1997 - 2008$.

Nuestro objetivo es ajustar el modelo de probabilidad conjunta propuesta por Villaseñor y González (2010),y la distribución Pareto Generalizada (DPG) a la base de datos mencionada anteriormente.

Para la estimación de algunos de los parámetros que intervienen en los dos modelos se hace uso de la metodología propuesta por Ferro y Segers (2003).

En base a pruebas estadísticas se obtuvo que el modelo de probabilidad conjunta, propuesto por Villaseñor y González (2010), ofrece un mejor análisis estadístico para este tipo de datos.

Las conclusiones son las siguientes.

1. Los dos modelos propuestos proporcionan un buen ajuste a la base de datos que se considera.
2. El modelo de probabilidad conjunta es mejor que el modelo de la DPG, ya que proporciona un análisis estadístico más detallado, debido al hecho de que se está considerando una variable adicional, lo cual es útil para la obtención de algunos resultados concretos.

Palabras clave: Índice extremo, Grupos de datos, Modelando probabilidades de datos.

Modelación de valores máximos de ozono

Abstract

Climate change has become an importante issue during the latest years. There is a great concern by many countries to constantly diminish the pollutant emissions to the earth atmosphere. Some the most developed cities and therefore the most contaminated ones, have their own monitoring system for assessing air quality. Such is the case in the metropolitan zone of Guadalajara, Mexico. It has its automatic net for atmospheric monitoring which consists of eight monitoring stations.

In this thesis work we are interested in the dairy maximum ozone readings coming from the Center monitoring station, located in the Guadalajara downtown zone. The considered data base contains the observations corresponding to the period $1997 - 2008$.

Our aim is to fit the joint probability model proposed by Villaseñor and González (2010), and the Generalizad Pareto distribution (GPD) to the database mentioned above.

To estimate some of the parameters involved in the models we make use of the methodology proposed by Ferro and Segers (2003).

On the basis of statistical tests, it turns out that the joint model proposed by Villaseñor and González (2010) offers a better statistical analysis for this type of data.

The conclusions are as follows.

1. Both proposed models provide a good fitting to the considered data base.
2. The joint probability model is better than the GPD model, since it provides a more detailed statistical analysis due to the fact that it considers an additional variable, which is useful for obtaining some specific results.

Key words: Extremal index, Data clusters, Probability data modeling.

AGRADECIMIENTOS

Al Consejo Nacional de Ciencia y Tecnología, por el apoyo económico brindado para los estudios de mi maestría en Estadística en el Colegio de Posgraduados campus Montecillo.

A la Línea Prioritaria de Investigación No. 15, Estadística, Modelado y Tecnologías de la información Aplicados a la Agricultura y al Medio Rural del CP, por el apoyo financiero recibido para la realización de mis estudios de maestría y el trabajo de tesis.

A todos los Doctores de la maestría en Estadística que con su conocimiento han contribuido a mi formación académica.

A los integrantes de mi Consejo Particular:

Dr. José A. Villaseñor Alva, por su tiempo, paciencia, y conocimiento compartido, gracias Doctor.

Dr. Javier Suarez Espinoza, por su apoyo, paciencia y por ser un gran ser humano, gracias Doctor.

Dr. Enrique Raúl Villa Diharce: por tomarse el tiempo para formar parte de mi consejo, por su apoyo en la revisión de este trabajo de tesis, gracias Doctor.

A la Dra. Elizabeth González Estrada por su apoyo incondicional.

A todos mis amigos y compañeros.

A mi Madre por darme la vida y a mi familia por hacer de esta vida un camino más alegre.

DEDICATORIA

A las personas que comparten su conocimiento teórico y práctico de la Estadística, a esas personas que en un papel de Maestros o investigadores, dan a este mundo la esperanza de que se puede ser mejor.
Cuando te encuentras a seres humanos que han entregado parte de su vida para difundir el conocimiento estadístico y cuando tu corazón arde en deseo por la ciencia, entonces los límites no existen; esto es lo que me llevo del Colegio de Posgraduados, y si era necesario salir de casa para aprenderlo, entones para mi ha valido la pena. Además me llevo la dicha de haber sido parte de la generación $2009 - 2010$.

Índice

Índice de Cuadros

Índice de figuras

Capítulo 1

Introducción

Guadalajara es la capital del estado de Jalisco y cabecera del zona Metropolitana de Guadalajara (ZMG). La ZMG, está integrada por 8 municipios del estado de Jalisco que son: Guadalajara, El Salto, Tlajomulco de Zúñiga, Tlaquepaque, Tonalá, Zapopan, Juanacatlán e Ixtlahuacán de los Membrillos. En al año del 2005 la ZMG agrupaba alrededor de $4,000,000$ habitantes, según censo del $INEGI$ 2005. La ZMG es la segunda en el país por su población después de la Zona Metropolitana de la Ciudad de México. Debido a la industria, entre otras fuentes de emisión, los contaminantes como el monóxido de carbono (CO), bióxido de nitrógeno (NO_2), ozono (O_3) y bióxido de azufre (SO_2), afectan a la ZMG (INEGI [1]). Las concentraciones de dichos contaminantes son evaluados por la Red Automática de Monitoreo Atmosférico de la zona Metropolitana de Guadalajara ($RAMAG$), la cual está constituida por ocho estaciones de monitoreo: Águilas, Atemajac, Centro, Loma Dorada, Miravalle, Oblatos, Tlaquepaque y Vallarta.

En este trabajo nos enfocamos solamente al estudio del ozono. El ozono es un gas que está presente tanto en la atmósfera superior de la Tierra (ozono estratosférico), como a nivel del suelo (ozono troposférico), y dependiendo de su ubicación puede ser bueno o malo.

El ozono estratosférico, se origina de forma natural en la estratósfera (entre 12 y 50 kms a partir del suelo), mediante la fotodisociación del oxígeno producida por la radiación solar ultravioleta. Éste, permite que se lleven a cabo diversos procesos en los ecosistemas naturales, a nivel celular; filtra y modera la intensidad de la radiación solar ultravioleta y otras partículas energéticas que inciden sobre la superficie terrestre.

El ozono troposférico, está a nivel de la tropósfera (de 0 a 12 kms a partir de la superficie terrestre). La reacción fotoquímica se produce cuando los óxidos de nitrógeno y los compuestos orgánicos volátiles reaccionan con la luz solar, lo que produce un átomo libre

[1] Publicación en línea, disponible en internet en el sitio http://www.inegi.org.mx/Sistemas/temasV2/Default.aspx?s=est&c=21385 [con acceso el $11-12-2010$].

de oxígeno (O), el cual puede adicionarse a una molécula de oxígeno (O_2) y formar una molécula de ozono (O_3). Este proceso es reversible y está condicionado por la intensidad de la radiación solar.

Los niveles del ozono troposférico para las ocho estaciones de monitoreo, se miden con el índice metropolitano de la calidad del aire ($IMECAS$), el cual consiste en una conversión de las concentraciones de los contaminantes a un número adimensional, que indica el nivel de la contaminación de una manera accesible para la población; de aquí en adelante cuando hablemos del ozono nos estaremos refiriendo al ozono troposférico. El *IMECA* se utiliza en todo el mundo, y tiene como propósito informar a la población de manera clara, oportuna y continua sobre los niveles de contaminación atmosférica, los probables daños a la salud y las medidas de protección que se pueden tomar. Otra medida que existe para medir las concentraciones de ozono en el medio ambiente son las partes por millón (ppm), la cual es la relación del volumen del ozono en un millón de volúmenes de aire.

La calidad del aire de una región está asociada al volumen, calidad y tipo de combustibles consumidos, equipos de combustión de las plantas industriales y de servicios, tecnologías de control y combustión de emisiones en vehículos, ubicación y condiciones meteorológicas, así como la interacción entre los diferentes contaminantes y los componentes del aire que modifican la química atmosférica.

El Cuadro 1.1 muestra la equivalencia entre *IMECAS* y *ppm* del ozono, además de su relación con la calidad del aire.

Cuadro 1.1: Intervalos de valores para el ozono.

Intervalo del *IMECA*	Intervalo de concentraciones (ppm)	Calidad del aire
$0 - 50$	$0.000 - 0.055$	buena
$51 - 100$	$0.056 - 0.110$	regular
$101 - 150$	$0.111 - 0.165$	mala
$151 - 200$	$0.166 - 0.220$	muy mala
> 200	> 0.220	extremadamente mala

Cuando la calidad del aire es *buena* se pueden llevar a cabo actividades al aire libre. Cuando es *regular* se pueden presentar posibles molestias, en los individuos sensibles, tales como efectos respiratorios debido a un prolongado esfuerzo al aire libre. Una *mala* calidad provoca asma, tos y dolor de cabeza en los niños y los adultos mayores con enfermedades cardiovasculares y (o) respiratorias. Si la calidad es *muy mala* los efectos adversos a la salud son mayores en la población en general, los individuos sensibles pueden experimentar tos y dolor agravados, además se reduce la función de los pulmones. Por último si la calidad es *extremadamente mala*, la población puede experimentan síntomas respiratorios severos y respiración débil, por lo cual se recomienda no salir de casa además de cerrar puertas y ventanas.

Para las ocho estaciones de monitoreo del área Metropolitana de Guadalajara, se cuenta con los registros máximos diarios de ozono, desde 01/11/95 hasta el 31/12/09 (Datos proporcionados por: Martínez, 2010).

La Figura 1.1, muestra el número total de los máximos diarios de ozono que rebasaron los 100 *IMECAS* en los últimos 14 años, para las ocho estaciones de monitoreo.

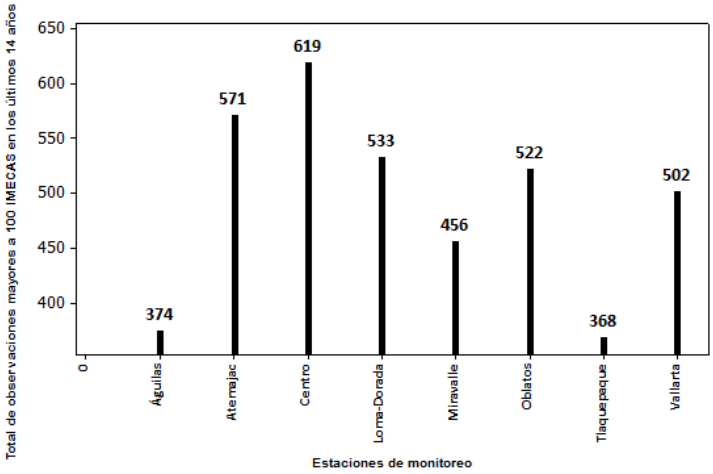

Figura 1.1: Estaciones de monitoreo que reportan niveles de ozono mayores a los 100 *IMECAS* en los últimos 14 años.

Nótese que la estación Centro es la que cuenta con más observaciones mayores a 100 *IMECAS*; seguida por la estación Atemajac. En la Figura 1.2 se observan las tendencias del ozono en *IMECAS* para la estación Centro en los últimos 14 años. Los niveles del ozono oscilan entre el 0 y 266 *IMECAS*. Nótese que antes del 13/03/98, los niveles del ozono rebasan los 200 *IMECAS*, mientras que después de esta fecha los niveles del ozono se mantienen por debajo de los 200 *IMECAS*.

En general es difícil tratar de predecir la tendencia del ozono para la estación Centro con la gráfica anterior. Estos datos tienden a estar correlacionados lo cual dificulta su análisis estadístico. Otro problema es la existencia de observaciones faltantes en la base de datos, debido a fallas en el equipo de monitoreo. En las ocho estaciones de monitoreo, existe el problema de datos faltantes.

1.1. Antecedentes

Sánchez (2001), hace un análisis de los datos de ozono para la estación Centro de la zona Metropolitana de Guadalajara, correspondiente al periodo 01/11/95 al 30/08/99. Fijando un umbral de 110 *ppm* ajusta una distribución Pareto Generalizada. Para estimar

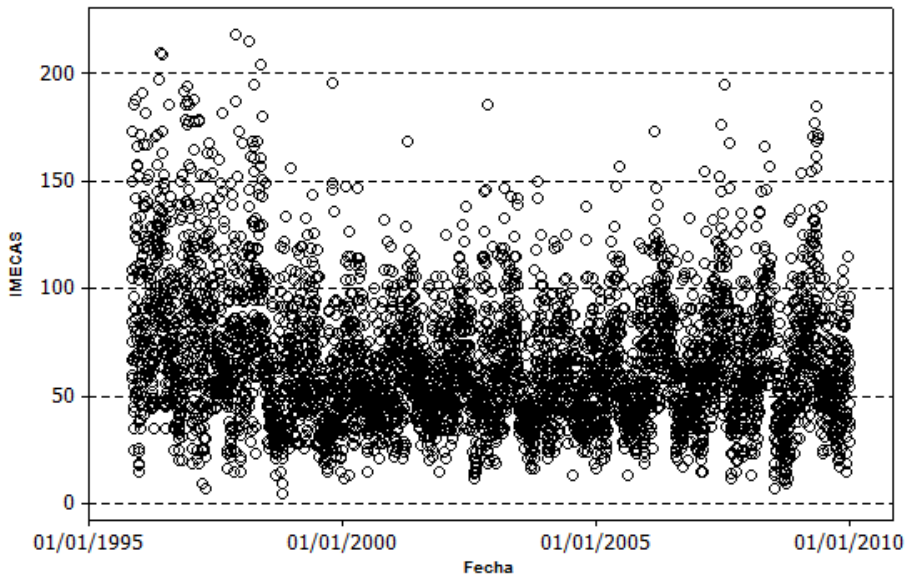

Figura 1.2: Registros de los máximos diarios del ozono durante los últimos 14 años para la estación Centro.

los parámetros de esta distribución hace una exploración entre los métodos de: máxima verosimilitud, de momentos, de probabilidades ponderadas, de elemental percentil y estimación beyesiana. Con base en el sesgo y la raíz del error cuadrado medio sugiere que por estimación bayesiana se obtienen las mejores estimaciones para los parámetros de la distribución Pareto Generalizada.

Hernández (2009), hace un análisis estadístico para el ozono para siete estaciones de monitoreo del área Metropolitana de Guadalajara. Los registros tomados en cuenta van desde el 6 de enero de 1997 al 31 de diciembre del 2006. En ese trabajo se consideran el máximo de ozono entre las 12 y 17 horas de toda la red de monitoreo de cada día. Además considera siete variables atmosféricas que son: promedio del mínimo del viento (vv), promedio máximo de temperatura (tem), promedio del mínimo en humedad (h), rango de velocidad del viento (rvv), rango de temperatura ($rtem$) y rango de humedad (rh). El modelo que emplea para el análisis de estos datos es un modelo autoregresivo de orden 6. Apoyándose de las gráficas de la densidad estimada de los residuales, cuantil contra cuantil y residuales contra tiempo, asegura que el ajuste del modelo es relativamente bueno.

Villaseñor y González (2010), analizaron los máximos diarios de ozono correspondientes a los años 2003 al 2007 para la estación de verano de la estación de monitoreo del Pedregal de la ciudad de México. Asumiéndose estacionaridad estricta para este conjunto de datos, se fijo un umbral $u = 110$ ppm y se agruparon los datos (Ferro y Segers, 2003a). De los grupos de excedencias, se obtuvieron los excesos máximos de cada uno estos grupos.

Villaseñor y González (2010), proponen una distribución de probabilidad conjunta, con la cual se pueda modelar los excesos máximos de grupo tomando en cuenta los tamaños de los grupos.

1.2. Objetivos

De los registros de ozono, que corresponden a los máximos diarios, monitoreados en el área Metropolitana de Guadalajara, se determinará una base de datos para ser analizada, tal que cumplan con la hipótesis de estacionaridad estricta. Luego se fijará un umbral u, y se le aplica la técnica de Ferro y Segers (2003a) para obtener los grupos. Una vez formados los grupos de excedencias se podrán determinar los excesos de grupo. En función de los grupos y los excesos máximos de cada grupo, se platean dos enfoques:

1. Tomar en cuenta los excesos máximos de cada grupo y los tamaños de los grupos, para ajustar la distribución de probabilidad conjunta propuesta por Villaseñor y González (2010).
2. Considerar los excesos máximos de cada grupo, y ajustar la distribución Pareto Generalizada

A los dos modelos que se obtengan, se les aplicarán las pruebas correspondientes para validar su bondad de ajuste a la base de datos en cuestión; y se discutirán las posibles ventajas que cada uno pueda ofrecer en la modelación de valores extremos.

1.3. Metodología

Al considerar los máximos diarios de ozono, y por el hecho de ser observaciones en el tiempo, tiende a existir dependencia entre las observaciones. La teoría de valores extremos sugiere que se formen grupos independientes y después se tomen los máximos de cada grupo para ayudar con dicho problema.

Considerando los máximos diarios de ozono durante los últimos 14 años, para las ocho estaciones de monitoreo, se determinará una base de datos que cumpla con la hipótesis de estacionaridad estricta.

Ferro y Segers (2003a) propone un esquema de agrupamiento automático. Nosotros emplearemos este enfoque para formar los grupos. Se fijara un umbral u y se determinarán los excesos máximos de cada grupo. En base a lo anterior se ajustará una distribución

Pareto Generalizada y la distribución de probabilidad conjunta (Villaseñor y González, 2010). En ambos análisis se aplicarán las pruebas estadísticas correspondientes para validar nuestras hipótesis.

1.3.1. Selección de la base de datos

En 1975 se iniciaron trabajos de monitoreo atmosférico en la ciudad de Guadalajara con equipo manual. En 1993 el gobierno del estado de Jalisco adquirió parte de la red de monitoreo atmosférico automático y en 1995 quedó finalmente a cargo de todos sus componentes.

Desde su integración, la *RAMAG* es operada por el gobierno del Estado, a través de la Secretaría de Medio Ambiente para el Desarrollo Sustentable (*SEMADES*).

Cada estación de monitoreo de la red, cuenta a su vez con 8 sistemas de monitoreo meteorológico que miden la dirección y velocidad del viento, así como la temperatura y humedad relativa, estos sistemas de monitoreo se localizan en los mismos sitios en que se ubican las casetas de monitoreo atmosférico, y la Comisión Estatal de Ecología del Gobierno del Estado de Jalisco (*COESE*), es la encargada de la operación y mantenimiento.

El Cuadro 1.2 muestra la ubicación de las ocho estaciones de monitoreo.

Cuadro 1.2: Red automática de monitoreo atmosférico de la *ZMG*.

Estación	Ubicación
Vallarta (VAL)	Zona poniente de Guadalajara
Centro (CEN)	Zona centro de Guadalajara
Miravalle (MIR)	Zona sur de Guadalajara
Oblatos (OBL)	Zona norte de Guadalajara
Atemajac (ATM)	Zona norte de Zapopan
Águilas (AGU)	Zona poniente de Zapopan
Loma Dorada (LDO)	Zona oriente de Tonalá
Tlaquepaque (TLA)	Zona oriente de Tlaquepaque

Se consideran los máximos diarios de ozono, desde 01/11/95 al 31/12/09, para las ocho estaciones de monitoreo. Luego dividimos a este conjunto de observaciones en las estaciones del año, es decir, primavera, verano, otoño e invierno; todo esto con la finalidad de observar en que época del año existe un mayor incremento de ozono.

El Cuadro 1.3, que corresponde a la estación de primavera, muestra el número total de observaciones y el total de datos faltantes cuya suma da un total de 1288. En el cuarto renglón del Cuadro, se muestra el total de observaciones que rebasan los 100 *IMECAS*.

Cuadro 1.3: Máximos diarios del ozono emitidos en primavera 1996 – 2009.

Total	AGU	ATM	CEN	LDO	MIR	OBL	TLA	VAL
Observaciones	1268	1263	1264	1279	1275	1000	1268	1259
Valores Perdidos	20	25	24	9	13	288	20	29
Valores > 100	150	202	270	276	230	224	178	194

El Cuadro 1.4 corresponde a la estación de verano. Las observaciones que rebasan los 100 *IMECAS*, las cuales se muestran en el cuarto renglón, disminuyen para las ocho estaciones, en comparación con el Cuadro anterior.

Cuadro 1.4: Máximos diarios del ozono emitidos en varano 1996 – 2009.

Total	AGU	ATM	CEN	LDO	MIR	OBL	TLA	VAL
Observaciones	1243	1198	1249	1276	1257	1039	1212	1208
Valores Perdidos	45	90	39	12	31	249	76	80
Valores > 100	69	42	58	52	52	33	40	72

El Cuadro 1.5 corresponde a la estación de otoño. Nótese que para otoño las ocho estaciones de monitoreo, cuentan con un mayor número de observaciones que rebasan los 100 *IMECAS* en comparación con la estación de verano. Los valores del segundo y tercer renglón dan un total de 1324 para cada una de las ocho estaciones de monitoreo.

Cuadro 1.5: Máximos diarios del ozono emitidos en otoño 1995 – 2009.

Total	AGU	ATM	CEN	LDO	MIR	OBL	TLA	VAL
Observaciones	1319	1279	1319	1316	1296	1049	1234	1277
Valores Perdidos	5	45	5	8	28	275	90	47
Valores > 100	91	147	138	97	89	130	78	125

Por último, el Cuadro 1.6 corresponde a la estación de invierno. La suma del segundo y tercer renglón dan un total de 1275 para cada estación de monitoreo.

Obsérvese que en el Cuadro 1.3 que corresponde a la estación de primavera, es el que cuenta con más observaciones que rebasan los 100 *IMECAS*; las estaciones de Loma Dorada y Centro son las que destacan con 276 y 270 observaciones respectivamente.

Para el desarrollo de esta tesis, se trabajará con los valores máximos diarios de ozono de la estación Centro. Es decir, la base de datos cuenta con un total de 1264 observaciones emitidas en primavera en el periodo 1996 al 2009. En la Figura 1.3, se muestra la gráfica de las 1264 observaciones.

En el Cuadro 1.3, se observo que son 270 observaciones que están por arriba de los 100 *IMECAS*.

Cuadro 1.6: Máximos diarios del ozono emitidos en invierno 1995 – 2009.

Total	AGU	ATM	CEN	LDO	MIR	OBL	TLA	VAL
Observaciones	1246	1259	1263	1268	1243	998	1236	1253
Valores Perdidos	29	16	12	7	32	277	39	22
Valores > 100	64	180	153	108	85	135	72	113

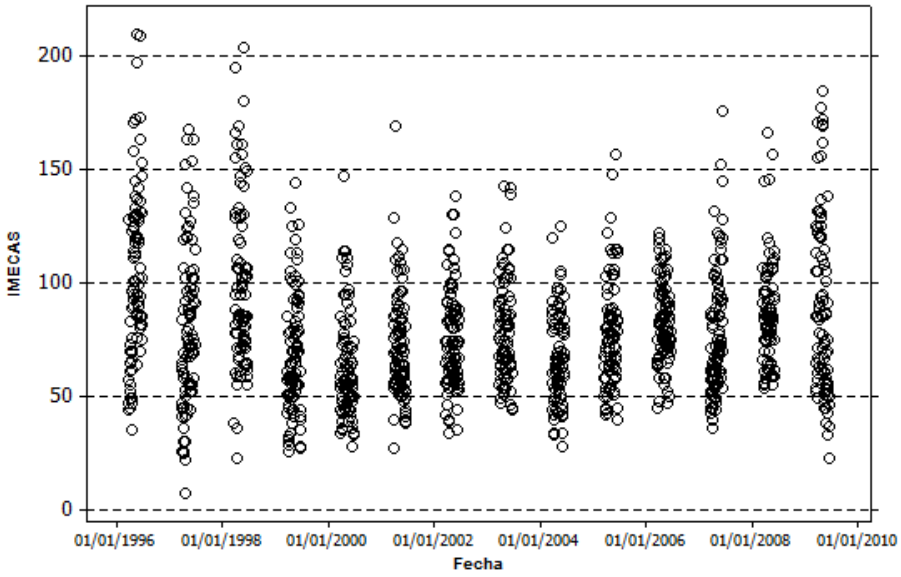

Figura 1.3: Valores máximos diarios de ozono emitidos en primavera en el periodo 1996 – 2009 de la estación Centro.

Capítulo 2

Distribución Pareto Generalizada

2.1. Distribución de Valores Extremos Generalizada

Sea $X_1, ..., X_n$ una secuencia de variables aleatorias independientes observadas en el tiempo, teniendo por función de distribución a F.

Sea

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

En teoría de distribuciones

$$\begin{aligned} P(M_n \leq z) &= P(X_1 \leq z, \ldots, X_n \leq z) \\ &= P(X_1 \leq z) \cdot \ldots \cdot P(X_n \leq z) \\ &= \prod_{i=1}^{n} P(X_i \leq z) \\ &= \{F(z)\}^n \end{aligned}$$

.

Nótese que F es desconocida, que si bien puede ser estimada de los datos observados, no es recomendable.

La distribución del valor extremo generalizada, es la distribución que se ajusta a los máximos de bloques independientes de observaciones adecuadamente normalizadas y surge a partir del desarrollo del teorema de Fisher-Tippet.

Teorema 1 *(Fisher-Tippett, Gnedenko) Sea $\{X_n\}$ una sucesión de variables aleatorias iid con una función de distribución común $F(x)$. Si existen sucesiones de constantes de*

normalización $\{a_n > 0\}$ *y* $\{b_n\} \in \mathbb{R}$

$yunafunciónde distribución Hnodegenerada, talque \lim_{n\longrightarrow\infty} P\left(\frac{M_n - b_n}{a_n} \leq z\right) = H(z),$

entonces $H(z)$ pertenece a alguna de las siguientes familias:

1. Gumbel (colas medias)

$$\Delta(z) \quad = \exp\left\{-\exp\left[-\left(\tfrac{z-b}{a}\right)\right]\right\}, \quad -\infty < z < \infty;$$

2. Fréchet (colas gruesas)

$$\Phi(z) = \begin{cases} \exp\left\{-\left(\tfrac{z-b}{a}\right)^{-\alpha}\right\}, & z > b, \\ 0, & z \leq b; \end{cases}$$

3. Weibull (colas cortas o suaves)

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

donde $a > 0$ y $\alpha > 0$.

Las tres familias pueden ser combinadas en un solo modelo, la Distribución de Valores Extremos Generalizada (*DVEG*) dada por

$$H(z) = \exp\left\{-\left[1 + \gamma\left(\frac{z-\mu}{\sigma}\right)^{-\frac{1}{\gamma}}\right]\right\}, \tag{2.1}$$

tal que $-\infty < \mu < \infty$, $\sigma > 0$, $-\infty < \gamma < \infty$, donde

$$\begin{aligned} \mu &: = \text{parámetro de localización}, \\ \sigma &: = \text{parámetro de escala}, \\ \gamma &: = \text{parámetro de forma}. \end{aligned}$$

Cuando $\gamma > 0$, se tiene la función de distribución Fréchet, cuando $\gamma < 0$ se tiene la función distribución Weibull. Con $\gamma = 0$ se obtiene un subconjunto de la *DVEG*, que es interpretado como el límite de $H(z)$ cuando $\gamma \longrightarrow 0$. Dicho subconjunto representa la familia Gumbel.

El resultado anterior se enuncia con el siguiente teorema

Teorema 2 *(Distribución de Valores Extremos Generalizada) Si existen secuencias de constantes* $\{a_n > 0\}$ *y* $\{b_n\}$ *tales que*

$$\lim_{n\longrightarrow\infty} P\left(\frac{M_n - b_n}{a_n} \leq z\right) = H(z),$$

donde $H(z)$ *es una función de distribución no degenerativa, entonces* H *es miembro de la familia de valores extremos generalizada, dada por*

$$H(z) = \exp\left\{-\left[1+\gamma\left(\frac{z-\mu}{\sigma}\right)^{-\frac{1}{\gamma}}\right]\right\},$$

definido sobre $\left\{z : 1+\gamma\left(\frac{z-\mu}{\sigma}\right) > 0\right\}$*, donde* $-\infty < \mu < \infty$*,* $\sigma > 0$*,* $-\infty < \gamma < \infty$*.*

La *DVEG* proporciona un modelo para la distribución de los máximos de bloques. El tamaño del bloque es un problema a considerar, pues bloques muy pequeños implicará tener sesgo y con bloques muy grandes sería tener demasiada varianza.

Definición 3 *Dada* X_n *una sucesión de variables aleatorias iid, con función de distribución acumulada dada por* F*, entonces* F *pertenece al dominio máximo de atracción de la distribución de valores extremos generalizada* H*, lo cual se denota como* $X \in MDA(H)$*, o de forma equivalente* $F \in MDA(H)$*, si existen constantes* $\{a_n > 0\}$ *y* $\{b_n\}$ *tal que*

$$\lim_{n\longrightarrow\infty} P\left(\frac{M_n - b_n}{a_n} \leq z\right) = H(z),$$

para toda z *en los puntos de continuidad de* H*.*

2.2. Distribución Pareto Generalizada

Una distribución relevante en la teoría de valores extremos es la distribución Pareto Generalizada (*DPG*) que surge a partir del método conocido como Peaks-Over-Threshold (*POT)*, que consiste en la modelización de los extremos que exceden un umbral dado.

Un aspecto importante a considerar para esta distribución es; determinar el valor del umbral u, cuya elección está sujeta al problema de la varianza y el sesgo. Un valor muy bajo para el umbral implicaría violar la base asintótica para el modelo y conducir a un mayor número de observaciones. Lo que puede disminuir la varianza del ajuste, pero también puede incrementar el sesgo al intentar modelar observaciones que no pertenecen a la cola de la distribución. Por otra parte, un valor muy grande para el umbral generaría

pocas excedencias para estimar el modelo, generando una alta varianza. Se reduciría el sesgo pero la estimación del índice γ sería más volátil al contar con un número menor de observaciones.

Sea X una variable aleatoria con función de distribución F y extremo derecho del soporte x_F dado por

$$x_F = \sup\{x \in \mathbb{R} \mid \mathbb{F}(x) < 1\}, \; x_F \leq \infty.$$

Para $u < x_F$, se dice que ha ocurrido una excedencia de u si $X > u$. Al valor de X se le llama excedencia y exceso a $X - u$. Por ejemplo véase la Figura 2.1, donde se muestra un umbral $u = 7$ y a partir de este se define la excedencia y el exceso.

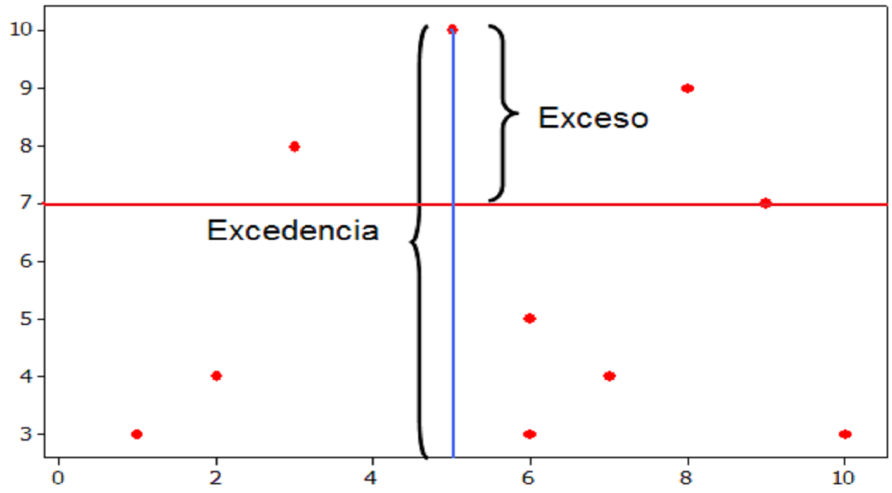

Figura 2.1: Excedencia y exceso dado un umbral u.

A continuación se define la distribución de los excesos.

Definición 4 *Dada una variable aleatoria X con función de distribución acumulada F, se define la distribución de excesos sobre el umbral u como*

$$\begin{aligned} F_u(x) &= P(X - u \leq x \mid X > u) \\ &= \frac{F(x+u) - F(u)}{1 - F(u)}, \end{aligned} \tag{2.2}$$

para $0 \leq x < x_F - u$.

A F_u se le conoce también como la distribución de la vida residual, del exceso de vida o del exceso de pérdida. La distribución de excesos, F_u, representa la probabilidad de que X exceda el umbral u a lo sumo en una cantidad x, condicionado a la información de que la X excedió el umbral.

El teorema de Pickands-Balkema De Haan (Balkema y De Haan, 1974), muestra que bajo condiciones de máximos dominios de atracción, la *DPG* es la distribución límite para los excesos sobre un umbral u cuando $u \longrightarrow \infty$.

Teorema 5 *(Pickands-Balkema.De Haan) Sea F una función de distribución acumulada con función de exceso F_u, para $u \geq 0$. Dado $\gamma \in \mathbb{R}$, $F \in MDA(H)$ si y sólo si existe una función medible positiva $\sigma(u)$ tal que*

$$\lim_{u \longrightarrow x_F} \sup_{0 \leq x \leq x_F - u} \left| F_u(x) - G_{\gamma,\sigma(u)}(x) \right| = 0,$$

donde G es la función de distribución acumulada de una variable aleatoria Pareto Generalizada, con parámetros γ y σ_u.

Nótese que la distribución límite del máximo es la *DVEG*, y que la distribución del límite de los valores que exceden un umbral es la distribución Pareto Generalizada, con parámetro de escala σ_u, y parámetro de forma γ, igual que la de *DVEG*.

El teorema anterior enuncia que para un umbral u lo suficiente elevado es posible encontrar valores de σ_u y γ tal que

$$F_u(x) \approx G(x), \text{ para } 0 \leq x \leq x_F - u,$$

donde

$$G(x) = \begin{cases} 1 - (1 + \gamma x)^{\frac{-1}{\gamma}}, & \text{si} \quad \gamma \neq 0, \\ 1 - e^{-x}, & \text{si} \quad \gamma = 0, \end{cases}$$

con $x \geq 0$ si $\gamma \geq 0$, y $0 \leq x \leq \frac{-1}{\gamma}$ si $\gamma < 0$.

Se puede introducir una familia de localización–escala, remplazando x con $(x - v)/\sigma_u$ para $v \in \mathbb{R}$, $\sigma_u > 0$, entonces la función de distribución de G, está dada por

$$G(x) = 1 - (1 + \frac{\gamma}{\sigma_u} x)^{\frac{-1}{\gamma}}, \; x \in D(\gamma, \sigma_u),$$

donde

$$D(\gamma, \sigma_u) = \begin{cases} [0, \infty), & \text{si} \quad \gamma \geq 0, \\ \left[0, \frac{-\sigma_u}{\gamma}\right], & \text{si} \quad \gamma < 0. \end{cases}$$

A continuación se define la *DPG*.

Definición 6 *La distribución Pareto Generalizada, está dada por*

$$G(x; \sigma_u, \gamma) = 1 - \left(1 + \frac{\gamma}{\sigma_u} x\right)_+^{\frac{-1}{\gamma}} \quad \textit{para} \quad \gamma \neq 0, \tag{2.3}$$

donde $A_+ = \text{máx}(0, A)$, $\sigma_u > 0$ *y* $\gamma \in \mathbb{R}$, *tal que*

$$\begin{array}{ll} x \geq 0 & si \quad \gamma \geq 0, \\ 0 \leq x \leq \frac{-\sigma_u}{\gamma} & si \quad \gamma < 0. \end{array}$$

La distribución está definida por el parámetro γ, o índice de cola, y por el parámetro de escala σ_u. Cuando mayor sea el parámetro γ más larga es la cola.

La elección correcta del umbral es indispensable, es elegir un valor lo suficientemente elevado como para que el teorema asintótico pueda ser considerado esencialmente exacto, y lo suficientemente bajo como para poder tener observaciones para la estimación de los parámetros σ_u y γ.

Al igual que la *DVEG*, la *DPG* engloba tres tipos de distribución: Pareto si $\gamma > 0$, Beta si $\gamma < 0$ y Exponencial si $\gamma = 0$.

2.2.1. Función media de los excesos

Una función que se usa para validar el uso de la *DPG* es la función media de los excesos (Davison y Smith, 1990). Dicha función es un diagnóstico gráfico para determinar el valor del umbral u. En Coles (2001) se enuncia la gráfica de la vida media residual para dicho propósito, revisemos la fundamentación y aplicación que puede tener dicha función.

Suponiendo que la *DPG* es válida como modelo para los excesos de un umbral u_0 generados por la serie $X_1, ..., X_n$, de los cuales un término arbitrario de la serie anterior es denotado por X, entonces

$$E(X - u_0 \mid X > u_0) = \frac{\sigma_{u_0}}{1 - \gamma},$$

para $\gamma < 1$.

Coles (2001) argumenta que si la distribución Pareto Generalizada es válida para los excesos de un umbral u_0, entonces ésta debería de ser igualmente válida para todos los umbrales $u > u_0$; luego para $u > u_0$ se tiene

$$\begin{aligned} E(X - u \mid X > u) &= \tfrac{\sigma_u}{1-\gamma} \\ &= \tfrac{\sigma_{u_0} + \gamma u}{1-\gamma}. \end{aligned}$$

que denota la función media de los excesos para del umbral u. Entonces para $u > u_0$, $E(X - u \mid X > u)$ es una función lineal de u.

Por lo tanto si Y tiene *DPG* con la media finita, entonces la gráfica de $E(X - u \mid X > u)$

vs u, para $u > 0$, debe parecerse a una línea recta con pendiente $\gamma/(1-\gamma)$, e intercepto $\sigma_u/(1+\gamma)$. Un problema con dicha gráfica, es que tiende a mostrar una alta variabilidad en los umbrales altos. Lo cual hace difícil discernir si una salida observada de linealidad es debido al fracaso de la *DPG* o a la variabilidad de la muestra.

Otro gráfico importante para validar si una muestra proviene de una *DPG*, es el que se genera con la expresión

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (X_{(i)} - u) \right) : u < X_{\text{máx}} \right\}, \tag{2.4}$$

donde $X_{(1)}, ..., X_{(n_u)}$ son n_u observaciones que exceden el umbral u, y $X_{\text{máx}}$ es el valor más grande de las X_i. Las parejas de valores de la ecuación (2.4) generan lo que se llama: gráfica de vida media residual. Si los excesos del umbral u_0 tienen distribución Pareto Generalizada, entonces la gráfica de vida media residual debería de ser aproximadamente lineal en u.

2.2.2. Prueba de bondad de ajuste bootstrap

Para una muestra aleatoria $X_1, ...X_n$ de una función de distribución F, definida en los números reales positivos, Villaseñor y González (2009) construyeron una prueba de bondad de ajuste para probar la hipótesis

$$H_0 : F \text{ tiene } DPG.$$

Para probar dicha hipótesis definen dos subclases de la distribución Pareto Generalizada:

$$A^+ = \{\text{Todas las } DPG \text{ con } \gamma \geq 0\}$$

y

$$A^- = \{\text{Todas las } DPG \text{ con } \gamma < 0\}.$$

Entonces para probar la hipótesis H_0, argumentan que es equivalente a probar

$$H_0 : F \in \left(A^+ \cup A^-\right).$$

En ese sentido proponen la prueba unión-intersección, que considera las pruebas

$$H_0^+ : F \in A^+$$

y

$$H_0^- : F \in A^-.$$

Para probar H_0^+ proponen la prueba estadística

$$R^+ = \begin{cases} R_1 & \text{si} \quad 0 \leq \hat{\gamma} < 0.5, \\ R_2 & \text{si} \quad \hat{\gamma} \geq 0.5, \end{cases}$$

donde R_1 y R_2 son los coeficientes de correlación que proponen para validar la bondad de ajuste de los modelos que se obtienen en cada caso.

Bajo H_0^+ se espera que los valores de R^+ estén cercanos a uno. Se rechaza H_0^+ si $R^+ < c_\alpha^+$, donde el valor crítico c_α^+ es el cuantil $100\alpha\,\%$ de la distribución de R^+ bajo H_0^+.

Para probar H_0^-, primero proponen a $|R^-|$, como un coeficiente de correlación para validar el modelo obtenido cuando se considera el caso $\gamma < 0$. Bajo H_0^- se espera que los valores de $|R^-|$ estén cercanos a uno. Se rechaza H_0^- si $|R^-| < c_\alpha^-$, donde el valor crítico c_α^- es el cuantil $100\alpha\,\%$ de la distribución de $|R^-|$ bajo H_0^-..

Para obtener los valores de c_α^- y c_α^+ usan el bootstrap paramétrico.

2.3. Estimación del índice extremo

Para la estimación del índice extremo, es decir $\theta \in [0,1]$ de $\{X_n\}_{n\geq 1}$ una secuencia estrictamente estacionaria de variables aleatorias con función de distribución marginal F, con punto final derecho del soporte finito o infinito $w = \sup\{x : F(x) < 1\}$ y la función $\bar{F} = 1 - F$, Ferro y Segers (2003a) consideran dos casos, según convenga. Se toma una muestra aleatoria $X_1, ..., X_n$, y un umbral u grande. Se define

$$N = N_u(u) = \sum_{i=1}^{n} I\left(X_i > u\right),$$

como el número de observaciones que exceden u, sea

$$1 \leq S_1 < \cdots < S_N \leq n,$$

los tiempos de las excedencias, y finalmente

$$T_i = S_{i+1} - S_i, \;\; i = 1, ..., N-1, \tag{2.5}$$

los tiempos entre las excedencias observadas.

Para el primer caso se define la variable aleatoria $T(u)$ como

$$T(u) \stackrel{d}{=} \text{mín}\left\{n \geq 1 : X_{n+1} > u\right\} \text{ dado } X_1 > u,$$

donde $\stackrel{d}{=}$ significa igual en distribución, y se define la convergencia en distribución de $\bar{F}(u)T(u)$ con

$$\bar{F}(u)T(u) \xrightarrow{d} T_\theta \text{ cuando } u \longrightarrow w.$$

donde T_θ es una variable aleatoria y $w = \sup\left\{x : F(x) < 1\right\}$.

Con los momentos de la variable aleatoria T_θ, Ferro y Segers (2003a) encuentran que

$$\hat{\theta}_n(u) = \frac{2\left(\sum_{i=1}^{N-1} T_i\right)^2}{(N-1)\sum_{i=1}^{N-1} T_i^2}.$$

En el segundo caso, si $T(u)$ denota una variable aleatoria en los enteros positivos, cuya distribución está dada por

$$P(T > n) = \theta p^{n\theta} \text{ para } n \geq 1,$$

donde $\theta \in (0, 1]$ y $p \in (0, 1)$. A partir de la ecuación anterior Ferro y Segers (2003a), proponen el segundo estimador para θ, dado por

$$\hat{\theta}_n^*(u) = \frac{2\left\{\sum_{i=1}^{N-1}(T_i - 1)\right\}^2}{(N-1)\sum_{i=1}^{N-1}(T_i - 1)(T_i - 2)}.$$

El $\hat{\theta}_n^*(u)$ asegura que las contribuciones de los tiempos más pequeños entre excedencias son 0, mientras que los tiempos más grandes entre las excedencias, raramente son afectadas. Como $\hat{\theta}_n^*(u)$ puede tomar valores más grandes que 1, y no está definido para tiempos entre excedencias más grandes que 2, entonces Ferro y Segers (2003a) definen

$$\tilde{\theta}_n(u) = \begin{cases} 1\hat{}\hat{\theta}_n(u) \text{ si } \text{máx}\left\{T_i : 1 \leq i \leq N-1\right\} \leq 2, \\ 1\hat{}\hat{\theta}_n^*(u) \text{ si } \text{máx}\left\{T_i : 1 \leq i \leq N-1\right\} > 2, \end{cases} \tag{2.6}$$

lo anterior permite estimar intervalos para el índice extremo θ.

El índice extremo mide la intensidad de dependencia en $\{X_n\}_{n\geq 1}$. Si $\theta = 0$, entonces la secuencia tiene una memoria prolongada, si $0 < \theta < 1$ la secuencia tiene una memoria corta, si $\theta = 1$ la secuencia no tiene memoria, y finalmente si $\theta > 0$, entonces la dependencia es débil.

2.4. Formación automática de grupos

Dada una muestra, $X_1, ..., X_n$ de una secuencia estacionaria de variables aleatorias, con función de distribución marginal F, función de supervivencia $\bar{F} = 1 - F$ y $w = \sup\{x : F(x) < 1\}$ el extremo derecho del soporte de F. El objetivo es identificar dentro de la muestra $X_1, ..., X_n$ grupos independientes.

Dos esquemas populares para la formación de grupos independientes son expuestos por Leadbetter et al. (1989) y Smith (1989); el problema con dichos esquemas es la elección ampliamente arbitraria de algunos parámetros involucrados en los esquemas de agrupamiento.

Ferro y Segers (2003*a*) proponen un esquema automático de agrupamiento para la formación de grupos independientes, mismo que será usado en este trabajo. El método se basa en el índice extremo θ, el cual es estimado antes de llevar a cabo el agrupamiento.

Ferro y Segers (2003*a*) clasifican los tiempos entre las excedencias (expresión (2.5)) en dos tipos: tiempos independientes entre grupos y tiempos independientes dentro de los grupos.

Asumiendo que se tiene $S_1 < \cdots < S_N$, N tiempos de las excedencias tal que $T_i = S_{i+1} - S_i$ son los tiempos entre las excedencias, para $i = 1, ..., N-1$, el método de agrupamiento es como sigue, dado un umbral, los tiempos entre llegadas de los excesos son clasificados en tiempos independientes entre grupos y tiempos independientes dentro de los grupos. Entonces se puede asumir que los tiempos entre las excedencias más grandes $C-1 = \lfloor \theta N \rfloor$ son tiempos aproximadamente independientes entre los grupos que divide al resto en conjuntos aproximadamente independientes de tiempos dentro de los grupos. Es decir si $T_{(C)}$ es el tiempo C-ésimo más grande de los T_i y T_{ij} es el tiempo j-ésimo de los T_i que excede a $T_{(C)}$, entonces $\{T_{ij}\}_{j=1}^{C-1}$ es un conjunto aproximadamente independiente de tiempos entre los grupos. En el caso de empates, se decrementa C hasta que $T_{(C-1)}$ sea estrictamente más grande que $T_{(C)}$. Dado $\mathcal{T}_j = \left\{T_{i_{j-1}+1}, ..., T_{i_{j-1}}\right\}$, donde $i_0 = 0$, $i_C = N$ y $\mathcal{T}_j = \emptyset$ si $i_j = i_{j-1} + 1$, entonces $\{T_{ij}\}_{j=1}^{C}$ es una colección de conjuntos aproximadamente independientes de tiempos dentro de los grupos. Cada conjunto $\mathcal{T}_j$ está asociado a un conjunto de excedencias dado un umbral, $C_j = \{X_k : k \in S_j\}$, donde $S_j = \left\{S_{i_{j-1}+1}, ..., S_{i_j}\right\}$.

Lo anterior justifica la descomposición del proceso observado en C grupos, donde el j-ésimo grupo incluye la excedencia C_j.

Este método (Ferro y Segers, 2003*a*), es equivalente al esquema expuesto por Smith (1989) con longitud de corrida igual a $k = T_{(C)}$, donde $C = \lfloor \theta N \rfloor + 1$, sin embargo el parámetro de agrupamiento ya no se toma de manera arbitraria pues el valor de k es ahora gobernado por el nivel de dependencia extrema en el proceso, el cual es cuantificado por θ. En la práctica C es remplazado por un estimador de θ, cualquier estimador de θ puede

ser usado. Empleando $\tilde{\theta}_n(u)$, se obtiene un procedimiento automático de agrupamiento, el cual está justificado por la teoría asintótica.

2.5. Pruebas de estacionaridad

Un proceso es estacionario de orden S, cuando sus momentos de orden S son independientes del tiempo. Esto equivale a decir, según la definición débil del proceso estocástico, que un proceso es estacionario si su media y varianza son iguales para cualquier tiempo t y si el valor de la covarianza entre dos periodos de tiempo depende solamente de la distancia o rezago entre esos dos periodos y no del tiempo en el cual se ha calculado la covarianza.

Una serie $\{X_n\}_{n\geq 1}$, es estacionaria de forma **débil**, si su media, varianza y covarianza son invariantes en el tiempo, es decir,

1. $E(X_t) = \mu < \infty$;
2. $Var(X_t) = E(X_t - \mu)^2 = \sigma^2 < \infty$;
3. $Cov(X_t, X_{t+k}) = E(X_t - \mu)(X_{t+k} - \mu) = \gamma_k < \infty$ (constante).
   La última ecuación trata sobre la covarianza con rezago k.

A la serie **estacionaria débil**, también se le conoce como serie de **covarianza estacionaria**.

Una serie es **estacionaria estricta**, si es de covarianza estacionaria y además la función de distribución es estacionaria.

A continuación se da la definición de estacionaridad estricta.

Definición 7 *Un proceso estocástico es estrictamente estacionario, si su ley de probabilidad no depende del tiempo. Es decir, si se toman cualquier par de subconjuntos consecutivos de una serie en el tiempo y su función de distribución conjunta es idéntica a cualquier subconjunto similar de la serie de tiempo dada.*

Estacionaridad de primer orden

Sea X_t, $t = 1, 2, ..., n$, una serie observada, para la cual se desea probar estacionalidad. Supóngase que se puede descomponer la serie en la suma de una tendencia deterministica,

una caminata aleatoria y un error estacionario, es decir

$$X_t = \beta t + r_t + \varepsilon_t,$$

donde r_t es una caminata aleatoria dada por

$$r_t = r_{t-1} + u_t,$$

tal que u_t es iid con media 0 y varianza σ_u^2. El valor inicial de r_0 es tratado como fijo y tomado como el intercepto. A partir de este planteamiento y permitiendo que β sea cero o distinto de cero, se trata de contrastar la hipótesis nula $\sigma_u^2 = 0$. Lo cual significaría $r_t = r_0$ para todo t, es decir, r sería constante. Obsérvese que la hipótesis nula implicaría estacionalidad de primer orden o de primer nivel si $\beta = 0$, o estacionalidad con respecto a una tendencia si $\beta \neq 0$.

Kwiatkowski et al. (1992) proponen la estadística de prueba $KPSS$, dada por

$$KPSS = \frac{1}{(\hat{w}n)^2} \sum_{k=1}^{n} \left(\sum_{j=1}^{k} (\hat{u}_j) \right)^2,$$

donde $(\hat{w})^2$ es un estimador no paramétrico de varianza grande, y $\hat{u}_j$ es la tendencia estimada de los datos. Para esta estadística de prueba, la hipótesis nula es estacionalidad de primer orden vs la alternativa raíz unitaria. Bajo la hipótesis nula

$$KPSS \longrightarrow \int_0^1 \left(W(\alpha) - \alpha W(1) \right)^2 d\alpha, \tag{2.7}$$

donde $W(\alpha) - \alpha W(1)$ es un puente browniano estándar. El símbolo $\longrightarrow$ en (2.7), significa convergencia débil de la medida de probabilidad asociada. Los valores críticos pueden ser encontrados en Kwiatkowski et al. (1992).

Estacionaridad estricta

Neri y Lima (2008) proponen una prueba para estacionaridad estricta.

Dado $\{X_t\}_{t=1}^n$ un conjunto de datos y $\tau \in [0, 1]$, se define

$$b(\tau) = \arg \max_{b \in \mathbb{R}} \sum_{t=1}^{n} \rho_\tau (X_t - b),$$

donde

$$\rho_\tau(u) = (1_{u<0} - \tau)\, u.$$

Por consiguiente $b(\tau)$ es el cuantil τ^{th} de la muestra $\{X_t\}_{t=1}^n$. El proceso empírico que proponen está dada por

$$S_n(r,\tau) := \frac{1}{\hat{\pi}(\tau)\sqrt{n}} \sum_{t=1}^{\lfloor n_r \rfloor} \psi_\tau(X_t - b(\tau)),$$

donde

$$\psi_\tau(u) = 1_{u<0} - \tau,$$

es el subgradiente de ρ_τ, $\tau \in [0,1]$ y $\hat{\pi}(\tau)^2$ es un estimador consistente no paramétrico de

$$\pi(\tau)^2 := \lim_{n\longrightarrow\infty} \mathbb{E}\left[\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_\tau(X_t - b_0(\tau))\right)^2\right],$$

$b_0(\tau)$ es el cuantil poblacional τ^{th} incondicional de $\{X_t\}_{t=1}^n$.

Para probar la estacionaridad estricta, Neri y Lima (2008), usan la métrica de Kolmogorv-Smirnoff para medir las fluctuaciones de $S_n(r,\tau)$ entre varios cuantiles $\tau \in \Gamma_w = [w, 1-w]$, para alguna $w \in \left(0, \frac{1}{2}\right)$, con lo cual proponen la estadística de prueba

$$SS = \max_{\tau\longrightarrow\Gamma_w} \max_{1\leq k\leq n} \frac{1}{\hat{\pi}(\tau)\sqrt{n}} \left| \sum_{t=1}^{k} \psi_\tau(X_t - b(\tau)) - \frac{k}{n} \sum_{t=1}^{n} \psi_\tau(X_t - b(\tau)) \right|. \tag{2.8}$$

Con la prueba anterior y fijando un nivel de significancia α, se puede aceptar o rechazar la hipótesis

$$\begin{aligned} H_0 &: \text{Estacionaridad estricta;} \\ H_1 &: \text{Heteroscedasticidad incondicional.} \end{aligned}$$

Neri y Lima (2008) proponen una metodología para calcular los valores críticos. Por ejemplo los valores críticos para los niveles de significancia de 10 %, 5 % y 1 % son 1.65, 1.77 y 2.01 respectivamente.

2.6. Prueba de Independencia

Dada la serie $\{X_t\}_{t=1}^n$, para verificar la hipótesis nula de ausencia de autocorrelación Box y Pierce (1970) proponen el estadístico

$$Q = n \sum_{k=t}^{L} \hat{r}_k^2 = n \sum_{k=t}^{L} \left(\frac{\sum_{t=k+1}^{n} \varepsilon_t \varepsilon_{t-k}}{\sum_{t=k+1}^{n} \varepsilon_t^2} \right), \tag{2.9}$$

donde n es el tamaño de la muestra y L es la longitud del rezago.

El estadístico Q, no es válido para muestras pequeñas.

Bajo la hipótesis nula de ausencia de autocorrelación, el estadístico Q se distribuye asintóticamente según una χ^2 con L grados de libertad.

Si la Q calculada excede al valor crítico de la tabla χ^2 al un nivel de significancia α seleccionado, entonces se rechaza la hipótesis nula

El problema al determinar la longitud del rezago, en estas pruebas de autocorrelación aún no ha sido resuelto.

2.7. Prueba de la chi-square

Lo importante en una prueba de bondad de ajuste es determinar la función de distribución subyacente que describe a la población, de la cual la muestra aleatoria fue tomada. Existe una gran variedad de pruebas de bondad de ajuste. La prueba que nosotros usaremos depende de la distribución χ^2, usualmente llamada prueba de la chi-cuadrada.

El estadístico está dado por

$$W = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{2.10}$$

donde las O_i son las frecuencias observadas y las E_i son las frecuencias esperadas (calculadas por la función de distribución acumulativa). El estadístico es usado para probar

$$\begin{aligned} H_0 &: \text{Los datos siguen la distribución especificada;} \\ H_a &: \text{Los datos no siguen la distribución especificada.} \end{aligned}$$

Se rechaza la hipótesis acerca de la función de distribución subyacente de los datos, si el valor calculado W excede el valor de $\chi^2_{\alpha,k-p-1}$ donde α es el área en la cola de la distribución χ^2, k es el número de clases en las cuales se han clasificado las observaciones de la muestra, y p es el número de parámetros estimados de la muestra para describir la distribución de probabilidad de la población

2.8. Prueba de Anderson-Darling

Scholz y Stephens (1987) proponen una prueba para ver si dada una muestra de datos, ésta viene de una población con una distribución específica, las hipótesis para la prueba Anderson-Darling son

$$\begin{aligned} H_0 &: \text{Los datos siguen la distribución específica;} \\ H_a &: \text{Los datos no siguen la distribución especificada.} \end{aligned}$$

Para el caso continuo introducen la estadística de bondad de ajuste

$$A_m^2 = m \int_{-\infty}^{\infty} \frac{\{F_m(x) - F_0(x)\}^2}{F_0(x)\{1 - F_0(x)\}} dF_0(x),$$

para probar la hipótesis de que la muestra aleatoria $X_1, ..., X_m$, con función de distribución empírica $F_m(x)$ viene de una población continua con función de distribución $F_0(x)$, completamente especificada. La función $F_m(x)$ está definida como la proporción de la muestra $X_1, ..., X_m$, la cual no es más grande que x.

La versión correspondiente para dos muestras es

$$A_{mn}^2 = \frac{mn}{N} \int_{-\infty}^{\infty} \frac{\{F_m(x) - G_n(x)\}^2}{H_N(x)\{1 - H_N(x)\}} dH_N(x), \tag{2.11}$$

donde $G_n(x)$ es la función de distribución empírica de la segunda muestra (independiente) $Y_1, ..., Y_n$, obtenida de una población continua con función de distribución continua $G(x)$ y

$$H_N(x) = \frac{\{mF_m(x) - nG_n(x)\}}{N},$$

con $N = m + n$, es la función de distribución empírica de la muestra agrupada. Si $H_N(x) = 1$, entonces la integral anterior es cero. En el caso de las dos muestras A_{nm}^2 es usado para probar la hipótesis que $F = G$. Sin especificar la función de distribución continua que tienen en común.

El estadístico de prueba de Anderson-Darling para k muestras (Scholz y Stephens, 1987) está dado por

$$A_{kn}^2 = \frac{1}{N} \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{N-1} \frac{(NM_{ij} - jn_i)^2}{j(N-j)},$$

donde M_{ij} es el número de observaciones en la i-ésima muestra, n_i, $i = 1, ..., k$ son las k muestras y $N = n_1 + \cdots + n_k$ es la muestra total.

El procedimiento de la prueba es el siguiente:

1. Se calcula A^2_{kn} y $\sigma^2_N = var(A^2_{kn})$ (Scholz y Stephens, 1987);
2. Se calcula
$$T_{kN} = \frac{A^2_{kn} - (k-1)}{\sigma^2_N};$$
3. De acuerdo a T_{kN}, los puntos porcentuales $t_{k-1}(\alpha)$ de la cola superior, están dados en Scholz y Stephens, (1987). Se rechaza H_0 en el nivel de significancia α si T_{kN} excede el punto dado $t_{k-1}(\alpha)$.

Capítulo 3

Modelo Bivariado

Dado $X_1, X_2, ...$, observaciones de una serie estacionaria con función de distribución F y un umbral u, los excesos se definen por

$$X_i - u > 0.$$

Aplicando la metodología de Ferro y Segers (2003a), se agrupan los datos, dando como resultado los grupos $C_1, C_2, ...$, y $N_1, N_2, ...$, los tamaños de dichos grupos.

También se define el máximo del grupo C_j por

$$Y_j = \text{máx}\left\{X_i - u : X_i \in C_j\right\},\ j = 1, 2, ...$$

Se asume que los excesos máximos $Y_1, Y_2, ...$, son variables aleatorias con función de distribución F_M; y los tamaños de los grupos $N_1, N_2, ...$, son variables aleatorias con función de densidad P_N y media $1/\theta$ (Leadbetter, 1983).

La distribución condicionada de M, dado que el tamaño del grupo N es k (Villaseñor y González, 2010) está dada por

$$P(M \leq y \mid N = k) = F_u^{n\theta}(y), \tag{3.1}$$

donde $\theta \in [0, 1]$, y F_u es la función de distribución de los excesos (ecuación (2.2)).

De acuerdo con la ecuación (3.1), la función de distribución marginal para M, está dada por

$$F_M(y) = \sum_{n=0}^{\infty} F_u^{n\theta}(y) P_N(n) = \varphi_N(F_u^{n\theta}(y)), \tag{3.2}$$

donde φ_N es la función generatriz de probabilidad de N.

Villaseñor y González (2010) sugieren que la variable aleatoria N, puede tener distribución Poisson $(1/\theta)$ o Binomial Negativa(γ, p).

3.1. Distribución condicionada de M dado que $N \sim BN(r,p)$

Si N tiene distribución Binomial Negativa (r,p), donde $r = \frac{p}{\theta q}$ para $0 < p < 1$, $q = 1-p$ y $\theta > 0$, entonces por la ecuación (3.2), la función de distribución de M, está dada por

$$F_M(y) = \left\{ \frac{p}{1 - qF_u^{\theta}(y)} \right\}^{\frac{p}{\theta q}}, \tag{3.3}$$

donde se propone que $F_u(y)$ sea la distribución Pareto Generalizada (ecuación (2.3)), sustituyendo la ecuación (2.3) en la ecuación (3.3) se tiene que

$$F_M(y) = \left\{ \frac{p}{1 - q\left[1 - \left(1 + \frac{\gamma}{\sigma_u} y\right)^{-1/\gamma}\right]^{\theta}} \right\}^{\frac{p}{\theta q}}. \tag{3.4}$$

3.1.1. Distribución de M cuando N es de tamaño k

Villaseñor y González (2010) en base al modelo (3.1), también proponen una distribución condicionada de M dado que $N = k$, siempre y cuando se cuente con la información suficiente para hacerlo. Dicha distribución está dada por

$$P(M \leq y \mid N = k) = G^{k\theta}(y; \sigma_u, \gamma), \tag{3.5}$$

donde G es la distribución Pareto Generalizada (ecuación (2.3)) para el caso cuando $0 < y < -\sigma_u/\gamma$, cuando $\gamma < 0$.

3.2. Estimación de parámetros

Tomando en cuenta los excesos máximos de los grupos y los tamaños de los grupos (Y_i, N_i) $i = 1, 2, ..., n$, se estiman los parámetros γ y σ_u, de la ecuación (3.4) (Villaseñor y González, 2010).

Si se considera el caso $0 < y < -\sigma_u/\gamma$ para $\gamma < 0$, entonces el estimador de máxima verosimilitud para $-\sigma_u/\gamma$ es $Y_{(n)}$, luego

$$\tilde{\sigma}_u = -\gamma Y_{(n)}. \tag{3.6}$$

En base a las estadísticas de orden $Y_{(1)}, Y_{(2)}, ..., Y_{(n-1)}$, el estimador de momentos de γ está dado por

$$\tilde{\gamma} = \frac{-1}{g(\theta, p)} \left(\frac{1}{n-1} \right) \sum_{i=1}^{n-1} \log \left(1 - \frac{Y_{(i)}}{Y_{(n)}} \right). \tag{3.7}$$

En efecto si Y tiene función de distribución F_M y haciendo $U = F_M(Y)$ para sustituirlo del lado izquierdo de la ecuación (3.3), y despejando para $F_u(y)$ se tiene

$$F_u(Y) = \left[\frac{1 - pU^{-\theta q/p}}{q} \right]^{1/\theta}, \tag{3.8}$$

tomando la parte derecha de la ecuación (2.3) e igualándola con la ecuación (3.8), se obtiene

$$1 - \left[\frac{1 - pU^{-\theta q/p}}{q} \right]^{1/\theta} = \left(1 + \frac{\gamma}{\sigma_u} Y \right)^{\frac{-1}{\gamma}},$$

luego aplicando la función logaritmo, se tiene que

$$\log \left(\left(1 + \frac{\gamma}{\sigma_u} Y \right)^{\frac{-1}{\gamma}} \right) = \log \left(1 - \left[\frac{1 - pU^{-\theta q/p}}{q} \right]^{1/\theta} \right),$$

de donde

$$V = \log \left(1 - \left[\frac{1 - pU^{-\theta q/p}}{q} \right]^{1/\theta} \right).$$

Finalmente aplicando el operador esperanza, se observa que

$$E\{V\} = g(\theta, p),$$

donde

$$g(\theta, p) = E\left(\log\left(1 - \left[\frac{1 - pU^{-\theta q/p}}{q}\right]^{1/\theta}\right)\right),$$

para valores dados de θ y p. La expresión $g(\theta, p)$ puede ser aproximada por simulación, ya que U tiene distribución uniforme $(a, 1)$, donde $a = p^{p/(\theta q)}$.

Prueba de bondad del ajuste

Una vez estimados los parámetros del modelo (3.4), es necesario evaluar su ajuste.

Del modelo (3.4), se obtiene que

$$G(Y; \sigma_u, \gamma) = \left[\frac{1 - pF_M(Y)^{-\theta q/p}}{q}\right]^{1/\theta},$$

entonces

$$\left[\frac{1 - pF_M(Y)^{-\theta q/p}}{q}\right]^{1/\theta} = 1 - \left(1 + \frac{\gamma}{\sigma}Y\right)^{-1/\gamma},$$

despejando

$$\left[1 - \left(\frac{1 - pF_M(Y)^{-\theta q/p}}{q}\right)^{1/\theta}\right]^{-\gamma} = 1 + \frac{\gamma}{\sigma}Y. \tag{3.9}$$

Ahora tomando la parte izquierda de la ecuación anterior y sustituyendo los parámetros θ, p y γ por sus estimaciones, $F_M(Y)$ por su función de distribución empírica $F_n(Y)$ (ecuación (A1)) y además basados en sus estadísticas de orden $Y_{([an])}, Y_{([an])+1}, ..., Y_{(n)}$ de los excesos máximos de los grupos, se obtiene

$$S_i = \left[1 - \left(\frac{1 - \hat{p}F_n(Y)^{-\frac{\hat{\theta}\hat{q}}{\hat{p}}}}{\hat{q}}\right)^{\frac{1}{\hat{\theta}}}\right]^{-\hat{\gamma}},$$

con $i = [an], [an] + 1, ..., n$.

Si el modelo (3.4) es correcto entonces por la ecuación (3.9) las parejas $\left(Y_{(i)}, S_i\right)$ deberían de caer aproximadamente en una línea recta.

También por una prueba gráfica se puede evaluar la bondad del modelo (3.5). De acuerdo a la ecuación (3.1) se desea probar que

$$P(M \leq y \mid N = k) = G^{k\theta}(y; \sigma_u, \gamma),$$

en base a $Y_{(1,k)}, Y_{(2,k)}, ..., Y_{(n_k,k)}$ obtenidas de los grupos de tamaño k.

A partir del modelo (3.5), es decir,

$$P(M \leq y \mid N = k) = \left\{1 - \left(1 + \frac{\gamma}{\sigma_u} y\right)^{-1/\gamma}\right\}^{k\theta},$$

se observa que

$$\left(1 + \frac{\gamma}{\sigma_u} y\right)^{-1/\gamma} = 1 - P(M \leq y \mid N = k)^{\frac{1}{k\theta}},$$

luego

$$1 + \frac{\gamma}{\sigma_u} y = \left[1 - P(M \leq y \mid N = k)^{\frac{1}{k\theta}}\right]^{-\gamma},$$

finalmente

$$Z_{j,k} = \left[1 - F_{n_k}^{1/k\hat{\theta}}(Y_{j,k})\right]^{-\gamma}.$$

Si F_{n_k} es la función de distribución empírica (ecuación (A1)) basada en las observaciones $Y_{(j,k)}$, $j = 1, 2, ..., n_k$, entonces F_{n_k} es un estimador de $P(M \leq y \mid N = k)$, luego las parejas $(Z_{j,k}, Y_{(j,k)})$ deberían caer aproximadamente en una línea recta siempre y cuando el modelo

$$P(M \leq y \mid N = k) = G_u^{k\theta}(y; \sigma_u, \gamma),$$

ajuste a los datos.

Capítulo 4

Aplicaciones del modelo Bivariado y la distribución Pareto Generalizada

Se ajustará la *DPG* y el modelo Bivariado (Villaseñor y González, 2010), al conjunto de datos, que mas adelante se detallará.

El objetivo es analizar semejanzas y diferencias entre ambos modelos, para obtener un análisis estadístico más completo. Empezaremos con el modelo Bivariado, haciendo las respectivas pruebas estadísticas; y de igual manera se procederá con el modelo de la *DPG.*

4.1. Estimación del modelo Bivariado

La base de datos a utilizar son los máximos diarios de ozono de la estación Centro, que corresponden a la estación de primavera de los años 1996 al 2009. La Figura 4.1 muestra las 1264 observaciones de la base de datos.

Antes de formar los grupos con la metodología de Ferro y Segers (2003a), es necesario probar la hipótesis de estacionaridad estricta en los datos, es decir

$$\begin{aligned} H_0 &: \text{Estacionaridad estricta;} \\ H_1 &: \text{Heterocedasticidad incondicionada.} \end{aligned}$$

Se efectuó la prueba (2.8) de estacionaridad estricta (Neri, 2008), obteniéndose un valor p de 0.001, si fijamos un $\alpha = 0.05$, entonces

$$valor\ p < \alpha,$$

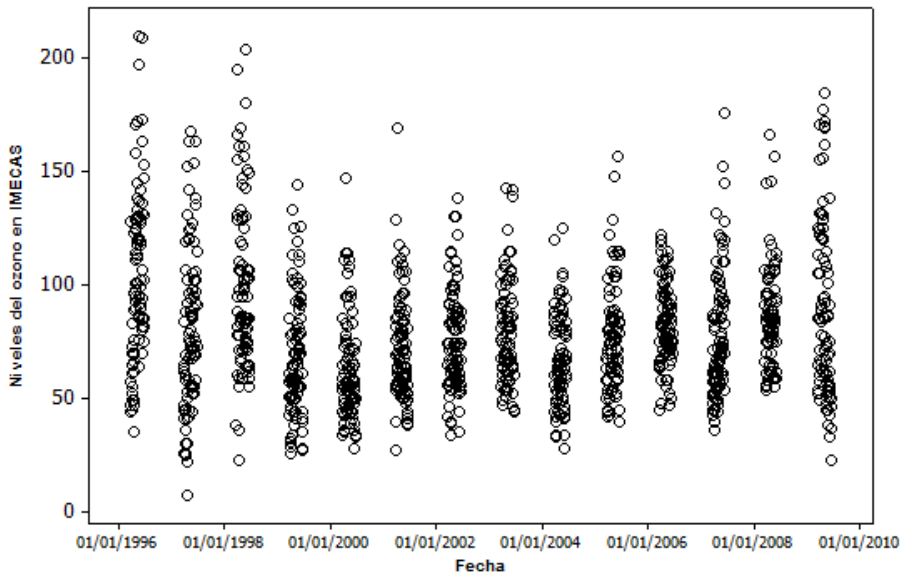

Figura 4.1: Niveles de ozono para la estación Centro en las estaciones de primavera $1996 - 2009$.

por lo tanto se rechaza la hipótesis nula.

Omitiendo de la base de datos las observaciones que corresponden a los años 1996 y 2009, queda una nueva base con 1080 observaciones, cuya gráfica se muestra en la Figura 4.2.

Nótese que ahora ninguna observación rebasa los 200 *IMECAS*. Efectuando nuevamente la prueba (2.8), se obtiene un valor p de 0.106, si fijamos un nivel de significancia de $\alpha = 0.05$, entonces

$$valor\ p > \alpha,$$

como el valor p es mayor que α, entonces no se rechaza H_0.

Ahora con un umbral de $u = 100$ (*IMECAS*), se ejecutan los programas *exi.intervals* y *decluster.runs* (Ferro y Segers, 2003*b*) para calcular el valor del índice extremo (ecuación (2.6)) y realizar el agrupamiento automático de los datos. Como resultados se obtiene un $\hat{\theta} = 0.5300$, y un total de 86 grupos.

El Cuadro 4.1, muestra los tamaños de los 86 grupos, el valor del exceso máximo por grupo y la fecha en la se rebaso el umbral.

El tamaño del grupo y las frecuencias de estos tamaños se muestran en el Cuadro 4.2.

Para verificar independencia entre los grupos $N_1, N_2, ..., N_{86}$, se aplica la prueba de Box y Pierce (ecuación (2.9)), como resultado se obtiene un valor p de 0.8662, con lo cual

Cuadro 4.1: Excesos máximos de cada uno de los 86 grupos.

N	Fecha	M	N	Fecha	M	N	Fecha	M
2	08/04/97	19	1	13/05/00	11	1	17/04/05	6
3	17/04/97	52	1	31/03/01	29	2	28/04/05	15
2	22/04/97	42	2	05/04/01	69	2	09/05/05	29
3	01/05/97	63	2	17/04/01	18	4	18/05/05	48
2	08/05/97	68	1	30/04/01	12	1	26/05/05	4
2	24/05/97	27	1	10/05/01	9	1	01/06/05	6
6	06/06/97	63	2	17/05/01	6	3	09/06/05	57
1	11/06/97	2	1	26/05/01	15	1	17/06/05	15
3	17/06/97	35	1	04/06/01	6	1	23/03/06	6
7	31/03/98	95	2	04/04/02	8	2	30/03/06	22
1	09/04/98	10	2	13/04/02	15	4	05/04/06	20
6	23/04/98	69	3	07/05/02	30	2	19/04/06	11
6	01/05/98	47	1	15/05/02	7	1	03/05/06	2
5	12/05/98	61	1	22/05/02	4	2	10/05/06	11
1	23/05/98	30	2	30/05/02	38	4	24/05/06	15
5	29/05/98	80	1	26/03/03	5	3	16/04/07	10
1	09/06/98	7	2	12/04/03	11	2	23/04/07	32
4	17/06/98	49	1	22/04/03	2	2	13/05/07	22
4	07/04/99	33	1	26/04/03	9	1	24/05/07	21
2	24/04/99	25	1	30/04/03	43	4	05/06/07	52
1	12/05/99	19	4	13/05/03	24	6	13/06/07	76
2	20/05/99	44	1	26/05/03	7	1	21/03/08	7
1	28/05/99	13	4	06/06/03	42	2	01/04/08	45
1	01/06/99	10	1	13/06/03	39	2	08/04/08	3
2	08/06/99	26	1	24/03/04	20	7	19/04/08	66
1	05/04/00	11	1	17/05/04	4	1	03/05/08	46
2	18/04/00	47	2	24/05/04	25	5	27/05/08	57
2	28/04/00	14	1	01/04/05	3	2	06/06/08	14
1	06/05/00	8	1	12/04/05	22			

Cuadro 4.2: Tamaños y frecuencias observadas de los tamaños de los grupos.

Tamaño del grupo	Frecuencia de los tamaños observados
1	36
2	27
3	6
4	8
5	3
6	4
7	2

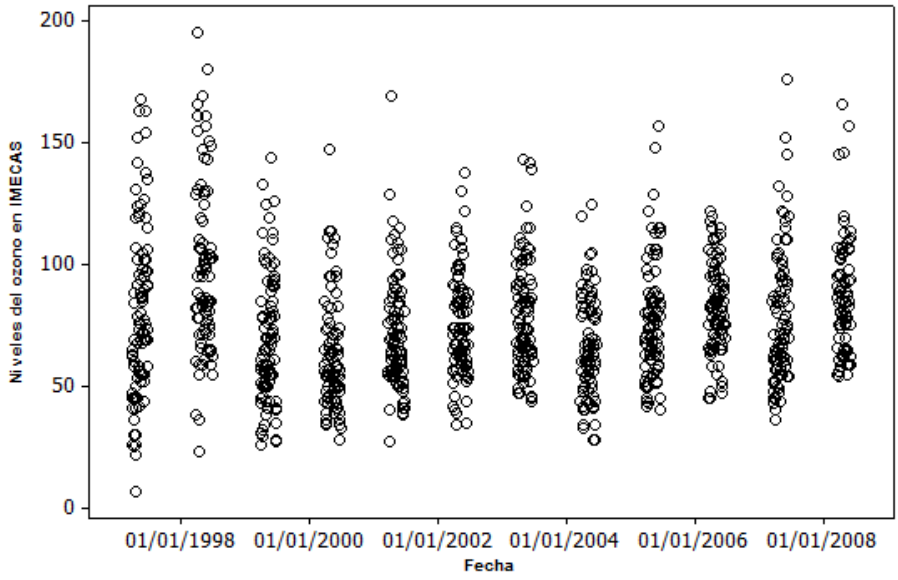

Figura 4.2: Niveles del ozono para la estación Centro de las estaciones de primavera 1997 – 2008.

podemos asegurar que los tamaños de los grupos no están correlacionados.

Calculando ahora la misma prueba (ecuación (2.9)) para los excesos máximos, se obtiene un valor p de 0.5175, con lo cual podemos asegurar que dichos excesos no están correlacionados. La Figura 4.3, no muestra la existencia de algún patrón sistemático para los excesos máximos en el tiempo.

4.1.1. Distribución de los tamaños de los grupos

Lo que ahora nos interesa es estimar la función de distribución de los tamaños de los grupos. En base a los resultados del Cuadro 4.2, se usará la prueba (2.10), para determinar si los tamaños de los grupos se distribuyen Poisson $(1/\theta)$ o Binomial Negativa (r, p).

La distribución Poisson está dada por

$$f_X(x;\mu) = \frac{e^{-\mu}\mu^x}{x!}, \ x = 0, 1, 2, ...$$

donde μ es estimada por

$$\hat{\mu} = \frac{1}{\hat{\theta}} = \frac{1}{0.5300666} = 1.8866 \approx 1.9.$$

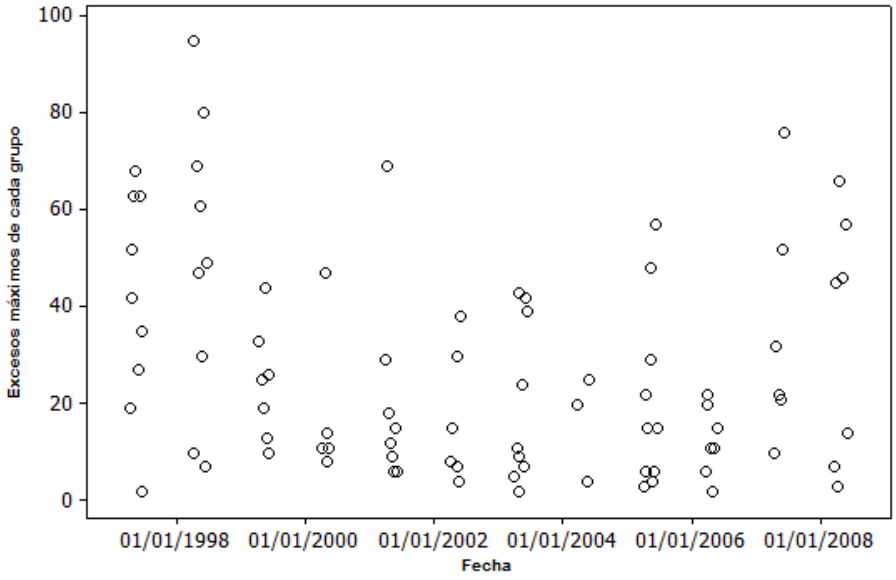

Figura 4.3: Excesos máximos de cada uno de los 86 grupos.

Conociendo el valor de $\hat{\mu}$ se calculan las frecuencias esperadas dadas por

$$f_N(x;\hat{\mu}) = \frac{e^{-1.9}(1.9)^x}{x!} = \frac{0.15(1.9)^x}{x!},\ x = 1, ..., 7$$

el Cuadro 4.3 muestra los resultados.

Cuadro 4.3: Frecuencias esperadas para los tama nos de los grupos con la distribución Poisson.

Tamaño de los grupos	Frecuencias observadas	Frecuencias esperadas
N	O	E
1	36	25
2	27	23
3	6	15
4	8	7
5	3	2.7
6	4	0.8
7	2	0.23

El estadístico de prueba es

$$\begin{aligned} W &= \sum_{i=1}^{7} \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(36-25)^2}{25} + \frac{(27-23)^2}{23} + \frac{(6-15)^2}{17} + \frac{(8-7)^2}{7} \\ &\quad + \frac{(3-2.7)^2}{2.7} + \frac{(4-0.8)^2}{0.8} + \frac{(2-0.23)^2}{0.23} \\ &= 36.898. \end{aligned}$$

Los grados de libertad son

$$gl = 7 - 1 - 1 = 5,$$

con $gl = 5$ y un $\alpha = 0.002$ se tiene el valor crítico de 18.9074, es decir,

$$\chi^2_{0.002,5} = 18.9074,$$

como el valor de W excede el valor de $\chi^2_{0.002,5}$, entonces se rechaza la hipótesis de que muestra provenga de la distribución Poisson $(\hat{\mu})$.

Por otra parte, la distribución Binomial Negativa está dada por

$$f_X(x; r, p) = \begin{pmatrix} r + x - 1 \\ x \end{pmatrix} p^r q^x, \ x = 0, 1, 2, ...,$$

donde $r > 0$, $0 < p \leq 1$, nótese que

$$\begin{aligned} E(x) &= \frac{rq}{p}, \\ V(x) &= \frac{rq}{p^2}. \end{aligned}$$

Villaseñor y González (2010), proponen que cuando N tiene distribución Binomial Negativa (r, p), donde $0 < p < 1$, $q = 1 - p$, $\theta > 0$, entonces

$$r = \frac{p}{\theta q}, \tag{4.1}$$

donde θ es el índice extremo.

Tomando en cuenta la ecuación (4.1) se tiene que

$$V(x) = \frac{rq}{p^2} = \frac{\left(\frac{p}{\theta q}\right)(q)}{p^2} = \frac{\frac{p}{\theta}}{p^2} = \frac{1}{\theta p},$$

simplificando

$$\hat{p} = \frac{1}{\hat{\theta}\hat{\sigma}^2_N},$$

donde $\hat{\sigma}_N^2$ es la varianza estimada de los 86 grupos, entonces

$$\hat{\sigma}_N^2 = (1.57133)^2 = 2.4691,$$

y como el valor de θ estimado es

$$\hat{\theta} = 0.53,$$

sustituyendo se tiene

$$\hat{p} = \frac{1}{\hat{\theta}\hat{\sigma}_N^2} = \frac{1}{(2.4691)\,(0.5300666)} = 0.76407,$$

entonces

$$\hat{q} = 1 - 0.76407 = 0.23593,$$

y finalmente

$$\hat{r} = \frac{\hat{p}}{\hat{\theta}\hat{q}} = \frac{0.76407}{(0.53)(0.23593)} = 6.1105 \approx 6.$$

Con estos parámetros estimados se calculan las frecuencias esperadas, dadas por

$$f_X(x;\hat{r},\hat{p}) = \begin{pmatrix} 5+x \\ x \end{pmatrix} (0.2)(0.236)^x,\ x = 1, ..., 7;$$

El Cuadro 4.4 muestra los resultados.

Cuadro 4.4: Frecuencias esperadas para los tama nos de los grupos con la distribución Binomial Negativa.

Tamaño de los grupos	Frecuencias observadas	Frecuencias esperadas
N_i	O_i	E_i
1	36	24
2	27	20
3	6	13
4	8	6.7
5	3	3.2
6	4	4.9
7	2	0.56

Haciendo los cálculos se tiene

$$\begin{aligned} W &= \sum_{i=1}^{7} \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(36-24)^2}{24} + \frac{(27-20)^2}{20} + \frac{(6-13)^2}{13} + \frac{(8-6.7)^2}{6.7} \\ &\quad + \frac{(3-3.2)^2}{3.2} + \frac{(4-4.9)^2}{4.9} + \frac{(2-0.56)^2}{0.56} \\ &= 16.352. \end{aligned}$$

Los grados de libertad son

$$gl = 7 - 2 - 1 = 4,$$

con $gl = 4$ y un $\alpha = 0.002$ se obtiene un valor critico de 16.9238, es decir,

$$\chi^2_{0.002,\ 4} = 16.9238,$$

como el valor de W no excede el valor de $\chi^2_{0.002,5}$, entonces no se rechaza de que la muestra proviene de la distribución $BN(\gamma, p)$. El valor de p calculado es

$$valor\ p = 0.002581,$$

como

$$\alpha < valor\ p,$$

se concluye de igual manera.

Como $N \sim BN(\gamma, p)$, entonces la distribución de la variable aleatoria M está dada por la ecuación (3.4), es decir

$$F_M(y) = \left\{ \frac{p}{1 - q\left[1 - \left(1 + \frac{\gamma}{\sigma} y\right)^{\frac{-1}{\gamma}}\right]^{\theta}} \right\}^{\frac{p}{\theta q}}.$$

Ahora bien, ordenando los excesos máximos del Cuadro 4.1 se estiman los parámetros γ y σ con las ecuaciones (3.6) y (3.7). Resultado

$$\begin{aligned} \hat{\gamma} &= -0.3566194, \\ \hat{\sigma} &= -\tilde{\gamma} Y_{(n)} = 35.51212, \end{aligned}$$

recordando que los otros parámetros estimados son

$$\begin{aligned} \hat{\theta} &= 0.5300666, \\ \hat{p} &= 0.77, \\ \hat{q} &= 0.23, \end{aligned}$$

entonces se puede dar una expresión más explícita para la distribución de la variable

aleatoria M cuando $N \sim BN(\gamma, p)$, es decir

$$F_M(y;\hat{\theta},\hat{p},\hat{q},\hat{\gamma},\hat{\sigma}) = \left\{ \frac{0.77}{1-(0.23)\left[1-(1-0.01y)^{2.804\,1}\right]^{0.5300}} \right\}^{6.316\,7}, \qquad (4.2)$$

la cual resulta de sustituir los parámetros correspondientes.

4.1.2. Evaluando la prueba de bondad del ajuste del modelo Bivariado

Para verificar la bondad de ajuste del modelo (4.2), necesitamos graficar las parejas de valores $(Y_{(i)}, S_i)$ donde

$$S_i = \left[1 - \left(\frac{1-\hat{p}F_n(Y_{(i)})^{-\frac{\hat{\theta}\hat{q}}{\hat{p}}}}{\hat{q}} \right)^{\frac{1}{\hat{\theta}}} \right]^{-\hat{\gamma}}.$$

Los valores que tomara $Y_{(i)}$, son las estadísticas de orden $Y_{([an])}, Y_{([an])+1}, ..., Y_{(n)}$, donde

$$a = p^{\frac{p}{\theta q}} = 0.1919996,$$

entonces

$$\begin{gathered}[an] = [(0.1944594)(86)] = [16.724] = 16,\\ [an]+1 = 17,\\ .\\ .\\ .\\ n = 86,\end{gathered}$$

entonces $i = 16, 17, ..., 86$. El resto de los parámetros estimados son

$$\begin{aligned}\hat{p} &= 0.77,\\ \hat{q} &= 0.23,\\ \hat{\theta} &= 0.5300,\\ \hat{\gamma} &= -0.3566194.\end{aligned}$$

La Figura 4.4 muestra la gráfica de los 70 valores, de las parejas $(Y_{(i)}, S_i)$.

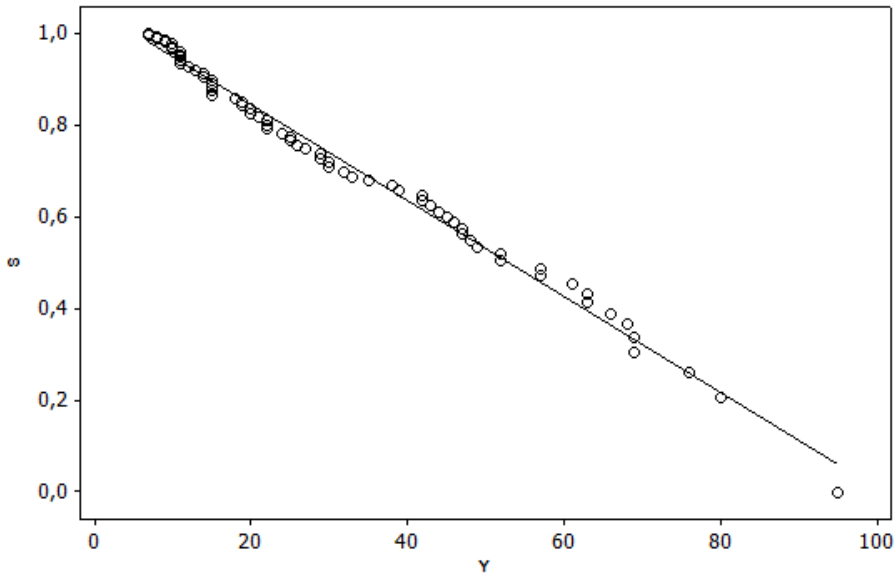

Figura 4.4: El coeficiente de correlaión de S y Y es de -0.996.

Como el coeficiente de correlación obtenido es de -0.996, entonces el modelo (4.2) produce un buen ajuste.

De acuerdo con el Cuadro 4.2, los valores que puede tomar k para estimar la función de distribución condicionada para M dado que $N = k$ son: $k = 1$ y $k = 2$.

Para $k = 1$, el modelo ajustado es

$$\begin{aligned} P(M \leq y \mid N = 1) &= (G(y; \hat{\sigma}_u, \hat{\gamma}))^{\hat{\theta}} \\ &= \left(1 - (1 - 0.01y)^{2.8041}\right)^{0.53}. \end{aligned} \tag{4.3}$$

La Figura 4.5 muestra las parejas de valores $(Y_{(j,1)}, Z_{j,1})$, para $j = 1, 2, ..., 36$, cuando $k = 1$.

Como el coeficiente de correlación de Pearson es de -0.978, entonces el modelo (4.3) produce un buen ajuste.

Para $k = 2$, el modelo ajustado es

$$\begin{aligned} P(M \leq y \mid N = 2) &= (G(y; \hat{\sigma}_u, \hat{\gamma}))^{2\hat{\theta}} \\ &= \left(1 - (1 - 0.01y)^{2.8041}\right)^{106}. \end{aligned} \tag{4.4}$$

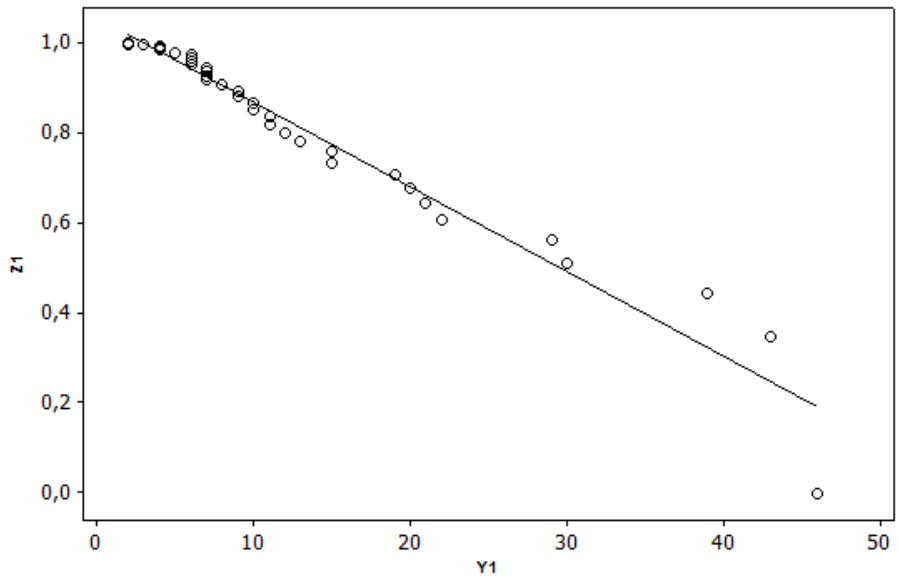

Figura 4.5: Cuando $k = 1$ el coeficiente de correlación es de -0.978.

La Figura 4.6 muestra las parejas de valores $(Y_{(j,2)}, Z_{j,2})$, para $j = 1, 2, ..., 27$ cuando $k = 2$.

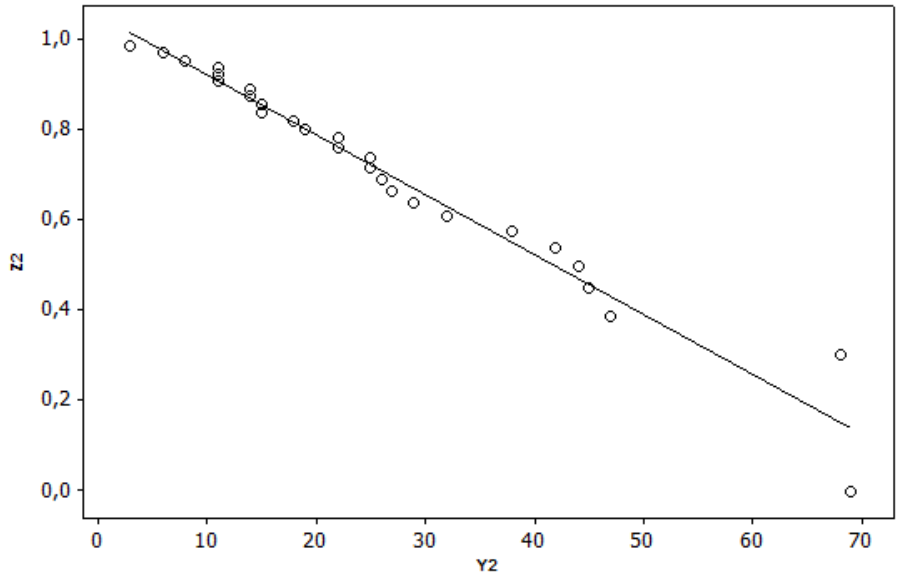

Figura 4.6: Cuando $k = 2$ el coeficiente de correlación es de -0.981.

Como el coeficiente de correlación de Pearson es de -0.981, entonces el modelo (4.4) produce un buen ajuste.

Por otra parte, Villaseñor y González (2010) asumen que si N es una variable aleatoria con distribución $BN(\gamma, p)$, entonces la esperanza del tamaño del grupo dado que el exceso máximo de éste es menor o igual a algún valor real y, está dada por

$$E\{N \mid M \leq y\} = \frac{pF_u^{\theta}(y)}{\theta\left(1 - qF_u^{\theta}(y)\right)}, \tag{4.5}$$

donde $F_u(y)$ es la ecuación (2.3); los parámetros estimados de la ecuación anterior son

$$\begin{aligned}
&\hat{p} = 0.77,\\
&\hat{q} = 0.23,\\
&\hat{\theta} = 0.5300,\\
&\hat{\gamma} = -0.3566194,\\
&\hat{\sigma}_u = \hat{\sigma}_N = (1.57133).
\end{aligned}$$

El Cuadro 4.5 muestra los resultados de la ecuación (4.5) tomando en cuenta el tamaño del grupo y el exceso máximo de cada grupo.

Nótese que los valores de la ecuación (4.5) convergen a

$$\frac{1}{\hat{\theta}} = \frac{1}{0.5300} = 1.8866.$$

También nótese que a medida que los excesos máximos de los grupos crecen, también crece el tamaño del grupo.

4.2. Estimación de la distribución Pareto Generalizada

Se ajustará la distribución Pareto Generalizada al conjunto de los 86 excesos máximos (Cuadro 4.1), que resultan después de agrupar las 193 excedencias obtenidas con el umbral $u = 100$ *IMECAS*.

Antes de comenzar, verifiquemos si es correcto el uso del modelo (2.3), con la gráfica de la vida media residual (ecuación (2.4)). Primero consideremos para el umbral un rango de valores, es decir, $u = 92, ..., 105$. Para cada uno de los umbrales, en el rango dado, se obtendrá un conjunto de excedencias, las cuales a su vez serán clasificadas en grupos

Cuadro 4.5: Estimación condicionada de la media del tama no del grupo.

N	y	$E\{N \mid M \leq y\}$	N	y	$E\{N \mid M \leq y\}$	N	$E\{N \mid M \leq y\}$	y
1	2	0.32932	2	11	0.83735	4	33	1.45231
1	2	0.32932	1	12	0.87739	3	35	1.48927
1	2	0.32932	1	13	0.91569	2	38	1.54045
2	3	0.41134	2	14	0.95241	1	39	1.55643
1	3	0.41134	2	14	0.95241	2	42	1.60131
1	4	0.48177	1	15	0.98767	4	42	1.60131
1	4	0.48177	2	15	0.98767	1	43	1.61529
1	4	0.48177	2	15	0.98767	2	44	1.62879
1	5	0.54463	1	15	0.98767	2	45	1.64184
2	6	0.60196	4	15	0.98767	1	46	1.65443
1	6	0.60196	2	18	1.08580	2	47	1.66657
1	6	0.60196	2	19	1.11623	6	47	1.66657
1	6	0.60196	1	19	1.11623	4	48	1.67829
1	6	0.60196	4	20	1.14562	4	49	1.68958
1	7	0.65500	1	20	1.14562	3	52	1.72099
1	7	0.65500	1	21	1.17404	4	52	1.72099
1	7	0.65500	2	22	1.20153	3	57	1.76565
1	7	0.65500	2	22	1.20153	5	57	1.76565
2	8	0.70456	1	22	1.20153	5	61	1.79505
1	8	0.70456	4	24	1.25388	3	63	1.80780
1	9	0.75122	2	25	1.27881	6	63	1.80780
1	9	0.75122	2	25	1.27881	7	66	1.82467
1	10	0.79538	2	26	1.30297	2	68	1.83449
1	10	0.79538	2	27	1.32637	6	69	1.83900
3	10	0.79538	2	29	1.37102	2	69	1.83900
1	11	0.83735	1	29	1.37102	6	76	1.86365
1	11	0.83735	3	30	1.39231	5	80	1.87295
2	11	0.83735	1	30	1.39231	7	95	1.88632
2	11	0.83735	2	32	1.43294			

(Ferro y Segers, 2003a), de cada uno de estos grupos se tomará la excedencia máxima, y se formará un conjunto de excedencias máximas; que a su vez se les restará su umbral u correspondiente, de dicho procedimiento se obtendrán los excesos máximos que a su vez serán promediados y graficados contra su umbral u correspondiente. Es decir, si se tiene $X_1, X_2, ..., X_n$ observaciones, $u = 92, ..., 105$, y también $C_{u,1}, C_{u,2}, ...,$ una secuencia de grupos de excedencias, entonces

$$Y_j = \text{máx}\left\{X_i - u : X_i \in C_{uj}\right\}, \ j = 1, 2, ..., \ u = 92, ..., 105$$

entonces

$$\frac{1}{n_u}\sum_{j=1}^{n_u}(Y_j) =: u < x_{\text{máx}}.$$

Bajo las condiciones anteriores, la Figura 4.7 muestra la gráfica de la de vida residual media.

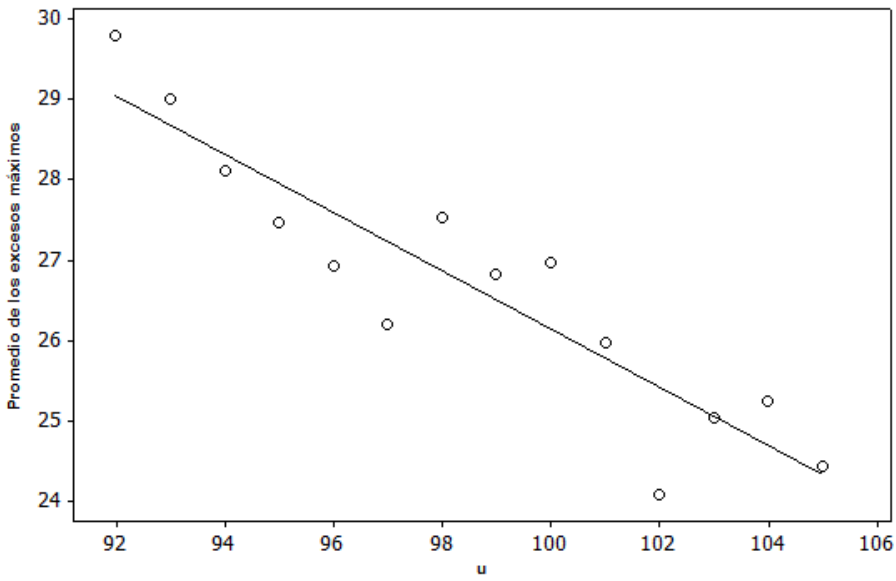

Figura 4.7: Gráfica de la vida media residual.

La recta ajustada, tiene por ecuación $ra = -0.363u + 62.4$, con una $R^2 = 0.84$, con lo cual podemos decir que la *DPG* es un buen modelo para las observaciones que exceden a $u = 100$. Nótese que la pendiente de la recta es -0.363, lo cual da una estimación aproximada para el parámetro γ. (Otra prueba para la *DPG* está dada en el Anexo A.)

4.2.1. Estimación de los parámetros γ y σ_u

Como sabemos la *DPG* está dada por

$$G(y;\sigma_u,\gamma) = 1 - \left(1 + \frac{\gamma}{\sigma_u} y\right)^{-\frac{1}{\gamma}},$$

donde $\sigma_u > 0$ y $\gamma \in \mathbb{R}$, tal que

$$\begin{array}{rl} y \geq 0 & \text{si} \quad \gamma \geq 0, \\ 0 \leq y \leq \frac{-\sigma_u}{\gamma} & \text{si} \quad \gamma < 0. \end{array}$$

Estimaremos los parámetros de la *DPG* para el caso donde $0 < y < -\sigma_u/\gamma$, si $\gamma < 0$.

Sea

$$U = G(y;\sigma_u,\gamma) = 1 - \left(1 + \frac{\gamma}{\sigma_u} y\right)^{-\frac{1}{\gamma}},$$

de donde

$$\left(1 + \frac{\gamma}{\sigma_u} y\right)^{-\frac{1}{\gamma}} = 1 - U,$$

aplicando logaritmo natural se tiene

$$-\frac{1}{\gamma} \log\left(1 + \frac{\gamma}{\sigma_u} y\right) = \log\left(1 - U\right),$$

se sabe que si U se distribuye uniforme en el intervalo $(0,1)$, entonces también $1 - U$ se distribuye uniforme en el intervalo $(0,1)$, luego se puede escribir

$$-\log\left(1 + \frac{\gamma}{\sigma_u} y\right) = (-\gamma)\left(-\log\left(U\right)\right),$$

nótese que $-\log U \sim \exp(1) \approx \Gamma(1,1)$, entonces $-\gamma \log U \approx \Gamma(1,-\gamma)$.

Ordenando los 86 excesos máximos, se obtiene el estimador de momentos

$$-\hat{\gamma} = -\frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \frac{\hat{\gamma}}{\hat{\sigma}_u} Y_{(i)}\right). \tag{4.6}$$

Por otra parte como estamos considerando el caso

$$0 \leq y \leq \frac{-\sigma_u}{\gamma} \text{ si } \gamma < 0,$$

entonces el estimador de máxima verosimilitud de

$$\frac{-\sigma_u}{\gamma},$$

es $Y_{(n)}$, luego el estimador de σ_u, está dado por

$$\hat{\sigma}_u = -\gamma Y_{(n)}. \tag{4.7}$$

Sustituyendo la ecuación (4.7), en la ecuación (4.6), y tomando la suma hasta $n-1$, se tiene que

$$\hat{\gamma} = \frac{1}{n-1} \sum_{i=1}^{n-1} \log\left(1 - \frac{Y_{(i)}}{Y_{(n)}}\right).$$

Haciendo los cálculos (Programa 6 del apéndice B) se tiene que

$$\hat{\gamma} = -0.3868249,$$
$$\hat{\sigma}_u = 36.74837,$$

serían los parámetros estimados, haciendo las sustituciones correspondientes se tiene que

$$G(y; \hat{\sigma}_u, \hat{\gamma}) = 1 - \left(1 + \left(\frac{-0.3868249}{36.74837}\right) y\right)^{\frac{-1}{-0.3868249}} = 1 - \left(1 + (-0.010526)\, y\right)^{2.5851} \tag{4.8}$$

4.2.2. Evaluando la bondad del ajuste de la distribución Pareto Generalizada

La bondad del ajuste de la *DPG* puede ser evaluada usando varios diagnósticos gráficos. Nosotros usaremos la estadística de prueba (2.11) y la gráfica QQ.

Para el modelo ajustado

$$G(y; \hat{\sigma}_u, \hat{\gamma}) = 1 - \left(1 + (-0.010526)\, y\right)^{2.5851},$$

donde $\sigma_u > 0$, $0 < y < -\sigma_u/\gamma$, si $\gamma < 0$, se calcula la estadística de prueba (2.11) y se obtiene un valor p de

$$valor\ p = 0.46438,$$

con un nivel de significancia de $\alpha = 0.05$, se concluye que la *DPG*, se ajusta bien a los datos.

La Figura 4.8 muestra la gráfica QQ, se puede observar que el conjunto de puntos sigue una tendencia lineal, lo cual implica que la ecuación (4.8) es un buen modelo para el conjunto de excesos máximos, mostrados en el Cuadro 4.1.

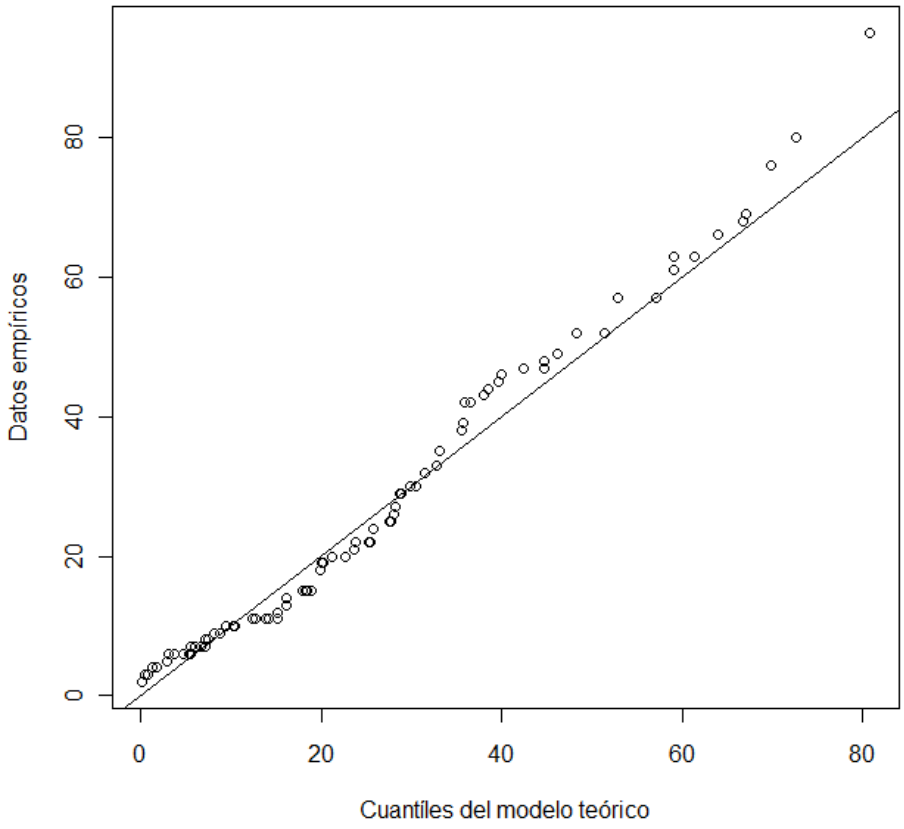

Figura 4.8: Gráfica QQ para la *DPG*.

Otra forma de evaluar el ajuste del modelo (4.8), es procediendo de la siguiente forma; dada la *DPG*

$$G(y; \hat{\sigma}_u, \hat{\gamma}) = 1 - \left(1 + \frac{\gamma}{\sigma_u} y\right)^{\frac{-1}{\gamma}},$$

se despeja el término $1 + \frac{\gamma}{\sigma_u} y$, resultando

$$1 + \frac{\gamma}{\sigma_u} y = (1 - G(y; \sigma_u, \gamma))^{-\gamma},$$

sustituyendo $G(y; \sigma_u, \gamma)$ por la función empírica (ecuación (A1)), se tiene

$$1 + \frac{\gamma}{\sigma_u} y = \left(1 - F_n(Y_{(i)})\right)^{-\gamma},$$

donde $Y_{(i)}$ $i = 1, ..., 86$ son los excesos máximos ordenados del Cuadro 4.1 ordenados.

Para evaluar la bondad del ajuste del modelo (4.8), definimos

$$Z_i = \left(1 - F_n(Y_{(i)})\right)^{-\hat{\gamma}}.$$

Si las parejas de valores $(Y_{(i)}, Z_i)$ caen en una línea recta, entonces el modelo (4.8) es correcto. La Figura 4.9 muestra la gráfica de las parejas $(Y_{(i)}, Z_i)$, estas tienen un coeficiente de correlación de -0.996, con lo cual podemos concluir que el ajuste del modelo (4.8) es bueno.

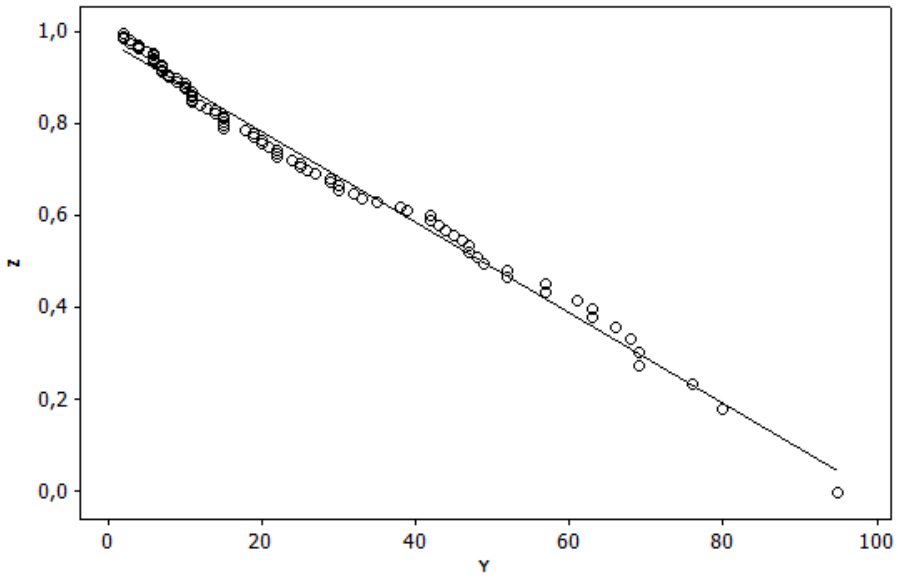

Figura 4.9: El coeficiente de correlaión de Z y Y es de -0.996.

Capítulo 5

Conclusiones

Dado que se consideró un umbral de $u = 100$ *IMECAS*, se obtuvieron los datos del Cuadro 4.1, a partir de éstos, los objetivos eran

1. Tomar en cuenta los excesos máximos de cada grupo y los tamaños de los grupos, para ajustar la distribución de probabilidad conjunta propuesta por Villaseñor y González (2010).
2. Considerar los excesos máximos de cada grupo, y ajustar la distribución Pareto Generalizada, cuyos parámetros serían estimados por el método de momentos.

De acuerdo con la sección 4.1 el modelo estimado para el primer objetivo fue

$$F_M(y;\hat{\theta},\hat{p},\hat{q},\hat{\gamma},\hat{\sigma}) = \left\{ \frac{0.77}{1-(0.23)\left[1-(1-0.01y)^{2.8041}\right]^{0.5300}} \right\}^{6.3167},$$

y considerando la prueba gráfica de la misma sección se concluyó que el modelo anterior tiene un buen ajuste para las $Y_i's$, pues se obtuvo un coeficiente de correlación de -0.996.

En el segundo objetivo, el modelo obtenido fue

$$G(y;\hat{\sigma}_u,\hat{\gamma}) = 1-(1-(0.010526)\,y)^{2.5851},$$

el cual de acuerdo con la prueba gráfica de la sección 4.2 se concluyó que también tiene un buen ajuste para las $Y_i's$, pues se obtuvo un coeficiente de correlación de -0.994.

Si evaluamos los dos modelos obtenidos en base al coeficiente de correlación pensaríamos que no existe gran diferencia entre ambos ajustes. Pero antes de quedarnos con esa impresión analicemos un poco más a fondo.

Aunque el modelo Bivariado, requiera en un principio de la estimación de más parámetros, su contribución al análisis de los datos es mayor, pues de éste se derivan la distribución condicionada de M cuando $N = 1$, es decir,

$$\begin{aligned} P(M \leq y \mid N = 1) &= (G(y; \hat{\sigma}_u, \hat{\gamma}))^{\hat{\theta}} \\ &= \left(1 - (1 - 0.01y)^{2.804\,1}\right)^{0.53}; \end{aligned}$$

y la distribución condicionada de M cuando $N = 2$, es decir,

$$\begin{aligned} P(M \leq y \mid N = 2) &= (G(y; \hat{\sigma}_u, \hat{\gamma}))^{2\hat{\theta}} \\ &= \left(1 - (1 - 0.01y)^{2.804\,1}\right)^{106}. \end{aligned}$$

Estos dos últimos modelos nos dan una idea de la distribución de los excesos máximos que no exceden un valor real y, y cuyo tamaño de grupo son 1 y 2 respectivamente.

Otra ecuación de suma importancia que se obtuvo a partir del modelo Bivariado fue la ecuación de la esperanza del tamaño del grupo dado que el exceso máximo de ese grupo es menor o igual que un número real y. Dicha ecuación es la que se muestra a continuación

$$E\{N \mid M \leq y\} = \frac{\hat{p}\left(1 - \left(1 + \frac{\hat{\gamma}}{\hat{\sigma}_u} y\right)^{\frac{-1}{\hat{\gamma}}}\right)^{\hat{\theta}}}{\hat{\theta}\left(1 - \hat{q}\left(1 - \left(1 + \frac{\hat{\gamma}}{\hat{\sigma}_u} y\right)^{\frac{-1}{\hat{\gamma}}}\right)^{\hat{\theta}}\right)},$$

donde los parámetros estimados son

$$\begin{aligned} &\hat{p} = 0.77, \\ &\hat{q} = 0.23, \\ &\hat{\theta} = 0.5300, \\ &\hat{\gamma} = -0.3566194, \\ &\hat{\sigma}_u = \hat{\sigma}_N = (1.57133). \end{aligned}$$

Algo que se observó fue que los valores de dicha ecuación convergen a

$$\frac{1}{\hat{\theta}} = \frac{1}{0.5300} = 1.8866,$$

y que a medida que los excesos máximos de los grupos crecían, también crecía el tamaño del grupo.

Como se puede observar las ventajas de usar el modelo Bivariado son más que las obtenidas con el uso de la distribución Pareto Generalizada.

Referencias

[1] Ancona-Navarrete M. A., and J.A. Tawn. 2000. *A Comparison of Methods for Estimating the Extremal Index*. Extremes 3(1) : 5 − 38.

[2] Balkema A.A., and L. De Haan. 1974. *Residual lifetime at great age*. Annals of Probability. 2(5) : 792 − 804.

[3] Beirlant J., Goegebeur, Y., Teugels, J., Segers, J., De Waal, D. and C. Ferro. 2004. *Statistics of Extremes, Theory and Applications*. Wiley Series in Probability and Statistics.

[4] Box G.E.P., and D. A. Pierce. 1970. *Distribution of residual correlations in autoregressive-integrated moving average time series models*. Journal of the American Statistical Association 65(332) : 1509 − 1526.

[5] Coles, S. 2001. *An introduction to Statistical Modeling of Extreme Values*. (ed.) Great Britain Springer-Verlag London. pp: 74 − 91.

[6] Davison A.C., and R.L. Smith. 1990. *Models for Exceedances over High Thresholds*. Journal of the Royal Statistical Society. Series B (Methodological) 52(3) : 393 − 442.

[7] Fawcett L., and D. Walshaw. 2007. *Beyesian inference for clustered extremes*. Extremes.

[8] Ferro C.A.T., and J. Segers. 2003*a*. *Inference for cluster of extreme values*. Royal Statistical Society 65(2): 545 − 556.

[9] Ferro C.A.T., and J. Segers. 2003*b*. *Index of /src/contrib/Archive/extRemes*. (Publicación en línea, disponible en internet en el sitio http://cran.r-project.org/src/contrib/Archive/extRemes/ [con acceso el 11 − 12 − 2010]).

[10] Hernández-Gallardo, Lorelie. 2009. *Modelado atmosférico para determinar niveles máximos diarios de ozono en la ciudad de Guadalajara.* Tesis para obtener el grado de Maestra en Ciencias. Universidad Autónoma Metropolitana Unidad-Iztapalapa.

[11] Hsing T. 1993. *Extremal index estimation for a weakly dependence in a stationary sequence*. The Annals of Statistics 21(4) : 2043 − 2070.

[12] Kwiatkowski D., Phillips, P.C.B., Schmidt P., and Y. Shin. 1992. *Testing the null hypothesis of stationarity against the alternative of unit root.* Journal of Econometrics $54 : 159 - 178$.

[13] Leadbetter M.R. 1983. *Extremes and local dependence in a stationary sequence.* Zeitschrift für Wahrscheinlichkeitshteorie und verwandte Gebiete $65 : 291 - 306$.

[14] Martínez-Cárdenaz M. A. 2010. Director del Cantro de Información Ambiental para el Desarrollo Sustentable. Poder Ejecutivo del Gobierno del Estado de Jalisco. http://semades.jalisco.bog.mx.

[15] Neri, B. 2008. R Code: Computes Several Stationarity Tests. (Publicación en línea, disponible en internet en el sitio http://homepages.nyu.edu/˜bpn207/Research/ [con acceso el $11 - 12 - 2010$]).

[16] Neri, B. y L.R. Lima. 2008. *A test for strict stationarity.* Department of Economics. New York University. (Publicación en línea, disponible en internet en el sitio http://homepages.nyu.edu/˜bpn207/Research/ [con acceso el $11 - 12 - 2010$]).

[17] Pickands J. 1975. *Statistical Inference Using Extreme Order Statistics.* The Annals of Statistics $3(1) : 119 - 131$.

[18] Robert C., Y., Segers J., and C.A.T. Ferro. 2009. *A sliding blocks estimator for the extremal index.* Electronic Journal of Statistics $3 : 993 - 1020$.

[19] Sánchez-Gómez, R. 2001. *Análisis de tendencia en excedencias sobre un umbral alto, con aplicaciones en ozono urbano.* Tesis para obtener el grado de Doctor en Ciencias. Colegio de Postgraduados, campus Montecillo.

[20] Scholz F.W. y M.A. Stephens. 1987. *K-Sample Anderson-Darling Tests.* Journal of the American Statistical Association $82(399) : 918 - 924$.

[21] Smith R. 1989. *Extreme Value Analysis of Enviromental Time Series: An Application to Trend Detection in Ground-Level Ozono.* Statistical Science $4(4) : 367 - 393$.

[22] Velasco-Luna, F. y Hernández-González, S. 2007. *Teoría de valores extremos: Una introducción.* Revista de Ciencias Básicas UJAT $6(1) : 10 - 16$.

[23] Villaseñor-Alva, J. A. y González-Estrada, E. 2009. *A bootstrap goodness of fit test for the generalizad Pareto distribution.* Computational Statistics and Data Analysis. $53 : 3835 - 3841$.

[24] Villaseñor-Alva, J. A. y González-Estrada, E. 2010. *On modeling cluster maxima with applications to ozone data from Mexico City.* Environmetrics. $21 : 528 - 540$.

Anexos

Anexo A

Función de distribución empírica

Una estimación para $F(x) = P(X \leq x)$, es la proporción de muestra de los puntos que caen dentro de $(-\infty,\ y]$. Dicha estimación es llamada la función de distribución empírica acumulativa o función de distribución empírica, para una muestra observada está definida por

$$F_n(y) = \begin{cases} 0, & y < y_{(1)}, \\ \frac{i}{n}, & y_{(i)} \leq y \leq y_{(i+1)}, \quad i = 1, ..., n-1, \\ 1, & y_{(n)} \leq y, \end{cases} \tag{A1}$$

donde $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$ es la muestra ordenada.

Prueba bootstrap

Otra prueba de bondad de ajuste para la distribución Pareto Generalizada, es propuesta por Villaseñor y González (2009). La prueba bootstrap que ellos proponen para evaluar el ajuste de la *DPG*, es la prueba unión-intersección, que propone los casos

$$\begin{aligned} H_0^- &: \text{La muestra aleatoria tiene } DPG \text{ con } \gamma < 0; \\ H_0^+ &: \text{La muestra aleatoria tiene } DPG \text{ con } \gamma \geq 0, \end{aligned}$$

Aplicando esta prueba al grupo de los 86 excesos máximos, se obtiene para H_0^- un valor p de 0.37337 y para H_0^+ un valor p de 0.005,con un nivel de significancia de $\alpha = 0.1$ la hipótesis, de que el grupo de los 86 excesos máximos tiene distribución Pareto Generalizada con $\gamma < 0$, no es rechazada

Anexo B: Código en R

Programa 1

Código para calcular el índice extremo y realizar el agrupamiento (Ferro y Segers, 2003a).

#—— Cargar los paquetes.

```
rm(list=ls())
library(extRemes)
library(RODBC)
library(vcd)
library(tseries)
library(MASS)
library(stats)
   #--------- El archivo de excel se llama "datos".
data=odbcConnectExcel(file.choose())
```

#—— Base de datos.

#—— LD1279 LD1276 LD1209 LD1190 CEN1264 CEN1080.

```
sqlTables(data)
mydat=sqlFetch(data,"CEN1080")
odbcClose(data)
a<-mydat$ld
length(a)
kpss.test(a)
stationarity(a)
```

#—— Primera forma de calcular theta.

```
x<-a
z<-x>100
exi.intervals(z)
```

#—— Calcula el número de grupos.

```
ozo <- decluster.intervals(z,exi.intervals(z))
```

#—— Goodness of fit Test.

#—— Tamaños de los grupos.

```
tcluster<-ozo$size
write.table(tcluster,file="N.xls",sep=";")
Box.test (tcluster,lag =1,type="Box-Pierce")
```

#—— Excesos máximos de cada uno de los grupos.

```
excemax<-c(19,52,42,63,68,27,63,2,35,95,10,69,47,61,30,80,7,49,33,25,19,44,
           13,10,26,11,47,14,8,11,29,69,18,12,9,6,15,6,8,15,30,7,4,38,5,11,
           2,9,43,24,7,42,39,20,4,25,3,22,6,15,29,48,4,6,57,15,6,22,20,11,2,
           11,15,10,32,22,21,52,76,7,45,3,66,46,57,14)
Box.test(excemax,lag=1,type="Box-Pierce")
**************************************************************************
**************************************************************************
```

Programa 2

Código para calcular

$$S_i = \left[1 - \left(\frac{1 - \hat{p}F_n(Y_{(i)})^{-\frac{\hat{\theta}\hat{q}}{\hat{p}}}}{\hat{q}}\right)^{\frac{1}{\hat{\theta}}}\right]^{-\hat{\gamma}}.$$

#—— Evaluando el ajuste del modelo cuando N tiene distribución Binomial Negativa.

```
rm(list=ls())
p<-0.77
q<-0.23
te<-0.53
ga<--0.3566194
sig<-35.51212
a<-(p)^(p/(te*q))
```

#—— Excesos máximos de cada uno de los 86 grupos.

```
#---- 2,2,2,3,3,4,4,4,5,6,6,6,6,6,7,
Y<-c(7,7,7,8,8,9,9,10,10,10,11,11,11,11,11,12,13,14,14,15,15,15,
    15,15,18,19,19,20,20,21,22,22,22,24,25,25,26,27,29,29,30,30,
    32,33,35,38,39,42,42,43,44,45,46,47,47,48,49,52,52,57,57,61,
    63,63,66,68,69,69,76,80,95)
m<-length(Y)
#----- Distribuci\'{o}n emp\'{i}rica.
FnY<-numeric(m)
for(j in 1:m)
   {
    FnY[j]<-(j+15)/86
   }
b<-(-te*q)/p
c<-1/te
S<-numeric(m)
Fn<-numeric(m)
for(i in 1:m)
   {
    Fn[i]<-(1-p*(FnY[i])^b)/q
    S[i]<-(1-(Fn[i])^c)^-ga
   }
```

#—— Guardar los grupos en excel.

```
write.table(S,file="Si.xls",sep=";")
**************************************************************************
**************************************************************************
```

Programa 3

Código para estimar

$$E\left\{N \mid M \leq y\right\} = \frac{pF_u^{\theta}(y)}{\theta\left(1 - qF_u^{\theta}(y)\right)},$$

donde

$$F_u(y) = 1 - \left(1 + \frac{\gamma}{\sigma}y\right)^{\frac{-1}{\gamma}}.$$

#—— La estimación del promedio de los tamaños de los grupos,

#—— condicionada a los excesos máximos de los grupos basados,

#—— en la distribución Binomial Negativa para los tamaños de los grupos.

```
rm(list=ls())
Y<-c(2,2,2,3,3,4,4,4,5,6,6,6,6,6,7,7,7,7,8,8,9,9,10,10,10,11,11,11,
    11,11,12,13,14,14,15,15,15,15,15,18,19,19,20,20,21,22,22,22,24,
    25,25,26,27,29,29,30,30,32,33,35,38,39,42,42,43,44,45,46,47,47,
    48,49,52,52,57,57,61,63,63,66,68,69,69,76,80,95)
n<-length(Y)
p<-0.77
q<-0.23
te<-0.53
ga<--0.3566194
sig<-35.51212
Fu<-numeric(n)
ec<-numeric(n)
ENM<-numeric(n)
for(r in 1:n)
   {
    Fu[r]<-1-(1+(ga/sig)*Y[r])^(-1/ga)
    ec[r]<-(Fu[r])^te
    ENM[r]<-p*(ec[r])/(te*(1-q*ec[r]))
   }
write.table(ENM, file="ENM.xls",sep=";")
**************************************************************************
**************************************************************************
```

Programa 4

Código para calcular

$$Z_{j,k} = \left[1 - F_{n_k}^{1/k\hat{\theta}}(Y_{j,k})\right]^{-\gamma},$$

cuando $k = 1$.

```
rm(list=ls())
te<-0.53
ga<--0.3566194
Y<-c(2,2,2,3,4,4,4,5,6,6,6,6,7,7,7,7,8,9,9,10,10,
    11,11,12,13,15,15,19,20,21,22,29,30,39,43,46)
n<-length(Y)
k<-1
```

#—— Distribución empírica.

```
Fnk<-numeric(n)
for(i in 1:n)
   {
    Fnk[i]<-(i)/36
   }
```

#—— Cálculo de Z1.

```
Z<-numeric(n)
FnkY<-numeric(n)
for(j in 1:n)
   {
    FnkY[j]<-(Fnk[j])^(1/(k*te))
    Z[j]<-(1-FnkY[j])^(-ga)
   }
write.table(Z,file="Zj1.xls",sep=";")
*****************************************************************************
*****************************************************************************
```

Programa 5

Y de manera similar cuando $k = 2$.

```
rm(list=ls())
te<-0.5300666
ga<--0.3566194
Y<-c(3,6,8,11,11,11,14,14,15,15,18,19,22,22,
     25,25,26,27,29,32,38,42,44,45,47,68,69)
n<-length(Y)
k<-2
```

#—— Distribución empírica.

```
Fnk<-numeric(n)
for(i in 1:n)
   {
    Fnk[i]<-i/27
   }
```

#—— Cálculo de Z2.

```
Z<-numeric(n)
FnkY<-numeric(n)
for(j in 1:n)
   {
    FnkY[j]<-(Fnk[j])^(1/(k*te))
    Z[j]<-(1-FnkY[j])^(-ga)
   }
write.table(Z,file="Zj2.xls",sep=";")
**************************************************************************
**************************************************************************
```

Programa 6

```
#—- Los datos a considerar son los máximos de los bloques


rm(list=ls())
library(RODBC)
library(extRemes)


#—- El archivo de excel se llama "datos "


data=odbcConnectExcel(file.choose())


#—- LD1279 LD1276 LD1209 LD1190 CEN1080 CEN1264


sqlTables(data)
mydat=sqlFetch(data,"CEN1080")
odbcClose(data)
v<-mydat$ld
promedio<-numeric(14)
u<-92
j<-1
while(u<=105)
     {
      z<-0
      Y<-0
      z <- v > u
      rg<- decluster.intervals( z, exi.intervals(z))$r
      dx<-dclust(v, u, rg, cluster.by = NULL)$xdat.dc
      dxx<-numeric(1080)
```

```
      for(h in 1:1080)
         {
          dxx[h]<-ifelse(dx[h]<u,NA,dx[h])
         }
      Y<-dxx[!is.na(dxx)]
      Y<-Y-u
      promedio[j]<-sum(Y)/length(Y)
      j<-j+1
      u<-u+1
     }
umbral<-seq(92,105, by=1)
plot(umbral,promedio)
**************************************************************************
**************************************************************************
```

Programa 7

Código para calcular

$$\tilde{\gamma} = \frac{-1}{g(\theta, p)} \left(\frac{1}{n-1} \right) \sum_{i=1}^{n-1} \log \left(1 - \frac{Y_{(i)}}{Y_{(n)}} \right),$$

y

$$\tilde{\sigma} = -\tilde{\gamma} Y_{(n)}.$$

#—— Estimación de los parámetros de gamma y sigma.

```
rm(list=ls())
n<-1000
p<-0.77
q<-0.23
te<-0.53
a<-p^(p/(te*q))
U<-runif(n,min=a,max=1)
V<-numeric(n)
for(i in 1:n)
   {
    b<-0
    b<-p*(U[i])^(-(te*q)/p)
    k<-1/te
    V[i]<-log(1-((1-b)/q)^k)
   }
```

#—— Cálculo de E{V}.

```
EV<-mean(V)
```

#—— Excesos máximos de cada uno de los 86 grupos.

```
Y<-c(2,2,2,3,3,4,4,4,5,6,6,6,6,6,7,7,7,7,8,8,9,9,10,10,10,11,11,
    11,11,11,12,13,14,14,15,15,15,15,15,18,19,19,20,20,21,22,22,
    22,24,25,25,26,27,29,29,30,30,32,33,35,38,39,42,42,43,44,45,
    46,47,47,48,49,52,52,57,57,61,63,63,66,68,69,69,76,80)
Yn<-95
m<-length(Y)
lgm<-numeric(m)
for(j in 1:m)
   {
    lgm[j]<-log(1-(Y[j]/Yn))
   }
su<-sum(lgm)
```

#—— Cálculo de gamma y sigma.

```
gamma<-(-1/EV)*(1/m)*su
sigma<-(-gamma)*Yn
**************************************************************************
**************************************************************************
```

Programa 8

Código para calcular

$$Z_i = \left(1 - F_n(Y_{(i)})\right)^{-\hat{\gamma}}.$$

```
rm(list=ls())
Y<-c(2,2,2,3,3,4,4,4,5,6,6,6,6,6,7,7,7,7,8,8,9,9,10,10,10,11,11,11,
     11,11,12,13,14,14,15,15,15,15,15,18,19,19,20,20,21,22,22,22,24,
     25,25,26,27,29,29,30,30,32,33,35,38,39,42,42,43,44,45,46,47,47,
     48,49,52,52,57,57,61,63,63,66,68,69,69,76,80,95)
n<-length(Y)
ga<--0.3868249
```

```
sig<-36.74837
G<-numeric(n)
FnY<-numeric(n)
for(i in 1:n)
   {
    G[i]<-(1-(i/n))^(-ga)
   }
cor(G,Y)
write.table(G,file="Ge.xls",sep=";")
*****************************************************************************
*****************************************************************************
```