COLEGIO DE POSTGRADUADOS

INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN
EN CIENCIAS AGRÍCOLAS

CAMPUS MONTECILLO

POSTGRADO DE SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA
ESTADÍSTICA

Modelación de Eventos Extremos Usando la Distribución Dagum

Benjamin Sexto Monroy

T E S I S

PRESENTADA COMO REQUISITO PARCIAL PARA
OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS

MONTECILLO,TEXCOCO, EDO. DE MÉXICO
2010

La presente tesis titulada: **Modelación de Eventos Extremos Usando la Distribución Dagum**, realizada por el alumno: **Benjamin Sexto Monroy**, 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. Humberto Vaquera Huerta

ASESOR _____
Dr. Javier Suárez Espinosa

ASESOR _____
Dr. Paulino Pérez Rodríguez

ASESOR _____
Dr. Barry C. Arnold

ASESOR _____
Dra. Elizabeth González Estrada

Montecillo, Texcoco, Estado de México, Octubre de 2010

Modelación de Eventos Extremos Usando la Distribución Dagum

Benjamin Sexto Monroy

Colegio de Postgraduados, 2010

En el presente trabajo se implementa la modelación de valores extremos con la metodología "block maxima", específicamente en niveles máximos diarios de ozono de la estación Pedregal de la Cd. de México de los años 2001 a 2008, usando la distribución Dagum que es de cola pesada y es usada generalmente como distribución de ingreso, pero tiene un antecedente en el campo meteorológico, donde fue usado para modelar la cantidad de precipitación pluvial. Una visualización general de los datos, hecha gráficamente, revela una tendencia a la baja de los niveles máximos diarios de ozono, por lo cual se decide realizar el análisis por año. Mediante la prueba de Kolmogorov-Smirnov se prueba si las observaciones de cada año siguen la distribución Dagum. El ajuste de la distribución Dagum se compara gráficamente y mediante un criterio de elección de modelos (Akaike) con el ajuste realizado por la distribución GEV que es la que se usa de manera estándar en valores extremos. Con el cálculo de cuantiles en un punto crítico para cada año y los cambios significativos observados gráficamente y comprobados con la técnica que denominamos modelo lineal generalizado vectorial, en los parámetros a de forma y b de escala, se comprueba la tendencia de la información con el paso del tiempo. En cuanto a la propuesta de prueba de bondad de ajuste, no se pudo construir una prueba general debido a que el estadístico de prueba mostró no ser invariante a los cambios en el parámetro p del cual depende, pero si se puede usar como una prueba computacional, una prueba bootstrap.

Palabras clave: Tendencia, Ozono, Valor Extremo, VGAM.

Modeling Extreme Events Using the Dagum Distribution

Benjamin Sexto Monroy

Colegio de Postgraduados, 2010

In this paper we implement the modeling of extreme values using the block maxima methodology in daily maximum ozone levels from the Pedregal station in Mexico City for the years 2001 to 2008. We used the Dagum distribution which has heavy tail and is generally used as income distribution but has a background in the field of meteorology. Also, it has been used to model the amount of rainfall. A graphical view of the data reveals a downward trend of daily maximum ozone levels so we decided to perform the analysis by year. We apply the Kolmogorov-Smirnov test to prove if the observations per year have aDagum distribution. To compare the fit of the Dagum distribution we made use of graphics and a criterion of choice of models (Akaike) with the adjustment made by the GEV distribution that is the one used as standard in extreme values. With the calculation of quantiles at a critical point for each year and the significant changes observed graphically and tested with the vector generalized linear model technique over the parameters a and b of scale and shape, respectively, we proved the trend of the observations over time. About the goodness of fit test proposed, we cannot built a general test because the test statistic was not as invariant to changes in the parameter p which depends. Although it can be used as a computational test, a bootstrap test.

Key words: Trends, Ozone, Extreme Value, VGAM.

AGRADECIMIENTOS

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

Al Colegio de Postgraduados, en particular al Programa de Estadística, por haberme brindado la oportunidad de seguir mi formación académica.

A los integrantes de mi Consejo Particular:

Dr. Humberto Vaquera Huerta, por su excelente dirección, su apoyo incondicional hizo posible la culminación de este trabajo.

Dr. Javier Suárez Espinosa, por revisar detalladamente el trabajo y por sus comentarios.

Dr. Paulino Pérez Rodríguez, por su ayuda desinteresada en la realización de este trabajo.

Dr. Barry C. Arnold, por su revisión detallada del trabajo, sus sugerencias y comentarios.

Dra. Elizabeth González Estrada, por revisar el trabajo y por sus sugerencias.

A cada uno de mis profesores que contribuyeron en mi formación, por su paciencia y ayuda brindada a lo largo de este recorrido.

A mis compañeros de clases y al personal administrativo por su amabilidad y atenciones que siempre me han brindado.

Este proyecto ha sido financiado parcialmente por la LPI 15: Estadística, Modelado y Tecnologías de Información Aplicadas a la Agricultura y al Medio Rural, Colegio de Postgraduados, México.

DEDICATORIA

A **Dios:** por prestarme vida, por estar a mi lado siempre y por su inmenso amor.

A mis **seres amados:** mi MADRE Francisca Monroy Eleuterio y mi PADRE Federico Sexto García, por ser siempre mis pilares en la vida y fuente inagotable de apoyo, amor y comprensión. A mis HERMANOS, por todo el cariño que me dan cada vez que convivimos. A mi NOVIA Patricia Estrada Drouaillet, por ser la luz y la felicidad en mi vida.

Índice

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

En todos los aspectos de la vida se dan fenómenos o procesos los cuales se quisieran tener siempre bajo control ya que un desequilibrio en ellas trae consecuencias no deseadas en la salud, en la economía, en el medio ambiente, etc.. Se llama valores extremos a valores inusuales de algún fenómeno causados por eventos raros, es decir, eventos que en el papel tienen baja probabilidad de ocurrir.

La importancia de estudiar los valores extremos radica en que su ocurrencia es de alto impacto y tienen consecuencias significativas. Dentro de la estadística se tiene a la teoría del valor extremo, que es la rama concerniente a deducciones en la cola de la distribución, es decir, el interés se centra en los valores más altos o más bajos de las variables en estudio y no en la parte central de la distribución como en gran parte de los estudios estadísticos. Las estadísticas de extremos sin duda pueden ser útiles en aplicaciones relacionadas con distribuciones de colas ligeras o acotadas, pero, su necesidad es más apremiante para variables que poseen una distribución de cola pesada (Katz *et al.*, 2002).

Los valores extremos se dan en muchos fenómenos, por ello la teoría de valores extremos tiene importantes aplicaciones en muchos campos, como la ciencia del medio ambiente (el nivel del mar, la velocidad del viento, niveles de contaminantes como el ozono), en la oceanografía (corrientes marinas extremas), en climatología (velocidades extremas de huracanes), en finanzas (compañías de seguros en riesgo de quiebra ante grandes siniestros), en hidrología (niveles de ríos o presas), en ingeniería (construcción de edificios resistentes a sismos), la ciencia del deporte, etc..

Este estudio centra su atención en un fenómeno del medio ambiente, específicamente en los niveles máximos diarios de ozono urbano registrados en la estación Pedregal de la Cd. de México de los años 2001 a 2008. El análisis de esta información es de suma importancia pues niveles muy altos de ozono provocan efectos adversos a la salud de la población expuesta a ella, causa síntomas como irritación ocular, de nariz

y garganta, tos, debilidad, náuseas, dolor de cabeza, etc., los cuales pueden ser más severos en ciertos grupos denominados susceptibles, como niños, ancianos, enfermos y personas que gustan de hacer ejercicio al aire libre (Ponce de Leon *et al.*, 1996).

El análisis de valores extremos se hace generalmente con la distribución de valores extremos generalizada (GEV, por sus siglas en inglés) cuando se modelan las observaciones más grandes de muestras de observaciones grandes, es decir, máximos de bloque ("block maxima"), y con la distribución Pareto cuando se modelan observaciones que rebasan un límite alto definido, es decir, picos sobre el umbral (peaks-over-threshold).

En el presente trabajo se propone usar a la distribución Dagum para modelar valores extremos (niveles máximos diarios de ozono) bajo la metodología "block maxima" o máximos de bloque. La distribución Dagum tiene 2 parámetros de forma y una de escala, es de cola pesada y se usa como una distribución del ingreso, sin embargo, la familia Dagum se propuso para modelar la cantidad de precipitación en la literatura meteorológica (Mielke, 1973), donde se le llamó distribución Kappa (de tres parámetros), la parametrización usada en aquel trabajo es diferente a la que usaremos, de hecho, Mielke y Johnson (1974) se refiere a ella como una generalización de dos distribuciones, la Singh-Maddala y la distribución de Dagum, y la llama distribución Beta-k.

Además de implementar la distribución Dagum para modelar valores extremos, haremos un análisis análogo con la distribución GEV y compararemos los resultados de ambas distribuciones, observaremos el comportamiento de la distribución de los niveles máximos de ozono a través de los años y como último punto del trabajo, propondremos una prueba de bondad de ajuste para la distribución Dagum ya que en la literatura existente no se ha tratado este tópico.

En el desarrollo del trabajo se tratarán los puntos mencionados en los párrafos anteriores. En el capítulo 2 se tienen los objetivos del trabajo. En el capítulo 3 se da la revisión bibliográfica en la se describe a detalle el origen y las características más importantes de la distribución Dagum así como de la distribución GEV, se mencionan los efectos adversos que provocan los niveles altos de ozono a la salud humana, se describen las herramientas que usaremos para determinar si la información se ajusta a la distribución Dagum y también se describe una herramienta usada como criterio para escoger un modelo sobre otro. En el capítulo 4 modelamos los niveles máximos de ozono usando la distribución Dagum mediante la técnica "block maxima" y la comparamos con la modelación usando la distribución GEV, y aplicamos herramientas estadísticas para verificar si se da una tendencia de la distribución de los niveles máximos de ozono al paso del tiempo. En el capítulo 5 se presenta la propuesta de prueba de bondad de ajuste para la distribución Dagum. El capítulo 6 contiene las conclusiones del trabajo y por último, se tiene un apéndice en el que se dispone de la información usada en nuestro trabajo, los niveles máximos de ozono por bloque.

Capítulo 2

Objetivos

- Implementar el uso de la distribución Dagum para modelar valores extremos.
- Proponer una prueba de bondad de ajuste para la distribución Dagum.

Capítulo 3

Revisión de Literatura

Este capítulo será dedicado a la presentación de las herramientas utilizadas para el desarrollo del trabajo de investigación, por ejemplo, el ozono (O_3) como contaminante y sus efectos en la salud, la distribución Dagum así como sus aspectos mas relevantes, la distribución de valor extremo generalizada (GEV por sus siglas en ingles), entre otras cosas.

3.1. El ozono

El ozono (O_3) es un gas altamente reactivo de color azul pálido, constituido por tres átomos de oxígeno en su estructura molecular. En la estratosfera (entre 12 y 50 Km a partir del suelo), existe un delgado escudo de gas (entre los 19 y los 23 kilómetros sobre la superficie terrestre), la capa de ozono, la cual rodea a la Tierra y la protege de los peligrosos rayos del sol. El ozono se produce mediante el efecto de la luz solar sobre el oxígeno y es la única sustancia en la atmósfera que puede absorber la dañina radiación ultravioleta (UV-B) proveniente del sol. Este delgado escudo hace posible la vida en la tierra.

Mas cerca de la superficie terrestre, a nivel de la troposfera (de 0 a 12 Km a partir de la superficie terrestre), el ozono no tiene el mismo desempeño que en la estratosfera, incluso es dañino para la salud humana. En esta zona el ozono se produce por la reacción fotoquímica de óxidos de nitrógeno (NO_x) y compuestos orgánicos volátiles (COVs) derivados del uso de combustibles fósiles, los cuales se denominan precursores del ozono. La reacción fotoquímica se produce cuando los (NO_x) y los (COVs) reaccionan con la luz solar, lo que produce un átomo libre de oxígeno (O). Este átomo libre 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.

El ozono que se forma como producto de la contaminación por emisiones se diferencia del que se encuentra en la atmósfera superior de la Tierra donde forma una capa que nos protege de los rayos dañinos ultravioletas del sol.

El ozono se encuentra de manera natural en la troposfera por la intrusión del ozono estratosférico aunado a las reacciones fotoquímicas de precursores biogénicos (compuestos emitidos por fuentes naturales como los volcanes, incendios forestales, etc.) y geogénicos (aquellos producidos por procesos geoquímicos). La concentración natural del ozono varía con la altitud, a mayor altitud se registra una concentración mayor. La concentración típica del ozono en ambientes naturales sin la influencia de las emisiones antropogénicas (son todas la directamente emitidas por el hombre o que son resultado de la actividad humana) se encuentra entre 10-40 partes por billón (ppb).

Las fuentes antropogénicas más importantes de los precursores de ozono son las emisiones vehiculares, emisiones industriales, plantas de energía, refinerías y los solventes químicos. A pesar de que estos precursores se originan en áreas urbanas, también pueden ser arrastrados por los vientos a lo largo de varios kilómetros provocando incrementos en la concentración de ozono en regiones menos pobladas.

En la Ciudad de México, las concentraciones ambientales de ozono tienen una marcada variación horaria, ya que sus precursores se emiten por la mañana y reaccionan conforme se incrementa la radiación solar, paulatinamente se incrementan también las concentraciones de ozono y disminuyen al atenuarse la radiación solar.

En México la Norma Oficial Mexicana NOM-020-SSA1-1993 establece un límite máximo permisible para ozono de 0.11 partes por millón (ppm), o lo que es equivalente a 216 mg/m^3, en una hora, para no rebasarse ninguna vez al año, para la protección a la salud de la población susceptible. Esta norma oficial mexicana coincide con la "US Environmental Protection Agency" (EPA) la cual considera especialmente importante estudiar los niveles de ozono por encima de 0.11 ppm por unidad de volumen debido a que son considerados peligrosos para la salud humana.

3.1.1. Efectos adversos del ozono en la salud y el ambiente

Debido a los efectos en la salud humana y también sobre el ambiente, el ozono ha recibido mucha atención científica y de las autoridades ambientales de las grandes ciudades como la Ciudad de México. El ozono reacciona activamente con los materiales expuestos a la intemperie, produce oxidación de metales y envejecimiento prematuro de materiales. Por otra parte, el ozono causa severos daños al follaje de algunas variedades de plantas y en otras reduce significativamente su crecimiento.

El ozono es un oxidante poderoso y puede reaccionar con varios componentes celulares y materiales biológicos, afecta a los tejidos vivos, se asocia con diversos padecimientos en la salud humana. Los individuos que viven en zonas donde se registran regularmente concentraciones altas de ozono, presentan diversos síntomas, por ejemplo: irritación ocular, de nariz y garganta, tos, dificultad y dolor durante la respiración profunda, dolor subesternal, opresión en el pecho, malestar general, debilidad, náusea y dolor de cabeza. Estos síntomas pueden ser más severos en ciertos grupos denominados susceptibles, como niños, ancianos, enfermos y personas que gustan de hacer ejercicio al aire libre (Ponce de Leon *et al.*, 1996).

Cuando el tracto respiratorio es expuesto al ozono se produce daño en el mismo, el alcance dependerá de la concentración de ozono, la duración de la exposición, los patrones de exposición y la ventilación. Este contaminante se asocia a síntomas respiratorios, especialmente tos. En asmáticos expuestos diariamente, se ha reportado un incremento en la incidencia de ataques asmáticos y síntomas respiratorios.

El ozono reduce la función pulmonar y hace más difícil la respiración profunda y vigorosa. También puede empeorar las enfermedades pulmonares crónicas tales como el enfisema y la bronquitis y reducir la capacidad del sistema inmunológico para defenderse de las infecciones bacterianas. El ozono puede causar daño permanente al pulmón. El daño en el corto plazo por causa del ozono en los pulmones de niños en desarrollo, puede resultar en una función pulmonar reducida en la edad adulta.

Debido a los daños a la salud que puede provocar el ozono, la Secretaria del Medio Ambiente de la Ciudad de México sugiere las siguientes medidas de protección; evitar exponerse al aire libre cuando la condición de calidad del aire sea “no satisfactoria”, realizar ejercicio y otras actividades al aire libre durante el período de horas que este contaminante tiene concentraciones bajas, antes del medio día y por la noche, permanecer en ambientes intramuros como la casa, oficina o escuela, con ventanas cerradas, durante el período de horas en que se incrementan las concentraciones de ozono, generalmente entre las 12:00 y las 16:00 horas, ingerir abundante agua y alimentos que contienen antioxidantes (frutas y verduras) y consultar el Índice Metropolitano de la Calidad del Aire (IMECA) de la zona donde se vive antes de realizar cualquier actividad física al aire libre.

3.2. Función de distribución empírica

Sean $X_1, ..., X_n$ una muestra aleatoria de tamaño n y sean $X_{(1)}, ..., X_{(n)}$ las mismas observaciones ordenadas de manera creciente. La función de distribución empírica se define como la frecuencia acumulada porcentual de los datos ordenados en forma ascendente y representa la probabilidad de ocurrencia de un valor menor o igual al dato considerado.

La función de distribución empírica es la función de $\mathbb{R}$ en [0,1], que denotamos por $F_n(x)$ y que toma los valores:

$$F_n(x) = \begin{cases} 0 & \text{si} \quad x < x_{(1)} \\ i/n & \text{si} \quad x_{(i)} \leq x < x_{(i+1)} \\ 1 & \text{si} \quad x > x_{(n)} \end{cases} \tag{3.1}$$

i es el numero de orden de la observación x. Entonces $F_n(x)$ es la proporción de los elementos de la muestra que son menores o iguales a x.

3.3. Distribución beta generalizada (GB)

Las distribuciones beta (existen dos tipos de esta distribución) son miembros del sistema de Pearson y han sido ampliamente utilizados en todas las ramas de las ciencias (Kleiber y Kotz, 2003). Están intrínsecamente relacionados con la función de distribución acumulativa de una variable aleatoria Beta:

$$I_x(p,q) = \frac{1}{B(p,q)} \int_0^x u^{p-1} (1-u)^{q-1} \, du, \;\; 0 \leq x \leq 1 \tag{3.2}$$

y la función beta incompleta

$$B_x(p,q) = \int_0^x u^{p-1} (1-u)^{q-1} \, du, \;\; 0 \leq x \leq 1 \tag{3.3}$$

Un recuento histórico de estas funciones fue proporcionada por Dutka (1981). La distribución beta generalizada (GB) es considerado como un modelo para la distribución del ingreso, dentro de sus casos especiales incluyen la distribución beta generalizada de primer tipo (GB1), la distribución beta generalizada de segundo tipo (GB2), la distribución Dagum (D), la distribución Singh-Maddala (SM), entre otros.

La función de densidad de la distribución beta generalizada (GB) es la siguiente:

$$GB(x;a,b,c,p,q) = \frac{a x^{ap-1} \left[1-(1-c)\,(x/b)^a\right]^{q-1}}{b^{ap} B(p,q) \left[1+c\,(x/b)^a\right]^{p+q}}, \;\; para \;\; 0 < x^a < \frac{b^a}{1-c} \tag{3.4}$$

cero de otra forma, donde $0 \leq c \leq 1$; $b,p,q > 0$; $B(p,q)$ es la función beta.

La distribución GB contiene las distribuciones beta generalizada de primer tipo y

beta generalizada de segundo tipo de cuatro parámetros y corresponden a los casos en que el parámetro $c = 0$ y $c = 1$, respectivamente.

3.3.1. Distribución beta generalizada de primer tipo (GB1)

La distribución beta generalizada de primer tipo (GB1) tiene la función de densidad siguiente:

$$GB1(x; a, b, p, q) = \frac{ax^{ap-1}\left[1-(x/b)^a\right]^{q-1}}{b^{ap}B(p,q)} \tag{3.5}$$

como mecionamos anteriormente; $GB1(x; a, b, p, q) = GB(x; a, b, c = 0, p, q)$

De la distribución beta generalizada se obtienen distribuciones como la distribución beta de primer tipo (B1) y la distribución Pareto. La distribución beta de primer tipo fue utilizada por Thurow (1970) para analizar los factores que contribuyen a la desigualdad de ingresos entre blancos y los negros, la función de densidad de dicha distribución es la siguiente:

$$B1(x; b, p, q) = \frac{x^{p-1}(b-x)^{q-1}}{b^p B(p,q)}, \quad 0 < x < b \tag{3.6}$$

en este caso; $B1(x; b, p, q) = GB1(x; a = 1, b, p, q)$.

La distribución Pareto fue de las primeras utilizadas para modelar ingreso:

$$Pareto(x; b, p) = \frac{px^{-p-1}}{b^{-p}}, \quad para \;\; b < y \tag{3.7}$$

en este caso; $Pareto(x; b, p) = GB1(x; a = -1, b, p, q = 1)$

3.3.2. Distribución beta generalizada de segundo tipo (GB2)

Para nuestros propósitos, la distribución fundamental en esta familia es la llamada distribución beta generalizada de segundo tipo. Debemos recordar la contribución de McDonald (1984) y sus asociados en el desarrollo de distribuciones GB2 como una distribución del ingreso y en la unificación de investigación en campos estrechamente relacionados.

Las interrelaciones entre los casos particulares de las distribuciones GB2 y otras distribuciones conocidas en la literatura son algo confusas, pero se ha dado una investigación coordinada independiente en los últimos 20-30 años (Kleiber y Kotz, 2003). Un indicio es el descubrimiento de que la distribución propuesta por Majumder y

3.3. Distribución beta generalizada (GB)

Chakravarty (1990) es simplemente una reparametrización de la distribución GB2. Esta observación escapó a investigadores durante al menos cinco años, a pesar del hecho de que los documentos no aparecieron en revistas independientes: uno orientado hacia las aplicaciones y la otra de tendencia más teórica.

La función de distribución acumulada (c.d.f., por sus siglas en ingles) de la distribución GB2 puede introducirse utilizando una expresión alternativa para la razón de la función beta incompleta que se obtiene haciendo $u := t/(1+t)$ en (3.2)

$$I_z(p,q) = \frac{1}{B(p,q)} \int_0^z \frac{t^{p-1}}{(1+t)^{p+q}} dt, \quad z > 0 \tag{3.8}$$

Introduciendo los parámetros a y b de escala y de forma respectivamente, y haciendo $z := (x/b)^a$ en (3.8), obtenemos una distribución con función de distribución acumulada:

$$F(x) = I_z(p,q), \quad donde \quad z = \left(\frac{x}{b}\right)^a, \quad x > 0 \tag{3.9}$$

con correspondiente densidad:

$$f(x) = \frac{a x^{ap-1}}{b^{ap} B(p,q) \left[1 + (x/b)^a\right]^{p+q}}, \quad x > 0 \tag{3.10}$$

donde b es es el parámetro de escala y a, p, q son parámetros de forma, todos los parámetros son positivos. La densidad GB2 es regularmente variante al infinito con índice $-aq-1$ y regularmente variante al origen con índice $-ap-1$, por lo que los tres parámetros de forma controlan el comportamiento de la cola del modelo. Sin embargo, estos tres parámetros no están en igualdad de condiciones, si consideramos la distribución $Y = logX$, con densidad

$$f(y) = \frac{a e^{ap(y-\log b)}}{B(p,q) \left[1 + e^{a(y-\log b)}\right]^{p+q}}, \quad -\infty < y < \infty \tag{3.11}$$

entonces, a resulta ser un parámetro de escala, mientras que p y q son todavía parámetros de forma. Vemos que cuanto mayor sea el valor de a, más delgada es la cola de la densidad (3.10),mientras que los valores relativos de p y q son importantes para determinar la asimetría de la distribución de (3.11).

La ecuación (3.10) fue propuesta como distribución del ingreso por McDonald (1984) y de forma independiente como un modelo para la distribución del monto de pérdidas en la ciencia actuaria por Venter (1983), quién la llamó distribución Beta transformada. Una década antes, fue discutido brevemente por Mielke y Johnson (1974) en una aplicación meteorológica como una generalización de dos distribuciones, la Singh-Maddala y la distribución de Dagum. También se puede considerar una distribución F generalizada, y aparece bajo este nombre, por ejemplo, Kalbfleisch y Prentice (1980). La distribución fue referida además, como distribución Feller-Pareto por Arnold (1983), quién presenta un parámetro adicional sobre localidad, y fue re-

descubierta, en una parametrización diferente, en la literatura de la econometría por Majumder y Chakravarty (1990). McDonald y Mantrala (1995) observaron que el modelo de Majumder Chakravarty es equivalente a la distribución GB2. Vamos a usar la notación de McDonald (1984).

A partir de la distribución beta generalizada de segundo tipo se pueden obtener otras distribuciones, por ejemplo, en el caso de que $a = 1$ se obtiene la distribución beta de segundo tipo (B2) cuya función de densidad es miembro de las distribuciones del sistema de Pearson, es nombrada distribución de Pearson VI.

Con los parámetros $a = b = p = q = 1$ en (3.11) la distribución GB2 puede ser considerada como una distribución Log-logística generalizada:

$$f(y) = \frac{e^y}{(1+e^y)^2}, \quad -\infty < y < \infty, \tag{3.12}$$

la cual es la densidad de una distribución logística estándar (Johnson *et al.*, 1995). Se ha referido a la densidad en (3.11) como la densidad de una distribución GB2 exponencial en McDonald y Xu (1995).

Las distribuciones de ingresos GB2 y B2 fueron derivadas también a partir de principios microeconómicos (un modelo neoclásico de optimizar el comportamiento de la empresa) por Parker (1999a,b). En su modelo los parámetros de forma p y q son funciones de salida de la elasticidad mano de obra y la elasticidad de las declaraciones de ingresos con respecto al capital humano, lo que permite hacerse una idea de las posibles causas de la desigualdad de las tendencias observadas.

El modelo GB2 es útil para la unificación de una parte sustancial del tamaño de distribuciones de la literatura. Contiene un gran número de distribuciones de ingreso y pérdida, como casos especiales están: la distribución Singh-Madala obtenida para $p = 1$, la distribución Dagum para $q = 1$, la distribución beta de segundo tipo (B2) para $a = 1$, la distribución Fisk (o Log-logística) para $p = q = 1$, y la distribución de Lomax (o Pareto tipo II) para $a = p = 1$. La Figura 3.1 muestra las interrelaciones de la distribución beta generalizada tipo 2.

Se incluyó la distribución de Lomax inversa en la Figura 3.1 para completar el cuadro, aunque apenas aparece algún trabajo que trata explícitamente esta distribución. Además, la distribución gamma generalizada emerge como un caso especial haciendo $b = q^{1/a}\beta$ y que $q \to \infty$. En consecuencia, las distribuciones gamma y Weibull son también casos de la GB2, ya que ambos son casos especiales de la distribución gamma generalizada (Kleiber y Kotz, 2003).

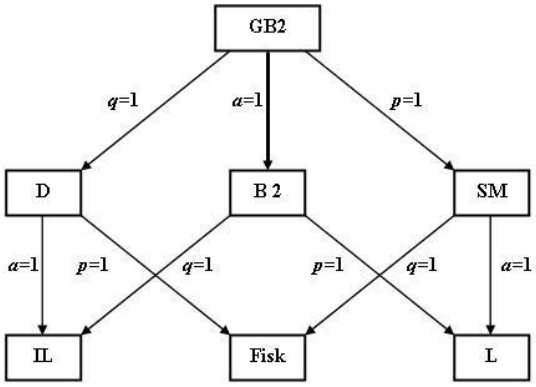

Figura 3.1: Distribución Beta Generalizada tipo II y sus interrelaciones: distribución beta generalizada de segundo tipo (GB2), distribución Dagum (D), distribución beta de segundo tipo (B2), distribución Singh-Madala (SM), Distribución Lomax inversa (IL), Distribución Fisk (Log-logística) (Fisk), distribución Lomax (L).

3.4. Distribución Dagum (D)

Camilo Dagum en 1970 no satisfecho con las distribuciones clásicas utilizadas para trabajar con datos de ingreso tales como la distribución de Pareto (Pareto, 1895)(desarrollada por el economista y sociólogo italiano Vilfredo Pareto en el siglo XIX), y la distribución log-normal (Gibrat, 1931) (popularizada por el ingeniero francés Robert Gibrat), se dedicó la búsqueda de una distribución estadística apropiada a ingresos empíricos y la distribución de la riqueza, el comenzó a buscar un modelo para tratar con colas pesadas presentes en la distribución de los ingresos empíricos y la riqueza, así como permitir una moda interior. El primer aspecto está bien capturado por la distribución Pareto, pero no por la distribución log-normal, este último por la distribución log-normal, pero no por la distribución Pareto (Kleiber, 2008). Dagum (1975) experimentando con un cambio en la distribución log-logística, una generalización de una distribución previamente considerado por Fisk (1961a,b), rápidamente se dio cuenta de que era necesario un parámetro adicional. Esto condujo a la distribución Dagum tipo I, una distribución de tres parámetros, y dos generalizaciones de cuatro parámetros (Dagum, 1977, 1980c).

Aunque se presenta como una distribución del ingreso sólo un año después del modelo Singh-Maddala (Singh y Maddala, 1976), la distribución de Dagum es menos conocida. Presumiblemente, esto se debe al hecho de que el trabajo Dagum se publicó en la revista francesa "Economie Appliquée", mientras que el documento de Singh-Maddala apareció en el más leído, "Econometrica". Sin embargo, en los últimos años hay indicios de que la distribución Dagum es, de hecho, una elección más apropiada en muchas aplicaciones, poco a poco se supo que la distribución de Dagum es a menudo preferible a la distribución Singh-Maddala en las aplicaciones de datos sobre ingresos.

Pasó más de una década hasta que la propuesta de Dagum comenzara a aparecer en la literatura econométrica y económica en estudios en idioma Inglés. El primer artículo en una revista importante de econometría que haga referencia a la distribución Dagum es de los autores Majumder y Chakravarty (1990). En la literatura estadística, la situación es más favorable, en la renombrada "Encyclopedia of Statistical Sciences" de Kotz *et al.* (1983) contiene, en el volumen 4, una entrada en los modelos de distribución del ingreso, como era de esperarse, el autor, Camilo Dagum (Dagum, 1983).

Dagum (1977) derivó su modelo de la observación empírica de que la elasticidad del ingreso $\eta(F,x)$ de la función de distribución acumulativa (c.d.f.) del ingreso es una función decreciente y acotada de F. A partir de la ecuación diferencial,

$$\eta(F,x) = \frac{d\log F(x)}{d\log x} = ap\{1-[F(x)]^{1/p}\}, \quad x>0 \tag{3.13}$$

sujeto a $p>0$ y $ap>0$ y se obtiene la distribución (3.17).

La forma más general de la distribución Dagum tiene la función de distribución acumulativa siguiente:

$$F(x) = \alpha + (1-\alpha)[1+(x/b)^{-a}]^{-p} \tag{3.14}$$

Las distribuciones Dagum Tipo I, II y III corresponden a los casos en que $\alpha = 0$, $0<\alpha<1$ y $\alpha<0$ respectivamente. La distribución Dagum tipo II se propuso como un modelo para la distribución del ingreso considerando ingresos nulos o negativos, pero sobre todo para ajustar los datos de riqueza, que a menudo presenta un gran número de unidades económicas con activos brutos nulos y activos netos nulos y negativos. La distribución Dagum Tipo III se asocia con un límite inferior positivo para X, x_0. En este documento trabajaremos la distribución Dagum tipo I y solo nos referiremos a ella como distribución Dagum.

La distribución Dagum es un caso particular de la distribución GB2 con el parámetro de forma $q=1$, por lo que su densidad es la siguiente:

$$f(x) = \frac{apx^{ap-1}}{b^{ap}\left[1+\left(\frac{x}{b}\right)^{a}\right]^{p+1}}, \quad x>0 \tag{3.15}$$

donde $a,b,p>0$.

Al igual que otra distribución propuesta para modelar ingresos, la distribución Singh-Maddala (SM), la distribución de Dagum fue redescubierta varias veces en diversos campos de la ciencia. Al parecer, esto ocurrió por primera vez con Burr (1942), como el tercer ejemplo de soluciones a la ecuación diferencial que define el sistema de distribución de Burr. Así, se le conoce también como la distribución Burr III. La distribución Dagum está estrechamente relacionada con la distribución Singh-Maddala,

específicamente:

$$X \sim D(a,b,p) \Longleftrightarrow \frac{1}{X} \sim SM\left(a, \frac{1}{b}, p\right) \tag{3.16}$$

Esta relación nos permite traducir varios resultados pertenecientes a la familia Singh-Maddala a los resultados correspondientes a las distribuciones de Dagum y viceversa. Singh-Maddala es conocida también como distribución Burr XII, o simplemente la distribución de Burr, por lo que no es de extrañar que la distribución Dagum se llama también la distribución Burr inversa, en particular, en la literatura actuaria (Klugman *et al.*, 1998). Al igual que la Singh-Maddala, la distribución Dagum se puede considerar una distribución log-logistica generalizada.

Antes de su uso como una distribución del ingreso, Mielke (1973) propuso la familia Dagum como un modelo para la cantidad de precipitación en la literatura meteorológica , donde se le llama distribución Kappa (de tres parámetros) la cual es equivalente a la distribución Dagum en una parametrización diferente. Mielke y Johnson (1974) la llamaron distribución Beta-k. En un desarrollo paralelo al tanto de Mielke (1973) pero probablemente consciente de Dagum (1977), Fattorini y Lemmi (1979) propusieron la distribución como una distribución del ingreso usando los parámetros$(\alpha, \beta, \theta) = (1/p, bp^{1/a}, ap)$. No obstante, la distribución se suele llamar la distribución Dagum en la literatura de la distribución del ingreso, y vamos a seguir este convenio a continuación.

3.4.1. Propiedades básicas

La función de distribución Dagum tiene forma cerrada:

$$F(x) = \left[1 + \left(\frac{x}{b}\right)^{-a}\right]^{-p}, \quad x > 0 \tag{3.17}$$

Recordemos que a, b, $p > 0$, b es un parámetro de escala y a y p son parámetros de forma, sin embargo, estos dos parámetros no están en igualdad de condiciones, por ejemplo, consideremos que X tiene distribución Dagum y sea $Y = \log X$ una distribución logística generalizada con densidad:

$$f(x) = \frac{ape^{ap(y-\log b)}}{\left[1 + e^{a(y-\log b)}\right]^{p+1}}, \quad -\infty < y < \infty \tag{3.18}$$

En este caso, solo p es el parámetro de forma o asimetría mientras que a pasa a ser un parámetro de escala y $\log b$ es un parámetro de localización. En la distribución Dagum se conocen los efectos que pueden causar las variaciones de los parámetros

de forma: cuando $ap < 1$, la densidad presenta un polo en el origen, cuando $ap = 1$, $0 < f(0) < \infty$, y cuando $ap > 1$ se tiene la moda, dicha moda se obtiene mediante la siguiente expresión:

$$x_{moda} = b\left(\frac{ap-1}{a+1}\right)^{1/a} \tag{3.19}$$

Ésta es una característica atractiva en que el modelo puede aproximar la distribución del ingreso, el cual es generalmente unimodal, y las distribuciones de riqueza, los cuales no tienen moda. Debemos destacar también que, a diferencia de varias distribuciones populares usadas para aproximar datos sobre ingresos, por ejemplo, la distribución lognormal, la distribución gamma y la distribución GB2, la distribución Dagum permite una expresión de forma cerrada para la función de distribución acumulativa.

La función cuantil también tiene forma cerrada:

$$F^{-1}(x) = b\left[u^{-1/p} - 1\right]^{-1/a}, \quad para \;\; 0 < u < 1 \tag{3.20}$$

Entonces se pueden generar con facilidad números aleatorios de una distribución Dagum utilizando el método de inversión.

El k-ésimo momento existe para $-ap < k < a$ y la expresión para obtenerlo es el siguiente:

$$E(X^k) = \frac{b^k B\left(p + k/a, 1 - k/a\right)}{B(p,1)} = \frac{b^k \Gamma\left(p + k/a\right)\Gamma\left(1 - k/a\right)}{\Gamma(p)} \tag{3.21}$$

donde $\Gamma()$ denota la función gamma y $B()$ la función beta.

Los casos particulares de la media y la varianza son:

$$E(X) = \frac{b\Gamma\left(p + 1/a\right)\Gamma\left(1 - 1/a\right)}{\Gamma(p)} \tag{3.22}$$

$$var(X) = \frac{b^2\left[\Gamma(p)\Gamma\left(p + 2/a\right)\Gamma\left(1 - 2/a\right) - \Gamma^2\left(p + 1/a\right)\Gamma^2\left(1 - 1/a\right)\right]}{\Gamma^2(p)} \tag{3.23}$$

La distribución Dagum tiene un número pequeño de momentos finitos (Dagum y Lemmi, 1989), generalmente tres o cuatro, en cuanto mayor sea la desigualdad de ingresos, menor será en número de momentos finitos. Al igual que la distribución

Dagum, la distribución de ingresos tiene un pequeño número de momentos finitos lo cual puede ser una gran ventaja para la distribución Dagum. Cuanto menor sea el número de momentos finitos, más gruesa o pesada es la cola derecha de la distribución Dagum. Dagum (1977) utilizó los parámetros $(\beta, \delta, \lambda) = (p, a, b^a)$.La parametrización que usamos aquí es la de McDonald (1984).

3.4.2. Estimación de parámetros

En lo referente a la estimación de los parámetros de la distribución Dagum, Dagum (1977) propuso un criterio de mínimos cuadrados no lineales basado en la distancia entre la función de distribución empírica dada en (3.1) y la función de distribución Dagum en (3.17),

$$\sum_{i=1}^{n} \left\{ F_n(x_i) - [1 + (x_i/b)^{-a}]^{-p} \right\}^2 \tag{3.24}$$

Mas tarde, Stoppa (1995) consideró un estimador tipo regresión utilizando la elasticidad en (3.13).

En la actualidad, generalmente se estiman los parámetros de la distribución Dagum por el método de máxima verosimilitud. En el caso en el que solo se dispone de datos agrupados, la función de verosimilitud $L(\boldsymbol{\theta})$ es multinomial (asumiendo independencia de datos):

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{m} \{F(x_j) - F(x_{j-1})\}, \;\; x_0 = 0, \;\; x_m = \infty \tag{3.25}$$

donde $\boldsymbol{\theta} = (a, b, p)^\top$. Por construcción, se dice que (3.25) es siempre acotada.

Hoy en día se tiene buena disponibilidad de observaciones individuales por lo cual la estimación por máxima verosimilitud para datos individuales es común en la actualidad. Sea $X_1, ..., X_n$ una muestra aleatoria de tamaño n de la distribución Dagum, el logaritmo de la función de verosimilitud se define como:

$$\ell = n \log a + n \log p + (ap - 1) \sum_{i=1}^{n} \log x_i - nap \log b - (p+1) \sum_{i=1}^{n} \log \left[1 + \left(\frac{x_i}{b} \right)^a \right] \tag{3.26}$$

Debemos obtener los estimadores $(\hat{a}, \hat{b}, \hat{p})$ que maximizan ℓ. Al derivar con respecto a cada parámetro e igualar a cero se obtienen las siguientes ecuaciones:

$$\frac{n}{a} + p\sum_{i=1}^{n}\log(x_i/b) = (p+1)\sum_{i=1}^{n}\frac{\log(x_i/b)}{1+(b/x_i)^a}, \tag{3.27}$$

$$np = (p+1)\sum_{i=1}^{n}\frac{1}{1+(b/x_i)^a}, \tag{3.28}$$

$$\frac{n}{p} + a\sum_{i=1}^{n}\log(x_i/b) = \sum_{i=1}^{n}\log\{1+(x_i/b)^a\} \tag{3.29}$$

las cuales deben resolverse numéricamente, afortunadamente esta tarea ya esta hecha en el programa estadístico R mediante el paquete "VGAM". En este trabajo, haremos uso de éste programa para estimar los parámetros de la distribución Dagum por el método de máxima verosimilitud.

3.5. Valores extremos

Se llama valores extremos a valores inusuales de algún fenómeno causados por eventos raros, es decir, eventos con baja probabilidad de ocurrir. La importancia de estudiar los valores extremos radica en que su ocurrencia es de alto impacto y tienen consecuencias significativas. Generalmente, a los valores extremos los identificamos como "outliers" en los análisis clásicos de información y en muchos de los casos son ignorados u omitidos.

Teniendo en cuenta que su aparición es, por definición, poco común, ha sido un reto para los estadísticos elaborar métodos adecuados para cuantificar la probabilidad y la intensidad de los fenómenos extremos. Por ejemplo, relacionado con el teorema central del límite para los promedios, una teoría estadística especializada ya está disponible para los extremos (Coles, 2001).

Es difícil dar con el origen preciso de la estadística de valores extremos, de los primeros indicios que se han encontrado refieren a Nicolas Bernoulli en 1709 cuando planteó el problema de la distancia media máxima desde el origen de "n" puntos distribuidos aleatoriamente en un línea recta de distancia fija t, Chaplin (1880) se planteó el problema del efecto del tamaño en la resistencia de materiales, es decir, un problema de mínimos. En la década de 1920, varios investigadores comenzaron a derivar al mismo tiempo la teoría estadística de los valores extremos, un temprano avance teórico fue producido por los estadísticos británicos R.A. Fisher y L. H. C. Tippett, quienes derivaron la forma límite de la distribución del valor máximo o mínimo de una muestra aleatoria (Fisher y Tippett, 1928).

En décadas posteriores, la teoría del valor extremo encontró aplicación en otras áreas en las que los eventos extremos, naturalmente, desempeñan un papel importante. En la década de 1930, se desarrollaron trabajos de gran interés al inicio de esta disciplina, Finetti (1932), Gumbel (1934, 1935a,b), von Mises (1936) y Rice (1939), trabajaron sobre la distribución de extremos (máximos y mínimos) de una muestra y culminaron cuando Gnedenko (1943) presentó en forma general el teorema de los tipos de distribución de extremos. El primer libro de importancia que trabajó con estadísticas de los extremos contempló una serie de aplicaciones, muchas relacionadas con el diseño de ingeniería, el libro se llamó "Statistics of extremes" de Gumbel (1958).

3.5.1. Teoría de valores extremos

Gran parte de los estudios estadísticos tienen como principal preocupación la parte central de la distribución, se ocupan de la modelación del promedio de la distribución de las variables en estudio, toman a la media muestral como estimador del promedio y el teorema del límite central proporciona un valioso resultado relacionado con el comportamiento asintótico de la media muestral. La teoría del valor extremo es la rama de la estadística concerniente a deducciones en la cola de la distribución, es decir, el interés se centra en los valores más altos o más bajos de las variables en estudio.

La teoría de valores extremos tiene importantes aplicaciones en muchos campos, como la ciencia del medio ambiente (el nivel del mar, la velocidad del viento, niveles de contaminantes como el ozono), en la oceanografía (corrientes marinas extremas), en climatología (velocidades extremas de huracanes), en finanzas (compañías de seguros en riesgo de quiebra ante grandes siniestros), en hidrología (niveles de ríos o presas), en ingeniería (construcción de edificios resistentes a sismos), la ciencia del deporte, etc.. Ahora hay una literatura en general sustancial en el área, podemos mencionar, por ejemplo, Beirlant (1996), Embrechts *et al.* (1997), Kotz y Nadarajah (2000), Coles (2001), Finkenstadt y Rootzén (2003), Smith (2003), Castillo *et al.* (2005), Reiss y Thomas (2007).

Aunque las estadísticas de extremos sin duda pueden ser útiles en aplicaciones relacionadas con colas ligeras o acotadas, su necesidad es más apremiante para variables que poseen una distribución de cola pesada (Katz *et al.*, 2002). Por ejemplo, los índices de propagación de la población han sido también descritos con distribuciones con colas pesadas (Clark *et al.*, 2001), la distribución de los daños económicos dados por disturbios como los huracanes, pueden ser también de colas pesadas (Katz, 2002b,a).

Por comodidad, los extremos se tratan sólo en términos de máximos de distribuciones (cola superior). Sin embargo, los mínimos (cola inferior) son considerados de manera

efectiva a través de la relación

$$min(X_1, X_2, ..., X_n) = -max(-X_1, -X_2, ..., -X_n)$$

para una muestra de observaciones $X_1, X_2, ..., X_n$.

3.6. Distribución de Valores Extremos Generalizada (GEV)

La distribución de Valores Extremos Generalizada (GEV, por sus siglas en inglés) surge del teorema del valor extremo (Fisher y Tippett, 1928) y Gnedenko (1943), como la distribución límite de máximos debidamente normalizados de una secuencia de variables aleatorias independientes e idénticamente distribuidas (i.i.d).

Supongamos que $X_1, X_2, ..., X_n$ son variables aleatorias i.i.d con función de distribución F y sea $M_n = max(X_1, X_2, ..., X_n)$, entonces:

$$\lim_{n \longrightarrow \infty} P(M_n \leq x) = \lim_{n \longrightarrow \infty} F^n(x) = 0$$

para cualquier x con $F(x) < 1$.

Ahora, si existen las constantes $a_n > 0$ y $b_n \in \mathbb{R}$, que después harán el papel de parámetros de escala y forma, y estandarizamos (en el sentido de la estandarización de las variables normales), entonces cuando n tiende a infinito;

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) = F^n(a_n x + b_n) \xrightarrow{d} G(x) \tag{3.30}$$

es decir,$(M_n - b_n)/a_n$ converge en distribución a G, donde G es una función de distribución límite no degenerada, entonces G puede ser una de las tres familias de funciones de distribución de valor extremo:

$$Gumbel : \Lambda(x) = \exp\left(-e^{-x}\right), \quad x \in \mathbb{R} \tag{3.31}$$

$$Fréchet : \Phi_\alpha(x) = \begin{cases} 0 & \text{si} \quad x \leq 0 \\ \exp\left(-x^{-\alpha}\right) & \text{si} \quad x > 0 \end{cases} \tag{3.32}$$

$$Weibull : \Psi_\alpha(x) = \begin{cases} \exp\left(-(-x)^\alpha\right) & \text{si} \quad x < 0 \\ \\ 1 & \text{si} \quad x \geq 0 \end{cases} \tag{3.33}$$

En las familias Fréchet y Weibull $\alpha > 0$. Las tres familias de funciones de distribución de valores extremos pueden ser combinadas en una distribución con parametrización común, dicho aporte fue hecho por von Mises (1936, 1954) y Jenkinson (1955), la construyen haciendo la reparametrización $\xi = 1/\alpha$ y la representación es conocida como función de distribución de valores extremos generalizada (GEV):

$$G(x; \mu, \sigma, \xi) = \exp\left\{ -\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}} \right\} \tag{3.34}$$

donde $\sigma > 0$, $-\infty < \mu < \infty$, $1 + \xi(x-\mu)/\sigma > 0$, $x_+ = max\{x, 0\}$.

La función de densidad es en consecuencia:

$$g(x; \mu, \sigma, \xi) = \frac{1}{\sigma}\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1} \exp\left\{ -\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} \right\} \tag{3.35}$$

para $1 + \xi(x-\mu)/\sigma > 0$.

Se tiene que μ es el parámetro de localidad, σ es el parámetro de escala y ξ es un parámetro de forma. Los casos en los que $\xi > 0$, $\xi < 0$ y $\xi = 0$, corresponden a los tipos de función de distribución Fréchet, Weibull y Gumbel respectivamente. De esta manera, dependiendo del valor de ξ, la forma de la distribución GEV asume tres posibles tipos:

1) Si $\xi = 0$ se tiene una distribución de cola ligera (Gumbel)

2) Si $\xi > 0$ se tiene una distribución de cola pesada (Fréchet)

3) Si $\xi < 0$ se tiene una distribución acotada (Weibull)

La distribución tipo Gumbel tiene una cola superior no acotada que disminuye a un ritmo relativamente rápido (por ejemplo, normal, exponencial, gamma, log-normal). Aunque la distribución tipo Fréchet también tiene una cola superior no acotada, disminuye tan lento (por ejemplo, Pareto, Cauchy, t-student) que sus momentos son infinitos para todas las ordenes mas grandes que $1/\xi$ (por ejemplo, la varianza es infinita si $\xi > 0.5$; la media es infinita si $\xi > 1$), esta distribución se dice que es de cola pesada (Katz *et al.*, 2005). La distribución tipo Weibull tiene un límite superior finito en $x = \mu - (\sigma/\xi)$.

Si una v.a. X tiene distribución GEV, entonces la variable estandarizada $(X-\mu)/\sigma$ tiene una distribución que no depende de μ ni de σ, sino únicamente de ξ. Al igual que la media y desviación estándar de la distribución normal que nos es más familiar, el parámetro de localización especifica donde la distribución esta "centrada", el parámetro de escala, su "propagación", el parámetro ξ está asociado al espesor de la cola de la distribución, en cuanto más grande sea el valor de ξ más gruesa es la cola de la distribución.

Realmente no es necesario suponer que las observaciones sean independientes, la distribución límite del máximo sigue siendo GEV bajo una amplia gama de condiciones de dependencia (por ejemplo, un proceso autoregresivo) y únicamente habría efectos sobre los parámetros de localidad y escala (Leadbetter *et al.*, 1983). La asunción de las observaciones idénticamente distribuidas también puede hacerse más flexible, con no estacionariedad llega a darse a través de covariables.

3.6.1. Función cuantil

Es natural que nos enfoquemos en los cuantiles extremos superiores de la distribución GEV. Específicamente, el "nivel de retorno" asociado con el "periodo de retorno" de $1/p$ es el $(1-p)$ésimo cuantil de la distribución GEV (por ejemplo, cuando modelamos máximos anuales, $p = 0.01$ correspondería a un periodo de retorno de 100 años) un concepto utilizado ampliamente en hidrología (Katz *et al.*, 2002).

Se puede obtener una expresión para los cuantiles invirtiendo la función de distribución GEV, es decir, $G^{-1}(1-p;\mu,\sigma,\xi)$, se tiene entonces;

$$x_p = \mu - \frac{\sigma}{\xi}\left[1 - \{-\log(1-p)\}^{-\xi}\right] \tag{3.36}$$

donde $G(x_p) = 1-p$ con $0 < p < 1$. En terminología de valores extremos, x_p es el nivel de retorno asociado con el periodo de retorno $1/p$. Para el cálculo de los cuantiles se emplean los estimadores de los parámetros.

3.6.2. Estimación de parámetros

Los parámetros de la distribución GEV pueden obtenerse por el método de máxima verosimilitud. El logaritmo de la función de verosimilitud es el siguiente:

$$\ell = -n\log\sigma - \left(\frac{1}{\xi}+1\right)\sum_{i=1}^{n}\log\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right] - \sum_{i=1}^{n}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} \quad (3.37)$$

En el método de máxima verosimilitud se establece que debemos escoger los estimadores $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ que maximicen ℓ, este cálculo puede hacerse numéricamente por computadora, en nuestro caso haremos uso del programa estadístico R mediante el paquete "evir" para lo obtención de dichos estimadores.

3.6.3. Máximos de bloques (Block maxima)

La distribución GEV es típicamente usada en la metodología llamada "block máxima"o máximos de bloque, que se aplica en muchas situaciones, por ejemplo, en la cantidad de precipitación máxima diaria durante todo un año, en niveles máximos diarios de algún contaminante como el ozono, etc., se pueden construir bloques por año, por meses o diarios dependiendo de la cantidad de información disponible. Este enfoque es a veces visto como ventajoso porque requiere sólo de un resumen simplificado de datos (Gaines y Denny, 1993).

La metodología se puede resumir en los siguientes pasos: supongamos que $X_1, X_2, ..., X_n$ son variables aleatorias i.i.d., construir los bloques de igual longitud y tomar los máximos de cada bloque, $M_n = max(X_1, X_2, ..., X_n)$, considerar los máximos de cada bloque como una muestra aleatoria, estimar los parámetros de la distribución GEV. Como complemento, se realiza la prueba de rachas de Wald-Wolfwitz y se calcula la función de autocorrelación para comprobar aleatoriedad e independencia de los máximos de bloque por año.

Una metodología alternativa para la construcción de los bloques es la presentada por Ferro *et al.* (2009), denominado "interval declustering", el cual construye grupos independientes de observaciones que han superado un nivel o umbral fijado (picos sobre el umbral), pero, en nuestro caso no trabajamos con niveles máximos que hayan superado un umbral.

3.7. Prueba de Kolmogorov-Smirnov

La prueba de Kolmogorov-Smirnov se emplea para probar el grado de concordancia entre la distribución de datos empíricos de una muestra aleatoria y alguna distribución teórica completamente especificada (F_0). La hipótesis a contrastar es la siguiente:

$\boldsymbol{H_0}$: los datos siguen una distribución F_0

$\boldsymbol{H_1}$: los datos no siguen una distribución F_0

El estadístico de prueba es:

$$D = \sup_{-\infty<x<\infty} |F_n(x) - F_0(x)|$$

donde;

- x_i es el i-ésimo valor observado en la muestra (cuyos valores han sido ordenados de menor a mayor)
- $F_n(x_i)$ es la función de distribución empírica (3.1)
- $F_0(x_i)$ es la función de distribución teórica (es la probabilidad de observar valores menores o iguales que x_i cuando H_0 es cierta)

De esta manera, D es la mayor diferencia absoluta entre la frecuencia acumulada observada $F_n(x)$ y la frecuencia acumulada teórica $F_0(x)$, obtenida a partir de la distribución de probabilidad que se especifica como hipótesis nula. Entonces, si los valores observados $F_n(x)$ son cercanos a los esperados en la distribución teórica $F_0(x)$, el valor del estadístico de prueba D será pequeño, y en cuanto mayor sea la diferencia entre $F_n(x)$ y $F_0(x)$ mayor será el valor del estadístico.

Una forma práctica de calcular es estadístico D consiste en obtener:

$$D^+ = \max_{1\leq i\leq n} \left\{ \frac{i}{n} - F_0(x_i) \right\} \tag{3.38}$$

$$D^- = \max_{1\leq i\leq n} \left\{ F_0(x_i) - \frac{i-1}{n} \right\} \tag{3.39}$$

y a partir de ellos elegimos D de la siguiente manera:

$$D = max \left\{ D^+, D^- \right\} \tag{3.40}$$

La regla de decisión en la siguiente: Rechazar H_0 si y solo si $D > D_\alpha$

Donde el valor de D_α se elige de tal manera que:

$P($ Rechazar $H_0|H_0$ es cierta $) = P(D > D_\alpha|$ los datos siguen la distribución $F_0(x)) = \alpha$

siendo α el nivel de significancia del contraste.

El valor D_α depende del tipo de distribución teórica a probar y se encuentra tabulado. En nuestro caso, usaremos es paquete estadístico R.

Podemos hacer uso también de un modo alternativo de realizar la prueba de Kolmogorov-Smirnov empleando el $p-value$ (nivel de significancia observado) asociado al estadístico D observado:

$$p - value = P(D > D_{obs}|H_0 \text{ es cierta })$$

Un $p - value$ grande quiere decir que, siendo cierta H_0, el valor del estadístico D era esperable y no hay razón para rechazar dicha hipótesis, en cambio, un $p - value$ pequeño indica que siendo cierta H_0, era muy difícil que se produjera el valor de D que se ha dado u observado por lo cual se debe rechazar la hipótesis nula. La regla de decisión para este contraste, con un nivel de significancia α, es la siguiente:

$$\text{Si } p - value \geq \alpha \Longrightarrow \text{ Aceptar } H_0$$

$$\text{Si } p - value < \alpha \Longrightarrow \text{ Rechazar } H_0$$

La obtención del $p - value$ requiere de conocer la distribución de D bajo la hipótesis nula y hacer el calculo correspondiente, en nuestro caso particular, el paquete estadístico R realiza este cálculo y proporciona el $p - value$.

3.8. Criterio de la información de Akaike (AIC)

El criterio de información de Akaike fue desarrollado por Hirotsugu Akaike bajo el nombre "An Information Citerion" (AIC) en 1971 y propuesta en Akaike (1974), es una medida de la calidad del ajuste del modelo estadístico estimado. El AIC no es una prueba en el modelo en el sentido de una prueba de hipótesis, es mas una herramienta para la selección del modelo. En el caso en el que en los modelos a comparar se utiliza el método de máxima verosimilitud para obtener los estimadores de los parámetros, el AIC se calcula mediante la siguiente expresión:

$$AIC(k) = 2k - 2log(L(\boldsymbol{\theta}_k)) \tag{3.41}$$

Donde, k es el número de parámetros del modelo obtenidos por el método de máxima

verosimilitud y $L(\boldsymbol{\theta}_k)$ es la función de verosimilitud del modelo. En general se dice que el estadístico de Akaike no tiene una interpretación tangible directa y es de utilidad cuando se comparan modelos. El modelo preferido sera aquel con menor AIC.

Capítulo 4

Modelación de niveles altos de ozono urbano usando la distribución Dagum

Este capitulo describe los recursos utilizados para llevar a cabo el trabajo de investigación (tipo de datos analizados, pruebas empleadas, software, etc.)

4.1. Registro de niveles de ozono en la Ciudad de México

El Sistema de Monitoreo Atmosférico de la Ciudad de México (SIMAT) es el sistema de la Secretaría del Medio Ambiente del Gobierno del Distrito Federal (SMA-GDF) empleado para la vigilancia y el monitoreo de la calidad del aire en la Zona Metropolitana del Valle de México.

El SIMAT realiza mediciones de ozono en partes por millón (ppm), esta unidad de medición es utilizada para conocer concentraciones diminutas de elementos presentes por unidad de volumen. En la actualidad el SIMAT está integrado por la Red Automática de Monitoreo Atmosférico (RAMA) que cuenta con 36 estaciones, este subsistema es el que realiza las mediciones de ozono; la Red Manual de Monitoreo Atmosférico (REDMA) que consta de 12 estaciones; la Red de Depósito Atmosférico (REDDA) que tiene 16 sitios de muestreo y la Red de Meteorología y Radiación Solar (REDMET) que opera con 16 estaciones.

Las técnicas para determinar la concentración de ozono son diversas. En la RAMA del SIMAT se realizan mediciones continuas y permanentes mediante equipo automático

que opera con base a las propiedades fisicoquímicas que identifican a cada contaminante. Los registros de concentraciones de ozono se obtienen cada minuto y se procesan como promedios horarios para su disposición al público en forma de "Bases de Datos". Con esta información se elabora y difunde oportunamente el IMECA para informar a la población sobre las condiciones de calidad del aire.

Para el desarrollo de este trabajo haremos uso de las mediciones del SIMAT de los niveles máximos diarios de ozono de los años 2001 a 2008 de la Estación Pedregal la cual se encuentra al Suroeste de la Ciudad de México en la Delegación Álvaro Obregón, dicha información la podemos encontrar en el siguiente enlace de internet www.sma.df.gob.mx/simat2/informaciontecnica/difusion/consultas/concentraciones.

Dado que este lapso de tiempo existen 2 años bisiestos debíamos contar con 2922 observaciones o registros, pero, el SIMAT presenta la leyenda o mensaje "nr" (no hay registro de dato) en 15 días (1 día en 2002, 5 días en 2006, 4 días en 2007 y 5 días en 2008), por lo cual contamos con 2907 observaciones o registros.

4.2. Ajuste de los datos de niveles máximos de Ozono

Contamos con información de los niveles máximos diarios de ozono de los años 2001 a 2008 de la Estación Pedregal, ubicada en la delegación Álvaro Obregón de la Ciudad de México. En general se tienen todas las mediciones del año (excepto por 15 días), presentamos todos los registros en el Apéndice A. El primer paso es realizar u organizar los bloques para cada año y obtener el máximo de cada bloque.

4.2.1. Construcción de bloques y obtención de máximos de bloque(Block Maxima)

En el análisis clusters se agrupan observaciones o información con características similares para formar grupos homogéneos, se tienen características similares dentro del cluster y hay diferencia entre clusters. La información con la que tratamos se refiere a máximos diarios de ozono, se puede considerar que no hay otro grupo o tipo de información, entonces agrupamos los datos en bloques de igual longitud y obtenemos el máximo de cada bloque.

Se decidió construir bloques de longitud 3 (días) por que consideramos que 72 horas es un lapso de tiempo suficiente de separación entre una observación y otra para lograr independencia de observaciones, además de obtener al menos 2 observaciones por semana. La construcción de los bloques así como la obtención del máximo de cada

bloque se realizó utilizando el paquete estadístico XTREMES. El procedimiento es simple; se cargan los datos de cada año de un archivo con extensión .DAT, se activa en el menú el dominio "MAX" (Extreme Value Distributions), accedemos a "Data", elegimos la opción "transform data" entre las opciones que aparecen optamos por "Save Blocks Maxima" y en "Block size" se introduce la longitud de cada bloque, en nuestro caso 3, finalmente, damos nombre al archivo resultante el cual contiene los máximos de cada bloque.

Ya que tenemos los máximos por bloque es necesario tener una visualización general de los datos, en este caso los presentamos en un gráfica de series de tiempo en la Figura 4.1.

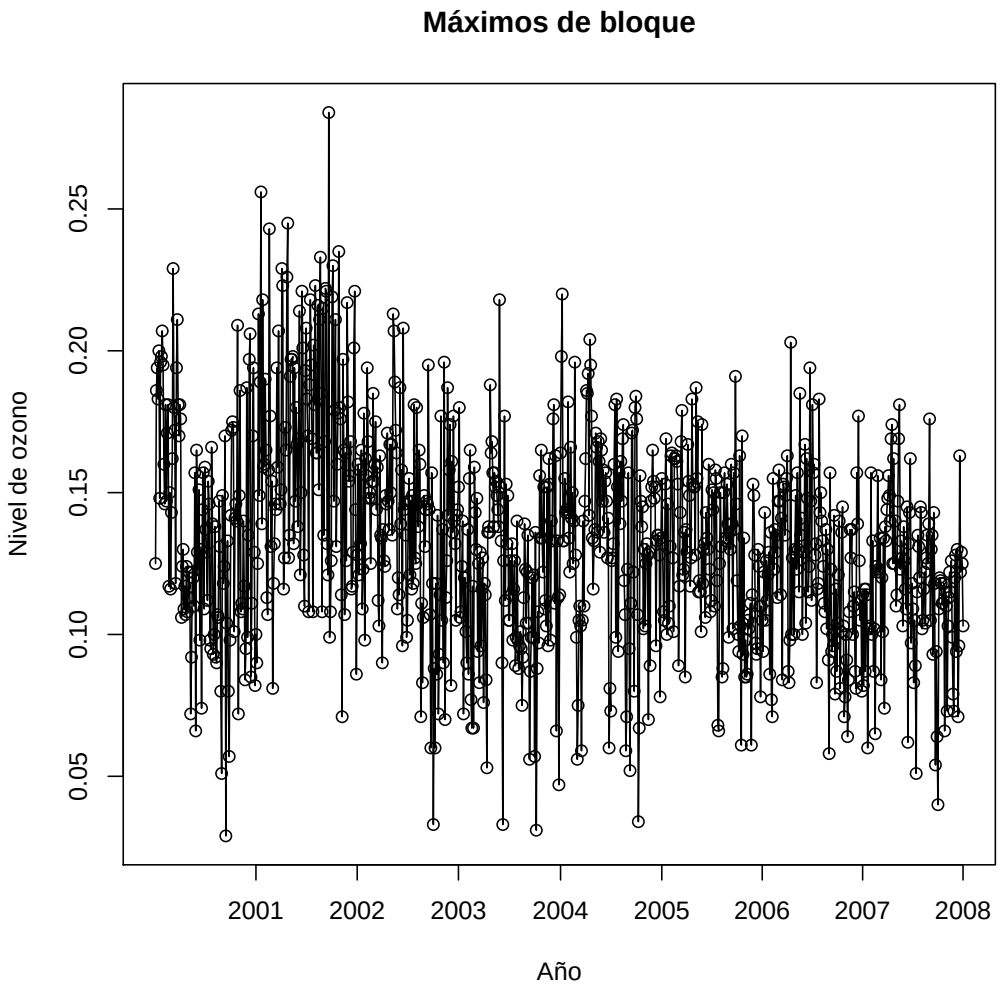

Figura 4.1: Serie de tiempo de los niveles de ozono máximos por bloque de los años 2001 a 2008

En la Figura 4.1 se observa que la distribución de los niveles máximos de ozono no

es el mismo cada año, se puede apreciar que, a excepción del año 2002 en los que se dan los niveles máximos de ozono más altos, se da una tendencia a la baja de los niveles máximos de ozono al paso de los años, por lo cual la serie no es estacionaria y podemos corroborarlo con su gráfica de autocorrelación en la Figura 4.2, por ello conviene hacer el análisis y ajuste de los niveles máximos de ozono para cada año y con ello eliminar dicho problema.

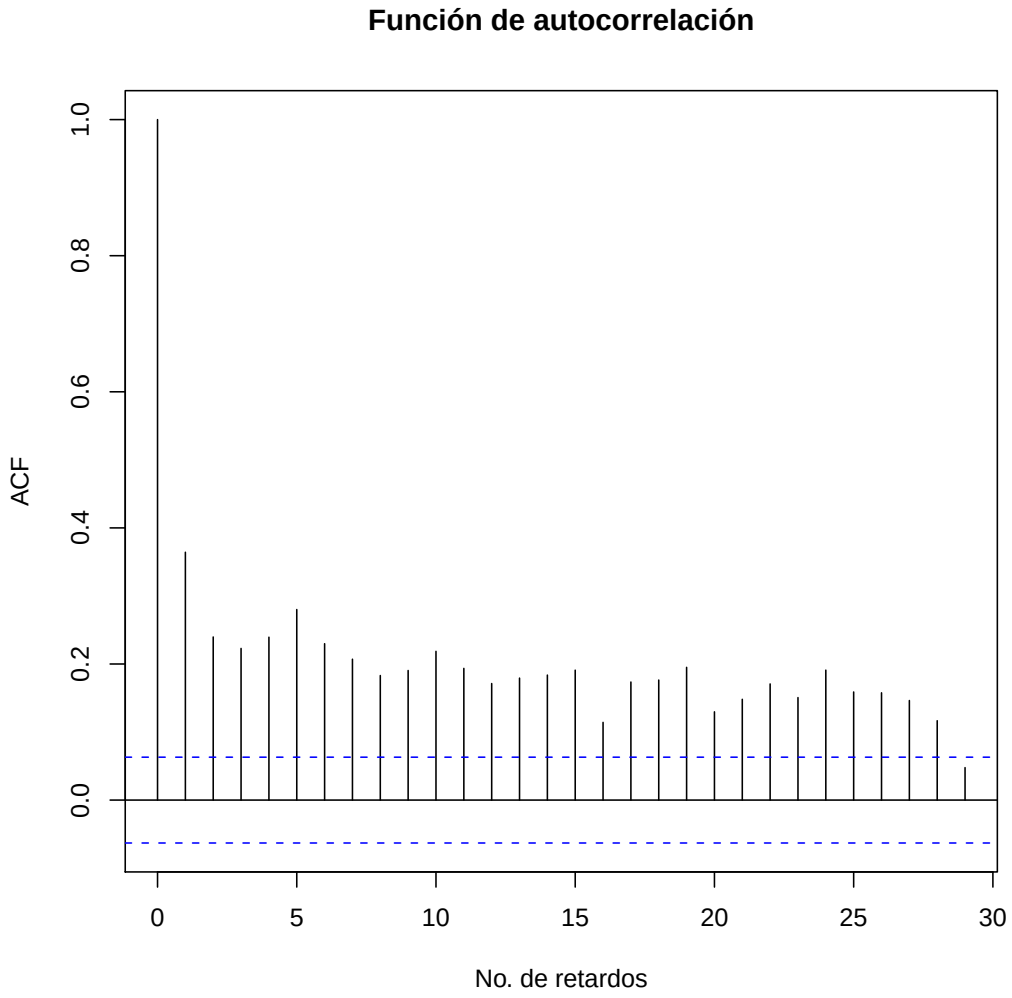

Figura 4.2: Función de autocorrelación de los niveles de ozono máximos por bloque de los años 2001 a 2008

Al realizar en análisis por año se tienen series estacionarias y observaciones aleatorias, tomemos como ejemplo los máximos por bloque del año 2002 y apliquemos la prueba de rachas de Wald-Wolfwitz y la gráfica de función de autocorrelación que se presenta en la Figura 4.3. Con relación a la prueba de rachas dada en la Tabla 4.1, el valor del estadístico de contraste es ($Z = -0.085$) y su nivel crítico (Sig. asintót. (bilateral)= 0.933), puesto que el nivel crítico es muy grande (mayor que 0.05), no se rechaza la

hipótesis de independencia y se dice que la secuencia de observaciones estudiada es aleatoria.

Tabla 4.1: Prueba de rachas a los máximos por bloque del año 2002

	Máximos por bloque 2002
Valor de prueba(media)	0.16693
Casos < Valor de prueba	62
Casos ≥ Valor de prueba	59
Casos en total	121
Número de rachas	61
Z	-0.085
Sig. asintót. (bilateral)	0.933

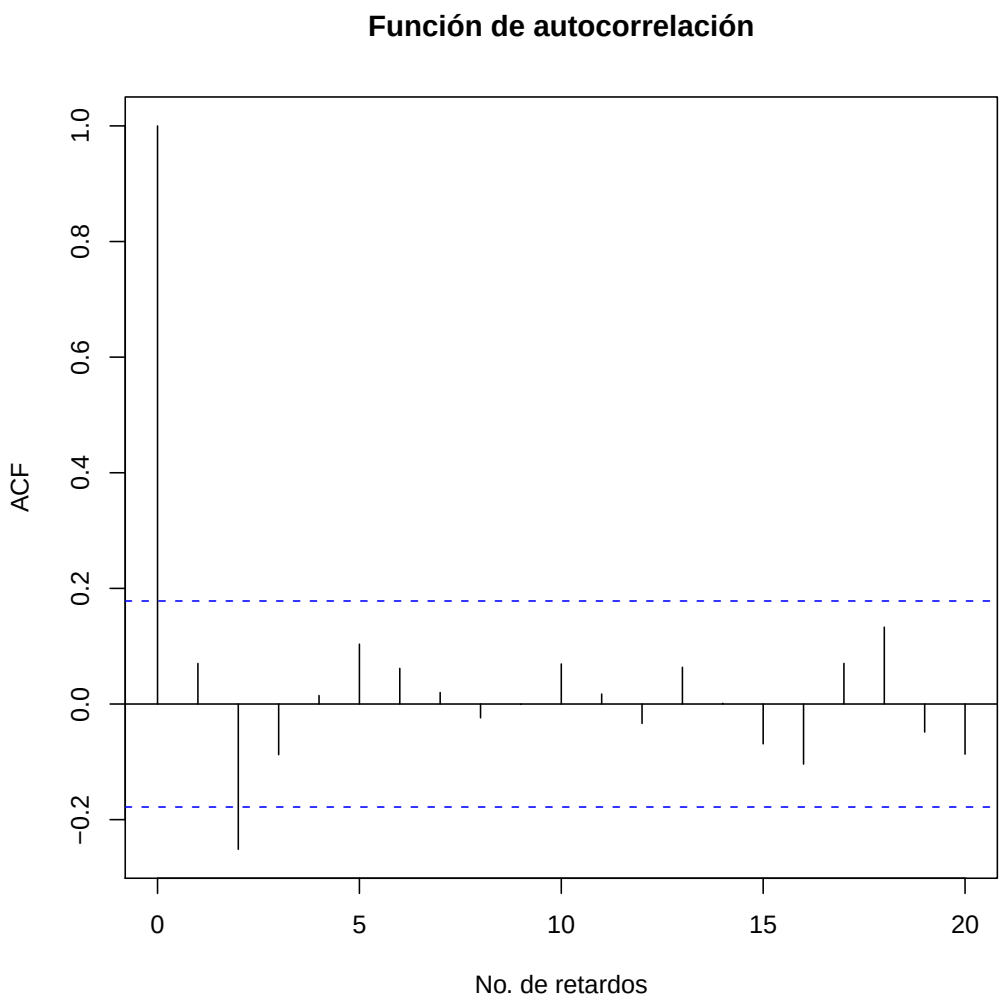

Figura 4.3: Función de autocorrelación de los niveles de ozono máximos por bloque del año 2002

Entonces además de implementar el ajuste de los niveles máximos de ozono por bloque

por año a la distribución Dagum y su comparación con la distribución de valores extremos generalizada (GEV) trataremos de verificar la presencia de tendencia en las observaciones al paso de los años.

El análisis de tendencia podremos hacerlo a través de la comparación de los parámetros y de los cuantiles de la distribución Dagum en un punto crítico obtenidos cada año y en base a los resultados verificaremos la tendencia observada.

4.2.2. Comparación gráfica del ajuste

Una vez obtenido los máximos de cada bloque (121 observaciones en los años 2001, 2002, 2003 y 2005, 122 observaciones en el año 2004, 120 observaciones en los años 2006, 2007 y 2008) los ajustamos a la distribución de nuestro interés, es decir, la distribución Dagum y la comparamos gráficamente con la distribución empírica de los datos y la distribución usada de manera estándar para valores extremos, la distribución de Valores Extremos Generalizada (GEV).

Antes de comparar las gráficas de la función de distribución empírica de los niveles máximos de ozono por bloque con las gráficas de la distribución Dagum ajustadas con los parámetros obtenidos para cada año conviene hacer el siguiente ejercicio; simulemos una muestra de la distribución $Dagum(a, b, p)$ con $a = 14$, $b = 1$ y $p = 0.5$ de tamaño $n = 121$ (este es el tamaño de muestra mas frecuente en nuestro trabajo), grafiquemos la función de distribución empírica de la muestra y grafiquemos también la curva de la distribución $Dagum(a, b, p)$ de la muestra. El resultado de este ejercicio esta en la Figura 4.4.

Se podría pensar que en la Figura 4.4 nos encontraríamos con un acoplamiento perfecto entre las gráficas de la función de distribución empírica de la muestra $Dagum(a, b, p)$ y la curva de la distribución $Dagum(a, b, p)$, pero no es así, aunque también el resultado depende de cada muestra, podríamos encontrarnos con ajustes casi perfectos así como con discrepancias mayores al observado en nuestro ejemplo siendo que la muestra si proviene de la distribución Dagum. Ahora tenemos una idea de como se pueden dar las gráficas de la función de distribución empírica de una muestra de distribución Dagum y la curva teórica de la distribución Dagum.

Ahora, para realizar el ajuste de nuestra información a las distribuciones Dagum y GEV para cada año, se necesitan los estimadores de los parámetros los cuales los obtenemos por el método de máxima verosimilitud con ayuda del programa estadístico R y los paquetes “VGAM” en el caso de la distribución Dagum y “evir” en el caso de la distribución GEV. Los gráficos de distribuciones obtenidos para cada año las tenemos en las Figuras 4.5-4.12.

En base a las gráficas podemos apreciar que en general el ajuste de ambas distribu-

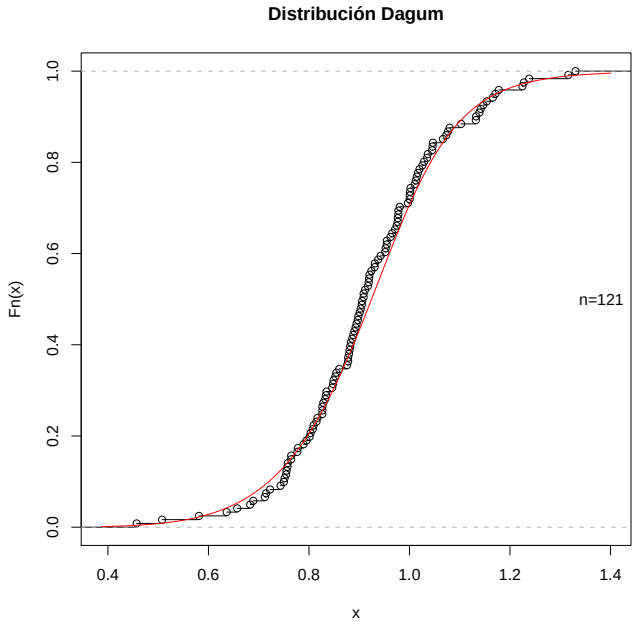

Figura 4.4: Función de distribución empírica de una muestra $Dagum(a, b, p)$ de tamaño $n = 121$ y curva de la función de distribución $Dagum(a, b, p)$ de tamaño $n = 121$, $a = 14$, $b = 1$ y $p = 0.5$

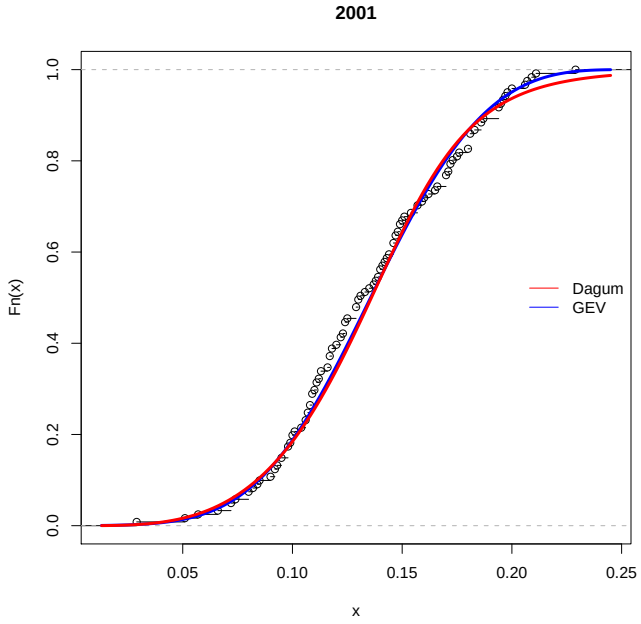

Figura 4.5: Comparación del ajuste de la distribución Dagum con las distribuciones empírica y GEV en el año 2001

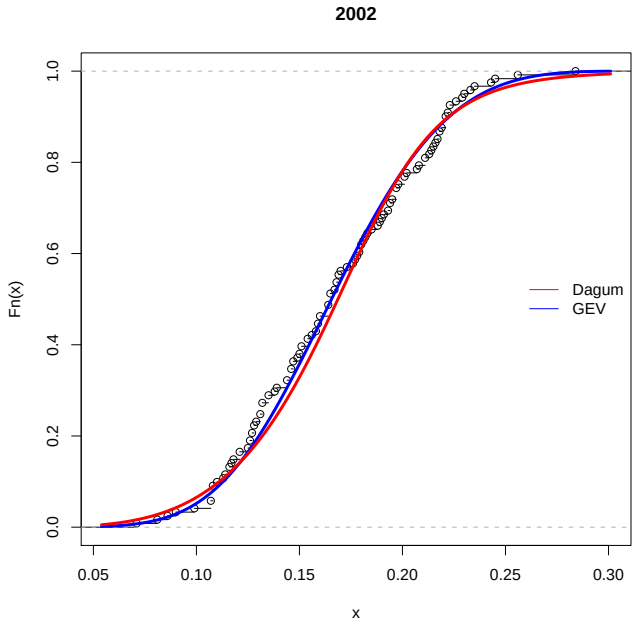

Figura 4.6: Comparación del ajuste de la distribución Dagum con las distribuciones empírica y GEV en el año 2002

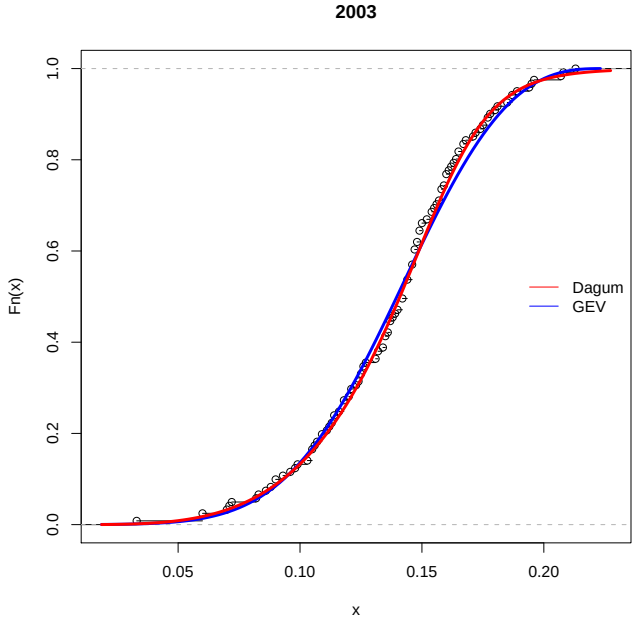

Figura 4.7: Comparación del ajuste de la distribución Dagum con las distribuciones empírica y GEV en el año 2003

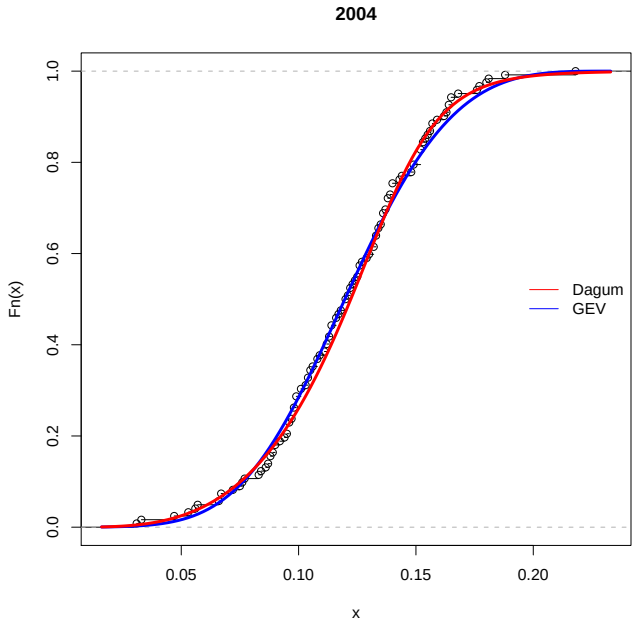

Figura 4.8: Comparación del ajuste de la distribución Dagum con las distribuciones empírica y GEV en el año 2004

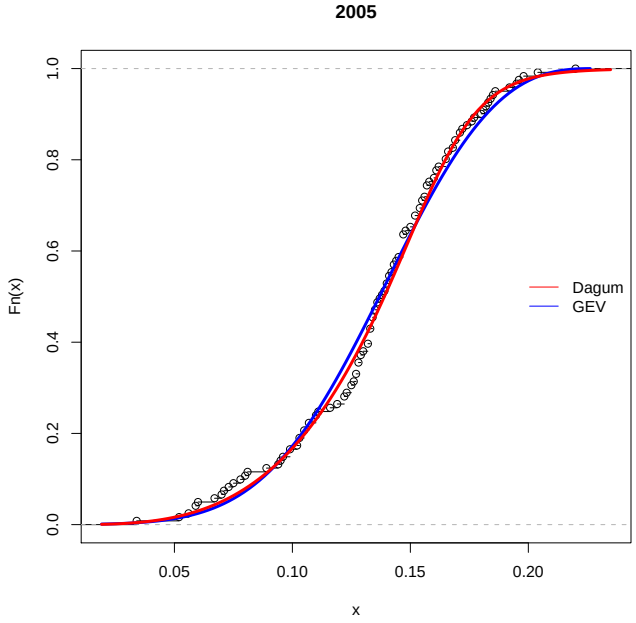

Figura 4.9: Comparación del ajuste de la distribución Dagum con las distribuciones empírica y GEV en el año 2005

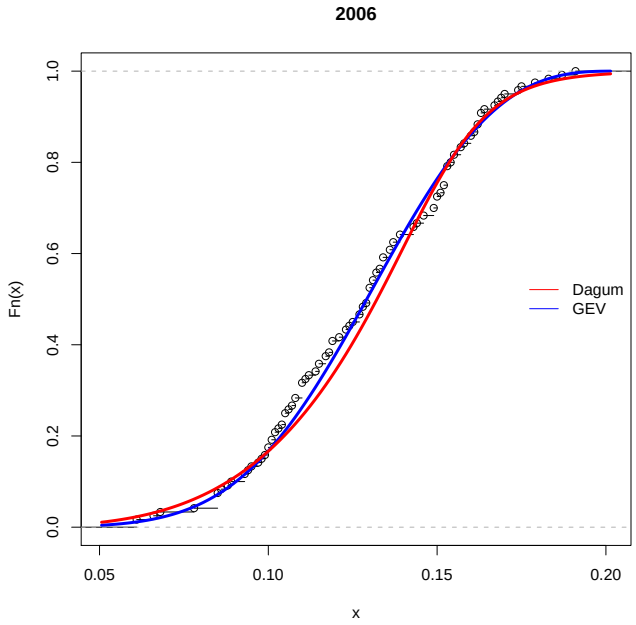

Figura 4.10: Comparación del ajuste de la distribución Dagum con las distribuciones empírica y GEV en el año 2006

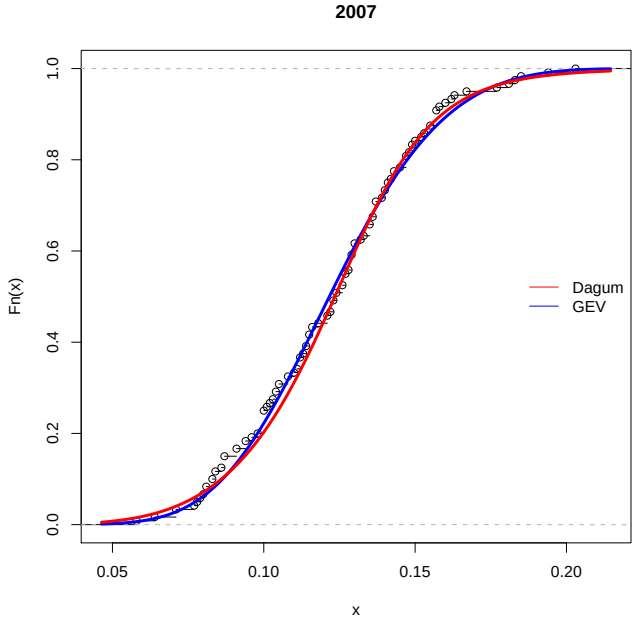

Figura 4.11: Comparación del ajuste de la distribución Dagum con las distribuciones empírica y GEV en el año 2007

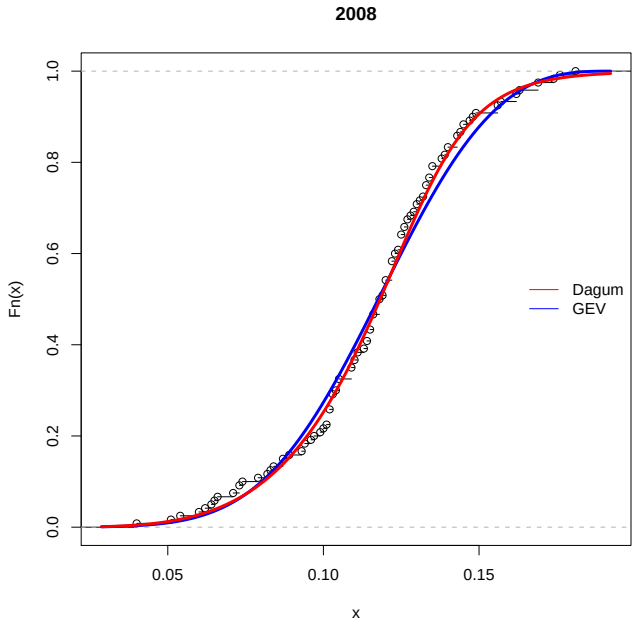

Figura 4.12: Comparación del ajuste de la distribución Dagum con las distribuciones empírica y GEV en el año 2008

ciones es muy semejante, pero si podemos decir que bajo este criterio se ve un mejor ajuste de la distribución Dagum para los datos de los años 2003, 2004, 2005 y 2008 y la distribución GEV presenta mejor ajuste para el resto de años en estudio, 2001, 2002, 2006 y 2007.

En la comparación gráfica vemos justificada la implementación de la distribución Dagum en valores extremos ya que compite con la distribución GEV e incluso tiene un mejor ajuste en la mitad de los casos. Sin embargo, es necesario usar otras técnicas que apoyen o refuten los resultados vistos en la comparación gráfica, por ejemplo, alguna prueba de bondad de ajuste.

4.2.3. Prueba de bondad de ajuste y criterio de información de Akaike (AIC)

Ademas de la comparación gráfica, probamos el ajuste de las observaciones a las distribuciones teóricas con la prueba de Kolmogorov-Smirnov con nivel de significancia $\alpha = 0.05$. Aplicamos la prueba de Kolmogorov-Smirnov y utilizamos el criterio del "p-value" para concluir si las observaciones máximas de bloques se ajustan a las distribuciones Dagum y Valor Extremo Generalizado.

Las pruebas de hipótesis a contrastar cada año son las siguientes:

$$\boldsymbol{H_0}: \text{ los datos siguen la distribución Dagum}$$

$$vs$$

$$\boldsymbol{H_1}: \text{ los datos no siguen la distribución Dagum}$$

De la misma forma:

$$\boldsymbol{H_0}: \text{ los datos siguen la distribución GEV}$$

$$vs$$

$$\boldsymbol{H_1}: \text{ los datos no siguen la distribución GEV}$$

La regla de decisión es la siguiente:

$$\text{Si } p-value \geq \alpha \implies \text{ Aceptar } H_0$$

$$\text{Si } p-value < \alpha \implies \text{ Rechazar } H_0$$

Ya que además de implementar la distribución Dagum para modelar valores extremos también la estamos comparando con la distribución GEV, incluimos también un criterio usado para la comparación de modelos, el criterio de información de Akaike o AIC. Tanto la prueba de bondad de ajuste como la obtención de el logaritmo de la función de verosimilitud usada para obtener el estadístico AIC los realizamos con ayuda del paquete estadístico R, los resultados para cada año se presentan en la Tabla 4.2.

Tabla 4.2: $p-values$ obtenidos en la prueba de Kolmogorov-Smirnov y el estadístico AIC

Año	$p-value$		AIC	
	Dagum	GEV	Dagum	GEV
2001	0.6371	0.7933	-426.6164	-434.7164
2002	0.5414	0.7623	-409.4166	-416.752
2003	0.9728	0.801	-475.4264	-474.6734
2004	0.9545	0.9743	-478.3866	-476.224
2005	0.6972	0.2813	-460.532	-456.9844
2006	0.5107	0.6088	-501.6258	-508.6648
2007	0.8596	0.9089	-502.811	-505.7758
2008	0.9318	0.5637	-518.5954	-516.0032

En la Tabla 4.2 podemos ver valores $p-value$ grandes para ambos contrastes realizados, es decir, en ningún caso se rechazan las hipótesis de que los niveles de ozono máximos por bloque siguen la distribución Dagum y la distribución GEV. Recordando la comparación gráfica en la que se aprecia mejor ajuste de la distribución Dagum

en los años 2003, 2004, 2005 y 2008 y mejor ajuste de la distribución GEV en el resto de los años, obtenemos los mismos resultados con la prueba de bondad de ajuste a excepción del año 2004, sin embargo, la diferencia de los $p - value$ obtenidos para ambas distribuciones en ese año es pequeña.

La Tabla 4.2 incluye también el estadístico de Akaike o AIC, el criterio para elegir un modelo sobre otro es seleccionar aquel con menor AIC, entonces bajo este criterio, la distribución Dagum es mejor modelo que la distribución GEV en los años 2003, 2004, 2005 y 2008, y la distribución GEV es mejor modelo en el resto de años considerados, estos resultados coinciden con nuestra comparación gráfica.

4.3. Análisis de tendencia

4.3.1. Comportamiento de los parámetros de la distribución Dagum

Los registros de los niveles máximos de ozono por bloque presentan tendencia a la baja en el lapso de tiempo considerado, por esta razón se decidió hacer el análisis y ajuste de la información por año, entonces deben darse características en las distribuciones de cada año como resultado de dicha tendencia, fijémonos por ejemplo, en los valores de los parámetros y analicemos su comportamiento. Los valores de los parámetros de cada año se citan en la Tabla 4.3.

Tabla 4.3: Parámetros de la distribución Dagum

Año	Parámetro		
	a	b	p
2001	8.3479	0.1615	0.4202
2002	9.7786	0.1959	0.4150
2003	13.5001	0.1669	0.2918
2004	11.2898	0.1486	0.2999
2005	14.1866	0.1701	0.2371
2006	14.9753	0.1563	0.2679
2007	10.3362	0.1394	0.4608
2008	12.3889	0.1369	0.3520

Además de la Tabla 4.3 se presenta también la representación gráfica de la evolución de los parámetros, en la Figura 4.13 se pueden notar tendencias bien marcadas de los parámetros a y b, a tiene tendencia creciente y b tendencia decreciente al paso del tiempo. En cambio, en la Figura 4.14 al parámetro p parece tener una tendencia decreciente aunque no es tan pronunciada, por lo tanto, parece necesaria otra herramienta o prueba para determinar si se da un cambio significativo en los parámetros.

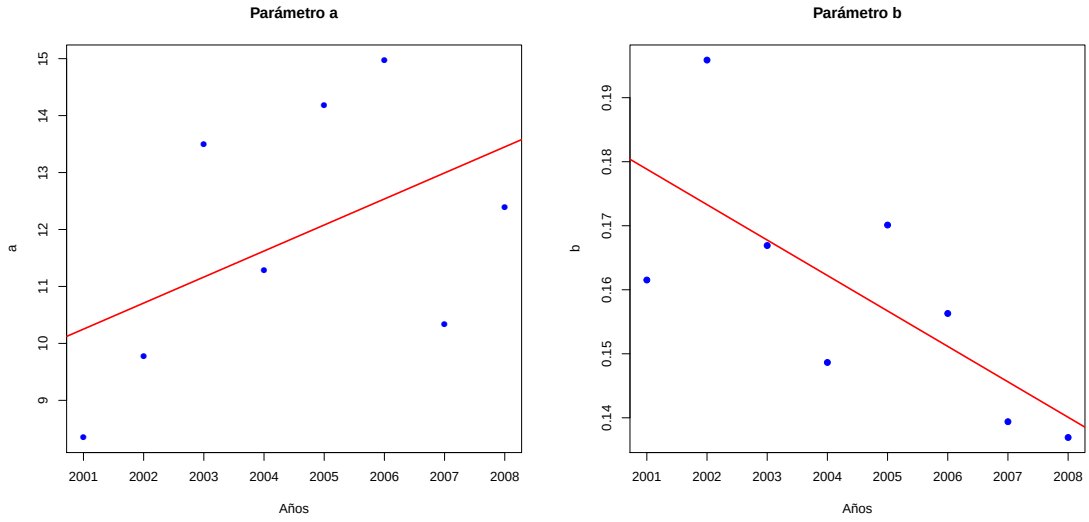

Figura 4.13: Parámetro a (izquierda) y parámetro b (derecha)

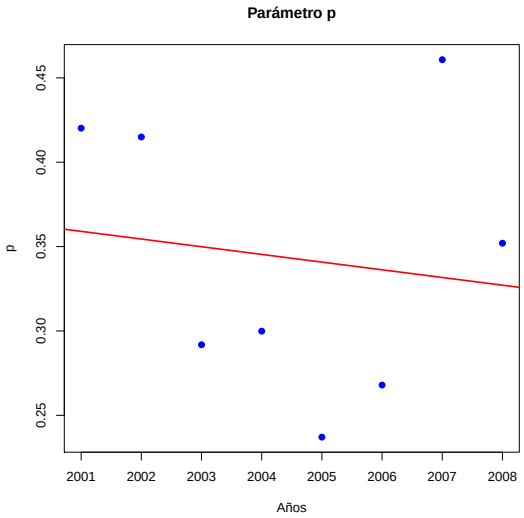

Figura 4.14: Parámetro p

4.3.2. Modelo Lineal Generalizado Vectorial

En esta sección se presenta una técnica a la que llamamos Modelo Lineal Generalizado Vectorial (Vector Generalized Linear Models, VGLM por sus siglas en inglés) que nos permite determinar si se da una relación lineal entre los parámetros de la distribución Dagum con una covariable, en nuestro caso, el tiempo (años). La técnica la haremos con la ayuda del programa estadístico R con el paquete VGAM (Vector Generalized

Additive Models) con la función $vglm()$, los VGLMs son un caso particular de los VGAMs. Describiremos brevemente los modelos VGLMs y VGAMs, más detalles se pueden encontrar en Yee y Hastie (2003) y Yee y Wild (1996), respectivamente.

En los modelos lineales generalizados (GLMs, por sus siglas en inglés) (Nelder y Wedderburn, 1972), donde y es una distribución de la familia exponencial, la media μ de y esta relacionada con p covariables $\boldsymbol{x} = (x_1, ..., x_p)^T$ por

$$g(\mu) = \eta(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

y han sido extendidos para formar la clase de modelos aditivos generalizados (GAMs, por sus siglas en inglés) (Hastie y Tibshirani, 1990) en los cuales

$$g(\mu) = \eta(\boldsymbol{x}) = \beta_0 + f_1 x_1 + \cdots + f_p x_p,$$

donde f_j son funciones suavizadas arbitrarias (pueden ser no lineales).

Las clases VGAM/VGLM son implementadas en el paquete VGAM (Yee, 2007), para el entorno del programa estadístico R (Ihaka y Gentleman, 1996). Los VGLMs y VGAMs permiten a todos los parámetros de la distribución ser modeladas como funciones lineales o funciones suavizadas de covariables.

Suponga que la respuesta observada $\boldsymbol{y}$ es un vector q-dimensional. Los VGLMs son definidos como un modelo para el cual la distribución condicional de $\boldsymbol{Y}$ dada la explicatoria $\boldsymbol{x}$ es de la forma:

$$f(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{B}) = h(\boldsymbol{y}, \eta_1, ..., \eta_M) \tag{4.1}$$

para alguna función conocida $h(\cdot)$, donde $\boldsymbol{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_M)$ es una matriz de orden $p \times M$ de los coeficientes de regresión desconocidos, y el j-ésimo predictor lineal es

$$\eta_j = \eta_j(\boldsymbol{x}) = \boldsymbol{\beta}_j^T\boldsymbol{x} = \beta_{(j)1}x_1 + \cdots + \beta_{(j)p}x_p = \sum_{k=1}^{p} \beta_{(j)k}x_k, \quad j = 1, ..., M \tag{4.2}$$

donde $\boldsymbol{x} = (x_1, ..., x_p)^T$ con $x_1 = 1$ si hay intercepto. La ecuación (4.2) nos indica que todos los parámetros pueden ser potencialmente modelados como funciones de la covariable $\boldsymbol{x}$, en nuestro caso la covariable es el tiempo. Note que los modelos VGLMs son como los GLMs (donde $M = 1$) pero permiten predictor lineal múltiple y abarcan modelos fuera de los confines de la familia exponencial.

En general no hay relación entre q y M, depende específicamente del modelo a ser ajustado, por ejemplo, para modelos simples de valores extremos M es el número de parámetros a estimar y q puedes ser un número entero positivo. Para nuestro caso, q=1 y M=3. Los VGAMs proporcionan extensiones de modelos aditivos a los VGLMs,

es decir, (4.2) es generalizada a:

$$\eta_j(\boldsymbol{x}) = \beta_{(j)1} + f_{(j)2}(x_2) + \cdots + f_{(j)p}(x_p) = \beta_{(j)1} + \sum_{k=2}^{p} f_{(j)k}(x_k), \quad j = 1, ..., M \quad (4.3)$$

una suma de funciones suavizadas de las covariables individuales. Las η_j son referidos como predictores aditivos. $\boldsymbol{f}_k = (f_{(1)k}(x_k), ..., f_{(M)k}(x_k))$ es centrada por unicidad, y se estiman de manera simultánea utilizando "vector smoothers".

En la práctica tal vez se quiera limitar el efecto de una covariable a ser el mismo para algunos de los η_j y no tener efecto en otras. Por ejemplo, en los VGAMs quisiéramos tener

$$\begin{aligned} \eta_1 &= \beta_{(1)1} + f_{(1)2}(x_2) + f_{(1)3}(x_3) \\ \eta_2 &= \beta_{(2)1} + f_{(1)2}(x_2), \end{aligned}$$

de modo que $f_{(1)2} \equiv f_{(2)2}$ y $f_{(2)3} \equiv 0$. Para los VGAMs, podemos representar estos modelos usando

$$\boldsymbol{\eta}(\boldsymbol{x}) = \boldsymbol{\beta}_{(1)} + \sum_{k=2}^{p} \boldsymbol{f}_k(x_k) = \mathbf{H}_1\boldsymbol{\beta}^*_{(1)} + \sum_{k=2}^{p} \mathbf{H}_k\boldsymbol{f}^*_k(x_k) \quad (4.4)$$

donde $\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_p$ se conocen como matrices de restricción de rango completo por columnas, $\boldsymbol{f}^*_k$ es un vector que contiene al posible conjunto reducido de funciones de los componentes y $\boldsymbol{\beta}^*_{(1)}$ es el vector de interceptos desconocido. Cuando no hay restricciones, $\mathbf{H}_1 = \mathbf{H}_2 = \cdots = \mathbf{H}_p = \mathbf{I}_M$ y $\boldsymbol{\beta}^*_{(1)} = \boldsymbol{\beta}_{(1)}$. Al igual que $\boldsymbol{f}_k$, $\boldsymbol{f}^*_k$ es centrada por unicidad. En los VGLMs, $\boldsymbol{f}_k$ es lineal, por lo tanto

$$\mathbf{B}^T = \begin{pmatrix} \mathbf{H}_1\boldsymbol{\beta}^*_{(1)} & \mathbf{H}_2\boldsymbol{\beta}^*_{(2)} & \cdots & \mathbf{H}_p\boldsymbol{\beta}^*_{(p)} \end{pmatrix}$$

Los VGLMs son estimados usualmente por máxima verosimilitud por Fisher Scoring o Newton-Raphson.

Ahora, en nuestro caso usaremos predictores lineales, es decir, el modelo lineal generalizado vectorial (VGLMs):

$$\begin{aligned} \log(a) &= \eta_1 = \beta_{(1)1}x_1 + \beta_{(1)2}x_2 \\ \log(b) &= \eta_2 = \beta_{(2)1}x_1 + \beta_{(2)2}x_2 \\ \log(p) &= \eta_3 = \beta_{(3)1}x_1 + \beta_{(3)2}x_2 \end{aligned}$$

donde $x_1 = 1$ ya que consideramos el intercepto, x_2 es el tiempo expresado en años, $q = 1$ y $M = 3$. Los resultados se presentan en la Tabla 4.4.

En la Tabla 4.4 se tienen los coeficientes obtenidos del modelo lineal generalizado vectorial (predictores lineales) y los $t - value$. Entonces considerando un nivel de

Tabla 4.4: Coeficientes de la regresión Dagum

Coefficients	Value	Std. Error	t-value
(Intercept):1	2.087559	0.1351299	15.44854
(Intercept):2	-1.676371	0.0367075	-45.66831
(Intercept):3	-0.920553	0.2083267	-4.41879
año:1	0.06091	0.0271134	2.24648
año:2	-0.036772	0.0065516	-5.61265
año:3	-0.022878	0.041165	-0.55575

significancia $\alpha = 0.05$ y el criterio de que si el valor absoluto del estadístico t ($t - value$) es mayor que 2 es significativo, concluimos que los coeficientes asociados a los parámetros a y b (año:1 y año:2) son significativos y por lo tanto éstos parámetros tiene relación lineal con nuestra covariable tiempo (años), se reafirma el signo de la pendiente observada gráficamente. Esta conclusión aclara la duda que se dió en la representación gráfica acerca de la tendencia lineal del parámetro p con respecto al tiempo, decimos entonces que su cambio no es significativo.

4.3.3. Cuantiles

Se han mencionado los efectos adversos a la salud que puede causar el ozono, un aspecto importante para beneficio de la salud fué notar que los niveles máximos diarios de ozono tienen tendencia a la baja al paso de los años. Una manera práctica de apreciar los cambios en la distribución de datos u observaciones es a través de la función cuantil, en nuestro caso, usaremos dicha función para observar el valor de la distribución en un punto crítico y debemos confirmar entonces la tendencia a valores mas bajos. Calculamos el cuantil $(1 - \alpha)100$ de la distribución para cada año con un nivel de significancia α=0.05, 0.10 y 0.50 y observamos su tendencia, los resultados se presentan en la Tabla 4.5.

Tabla 4.5: Cuantiles $(1 - \alpha)100$ Dagum

Año	α		
	0.05	0.10	0.5
2001	0.2062722	0.1877332	0.1359919
2002	0.2410239	0.2223855	0.168677
2003	0.188601	0.177531	0.140998
2004	0.1724811	0.1604845	0.1222518
2005	0.1880305	0.1772189	0.1389722
2006	0.1734096	0.1641078	0.1321886
2007	0.1714448	0.1589807	0.1234787
2008	0.1590109	0.1490738	0.1182318

Ya que hemos comparado anteriormente el ajuste de los datos de niveles máximos de ozono a la distribución Dagum con el ajuste a la distribución GEV conviene presentar

también los cuantiles de la distribución GEV, los tenemos en la Tabla 4.6.

Tabla 4.6: Cuantiles $(1-\alpha)100$ GEV

Año	α		
	0.05	0.10	0.5
2001	0.1996686	0.1865772	0.1356007
2002	0.2384297	0.2226471	0.1655825
2003	0.1914159	0.181634	0.1397398
2004	0.1774077	0.1651994	0.1197193
2005	0.192768	0.182186	0.1368794
2006	0.1733466	0.1648492	0.1291525
2007	0.1724634	0.1611932	0.1216584
2008	0.1615318	0.1529513	0.1176288

En las Figuras 4.15 y 4.16 podremos comparar mejor los cuantiles de ambas distribuciones. En general, la diferencia es pequeña entre los cuantiles de ambas distribuciones y tienen la misma tendencia. Pero, centremos la atención en la Figura 4.15, en la Figura a la izquierda en donde se tiene en cuantil $(1-\alpha)100$ con $\alpha = 0.05$ el cual tomamos como punto crítico, observamos la tendencia a la baja de los valores de este cuantil, lo cual quiere decir que al paso de los años los niveles máximos de ozono muy altos han sido menos frecuentes o disminuido.

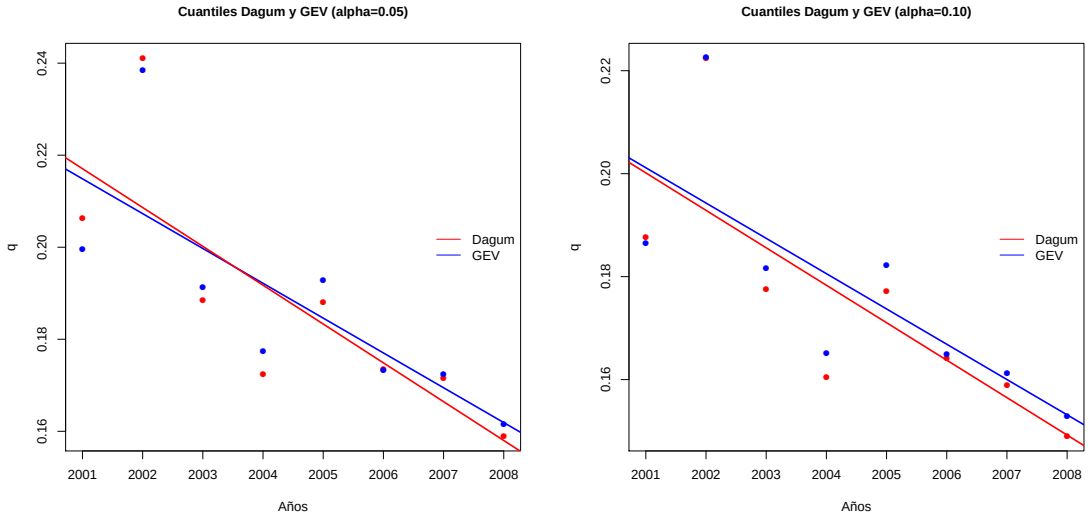

Figura 4.15: Comparación de los cuantiles $(1-\alpha)100$ de las distribuciones Dagum y GEV, $\alpha = 0.05$ (izquierda) y $\alpha = 0.10$ (derecha)

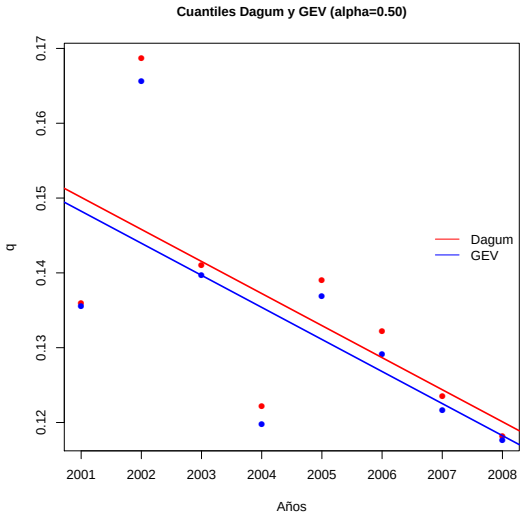

Figura 4.16: Comparación de los cuantiles $(1-\alpha)100$ de las distribuciones Dagum y GEV, $\alpha = 0.50$

Capítulo 5

Prueba de bondad de ajuste

Una prueba de bondad de ajuste tiene como objetivo medir la concordancia o compatibilidad de algunos datos con una distribución teórica determinada. Para el caso de la distribución Dagum no se ha encontrado en la literatura una prueba de bondad de ajuste, de esta manera, proponemos una prueba de bondad de ajuste para la distribución Dagum con parámetros desconocidos basada en el coeficiente de correlación, a continuación se describe la prueba.

Sean $X_1, ...X_n$ una muestra aleatoria con funciones de distribución F y densidad f desconocidos con soporte en $\mathbb{R}^+$. Dada la muestra aleatoria, quisiéramos saber si dicha muestra tiene o sigue la distribución Dagum, es decir, estamos interesados en contrastar la siguiente prueba de hipótesis:

$$\begin{aligned} H_0 : f(x;\ \cdot\ ) = f_0(x; a, b, p) &= \frac{apx^{ap-1}}{b^{ap}\left[1 + \left(\frac{x}{b}\right)^a\right]^{p+1}} \\ &\text{vs} \\ H_1 : f(x;\ \cdot\ ) &\neq f_0(x; a, b, p) \end{aligned} \tag{5.1}$$

5.1. Prueba propuesta

Haremos uso de la función de distribución empírica definida en (3.1):

$$F_n(x) = \begin{cases} 0 & \text{si} \quad x < x_{(1)} \\ i/n & \text{si} \quad x_{(i)} \leq x < x_{(i+1)} \\ 1 & \text{si} \quad x > x_{(n)} \end{cases}$$

y de la función de distribución Dagum dada en (3.17):

$$F(x) = \left[1 + \left(\frac{x}{b}\right)^{-a}\right]^{-p}, \quad x > 0.$$

Ahora, podemos referirnos al teorema de Glivenko-Cantelli, que establece una convergencia casi segura, cuando $n \to \infty$, entre las distribuciones empírica $F_n(x)$ y la distribución teórica $F(x)$, en este caso la distribución Dagum, por lo tanto, es factible establecer:

$$\begin{aligned}
F_n(x) &\approx F(x) \\
F_n(x) &\approx \left[1 + \left(\frac{x}{b}\right)^{-a}\right]^{-p} \\
[F_n(x)]^{-1/p} &\approx 1 + \left(\frac{x}{b}\right)^{-a} \\
[F_n(x)]^{-1/p} - 1 &\approx \left(\frac{x}{b}\right)^{-a} \\
\log\left\{[F_n(x)]^{-1/p} - 1\right\} &\approx -a\log(x) + a\log(b) \qquad (5.2)
\end{aligned}$$

Note que (5.2) podemos escribirla como:

$$Y = \beta + \alpha Z \qquad (5.3)$$

donde

$$Y = \log\left\{[F_n(x)]^{-1/p} - 1\right\} \qquad (5.4)$$

$$Z = \log x \qquad (5.5)$$

$\beta = a\log(b)$ y $\alpha = -a$.

Entonces, se espera una fuerte relación lineal entre $\log\left\{[F_n(x_i)]^{-1/p} - 1\right\}$ y $\log x_i$ dadas en (5.2) baja la hipótesis nula en (5.1) y se espera que la relación lineal exista aún si se sustituye p por un estimador consistente($\hat{p}$) lo cual se tratará de probar por medio de alguna metodología o procedimiento, por ejemplo, por medio del coeficiente de correlación.

El coeficiente de correlación es el siguiente:

$$r_n = Corr(Y, Z) = \rho_{Y,Z} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(Z_i - \bar{Z})}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sum_{i=1}^{n}(Z_i - \bar{Z})^2}}. \qquad (5.6)$$

Note que para el cálculo del estadístico solo se depende del parámetro p y lo susti-

tuiremos con su estimador ($\hat{p}$) obtenido por el método de máxima verosimilitud. La regla de decisión es rechazar H_0 en (5.1) con un nivel de significancia α si $r_n \leq C_n(\alpha)$, donde $C_n(\alpha)$ es tal que:

$$\alpha = \max_p P(\text{ Rechazar } H_0|H_0) = \max_p P(r_n \leq C_n(\alpha)) \tag{5.7}$$

Podemos obtener la distribución de r_n bajo la hipótesis nula para valores fijos del parámetro de forma p a través de simulación Monte Carlo, el procedimiento se describe a continuación:

1) Fijar n, a, b y p.
2) Simular una muestra de tamaño n, $X_1, ..., X_n$, de la distribución Dagum, $D(a, b, p)$.
3) Calcular el estimador de máxima verosimilitud del parámetro p usando en paquete VGAM del programa R.
4) Ordenar las $x_i's$ en forma ascendente.
5) Calcular la funcion de distribución empírica dada en (3.1).
6) Calcular y_i.
7) Calcular z_i.
8) Calcular en coeficiente de correlación (r_n) en (5.6) entre las y_i y las z_i.
9) Repetir los pasos 2 a 8, B veces.

Bajo H_0 la distribución de r_n debe estar concentrada muy cerca a 1, en cambio, si la muestra aleatoria no viene de una distribución Dagum los valores del coeficiente de correlación no serán cercanos a uno y serán mas pequeños que el valor crítico. Pongamos en marcha el procedimiento, por ejemplo, consideremos 2 muestras aleatorias con distribución Dagum con $n = 121$ ya que esas son nuestras observaciones por año, $p = 0.35, 1$ y B=10000. Los valores críticos $C_n(\alpha)$ se obtienen con los cuantiles $100 \times \alpha$ de la distribución empírica de r_n.

La distribución de r_n con los valores de p considerados se muestran en la Figura 5.1. La finalidad de obtener la distribución de r_n con dos valores diferentes de p es mostrar que dicha distribución no es invariante a los cambios en p, esta situación se puede apreciar en la Figura 5.1 en la que se ve un aumento de las frecuencias de valores más cercanos a 1 del coeficiente de correlación considerando $p = 1$ (Figura a la derecha) con respecto a $p = 0.35$ (Figura de la izquierda).

En la Tabla 5.1 se presentan los valores críticos de la distribución de r_n considerando además los valores de $p = 0.5, 2$ y se refleja la situación vista en la Figura 5.1, la

distribución de r_n cambia con los valores de p y se refleja en los valores críticos $(C_n(\alpha))$ con $\alpha = 0.05$. Note que valores más grandes de p inciden en valores del coeficiente de correlación más cercanos a 1, situación que se observa en la Figura 5.1.

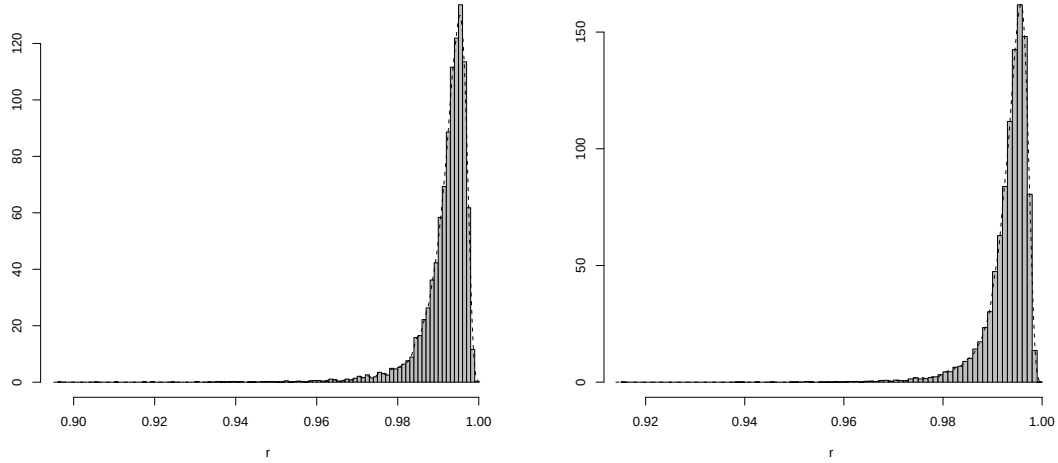

Figura 5.1: Distribución de r_n, $p = 0.35$ (izquierda) y $p = 1$ (derecha)

Tabla 5.1: Valores críticos $(C_n(\alpha))$ de r_n con $n = 121$, $\alpha = 0.05$ y diferentes valores de p

n=121	p			
	0.35	0.5	1	2
$C_n(\alpha)$	0.9814683	0.9828373	0.984775	0.985119

Dado que la distribución de r_n no es invariante a los cambios en el parámetro p, no se pueden construir tablas generales como en los casos de pruebas de bondad de ajuste existentes en las que se tienen tablas con valores críticos generales que dependen del tamaño de muestra n y el nivel de significancia (α). Debido a la situación prevaleciente, la propuesta de prueba solo puede aplicarse a muestras específicas, es decir, no puede ser una prueba general pero sí como una prueba computacional, una prueba bootstrap (bootstrap paramétrico).

5.1.1. Aplicación de la prueba propuesta a nuestra información

Ahora, aplicamos la prueba a la información con la que se ha tratado en este trabajo, niveles máximos de ozono por bloque de los años 2002 a 2008, el procedimiento es muy

semejante al descrito anteriormente, solo se tienen pequeños cambios, el procedimiento de la prueba bootstrap el como sigue:

1) Dadas las observaciones $x_1, ..., x_n$, calcular su función de distribución empírica y los estimadores de máxima verosimilitud de a, b, p, digamos $\hat{a}$, $\hat{b}$, $\hat{p}$.

2) Calcular es eltadístico r_n usando $\hat{p}$.

3) Generar una muestra boostrap de tamaño n de la distribución Dagum, $D(\hat{a}, \hat{b}, \hat{p})$.

4) Calcular la funcion de distribución empírica de la muestra bootstrap.

5) Calcular el estimador de máxima verosimilitud de p, digamos $\hat{p}^*$, de la muestra bootstrap dada.

6) Calcular el estadístico de prueba denotado ahora por r_n^* usando $\hat{p}^*$ y la muestra bootstrap.

7) Repetir los pasos 3 a 6, $B = 10000$ veces para obtener r_{nj}^*, $j = 1, ..., 10000$

8) Obtener $(C_n(0.05))$ como $r_{n(500)}^*$, donde r_{nj}^*, $j = 1, ..., 10000$ denota los valores ordenados de r_n^*.

La regla de decisión es rechazar H_0 en (5.1) con un nivel de significancia α si $r_n \leq C_n(\alpha)$. En nuestro caso usamos $\alpha = 0.05$ y los resultados para cada año se tienen en la Tabla 5.2. En base a los resultados en la Tabla 5.2 podemos concluir que con la

Tabla 5.2: Valores críticos $(C_n(\alpha))$ de r_{nj}^*, $\alpha = 0.05$ de los años 2002 a 2008

Año	r_n	$(C_n(\alpha))$
2001	0.9907647	0.9819275
2002	0.9841725	0.9814643
2003	0.9958358	0.978694
2004	0.9960264	0.979012
2005	0.9949232	0.978067
2006	0.9823936	0.9783751
2007	0.9889576	0.9822448
2008	0.9959603	0.9806644

prueba propuesta no se rechaza la hipótesis de que los niveles máximos de ozono por bloque siguen la distribución Dagum en cada año.

Capítulo 6

Conclusiones

En base al ejercicio realizado en este trabajo podemos nombrar varias conclusiones. En lo referente a la implementación de la distribución Dagum para modelar valores extremos en niveles de ozono, podemos concluir que:

- En base a la comparación gráfica, la distribución Dagum presenta un ajuste similar e incluso mejor en la mitad de los casos (años 2003, 2004, 2005 y 2008) al ajuste realizado por la distribución GEV la cual es la distribución usada generalmente para trabajar con valores extremos.
- Con la prueba de Kolmogorov-Smirnov, con nivel de significancia $\alpha = 0.05$, no se rechaza en ningún caso la hipótesis de que los niveles máximos de ozono por bloque siguen la Distribución Dagum o la distribución GEV. Para la distribución Dagum se obtienen $p-values$ mayores a los obtenidos con la Distribución GEV en los mismos casos en los que se percibió gráficamente un mejor ajuste de la distribución Dagum, a excepción del año 2004, sin embargo, la diferencia de los $p - value$ obtenidos para ambas distribuciones en ese año es pequeña.
- Con el criterio de Akiake o AIC, se considera mejor modelo el de Dagum en los mismos casos que la comparación gráfica y la prueba de bondad de ajuste, es decir, en los años 2003, 2004, 2005 y 2008.
- Con los resultados dados en este trabajo, se justifica la implementación de la distribución Dagum para modelar valores extremos y se presenta como una opción a considerar cuando se trabaja con valores extremos por las facilidades que representa al tener programadas sus características principales en el paquete R.

La tendencia a la baja observada de los niveles máximos de ozono fué constatada con diversas técnicas:

- Se observaron gráficamente cambios en los parámetros al paso de los años como consecuencia de la tendencia, sin embargo, la tendencia en el parámetro p no era tan clara por lo que generó dudas.

- Con el Modelo Lineal Generalizado Vectorial (VGLM, por sus siglas en inglés) se verificaron los resultados observados gráficamente acerca de los parámetros, con un nivel se significancia $\alpha = 0.05$, hay cambios significativos (tendencia) en los parámetros a y b con pendientes positiva y negativa, respectivamente. En el parámetro p no se da cambio significativo.

- En base a los cuantiles $(1 - \alpha)100$ con $\alpha = 0.05$ se confirma la tendencia a la baja de los niveles máximos de ozono, es decir, al paso de los años los niveles máximos de ozono muy altos se han dado con menor frecuencia. Los cuantiles $(1 - \alpha)100$ con $\alpha = 0.05$ de las distribuciones Dagum y GEV son muy similares y presentan la misma tendencia.

En relación a la prueba de bondad de ajuste que se propone en el presente trabajo, la conclusión es la siguiente:

- Debido a que la distribución del estadístico r_n no es invariante a los cambios en el parámetro p, no se puede construir una prueba general que dependa únicamente del tamaño de muestra n y del nivel de significancia α como algunas pruebas de bondad de ajuste existentes, pero sí se puede dejar como una prueba computacional, una prueba bootstrap (bootstrap paramétrico). En base a la prueba propuesta, no se rechaza la hipótesis de que los niveles máximos de ozono por bloque siguen la distribución Dagum en cada año estudiado considerando un nivel de significancia $\alpha = 0.05$.

Referencias

Akaike, H. (1974). A new look at the statistical model indentification. *IEEE Transactions on Automatic Control*, 19, 716–722.

Arnold, B. C. (1983). *Pareto Distributions.*. Fairland, MD: International Co-operative Publishing House, Fairland, Maryland.

Beirlant, J. (1996). Tail index estimation, Pareto quantile plots, and regression diagnostics. *J. Amer. Statist. Assoc.*, 91, 1659–1667.

Burr, I. W. (1942). Cumulative frequency functions. *Annals of Mathematical Statistics*, 13, 215–232.

Castillo, E., Hadi, A., Balakrishnan, N. y Sarabia, J. (2005). *Extreme Value and Related Models with Applications in Engineering and Science*. John Wiley & Sons, Hoboken, New Jersey.

Chaplin, W. S. (1880). The relation between the tensile strengths of long and short bars. *Van Nostrand's Engineering Magazine*, 23, 441–444.

Clark, J. S., Lewis, M. y Horvath, L. (2001). Invasion by extremes: population spread with variation in dispersal and reproduction. *American Naturalist*, 157, 537–554.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Verlag, London.

Dagum, C. (1975). A model of income distribution and the conditions of existence of moments of finite order. *Bulletin of the International Statistical Institute*, 46(Proceedings of the 40th Session of the ISI, Warsaw, Contributed Papers), 199–205.

Dagum, C. (1977). A New Model for Personal Income Distribution: Specification and Estimation. *Economie Appliquée*, 30, 413–437.

Dagum, C. (1980c). Sistemas generadores de distribucio´n de ingreso y la ley de Pareto. *El Trimestre Economico*, 47, 877–917.

Dagum, C. (1983). *Income Distribution Models*, cap. Vol. 4. John Wiley.

Dagum, C. y Lemmi, A. (1989). A contribution to the analysis of income distribution and income inequality, and a case study: Italy. *Research on Economic Inequality*, 1, 123–157.

Dutka, J. (1981). The incomplete beta function-a historical profile. *Archive for the History of Exact Sciences*, 24, 11–29.

Embrechts, P., Klüppelberg, C. y Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer, Berlin.

Fattorini, L. y Lemmi, A. (1979). Proposta di un modello alternativo per lánalisi della distribuzione personale del reddito. *Atti Giornate di Lavoro AIRO*, 28, 89–117.

Ferro, C., Segers, J. y Robert, C. (2009). A sliding blocks estimator for the extremal index. *Electronic Journal of Statistics*, 3, 993–1020.

Finetti, B. D. (1932). Sulla legge di probabílitá degli estremí. *Metron*, 9, 127–138.

Finkenstadt, B. y Rootzén, H. (2003). *Extreme Values in Finance, Telecommunications and the Environment*. Chapman and Hall/CRC Press, London.

Fisher, R. y Tippett, L. (1928). Limiting forms of the frequency distributions of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 24, 180–190.

Fisk, P. R. (1961a). The graduation of income distributions. *Econometrica*, 29, 171–185.

Fisk, P. R. (1961b). Estimation of location and scale parameters in a truncated grouped sech-square distribution. *Journal of the American Statistical Association*, 56, 692–702.

Gaines, S. D. y Denny, M. W. (1993). The largest, smallest, highest, lowest, longest, and shortest: extremes in ecology. *Ecology*, 74, 1677–1692.

Gibrat, R. (1931). *Les Inégalités Économiques*. Librairie du Recueil Sirey, Paris.

Gnedenko, B. V. (1943). Sur la distribution limite du terme maximum dúne série aléatoire. *Ann. Math*, 44, 423–453.

Gumbel, E. J. (1934). Les moments des distributions finales de la premiére et de la derniére valeur. *Camptes Rendus de lÁcadémie des Sciences*, 198, l 41–143.

Gumbel, E. J. (1935a). Les valeurs extremes des distributions statistiques. *Annales de línstitut Henri Poincaré*, 5, 1 15–158.

Gumbel, E. J. (1935b). La plus grande valeur. *Aktuárské Vedy*, 5, 83–39, 133–143 and 145–160.

Gumbel, E. J. (1958). *Statistics of extremes*. Columbia Univ. Press, New York.

Hastie, T. y Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Ihaka, R. y Gentleman, R. (1996). R: A language for data analysis and graphics. *J. Comput. Graph. Stat.*, 5, 299–314.

Jenkinson, A. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quart. J. Roy. Meteo. Soc.*, 81, 158–171.

Referencias

Johnson, N. L., Kotz, S. y Balakrishnan, N. (1995). *Continuous Univariate Distributions*, tomo 2. John Wiley & Sons, New York, segunda edición.

Kalbfleisch, J. D. y Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. John Wiley, New York.

Katz, R. (2002b). Stochastic modeling of hurricane damage. *Journal of Applied Meteorology*, 41, 754–762.

Katz, R. W. (2002a). Techniques for estimating uncertainty in climate change scenarios and impact studies. *Climate Research*, 20, 167–185.

Katz, R. W., Brush, G. S. y Parlange, M. B. (2005). Statistics of Extremes: Modeling Ecological Disturbances. *Ecological Society of America*, 86, 1124–1134.

Katz, R. W., Parlange, M. B. y Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in Water Resources*, 25, 1287–1304.

Kleiber, C. (2008). *Modeling Income Distributions and Lorenz Curves*, cap. 6, A Guide to the Dagum Distributions, 97–117. Springer New York.

Kleiber, C. y Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. John Wiley, Hoboken, New Jersey.

Klugman, S. A., Panjer, H. H. y Willmot, G. E. (1998). *Loss Models*. John Wiley, New York.

Kotz, S., Johnson, N. L. y Read, C. (1983). *Encyclopedia of Statistical Sciences*. John Wiley, New York.

Kotz, S. y Nadarajah, S. (2000). *Extreme Value Distributions: Theory and Applications*. Imperial College Press, London.

Leadbetter, M. R., Lindgren, G. y Rootzen, H. (1983). *Extremes and related properties of random sequences and processes*. Springer-Verlag, New York.

Majumder, A. y Chakravarty, S. R. (1990). Distribution of personal income: Development of a new model and its application to U.S. income data. *Journal of Applied Econometrics*, 5, 189–196.

McDonald, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica*, 52, 647–663.

McDonald, J. B. y Mantrala, A. (1995). The distribution of income: Revisited. *Journal of Applied Econometrics*, 10, 201–204.

McDonald, J. B. y Xu, Y. J. (1995). A generalization of the beta distribution with applications. *Journal of Econometrics*, 66, 133–152.

Mielke, P. W. (1973). Another family of distributions for describing and analyzing precipitation data. *Journal of Applied Meteorology*, 12, 275–280.

Mielke, P. W. y Johnson, E. S. (1974). Some generalized beta distributions of the second kind having desirable application features in hydrology and meteorology. *Water Resources Research*, 10, 223–226.

Nelder, J. A. y Wedderburn, R. W. M. (1972). Generalized linear models. *J. R. Statist. Soc.*, A, 135, 370–384.

Pareto, V. (1895). La Legge della Domanda. *Giornale degli Economisti, English Translation in Rivista di Politica Economica*, 10, 87(1997), 59–68, 691–700.

Parker, S. C. (1999a). The generalized beta as a model for the distribution of earnings. *Economics Letters*, 62, 197–200.

Parker, S. C. (1999b). The beta as a model for the distribution of earnings. *Bulletin of Economic Research*, 53, 243–251.

Ponce de Leon, A., Anderson, H., Bland, J. y Bower, J. (1996). Effects of air pollution on daily hospital admissions for respiratory disease in London between 1987-88 and 1991-92. *J Epidemiol Comm Health*, Vol. 50 (Supplement 1), S63–S70.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Reiss, R. y Thomas, M. (2007). *Statistical Analysis of Extreme Values, with Application to Insurance, Finance, Hydrology and Other Fields*. Birkhuser Verlag, Basel.

Rice, S. O. (1939). The distribution of the maxima of a random curve. *Amer. J. Math.*, 61, 409–416.

Singh, S. K. y Maddala, G. S. (1976). A function for the size distribution of incomes. *Econometrica*, 44, 963–970.

Smith, R. (2003). *Extreme Values in Finance, Telecommunications and the Environment*, cap. 1, Statistics of extremes, with applications in environment, insurance and finance, 1–78. Chapman and Hall/CRC Press.

Stoppa, G. (1995). Explicit Estimators for Income Distributions, in C. Dagum and A. Lemmi (eds.). *Research on Economic Inequality*, 6: Income Distribution, Social Welfare, Inequality and Poverty, Greenwich, CT: JAI Press, 393–405.

Thurow, L. (1970). Analyzing the American income distribution. *American Economic Review (Papers and Proceedings)*, 48, 261–269.

Venter, G. (1983). Transformed beta and gamma distributions and aggregate losses. *Proceedings of the Casualty Actuarial Society*, 70, 156–193.

von Mises, R. (1936). La distribution de la plus grande de n valeurs. *Revue Mathématique de lÚnion Interbalkanique (Athens)*, 1, 141–160.

von Mises, R. (1954). La distribution de la plus grande de n valeurs. *American Mathematical Society*, Selected Papers Volumen II, 271–294.

Referencias

Yee, T. (2007). *A User´s Guide to the vgam Package*.

Yee, T. y Hastie, T. (2003). Reduced-rank Vector Generalized Linear Models. *Statistical Modelling*, 3(1), 15–41.

Yee, T. y Wild, C. (1996). Vector Generalized Additive Models. *Journal of the Royal Statistical Society B*, 58(3), 481–493.

Apéndices

Apéndice A: Máximos de bloques de niveles de ozono de la estación Pedregal, Delegación Álvaro Obregón, Ciudad de México

Máximos de bloques

Tabla .1: Máximos de bloques de los años 2001-2008

Año							
2001	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**
0.125	0.09	0.16	0.18	0.198	0.132	0.094	0.082
0.186	0.125	0.121	0.106	0.22	0.108	0.122	0.116
0.194	0.213	0.158	0.108	0.133	0.105	0.143	0.087
0.183	0.149	0.132	0.124	0.133	0.153	0.105	0.116
0.2	0.189	0.165	0.14	0.155	0.169	0.126	0.114
0.148	0.256	0.109	0.072	0.143	0.1	0.112	0.06
0.196	0.139	0.123	0.12	0.144	0.146	0.121	0.102
0.198	0.218	0.178	0.12	0.143	0.104	0.129	0.102
0.207	0.159	0.098	0.101	0.182	0.11	0.086	0.103
0.195	0.16	0.162	0.09	0.134	0.131	0.136	0.157
0.16	0.19	0.156	0.137	0.122	0.163	0.077	0.1
0.146	0.165	0.194	0.086	0.166	0.164	0.071	0.133
0.181	0.113	0.168	0.155	0.141	0.127	0.155	0.087
0.147	0.107	0.148	0.165	0.15	0.101	0.123	0.102
0.171	0.158	0.15	0.077	0.125	0.162	0.137	0.065
0.181	0.243	0.125	0.067	0.14	0.162	0.129	0.124
0.117	0.177	0.148	0.117	0.196	0.163	0.135	0.122
0.15	0.131	0.154	0.067	0.128	0.161	0.113	0.156
0.116	0.154	0.185	0.159	0.099	0.153	0.147	0.123
Sigue ...							

Tabla .1: Máximos de bloques 2001-2008 (continuación)

Año							
2001	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**
0.143	0.081	0.154	0.143	0.056	0.089	0.158	0.122
0.162	0.118	0.16	0.13	0.075	0.117	0.147	0.134
0.229	0.132	0.175	0.148	0.105	0.143	0.114	0.084
0.18	0.146	0.144	0.094	0.11	0.168	0.141	0.12
0.172	0.144	0.159	0.125	0.103	0.179	0.084	0.101
0.118	0.194	0.112	0.083	0.059	0.121	0.153	0.134
0.194	0.159	0.103	0.129	0.105	0.123	0.152	0.074
0.211	0.207	0.163	0.096	0.14	0.136	0.135	0.138
0.181	0.146	0.135	0.116	0.11	0.085	0.149	0.133
0.17	0.146	0.134	0.127	0.147	0.137	0.155	0.144
0.181	0.151	0.09	0.076	0.162	0.129	0.163	0.148
0.176	0.229	0.137	0.118	0.186	0.167	0.087	0.156
0.106	0.223	0.126	0.114	0.185	0.154	0.083	0.149
0.124	0.116	0.124	0.084	0.192	0.149	0.098	0.14
0.13	0.127	0.149	0.053	0.142	0.117	0.203	0.169
0.109	0.173	0.146	0.136	0.204	0.127	0.1	0.174
0.117	0.165	0.171	0.138	0.195	0.183	0.127	0.125
0.107	0.226	0.147	0.136	0.177	0.153	0.127	0.162
0.124	0.245	0.137	0.188	0.134	0.152	0.1	0.125
0.108	0.127	0.167	0.164	0.116	0.152	0.13	0.14
0.113	0.135	0.15	0.168	0.133	0.128	0.149	0.11
0.122	0.191	0.167	0.157	0.137	0.187	0.123	0.114
0.113	0.197	0.135	0.138	0.171	0.155	0.157	0.147
0.072	0.197	0.213	0.157	0.166	0.175	0.13	0.169
0.092	0.198	0.207	0.152	0.136	0.115	0.115	0.181
0.122	0.132	0.189	0.154	0.168	0.119	0.185	0.123
0.12	0.194	0.172	0.149	0.161	0.115	0.135	0.133
0.11	0.147	0.164	0.144	0.129	0.101	0.141	0.127
0.157	0.18	0.109	0.138	0.169	0.174	0.1	0.103
0.066	0.167	0.12	0.218	0.165	0.118	0.129	0.138
0.165	0.138	0.114	0.152	0.147	0.13	0.129	0.126
0.129	0.169	0.187	0.133	0.158	0.132	0.167	0.116
0.123	0.214	0.139	0.09	0.136	0.106	0.104	0.109
0.151	0.121	0.158	0.033	0.154	0.143	0.162	0.145
0.098	0.15	0.096	0.126	0.157	0.134	0.148	0.062
0.13	0.221	0.208	0.177	0.147	0.13	0.115	0.132
0.074	0.201	0.144	0.112	0.127	0.16	0.124	0.143
0.144	0.128	0.146	0.153	0.141	0.123	0.194	0.162
0.157	0.11	0.135	0.132	0.06	0.108	0.147	0.097
0.109	0.193	0.099	0.149	0.081	0.112	0.112	0.118
0.159	0.208	0.105	0.105	0.073	0.151	0.181	0.109
0.137	0.183	0.147	0.105	0.126	0.144	0.157	0.083

Sigue ...

Tabla .1: Máximos de bloques 2001-2008 (continuación)

Año							
2001	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**
0.129	0.17	0.155	0.126	0.127	0.15	0.16	0.105
0.112	0.108	0.121	0.114	0.132	0.11	0.128	0.089
0.154	0.188	0.147	0.132	0.16	0.158	0.132	0.051
0.146	0.218	0.116	0.098	0.181	0.155	0.083	0.115
0.098	0.195	0.149	0.116	0.099	0.119	0.116	0.135
0.095	0.169	0.118	0.126	0.183	0.068	0.118	0.131
0.166	0.108	0.181	0.089	0.152	0.066	0.183	0.12
0.139	0.202	0.127	0.099	0.094	0.125	0.143	0.145
0.093	0.181	0.125	0.14	0.157	0.15	0.15	0.143
0.1	0.223	0.18	0.098	0.139	0.131	0.14	0.102
0.138	0.164	0.146	0.088	0.161	0.085	0.133	0.105
0.092	0.164	0.138	0.096	0.147	0.088	0.124	0.116
0.09	0.216	0.165	0.101	0.169	0.15	0.111	0.122
0.107	0.151	0.16	0.095	0.174	0.157	0.108	0.104
0.106	0.211	0.071	0.075	0.119	0.134	0.137	0.127
0.147	0.233	0.111	0.092	0.107	0.133	0.102	0.135
0.131	0.185	0.083	0.113	0.059	0.153	0.13	0.125
0.08	0.108	0.106	0.139	0.071	0.139	0.091	0.139
0.051	0.215	0.149	0.123	0.123	0.099	0.058	0.176
0.149	0.135	0.131	0.104	0.157	0.136	0.157	0.105
0.118	0.168	0.144	0.122	0.095	0.13	0.123	0.13
0.124	0.222	0.147	0.135	0.052	0.16	0.094	0.135
0.17	0.221	0.146	0.104	0.111	0.139	0.1	0.093
0.029	0.132	0.195	0.056	0.171	0.102	0.096	0.143
0.104	0.121	0.144	0.087	0.172	0.157	0.142	0.118
0.133	0.284	0.107	0.099	0.122	0.137	0.079	0.054
0.08	0.099	0.06	0.12	0.08	0.191	0.116	0.094
0.057	0.108	0.157	0.099	0.18	0.146	0.087	0.064
0.098	0.126	0.118	0.121	0.184	0.119	0.104	0.04
0.142	0.219	0.033	0.057	0.176	0.1	0.121	0.109
0.172	0.23	0.088	0.136	0.107	0.094	0.14	0.12
0.175	0.147	0.06	0.031	0.034	0.163	0.081	0.119
0.173	0.179	0.118	0.088	0.067	0.103	0.101	0.118
0.101	0.211	0.086	0.108	0.156	0.061	0.136	0.111
0.141	0.131	0.142	0.097	0.147	0.17	0.145	0.111
0.146	0.16	0.072	0.156	0.138	0.093	0.103	0.118
0.14	0.178	0.093	0.134	0.145	0.134	0.071	0.066
0.209	0.235	0.114	0.165	0.102	0.085	0.078	0.115
0.072	0.18	0.177	0.134	0.13	0.098	0.1	0.11
0.149	0.176	0.105	0.122	0.103	0.085	0.091	0.073
0.186	0.114	0.137	0.152	0.125	0.086	0.064	0.103
0.108	0.071	0.09	0.109	0.132	0.105	0.084	0.113
Sigue ...							

Tabla .1: Máximos de bloques 2001-2008 (continuación)

Año							
2001	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**
0.111	0.197	0.196	0.112	0.128	0.102	0.108	0.122
0.109	0.165	0.07	0.103	0.07	0.107	0.137	0.115
0.14	0.107	0.113	0.152	0.129	0.111	0.127	0.126
0.117	0.164	0.126	0.096	0.089	0.061	0.137	0.099
0.084	0.126	0.187	0.162	0.147	0.114	0.1	0.079
0.095	0.217	0.142	0.153	0.152	0.153	0.11	0.073
0.187	0.182	0.158	0.098	0.165	0.149	0.115	0.12
0.099	0.154	0.174	0.14	0.154	0.128	0.087	0.128
0.135	0.156	0.082	0.133	0.148	0.097	0.081	0.094
0.197	0.168	0.161	0.176	0.152	0.093	0.157	0.13
0.206	0.116	0.136	0.181	0.096	0.095	0.139	0.071
0.085	0.117	0.177	0.111	0.133	0.13	0.177	0.096
0.111	0.129	0.14	0.163	0.135	0.124	0.126	0.163
0.17	0.201	0.132	0.066	0.135	0.11	0.105	0.122
0.194	0.221	0.105	0.133	0.128	0.078	0.112	0.129
0.129	0.144	0.142	0.113	0.078	0.11	0.08	0.125
0.082	0.086	0.144	0.047	0.134	0.105	0.114	0.103
0.1	0.128	0.152	0.114	0.155			
			0.164				
Fin							

Apéndice B: Programas de R usados en el trabajo

Prueba bootstrap

Haremos uso del paquete VGAM del programa R, es necesario instalar dicho paquete el cual puede obtenerse gratuitamente en la página web del proyecto R, http://www.r-project.org. Las rutinas que se presentan funcionan en el paquete R-2.8.0 ó superior.

```
#Función para calcular del coeficiente de correlación la muestra entre y y z.
library(VGAM)
ccrd=function(x)
{
  d=vglm(x~1,dagum)
```

```
  p=exp(coef(d))
  s=sort(x)
  n=length(s)
  f=((1:n)-0.5)/n
  y=-log((f^(-1/p[3]))-1)
  z=log(s)
  cor(y,z)
}
ccrd(x)

# Programa de la prueba bootstrap
library(VGAM)
x<-c() #vector de datos u observaciones (muestra)

#Función para obtener el valor crítico

corrwz=function(x,B,alpha)
{
  ajustado=vglm(x~1,dagum)
  m=exp(coef(ajustado))#estimadores de la muestra
  n=length(x)
  i=1;
correlacion=rep(NA,B)
while(i<=B)
  {
     print(i)
     y_bootstrap=rdagum(n,m[1],m[2],m[3]) #Generar muestra bootstrap
     out=try(vglm(y_bootstrap~1,dagum),silent=TRUE)
     if(class(out)[1]=="vglm")
     {
e=exp(coef(out))  #estimadores de la muestra bootstrap
s=sort(y_bootstrap)
f=((1:n)-0.5)/n
y=-log((f^(-1/e[3]))-1)
z=log(s)
correlacion[i]=cor(y,z)#coeficiente de correlación
   i=i+1;
     } else
     {
     }
  }
correlacion
t=sort(correlacion)
q=quantile(t,alpha) #valor crítico
q
}
```

```
corrwz(x,10000,0.05)

#Regla de decisión; Rechazar la hipótesis nula con nivel de significancia alpha
#si el valor obtenido en ccrd(x) es menor o igual al valor crítico
#obtenido en corrwz(x,10000,0.05).
```

Ajuste de las distribuciones Dagum y GEV

Para realizar el ajuste de la distribución Dagum es necesario instalar el paquete VGAM en el programa R. Las rutina funciona en el paquete R-2.8.0 ó superior. Para realizar el ajuste de la distribución GEV es necesario instalar el paquete *evir* en el programa R.

```
#Estimadores de la distribución Dagum
library(VGAM)
x<-c()#Datos
fit=vglm(x~ 1,dagum)#Ajuste de la distribución Gagum
summary(fit)
p=exp(coef(fit))#Estimadores de máxima verosimilitud de la distribución Dagum
p

#Estimadores de la distribución GEV
library(evir)
x<-c()#Datos
out <- gev(x,1)#Ajuste de la distribución GEV
               #El valor 1 es por que ya hemos obtenido los máximos por bloque
               #con el programa Xtremes
out            #Estimadores de máxima verosimilitud de la distribución GEV
```

Modelo Lineal Generalizado Vectorial

Es necesario cargar el paquete VGAM en el programa R. Las rutina funciona en el paquete R-2.8.0 ó superior.

```
library(VGAM)
x<-c()#Datos, respuesta observada (en nuestro caso, niveles
      #de ozono máximos por bloque)
t<-c()#Datos, de la(s) covariable(s) (en nuestro caso, el tiempo en años)
f<-vglm(x~t,family=dagum)#ajuste del modelo
summary(f)#coeficientes del modelo
```