



# COLEGIO DE POSTGRADUADOS

---

---

INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN  
EN CIENCIAS AGRÍCOLAS

**CAMPUS MONTECILLO**

POSTGRADO DE SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA  
ESTADÍSTICA

**Análisis de Valores Extremos en Presencia de  
Censura Aleatoria y Estacionalidad**

Benigno Estrada Drouaillet

T E S I S

PRESENTADA COMO REQUISITO PARCIAL PARA  
OBTENER EL GRADO DE:

DOCTOR EN CIENCIAS

MONTECILLO, TEXCOCO, EDO. DE MÉXICO  
2014

---

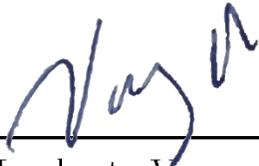
La presente tesis titulada: **Análisis de Valores Extremos en Presencia de Censura Aleatoria y Estacionalidad**, realizada por el alumno: **Benigno Estrada Drouaillet**, bajo la dirección del Consejo Particular indicado ha sido aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:

## DOCTOR EN CIENCIAS

### SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA ESTADÍSTICA

#### CONSEJO PARTICULAR

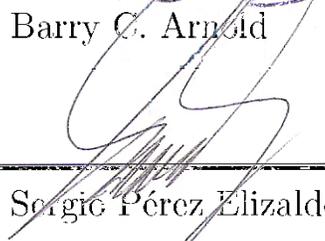
CONSEJERO

  
\_\_\_\_\_  
Dr. Humberto Vaquera Huerta

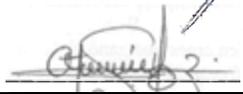
ASESOR

  
\_\_\_\_\_  
Dr. Barry C. Arnold

ASESOR

  
\_\_\_\_\_  
Dr. Sergio Pérez Elizalde

ASESOR

  
\_\_\_\_\_  
Dr. Francisco J. Ariza Hernández

ASESOR

  
\_\_\_\_\_  
Dr. Gerardo H. Terrazas González

Montecillo, Texcoco, Estado de México,  
2014

# Análisis de Valores Extremos en Presencia de Censura Aleatoria y Estacionalidad

Benigno Estrada Drouaillet

Colegio de Postgraduados, 2014

En este trabajo se desarrolla un modelo de regresión a través del parámetro de localidad de la distribución de Valores Extremos Generalizada (VEG) bajo el esquema de censura aleatoria para analizar la tendencia en el tiempo de los valores máximos diarios de partículas suspendidas ( $PM_{10}$ ) del año 1995 al 2013 en la Zona Metropolitana de la Ciudad de México (ZMCM). Los registros que se estudian pertenecen a 11 estaciones del Sistema de Monitoreo Atmosférico de la Ciudad de México (SIMAT). Los datos presentan niveles de censura entre 10 % y 15 % debido a fallas en los equipos de medición. Además, los valores máximos muestran un patrón estacional por lo que se incorpora una senoide en el parámetro de localidad. La estimación de los parámetros de la distribución VEG se realiza con el método de máxima verosimilitud asumiendo censura aleatoria. Los parámetros estimados de la distribución VEG se utilizan para calcular niveles de retorno y generar mapas de contorno para estudiar el patrón espacial de las  $PM_{10}$ . También, se realiza un estudio Monte Carlo para analizar el efecto del nivel de censura y el tamaño de muestra en el sesgo de los parámetros y de los niveles de retorno estimados. Finalmente, se emplea una modificación del estadístico Anderson-Darling, y generalizaciones de los estadísticos Kolmogorov-Smirnov y Cramér-von Mises para probar bondad de ajuste e identificar de entre las pruebas empleadas la de mayor potencia y la que conserve de mejor manera el tamaño de la prueba.

**Palabras clave:** distribución VEG, censura aleatoria, partículas suspendidas, pruebas de bondad de ajuste.

# Analysis of Extreme Values in Presence of Random Censoring and Seasonality

Benigno Estrada Drouaillet

Colegio de Postgraduados, 2014

In this research work, a regression model was developed through the locality parameter of Generalized Extreme Value distribution (GEV) under random censoring scheme to analyze the trend over time of maximum daily values of air particle matter for the period of years 1995-2013 in the Metropolitan Zone of Mexico City (ZMCM). The records are studied belong to 11 stations Atmospheric Monitoring System of the City of Mexico (SIMAT). Data present levels of censoring between 10 % and 15 % due to faulty measuring equipment. Furthermore, the maximum values show a seasonal pattern so a sinusoid is incorporated location parameter. The estimation of the parameters of the GEV distribution is performed with the maximum likelihood method assuming random censorship. The estimated parameters of the GEV distribution is used to calculate return levels and generate contour maps to study the spatial pattern of  $PM_{10}$ . Monte Carlo study is also performed to analyze the effect of the level of censoring and the sample size in the bias parameters and return levels estimated. Finally, a modification of the Anderson-Darling statistic, and generalizations of the Kolmogorov-Smirnov and Cramér-von Mises statistics are used to test goodness of fit and identify among the tests used the more powerful and better to keep the size of the test.

**Key words:** GEV distribution, random censoring, air particle matter, goodness of fit tests.

## AGRADECIMIENTOS

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

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 e infinita paciencia, su apoyo incondicional hizo posible la culminación de este trabajo.

Dr. Sergio Pérez Elizalde, por su gran ayuda, su paciencia y su valioso tiempo brindados durante toda mi estancia.

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

Dr. Francisco J. Ariza Hernández, por sus sugerencias y comentarios, su apoyo brindado para la culminación de mis estudios.

Dr. Gerardo H. Terrazas González, por sus observaciones y tiempo dedicados de manera desinteresada en la revisión del presente trabajo.

Dr. David del Valle Paniagua, por su apoyo brindado para la culminación de mis estudios.

A cada uno de mis profesores que contribuyeron en mi formación, por su paciencia, motivación 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.

## DEDICATORIA

A **DIOS** por sus bendiciones, porque estas siempre a mi lado.

A mis PADRES *Irma* y *Benigno*, por su apoyo incondicional, amor y comprensión.  
A mis HERMANOS *Patricia* y *René* por su compañía, sus cuidados y apoyo en los momentos difíciles. A mi ESPOSA *Lucero*, por llegar en el mejor momento a mi vida y ayudarme a alcanzar mis metas.

# Índice

<b>1. Introducción</b>	<b>1</b>
<b>2. Objetivos</b>	<b>3</b>
<b>3. Marco teórico</b>	<b>4</b>
3.1. Teoría de Valores Extremos . . . . .	4
3.1.1. Máximos de bloques (Block maxima) . . . . .	5
3.2. Distribución de Valores Extremos Generalizada . . . . .	6
3.2.1. Función cuantil . . . . .	12
3.2.2. Estimación de parámetros . . . . .	12
3.2.3. Relación entre la distribución VEG y la distribución Pareto Generalizada (PG) . . . . .	14
3.3. Tipos de censura . . . . .	15
3.3.1. Censura por la derecha . . . . .	16
3.3.2. Censura aleatoria . . . . .	16
3.3.3. Censura por la izquierda . . . . .	17
3.3.4. Censura por intervalo . . . . .	17
<b>4. Estimación de la los parámetros de la distribución VEG bajo censura</b>	

<b>aleatoria</b>	<b>19</b>
4.1. Incorporando un patrón estacional . . . . .	21
4.2. Elección del modelo . . . . .	22
<b>5. Pruebas de bondad de ajuste para la distribución VEG bajo censura aleatoria</b>	<b>23</b>
5.1. Juego de hipótesis . . . . .	24
5.2. Estadísticos de prueba . . . . .	25
5.2.1. Prueba de Anderson-Darling modificada . . . . .	25
5.2.2. Generalización de la prueba de Cramér-von Mises . . . . .	26
5.2.3. Generalización de la prueba de Kolmogorov-Smirnov . . . . .	26
5.3. Algoritmo para generar una muestra artificial . . . . .	27
5.4. Aproximación de los valores críticos de los estadísticos de prueba . . . . .	28
5.5. Comparación de la potencia de las pruebas $A_U$ , $Z_C$ y $Z_K$ por simulación	32
5.6. Comparación del tamaño de las pruebas $A_U$ , $Z_C$ y $Z_K$ por simulación	36
<b>6. Modelación de máximos por bloque de <math>PM_{10}</math> en ZMCM usando la distribución VEG</b>	<b>39</b>
6.1. Partículas suspendidas ( $PM_{10}$ ) . . . . .	40
6.1.1. Efectos adversos de las partículas suspendidas ( $PM_{10}$ ) en la salud	41
6.2. Registro de niveles de $PM_{10}$ en ZMCM . . . . .	43
6.3. Modelación de las concentraciones máximas de $PM_{10}$ en ZMCM . . . . .	44
6.4. Estudio Monte Carlo . . . . .	47
<b>7. Conclusiones</b>	<b>49</b>

# Índice

---

<b>Referencias</b>	<b>50</b>
<b>Apéndice</b>	<b>56</b>
Apéndice A: Valores críticos de las pruebas de bondad de ajuste . . . . .	56
Apéndice B: Potencia de las pruebas de bondad de ajuste . . . . .	57
Apéndice C: Tamaño de las pruebas de bondad de ajuste . . . . .	61
Apéndice D: Códigos en R . . . . .	62

# Índice de tablas

6.1. Ubicación en de las estaciones de monitoreo analizadas . . . . .	39
6.2. Parámetros estimados del modelo VEG con covariable <i>year</i> para cada una de las estaciones de monitoreo . . . . .	45
6.3. Parámetros estimados del modelo VEG sin la covariable <i>year</i> para cada una de las estaciones de monitoreo . . . . .	45
6.4. Estimación del cuantil 0.95 del modelo VEG con la covariable <i>year</i> para cada una de las estaciones de monitoreo . . . . .	46
.1. Valores críticos ( $C_{n,lc}^{A_U}(\alpha)$ ) de la prueba $A_U$ con $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ ) . . . . .	56
.2. Valores críticos ( $C_{n,lc}^{Z_K}(\alpha)$ ) de la prueba $Z_K$ con $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ ) . . . . .	56
.3. Valores críticos ( $C_{n,lc}^{Z_C}(\alpha)$ ) de la prueba $Z_C$ con $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ ) . . . . .	56
.4. Potencia de la prueba $A_U$ para alternativa <i>Dagum</i> (15.26, 12.41, 18.98) con $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ ) . . . . .	57
.5. Potencia de la prueba $Z_K$ para alternativa <i>Dagum</i> (15.26, 12.41, 18.98) con $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ ) . . . . .	57
.6. Potencia de la prueba $Z_C$ para alternativa <i>Dagum</i> (15.26, 12.41, 18.98) con $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ ) . . . . .	57

.7.	Potencia de la prueba $A_U$ para alternativa $Gamma(121.99, 7.78)$ con $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ ) . . . . .	58
.8.	Potencia de la prueba $Z_K$ para alternativa $Gamma(121.99, 7.78)$ con $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ ) . . . . .	58
.9.	Potencia de la prueba $Z_C$ para alternativa $Gamma(121.99, 7.78)$ con $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ ) . . . . .	58
.10.	Potencia de la prueba $A_U$ para alternativa $Log - normal(2.74, 0.08)$ con $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ ) . . . . .	59
.11.	Potencia de la prueba $Z_K$ para alternativa $Log - normal(2.74, 0.08)$ con $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ ) . . . . .	59
.12.	Potencia de la prueba $Z_C$ para alternativa $Log - normal(2.74, 0.08)$ con $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ ) . . . . .	59
.13.	Potencia de la prueba $A_U$ para alternativa $Weibull(7.90, 16.39)$ con $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ ) . . . . .	60
.14.	Potencia de la prueba $Z_K$ para alternativa $Weibull(7.90, 16.39)$ con $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ ) . . . . .	60
.15.	Potencia de la prueba $Z_C$ para alternativa $Weibull(7.90, 16.39)$ con $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ ) . . . . .	60
.16.	Tamaño de la prueba $A_U$ con $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ ) . . . . .	61
.17.	Tamaño de la prueba $Z_K$ con $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ ) . . . . .	61
.18.	Tamaño de la prueba $Z_C$ con $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ ) . . . . .	61

# Índice de figuras

3.1. (a) máximo de bloque y (b) máximo con censura aleatoria de $PM_{10}$ . . . . .	6
5.1. Distribución de la estadística de prueba $A_U$ con $n = 350$ y diferentes niveles de censura . . . . .	30
5.2. Distribución de la estadística de prueba $Z_C$ con $n = 350$ y diferentes niveles de censura . . . . .	30
5.3. Distribución de la estadística de prueba $Z_K$ con $n = 350$ y diferentes niveles de censura . . . . .	31
5.4. Potencia de la prueba $A_U$ , $Z_C$ y $Z_K$ con $H1 : Dagum(15.26, 12.41, 18.98)$	33
5.5. Potencia de la prueba $A_U$ , $Z_C$ y $Z_K$ con $H1 : Gamma(121.99, 7.78)$ . . . . .	34
5.6. Potencia de la prueba $A_U$ , $Z_C$ y $Z_K$ con $H1 : log - normal(2.74, 0.08)$	35
5.7. Potencia de la prueba $A_U$ con $H1 : Weibull(7.90, 16.39)$ . . . . .	36
5.8. Tamaño de la prueba $A_U$ . . . . .	37
5.9. Tamaño de la prueba $Z_C$ . . . . .	37
5.10. Tamaño de la prueba $Z_K$ . . . . .	38
6.1. Tamaño de las fracciones de material particulado . . . . .	40
6.2. Frecuencias por hora de los niveles máximos de $PM_{10}$ . . . . .	41
6.3. Material particulado en el sistema respiratorio . . . . .	42

## Índice de figuras

---

6.4. Máximos de bloque de $PM_{10}$ registrados en las estaciones de monitoreo localizadas en ZMCM . . . . .	44
6.5. Estimación del cuantil 0.95 del modelo VEG con covariable <i>year</i> para cada una de las estaciones de monitoreo para los años 1995, 2005 y 2013	46
6.6. Porcentaje de sesgo de EMV en relación a sus valores verdaderos $\mu = 150$ , $\sigma = 70$ y $\xi = 0.25$ bajo el esquema de censura aleatoria . . . . .	47
6.7. Porcentaje de sesgo de los cuantiles estimados en relación a sus valores verdaderos $q_{90} = 361.46$ , $q_{95} = 458.35$ y $q_{99} = 754.32$ bajo el esquema de censura aleatoria . . . . .	48

# Capítulo 1

## Introducción

En muchas aplicaciones estadísticas, el interés se enfoca en estimar características centrales de la población (temperatura promedio, ingreso medio etc.) basadas en muestras aleatorias de una población bajo estudio. Sin embargo, en algunas áreas de estudio no se tiene interés en estimar promedios, sino que este se centra en estimar características máximas o mínimas (extremas) de la población o de algún fenómeno. Enseguida se muestran aplicaciones en distintas áreas de estudio que motivan el análisis de los valores extremos.

En el área de ingeniería oceánica, se sabe que la altura de las olas es el factor principal a considerar para propósitos de diseño. Así, el diseño de plataformas marinas, diques, y otras obras portuarias dependen del conocimiento de la distribución de probabilidad de las olas más altas. En ingeniería estructural, un edificio debe resistir al temblor más intenso que ocurra durante el periodo para el cual se diseña. Por lo que se requieren de estimaciones precisas del evento sísmico de mayor magnitud para designar márgenes de seguridad realistas en el diseño de la estructura. Por otro lado, se sabe que las condiciones meteorológicas extremas pueden intervenir en aspectos que influyen la vida humana como el desarrollo próspero de la agricultura y ganadería, el desempeño y operación de algunas máquinas, los tiempos de vida de ciertos materiales, de tal forma que en lugar de estudiar los valores medios (temperatura, lluvia, velocidad del viento, etc.), se está interesado en la ocurrencia de eventos extremos (temperaturas muy altas o muy bajas, precipitaciones extremas, ciclones, etc.). Más aún, con la existencia de grandes concentraciones de personas o la aparición de nuevas industrias, la contaminación del aire, ríos y costas se ha convertido en un problema común para varios países. La concentración de los contaminantes, que se expresa como la cantidad del contaminante por unidad de volumen, se encuentra regulada por normas gubernamentales para que se mantenga por debajo de cierto nivel crítico. Así, las regulaciones sólo son satisfechas, si la concentración más alta del contaminante durante el periodo de interés no sobrepasa el nivel crítico. En consecuencia, el comportamiento de los valores máximos del contaminante es fundamental en los estudios de contaminación.

## 1. Introducción

---

En [Laurens de Hann \(2006\)](#) y [Beirlant \*et al.\* \(2004\)](#) puede encontrar otros ejemplos de interés.

Cuando se tienen tamaños de muestra grandes, en la práctica por lo general el análisis de los valores extremos se suele realizar a través de dos metodologías. La primera se refiere al uso de la distribución de Valores Extremos Generalizada (VEG), en este caso se modelan las observaciones más grandes de muestras del mismo tamaño, es decir, máximos de bloque (en inglés *block maxima*). La segunda es mediante el uso de la distribución Pareto Generalizada (PG), la cual se utiliza para modelar las observaciones que sobrepasan un límite previamente definido, es decir, picos sobre el umbral (*peaks-over-threshold*).

Este trabajo se enfoca en un estudio del medio ambiente, específicamente se analizan las concentraciones máximas de partículas suspendidas con diámetro menor a 10 micrómetros ( $\mu m$ ),  $PM_{10}$ , de 11 estaciones de monitoreo que pertenecen a la RAMA (Red Automática de Monitoreo Atmosférico). El análisis se lleva a cabo con la distribución VEG, el tamaño de bloque se determina con la función de autocorrelación parcial (FAP) de manera que está sea prácticamente cero en todos los rezagos. Los datos que se analizan presentan un comportamiento estacional por lo que se incorpora una componente sinusoidal en el modelo propuesto, además existen lecturas perdidas que impiden determinar el valor máximo de algunos bloques por lo que el proceso de estimación se realiza bajo el esquema de censura aleatoria.

Las distribuciones límite para el análisis de valores extremos son propuestas por [Fréchet \(1927\)](#), [Fisher y Tippett \(1928\)](#). Sin embargo, [Gnedenko \(1943\)](#) desarrolla una demostración formal para el teorema de Fisher y Tippett, la cual proporciona las condiciones necesarias y suficientes para obtener las leyes límite. Posteriormente, [von Mises \(1954\)](#) y [Jenkinson \(1955\)](#) proponen de manera independiente la distribución VEG que contiene las tres distribuciones límite de valores extremos, Weibull, Gumbel y Fréchet.

La distribución VEG tiene tres parámetros:  $\mu$ , el parámetro de localización;  $\sigma$ , el parámetro de escala; y  $\xi$ , el parámetro de forma. En este análisis, se desarrolla un modelo de regresión a través del parámetro de localización para estudiar la tendencia en el tiempo de las concentraciones máximas de  $PM_{10}$  en la Zona Metropolitana de la Ciudad de México (ZMCM) e identificar las zonas con problemas más severos de altas concentraciones de  $PM_{10}$ . Posteriormente se analiza el efecto del nivel de censura y el tamaño de muestra en el sesgo de los parámetros y de los niveles de retorno estimados. Finalmente, se emplea una modificación del estadístico de Anderson-Darling, y generalizaciones de los estadísticos Kolmogorov-Smirnov y Cramér-von Mises para probar bondad de ajuste e identificar de entre las pruebas empleadas la de mayor potencia y la que conserve de mejor manera el tamaño de la prueba.

# Capítulo 2

## Objetivos

- Proponer una metodología estadística para investigar tendencias temporales en valores extremos bajo el esquema de censura aleatoria.
- Estudiar el sesgo de los parámetros y niveles de retorno en función del nivel de censura y el tamaño de muestra.
- Proponer una prueba de bondad de ajuste basada en los estadísticos de Anderson-Darling, Kolmogorov-Smirnov y Cramér-von Mises para la distribución de Valores Extremos Generalizada bajo censura aleatoria, así como estudiar la potencia y el tamaño de la prueba.
- Identificar la tendencia de las concentraciones máximas de  $PM_{10}$  en el Valle de México así como determinar las zonas que presentan mayores problemas de altas concentraciones de  $PM_{10}$ .

# Capítulo 3

## Marco teórico

### 3.1. Teoría de Valores Extremos

Gran parte de los estudios estadísticos tienen como principal interés hacer inferencia sobre 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 que se encarga de realizar 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.

Los valores extremos son valores que se observan con poca frecuencia en diversos estudios y fenómenos naturales. Si bien es cierto que su probabilidad de ocurrencia es relativamente baja, el alto impacto que generan en el entorno han motivado desde hace tiempo su estudio. Generalmente, a los valores extremos se les identifica como “outliers” en los análisis clásicos de información y en muchos de los casos son ignorados u omitidos.

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 una 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. Un primer avance teórico lo desarrollan [Fréchet \(1927\)](#), [Fisher y Tippett \(1928\)](#) quienes derivan las distribuciones límite para el análisis de valores extremos. En la década de 1930, se desarrollaron trabajos de gran interés, [Finetti \(1932\)](#), [Gumbel \(1934, 1935a,b\)](#) y [Rice \(1939\)](#), sobre la distribución de extremos de una muestra. Sin embargo, es gracias al trabajo de [Gnedenko \(1943\)](#) que se tiene

### 3.1. Teoría de Valores Extremos

---

una demostración formal para el teorema de Fisher y Tipett, la cual proporciona las condiciones necesarias y suficientes para obtener las leyes límite. Posteriormente, [von Mises \(1954\)](#) y [Jenkinson \(1955\)](#) proponen de manera independiente la distribución VEG que contiene las tres distribuciones límite de valores extremos, Weibull, Gumbel y Fréchet. El primer libro de importancia que trabajó con estadísticas de extremos y que por varios años fue la principal referencia sobre el tema, es obra de [Gumbel \(1954\)](#).

La teoría de valores extremos tiene importantes aplicaciones en diversas áreas del conocimiento, como la ciencia del medio ambiente (nivel de la precipitación, velocidad del viento, concentración de contaminantes), en oceanografía (altura de olas, velocidad de corrientes marinas), en climatología (temperaturas extremas, velocidad de huracanes), en finanzas (riesgo de aseguradoras ante grandes siniestros), en hidrología (nivel de ríos o presas), en ingeniería (resistencia de construcciones ante sismos de alta intensidad), etc.. Hoy en día está disponible una amplia literatura sobre la teoría y aplicaciones de valores extremos, por ejemplo, [Smith \(1984\)](#), [Castillo y Hadi \(1994\)](#), [Kotz y Nadarajah \(2000\)](#), [Kalbfleisch y Prentice \(2002\)](#), [Beirlant \*et al.\* \(2004\)](#), [Coles \(2001\)](#), [Nakajima \*et al.\* \(2012\)](#).

#### 3.1.1. Máximos de bloques (Block maxima)

La distribución VEG es típicamente usada en la metodología llamada “block maxima” 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, etc., se pueden construir bloques por año, por meses o diarios dependiendo de la cantidad de información disponible y de la dependencia entre las observaciones.

La metodología consiste en dividir la serie de tiempo que contiene a las observaciones en periodos (bloques) de igual tamaño y que no se traslapen, dentro de cada bloque se calcula el máximo (mínimo), el método es descrito por [Gaines y Denny \(1993\)](#). Como complemento, se calcula la función de autocorrelación parcial (FACP) para comprobar aleatoriedad e independencia de los máximos de bloque obtenidos.

En la Figura [3.1\(a\)](#) se calcula el máximo de un bloque con datos de concentraciones de PM<sub>10</sub>. Sin embargo, en ocasiones no es posible calcular el máximo de bloque por falta de información como se muestra en la Figura [3.1\(b\)](#) en tal caso se calcula un máximo parcial y se dice que se tiene una observación censurada.

## 3.2. Distribución de Valores Extremos Generalizada

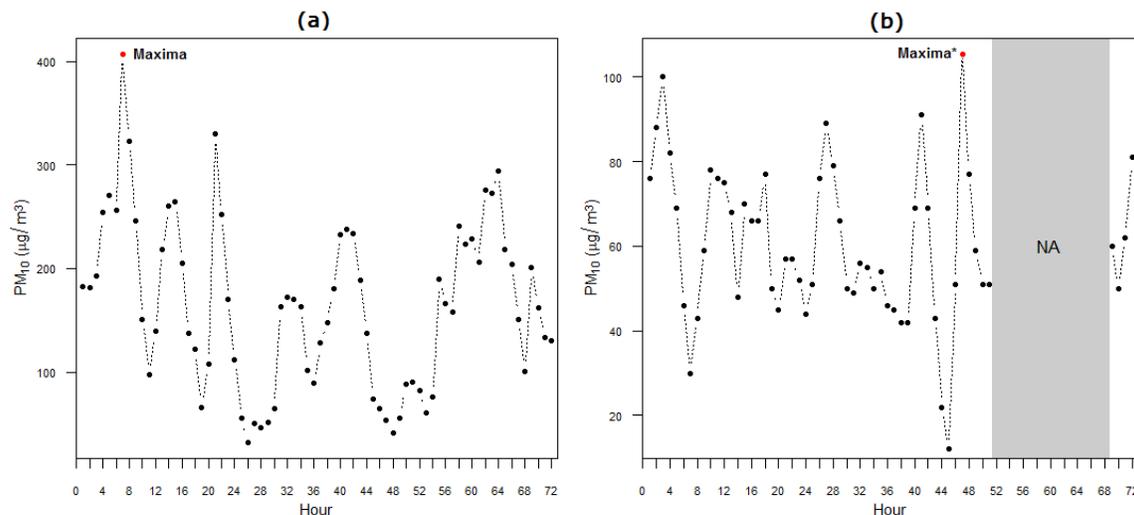


Figura 3.1: (a) máximo de bloque y (b) máximo con censura aleatoria de  $PM_{10}$

## 3.2. Distribución de Valores Extremos Generalizada

De entre los estadísticos de orden, el mínimo y el máximo son los de mayor interés en las aplicaciones. Sin embargo, es posible analizar los valores extremos sólo en términos de máximos de distribuciones (cola superior) debido a que los mínimos (cola inferior) obedecen la siguiente relación

$$\min(Y_1, Y_2, \dots, Y_n) = -\max(-Y_1, -Y_2, \dots, -Y_n)$$

para una muestra de observaciones  $Y_1, Y_2, \dots, Y_n$ .

En teoría la distribución del máximo  $M_n$  se puede derivar de manera exacta para todos los valores de  $n$ :

$$\begin{aligned} P[M_n \leq y] &= P[Y_1 \leq y, \dots, Y_n \leq y] \\ &= [F(y)]^n. \end{aligned} \tag{3.1}$$

Sin embargo, este resultado no es de mucha ayuda en la práctica, ya que la función de distribución  $F$  se desconoce. Una posibilidad es estimar  $F$  de los datos observados y sustituir este estimador en (3.1). Desafortunadamente, diferencias muy pequeñas en la estimación de  $F$  pueden conducir a diferencias considerables de  $F^n$ .

## 3.2. Distribución de Valores Extremos Generalizada

---

Un enfoque alternativo es aceptar que  $F$  es desconocida y buscar familias de modelos para aproximar a  $F^n$ , la cual se puede estimar a partir de los datos extremos solamente. Esto es similar a la práctica usual de aproximar la distribución de la media muestral por la distribución normal, justificada por el teorema de límite central. En esta sección se utilizan argumentos análogos a los de la teoría de límite central para encontrar las distribuciones límite  $F^n$ .

Cuando  $n$  tiende a infinito, se tiene

$$\lim_{n \rightarrow \infty} F^n(y) = \lim_{n \rightarrow \infty} [F(y)]^n = \begin{cases} 1, & \text{si } F(y) = 1 \\ 0, & \text{si } F(y) < 1. \end{cases}$$

Esto significa que la distribuciones límite son degeneradas (sólo toma valores 0 y 1). Para evitar este problema se utiliza una normalización lineal de la variable  $M_n$ :

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

para sucesiones de constantes  $\{a_n \geq 0\}$  y  $\{b_n\}$ . Elegir apropiadamente las  $\{a_n\}$  y  $\{b_n\}$  estabiliza la localidad y escala de  $M_n^*$  cuando  $n$  se incrementa, evitando las dificultades que surgen con la variable  $M_n$ . Por lo tanto, el objetivo es encontrar las distribuciones límite para  $M_n^*$ , con elecciones apropiadas de  $\{a_n\}$  y  $\{b_n\}$ .

Todas las posibles distribuciones límite para  $M_n^*$  están dadas por el Teorema 3.1.

**Teorema 3.1** *Si existen sucesiones de constantes  $\{a_n \geq 0\}$  y  $\{b_n\}$  tal que*

$$P \left[ \frac{M_n - b_n}{a_n} \leq y \right] \rightarrow G(y) \text{ cuando } n \rightarrow \infty,$$

*donde  $G$  es una función de distribución no degenerada, entonces  $G$  pertenece a una de las siguientes familias:*

## 3.2. Distribución de Valores Extremos Generalizada

---

$$\text{I : } G(y) = \exp \left\{ - \exp \left[ - \left( \frac{y-b}{a} \right) \right] \right\}, \quad -\infty < y < \infty; \quad (3.2)$$

$$\text{II : } G(y) = \begin{cases} 0, & y \leq b, \\ \exp \left[ - \left( \frac{y-b}{a} \right)^{-\alpha} \right], & y > b; \end{cases} \quad (3.3)$$

$$\text{III : } G(y) = \begin{cases} \exp \left\{ - \left[ - \left( \frac{y-b}{a} \right)^\alpha \right] \right\}, & y < b, \\ 1, & y \geq b, \end{cases} \quad (3.4)$$

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

En palabras, el Teorema 3.1 indica que la muestra de máximos reescalados  $(M_n - b_n)/a_n$  converge en distribución a una variable que se distribuye de acuerdo a una de las familias I, II o III. En conjunto, a estas tres clases de distribuciones se les denomina como las **distribuciones de valores extremos** tipo I, II y III ampliamente conocidas como las familias **Gumbel**, **Fréchet** y **Weibull** respectivamente. Cada familia tiene un parámetro de localidad y escala,  $b$  y  $a$  respectivamente; además, las familias Fréchet y Weibull tienen un parámetro de forma  $\alpha$ .

El Teorema 3.1 implica que, cuando  $M_n$  se puede estabilizar con sucesiones apropiadas  $\{a_n\}$  y  $\{b_n\}$ , la correspondiente variable normalizada  $M_n^*$  tiene una distribución límite que pertenece a uno de los tres tipos de distribuciones de valores extremos. La característica a resaltar de este resultado es que los tres tipos de distribuciones de valores extremos son los únicos límites posibles para las distribuciones de  $M_n^*$ , sin importar la distribución  $F$  de la población. Es en este sentido que el Teorema 3.1 proporciona una analogía del teorema de límite central.

Los tres tipos de distribuciones límite que resultan del Teorema 3.1 tienen distintos comportamientos de la cola. Sea  $y_+$  el valor más pequeño de  $y$  tal que  $G(y) = 1$ . Entonces, para la distribución Weibull  $y_+$  es finito, mientras que para las distribuciones Fréchet y Gumbel  $y_+ = \infty$ . Sin embargo, la densidad de  $G$  decae exponencialmente para la distribución Gumbel y polinomialmente para la distribución Fréchet. Es así, que en las aplicaciones las tres familias pueden cubrir un amplio rango de comportamientos de valores extremos. En las primeras aplicaciones de la teoría de valores extremos, era usual adoptar una de las tres familias, y entonces estimar los parámetros de esa distribución. Pero hay dos desventajas; primero, se requiere de una técnica para elegir cual de las tres familias es la más apropiada para los datos que se analizan; segundo, una vez que se ha tomado una decisión, la inferencia posterior supone que esta elección es correcta, y no da lugar a la incertidumbre que implica una selección de este tipo, a pesar de que esta incertidumbre pueda ser sustancial.

Se tiene un mejor análisis al reformular los modelos del Teorema 3.1. [von Mises \(1954\)](#) y [Jenkinson \(1955\)](#) se dieron a la tarea de encontrar una familia que pudiera incluir

## 3.2. Distribución de Valores Extremos Generalizada

---

a las familias Gumbel, Fréchet y Weibull. Esta familia más general tiene función de distribución de la forma

$$G(y; \mu, \sigma, \xi) = \exp \left\{ - \left[ 1 + \xi \left( \frac{y - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}} \right\}, \quad (3.5)$$

que se define en el conjunto  $\{y : 1 + \xi(y - \mu)/(\sigma) > 0\}$ , donde los parámetros satisfacen  $\mu \in \mathbb{R}$ ,  $\sigma > 0$ ,  $\xi \in \mathbb{R}$ . Esta familia se conoce como la distribución de **valores extremos generalizada** (VEG). En adelante si una variable aleatoria (v.a.)  $Y$  sigue esta distribución entonces se denotará por  $Y \sim G(y; \mu, \sigma, \xi)$ . El modelo tiene tres parámetros: un parámetro de localización,  $\mu$ ; un parámetro de escala,  $\sigma$ ; y un parámetro de forma,  $\xi$ . Las distribuciones de valores extremos tipo II y tipo III corresponden a los casos  $\xi > 0$  y  $\xi < 0$  respectivamente. Entonces, bajo esta parametrización la distribución Weibull se encuentra acotada en  $y_+ = \mu - (\sigma/\xi)$ . El subconjunto de la familia VEG con  $\xi = 0$  se interpreta como el límite de (3.5) cuando  $\xi \rightarrow 0$ , que conduce a la **familia Gumbel** con función de distribución

$$G(y; \mu, \sigma) = \exp \left[ - \exp \left\{ - \left( \frac{y - \mu}{\sigma} \right) \right\} \right], \quad y \in \mathbb{R}. \quad (3.6)$$

La agrupación de las tres familias de distribuciones de valores extremos en una sola familia simplifica la implementación estadística. Por medio de la inferencia sobre  $\xi$ , los datos por sí mismos determinan el tipo de comportamiento de la cola más apropiado, y no hay necesidad de realizar juicios a priori respecto a cual de las tres familias de valores extremos se debería elegir. Más aún, la incertidumbre en el valor inferido  $\xi$  mide la falta de certeza en cuanto a cuál de las tres familias es más apropiada para un conjunto de datos dado.

Por conveniencia se replantea el Teorema 3.1 de la siguiente manera:

**Teorema 3.2** *Si existen sucesiones de constantes  $\{a_n \geq 0\}$  y  $\{b_n\}$  tal que*

$$P \left[ \frac{M_n - b_n}{a_n} \leq y \right] \rightarrow G(y) \quad \text{cuando } n \rightarrow \infty, \quad (3.7)$$

*donde  $G$  es una función de distribución no degenerada, entonces  $G$  es miembro de la familia VEG*

### 3.2. Distribución de Valores Extremos Generalizada

---

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

que se define en  $\{y : 1 + \xi(y - \mu)/(\sigma) > 0\}$ , donde  $\mu \in \mathbb{R}$ ,  $\sigma > 0$ ,  $\xi \in \mathbb{R}$ .

Interpretar el límite en el Teorema 3.2 como una aproximación para valores grandes de  $n$  sugiere el uso de la familia VEG para modelar la distribución de secuencias grandes de máximos. La aparente dificultad sobre el desconocimiento de las constantes de normalización en la práctica se resuelve fácilmente. Permita la relación (3.7),

$$P \left[ \frac{M_n - b_n}{a_n} \leq y \right] \approx G(y)$$

para  $n$  suficientemente grande. De manera equivalente,

$$\begin{aligned} P[M_n \leq y] &\approx G\{(y - b_n)/a_n\} \\ &= G^*(y), \end{aligned} \tag{3.8}$$

donde  $G^*$  es otro miembro de la familia VEG. En otras palabras, si el Teorema 3.2 permite aproximar la distribución de  $M_n^*$  por un miembro de la familia VEG para  $n$  grande, la distribución de  $M_n$  también se puede aproximar por otro miembro de la misma familia. Debido a que los parámetros de la distribución se tienen que estimar, es irrelevante en la práctica que los parámetros de la distribución  $G$  sean diferentes a los de la distribución  $G^*$ .

Este argumento permite el siguiente enfoque para modelar extremos de una serie de observaciones independientes  $X_1, X_2, \dots$ . Los datos son agrupados en sucesiones de observaciones de longitud  $n$ , para algún valor grande de  $n$ , generando una serie de máximos de bloque,  $M_{n,1}, \dots, M_{n,m}$ , a los cuales se les puede ajustar la distribución de VEG. Con frecuencia los bloques se escogen de tal manera que correspondan a un periodo de longitud de un año, en tal caso  $n$  es el número de observaciones en un año y  $m$  es el número de máximos de bloque que corresponden a máximos anuales.

De acuerdo a Monroy (2010) si una v.a.  $Y$  tiene distribución VEG, entonces la variable estandarizada  $(Y - \mu)/\sigma$  tiene una distribución que no depende de  $\mu$  ni de  $\sigma$ , sino únicamente de  $\xi$ . El parámetro de localidad determina donde está centrada la distribución, el parámetro de escala su propagación, el parámetro de forma está asociado al espesor de la cola de la distribución, en cuanto más grande sea el valor de  $\xi$  más pesada es la cola de la distribución.

### 3.2. Distribución de Valores Extremos Generalizada

---

Otras características de la distribución VEG son las siguientes:

- Esperanza matemática

$$E[Y] = \begin{cases} \mu + \sigma \frac{\Gamma(1-\xi)-1}{\xi} & , \text{ si } \xi \neq 0, \xi < 0 \\ \mu + \sigma\gamma & , \text{ si } \xi = 0 \\ \text{no existe} & , \text{ si } \xi \geq 1 \end{cases} \quad (3.9)$$

donde  $\Gamma(\cdot)$  es la función Gamma, y  $\gamma$  es la constante de Euler.

- Varianza

$$Var[Y] = \begin{cases} \sigma^2 \frac{g_2 - g_1^2}{\xi^2} & , \text{ si } \xi \neq 0, \xi < 1/2 \\ \sigma^2 \frac{\pi^2}{6} & , \text{ si } \xi = 0 \\ \text{no existe} & , \text{ si } \xi \geq 1/2 \end{cases} \quad (3.10)$$

donde  $g_k = \Gamma(1 - k\xi)$ .

- Mediana

$$M_e[Y] = \begin{cases} \mu + \sigma \frac{\ln(2)^{-\xi} - 1}{\xi} & , \text{ si } \xi \neq 0 \\ \mu - \sigma \ln(\ln(2)) & , \text{ si } \xi = 0 \end{cases} \quad (3.11)$$

- Moda

$$M[Y] = \begin{cases} \mu + \sigma \frac{(1+\xi)^{-\xi} - 1}{\xi} & , \text{ si } \xi \neq 0 \\ \mu & , \text{ si } \xi = 0 \end{cases} \quad (3.12)$$

- Coeficiente de asimetría

$$C_a[Y] = \begin{cases} \frac{g_3 - 3g_1g_2 + 2g_1^3}{(g_2 - g_1^2)^{3/2}} & , \text{ si } \xi \neq 0 \\ \frac{12\sqrt{6}\zeta(3)}{\pi^3} & , \text{ si } \xi = 0 \end{cases} \quad (3.13)$$

donde  $\zeta$  es la función zeta de Riemann.

- Coeficiente de curtosis

$$C_c[Y] = \begin{cases} \frac{g_4 - 4g_1g_3 + 6g_2g_1^2 - 3g_1^4}{(g_2 - g_1^2)^2} - 3 & , \text{ si } \xi \neq 0 \\ 12/5 & , \text{ si } \xi = 0 \end{cases} \quad (3.14)$$

En lo que se refiere a la dependencia entre observaciones, [Leadbetter et al. \(1983\)](#) indica que la distribución límite del máximo sigue siendo VEG bajo una amplia gama de condiciones de dependencia (por ejemplo, un proceso autoregresivo) y únicamente habría efectos sobre los parámetros de localidad y escala.

## 3.2. Distribución de Valores Extremos Generalizada

---

### 3.2.1. Función cuantil

Es natural que exista interés en los cuantiles extremos de la distribución VEG, estos se obtienen por medio de invertir (3.5):

$$y_p = \begin{cases} \mu - \frac{\sigma}{\xi} [1 - \{-\log(1-p)\}^{-\xi}] & , \xi \neq 0, \\ \mu - \sigma \log \{-\log(1-p)\} & , \xi = 0, \end{cases} \quad (3.15)$$

y reemplazar los parámetros desconocidos por sus respectivos estimadores, donde  $G(y_p) = 1 - p$  con  $0 < p < 1$  (Beirlant *et al.*, 2004). En terminología común,  $y_p$  es el **nivel de retorno** asociado con el **periodo de retorno**  $1/p$ , para un razonable valor de precisión, se espera que el nivel  $y_p$  sea rebasado en promedio una vez cada  $1/p$  años. Es decir,  $y_p$  es superado por el máximo anual en un año en particular con probabilidad  $p$ .

Debido a que los cuantiles permiten expresar los modelos de probabilidad en la escala de los datos, la relación entre el modelo VEG y sus parámetros se interpreta con mayor facilidad por medio de la expresión de cuantiles (3.15). En particular, si se define  $z_p = -\log(1-p)$ , de modo que

$$y_p = \begin{cases} \mu - \frac{\sigma}{\xi} [1 - z_p^{-\xi}] & , \xi \neq 0, \\ \mu - \sigma \log z_p & , \xi = 0, \end{cases} \quad (3.16)$$

entonces, si se grafica  $y_p$  contra  $z_p$  en escala logarítmica - o equivalentemente, si se grafica  $y_p$  contra  $z_p$  - el gráfico es lineal cuando  $\xi = 0$ . Si  $\xi < 0$  el gráfico es convexo con límite asintótico en  $\mu - \sigma/\xi$  cuando  $p \rightarrow 0$ ; si  $\xi > 0$  el gráfico es cóncavo y no tiene un límite finito. Este gráfico corresponde a un **gráfico de nivel de retorno**. Debido a la simplicidad de la interpretación, y debido a que la elección de la escala comprime la cola de la distribución de manera que el efecto de extrapolación se resalta, los gráficos de nivel de retorno son apropiados para la representación y la validación del modelo.

### 3.2.2. Estimación de parámetros

Motivados por el Teorema 3.2, la distribución VEG proporciona un modelo para la distribución de los máximos de bloque. Su aplicación consiste en construir bloques de igual longitud a partir de los datos originales. Pero al implementar este modelo para cualquier conjunto de datos, la elección del tamaño de bloque puede ser crítica. La elección del tamaño equivale a un compromiso entre el sesgo y la varianza: bloques muy pequeños dan lugar a que la aproximación por el modelo límite del Teorema 3.2

## 3.2. Distribución de Valores Extremos Generalizada

---

sea en general pobre, lo cual genera estimaciones y extrapolaciones sesgadas; bloques grandes producen pocos máximos de bloque, que se traduce en una sobreestimación de la varianza.

Por simplicidad los máximos de bloque se denotarán por  $Y_1, \dots, Y_n$ . Si las variables originales  $X_i$  son independientes entonces los  $Y_i$  también son independientes. Sin embargo, la independencia de los  $Y_i$  es razonable aún si las  $X_i$  constituyen una serie dependiente. En este caso, aunque no considerado por el Teorema 3.2, la resolución de que los  $Y_i$  tiene una distribución VEG es todavía razonable; ver Coles (2001).

Los parámetros de la distribución VEG se pueden obtener por el método de máxima verosimilitud. Sin embargo, este método tiene una dificultad potencial en relación a las condiciones de regularidad que son necesarias para que las propiedades asintóticas asociadas a los estimadores de máxima verosimilitud (EMV) sean válidas. Estas condiciones no son satisfechas por el modelo VEG debido a que los puntos extremos de la distribución VEG son funciones de los parámetros:  $\mu - \sigma/\xi$  es el extremo derecho de la distribución cuando  $\xi < 0$ , y es el extremo izquierdo cuando  $\xi > 0$ . Esta violación de las condiciones de regularidad provoca que los resultados asintóticos de la verosimilitud no se puedan aplicar en automático. Smith (1985) estudió este problema en detalle y obtuvo los siguientes resultados:

- cuando  $\xi > -0.5$ , los EMV son regulares, en el sentido de que tienen las propiedades asintóticas usuales;
- cuando  $-1 < \xi < -0.5$ , los EMV en general se pueden obtener, pero no tienen las propiedades asintóticas;
- cuando  $\xi < -1$ , los EMV no se pueden obtener.

El caso  $\xi \leq -0.5$  corresponde a distribuciones con la cola derecha muy corta. Esta situación es difícil de encontrar en aplicaciones de valores extremos, así las limitaciones teóricas del método de máxima verosimilitud no son un obstáculo en la práctica.

Bajo el supuesto de que  $Y_1, \dots, Y_n$  son v.a. independientes que tienen distribución VEG, la log-verosimilitud para los parámetros del modelo VEG cuando  $\xi \neq 0$  es

$$\ell(\mu, \sigma, \xi) = -n \log \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^n \log \left[1 + \xi \left(\frac{y_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^n \left[1 + \xi \left(\frac{y_i - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}, \quad (3.17)$$

siempre que

$$1 + \xi \left(\frac{y_i - \mu}{\sigma}\right) > 0, \quad \text{para } i = 1, \dots, n. \quad (3.18)$$

## 3.2. Distribución de Valores Extremos Generalizada

---

Las combinaciones de parámetros que no satisfacen (3.18), corresponden a una configuración para la cual al menos uno de los datos observados se encuentra fuera del soporte de la distribución, la verosimilitud es cero y la log-verosimilitud es  $-\infty$ .

El caso  $\xi = 0$  requiere un tratamiento por separado utilizando el límite Gumbel de la distribución VEG. Esto conduce a la log-verosimilitud

$$\ell(\mu, \sigma) = -n \log \sigma - \sum_{i=1}^n \left( \frac{y_i - \mu}{\sigma} \right) - \sum_{i=1}^n \exp \left\{ - \left( \frac{y_i - \mu}{\sigma} \right) \right\}. \quad (3.19)$$

Por medio de maximizar (3.17) y (3.19) con respecto al vector de parámetros  $(\mu, \sigma, \xi)$  se obtienen los EMV de la familia completa de VEG. No hay una solución analítica, pero dado un conjunto de datos la maximización es sencilla por medio de utilizar algoritmos de optimización numérica. En el programa estadístico R esta disponible la librería VGAM por medio de la cual se pueden obtener dichos estimadores de manera numérica.

### 3.2.3. Relación entre la distribución VEG y la distribución Pareto Generalizada (PG)

**Teorema 3.3** Sean  $X_1, X_2, \dots$  una sucesión de v.a. independientes con función de distribución común  $F$ , y sea

$$M_n = \max\{X_1, \dots, X_n\}.$$

Suponga que  $F$  satisface el Teorema 3.2, así que para  $n$  grande,

$$P[M_n \leq y] \approx G(y),$$

donde

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

para  $\mu \in \mathbb{R}$ ,  $\sigma > 0$ ,  $\xi \in \mathbb{R}$ . Entonces, para  $u$  suficientemente grande, la función de distribución de  $(X - u)$ , condicionada en  $X > u$ , se aproxima por

### 3.3. Tipos de censura

---

$$H(z) = 1 - \left(1 + \frac{\xi(z - \mu)}{\tilde{\sigma}}\right)^{-1/\xi} \quad (3.20)$$

con soporte en el conjunto  $\{z : z > 0 \text{ y } (1 + \xi z/\tilde{\sigma}) > 0\}$ , donde

$$\tilde{\sigma} = \sigma + \xi(u - \mu). \quad (3.21)$$

El Teorema 3.3 puede ser más preciso, estableciendo a (3.20) como una distribución límite cuando  $u$  se incrementa.

La familia de distribuciones definidas en (3.20) se conoce como la **familia Pareto generalizada** (PG). El Teorema 3.3 implica que, si los máximos de bloque tienen distribución asintótica  $G$ , entonces los excedentes sobre el umbral tienen una correspondiente distribución asintótica dentro de la familia PG. Más aún, los parámetros de la distribución PG se pueden determinar por los parámetros asociados a la distribución VEG. En particular, el parámetro  $\xi$  en (3.20) es igual al parámetro correspondiente de la distribución VEG. Escogiendo un diferente, pero todavía grande, tamaño de bloque  $n$  debería afectar los valores de los parámetros de la distribución VEG, pero no los parámetros correspondientes a la distribución PG:  $\xi$  es invariante a los tamaños de bloque, mientras que el cálculo de  $\tilde{\sigma}$  en (3.21) es imperturbable por los cambios en  $\mu$  y  $\sigma$  los cuales se autocompensan.

La dualidad entre las familias VEG y PG significa que el parámetro de forma  $\xi$  es dominante en la determinación del comportamiento cualitativo de la distribución PG, tal como sucede con la distribución VEG. Si  $\xi < 0$  la distribución de los excedentes tiene acotada la cola derecha en  $u - \tilde{\sigma}/\xi$ ; si  $\xi > 0$  la cola derecha no está acotada. La distribución tampoco está acotada si  $\xi = 0$ , este caso nuevamente se debe interpretar como el límite cuando  $\xi \rightarrow 0$  en (3.20), el cual conduce al siguiente resultado

$$H(z) = 1 - \exp\left(-\frac{z}{\tilde{\sigma}}\right), \quad z > 0, \quad (3.22)$$

que corresponde a la distribución exponencial con parámetro  $1/\tilde{\sigma}$ .

### 3.3. Tipos de censura

Existen varios tipos (y mecanismos) de censura, tales como censura por la derecha (Tipo I y Tipo II), censura por la izquierda, censura por intervalo y censura aleatoria,

### 3.3. Tipos de censura

---

las cuales serán discutidas en esta sección. Para una revisión más completa y ejemplos de cada una ver [Klein y Moeschberger \(2003\)](#).

#### 3.3.1. Censura por la derecha

En la **censura por la derecha Tipo I** el evento es observado solo si ocurre antes de algún tiempo pre-especificado (tiempo de censura o límite de observación), el cual puede ser fijo para todas las unidades bajo estudio o puede variar de unidad a unidad.

Sean  $X_1, \dots, X_n$  variables aleatorias que representan los tiempos de vida de  $n$  unidades bajo estudio. Las  $X$ 's se asumen independientes e idénticamente distribuidas (i.i.d.) con función de densidad de probabilidad  $f(x)$  y función de supervivencia  $S(x) = 1 - F(x)$ . Sean  $L_i > 0, i = 1, \dots, n$  los tiempos de censura especificados al inicio del estudio. Si la censura es fija para todas las unidades, significa que  $L_1 = \dots = L_n$ .

El tiempo de vida exacto de una unidad bajo estudio será conocido si, y sólo si,  $X_i \leq L_i$ . Si  $X_i > L_i$ , se trata de una unidad sobreviviente y su tiempo al evento es censurado en el tiempo  $L_i$ . Entonces, en lugar de observar los tiempos de vida  $X_1, \dots, X_n$  se observan los tiempos dados por las variables  $T_1, \dots, T_n$ , con  $T_i = \min(X_i, L_i)$ . Los datos de tiempos de vida observados pueden ser convenientemente representados por pares de v.a. de la forma  $(T_i, \delta_i)$ , donde  $\delta = 1$  si  $T_i = X_i$  y  $\delta = 0$  si  $T_i = L_i$ .

Un segundo tipo de censura por la derecha es la **censura Tipo II**, en la cual el estudio continúa solo hasta que las primeras  $r$  unidades experimentan el evento de interés, donde  $r$  es algún valor entero determinado al inicio del estudio, tal que  $r < n$ . En esta situación, solo se observan los  $r$  tiempos de vida menores. Entonces, los datos observados consistirán de  $(T_i, \delta_i)$ , donde  $T_i = X_{(r)}$  para aquellas unidades censuradas y  $T_i = X_{(i)}$  con  $\delta = 1$ , siendo las  $X_{(i)}$  estadísticas de orden.

#### 3.3.2. Censura aleatoria

En algunas ocasiones ocurre que un evento independiente al evento de interés es causa de que una unidad bajo estudio sea censurada de manera aleatoria. Por ejemplo, en estudios médicos puede ocurrir que muertes accidentales, migraciones de las personas, pacientes que abandonen el experimento clínico, muerte por alguna causa distinta a la de interés, etc., sean la razón por la cual un individuo es removido del estudio (censurado). El mecanismo de censura que sigue a este ejemplo es denominado **censura aleatoria**.

Sean  $X_1, \dots, X_n$  una muestra aleatoria de una función de distribución continua  $F$ ,

### 3.3. Tipos de censura

---

y sean  $L_1, \dots, L_n$  una muestra aleatoria independiente con función de distribución de censura  $G$ . Como antes, las  $X$ 's denotan tiempos de vida y las  $L$ 's tiempos de censura. Entonces, para datos con censura aleatoria por la derecha, las observaciones consisten de las parejas  $(T_i, \delta_i)$  para  $i = 1, \dots, n$ , donde  $T_i = \min(X_i, L_i)$  y  $\delta_i = 1$  si  $X_i \leq L_i$ .

Note que, a diferencia de la censura por la derecha Tipo I, las  $L$ 's son valores de otras variables aleatorias no conocidas de antemano, las cuales se asumen independientes de las  $X$ 's.

#### 3.3.3. Censura por la izquierda

Se considera que un tiempo de vida  $X_i$  asociado con una unidad específica bajo estudio es censurado por la izquierda si es menor que un tiempo de censura  $C$ , es decir, el evento de interés ya ha ocurrido antes de que la unidad sea observada en el estudio al tiempo  $C$ . Para esta unidad, se sabe que ya ha experimentado el evento antes del tiempo  $C$ , pero el tiempo exacto al cual ocurrió es desconocido. El tiempo de vida exacto  $X_i$  será conocido si, y solo si,  $X_i \geq C$ . Los datos censurados por la izquierda pueden también ser representados por pares de variables aleatorias  $(T_i, \delta_i)$ , donde como antes  $T_i = X_i$  si se observa el tiempo de vida exacto y  $\delta_i$  indica si se trata de un tiempo censurado o no.

#### 3.3.4. Censura por intervalo

Un tipo más general de censura ocurre cuando solo se conoce que el tiempo de vida ocurre dentro de un intervalo, y se denomina censura por intervalo. Por ejemplo, esta censura ocurre cuando los pacientes en un estudio clínico o longitudinal se revisan periódicamente y solo se sabe que el tiempo al evento de los pacientes ocurre en un intervalo  $(U_i, V_i]$ . Este tipo de censura también puede ocurrir en experimentos industriales donde hay una inspección periódica del adecuado funcionamiento de algún componente físico.

Note que la censura por intervalo es una generalización de las censuras por la derecha y por la izquierda. Cuando el punto izquierdo del intervalo es cero y el punto derecho es  $C$  definido como en la Sección 3.3.3, entonces se tiene censura por la izquierda; cuando el punto izquierdo es  $L_i$  definido en la Sección 3.3.1 y el punto derecho es infinito, se tiene censura por la derecha.

En todos los casos una censura, por definición, siempre impide el conocimiento de  $X_i$ . Por otro lado, el evento de interés, tal como una muerte o una falla, no siempre evita el conocimiento del correspondiente  $L_i$  o  $C_i$ , en caso de que los límites de las

### 3.3. Tipos de censura

---

observaciones sean no aleatorios y previsibles.

Para los distintos mecanismos de censura (cuando sea aplicable), es deseable suponer que la muerte o falla (o cualquier otro que sea el evento de interés) de una unidad y la pérdida (censura) de la misma, o de cualquier otra, nunca ocurren al mismo tiempo. De otro modo, cuando el evento de interés para una unidad bajo estudio ocurre a un tiempo  $X_i = L_i$ , el tiempo de vida exacto de esa unidad es efectivamente conocido, por lo que es tratado como tal y no será censurado. En notación, esta unidad será registrada como  $(T_i, \delta_i)$ , donde  $T_i = X_i$  y  $\delta_i = 1$ .

# Capítulo 4

## Estimación de la los parámetros de la distribución VEG bajo censura aleatoria

Los procesos con censura aleatoria se encuentran frecuentemente en investigaciones ambientales, en el contexto de este trabajo, es aquel en el cual para cada bloque bajo estudio se asume que hay un valor máximo  $M$  y un valor de censura aleatoria  $C$ .

Sean  $M$  y  $C$  variables aleatorias independientes. Sean  $Y = \min(M, C)$  y  $\delta$  una variable que indica si el valor máximo  $M$  es censurado  $\delta = 0$  o no  $\delta = 1$ . Además, sea  $\mathbf{x}$  el vector de covariables asociado a  $M$  tal que

$$Y|\mathbf{x} \sim G(\mu(\mathbf{x}, \theta), \sigma, \xi), \quad (4.1)$$

donde  $G$  es la distribución VEG y  $\theta$  es el vector de parámetros que corresponde a  $\mathbf{x}$ .

La función de densidad de probabilidad (f.d.p.) y la función de sobrevivencia de  $M$  se denotan por  $g(m; \mu(\mathbf{x}, \theta), \sigma, \xi)$  y  $G^*(m; \mu(\mathbf{x}, \theta), \sigma, \xi) = 1 - G(m; \mu(\mathbf{x}, \theta), \sigma, \xi)$ , respectivamente. Por otro lado, la f.d.p. y la función de sobrevivencia de  $C$  se denotan por  $f(c)$  y  $F(c)$ , respectivamente. La función de densidad conjunta de  $Y$  y  $\delta$ ,  $g_{Y,\delta}(y, \delta)$ , se puede obtener de la función de densidad conjunta de  $M$  y  $C$ ,  $g_{M,C}(m, c)$ , de la siguiente manera

$$\begin{aligned} P[Y = y, \delta = 0] &= P[M > c, C = y] \\ &= \frac{d}{dy} \int_0^y \int_v^\infty g_{M,C}(u, v) dudv. \end{aligned} \quad (4.2)$$

#### 4. Estimación de la los parámetros de la distribución VEG bajo censura aleatoria

---

Puesto que  $M$  y  $C$  son independientes con densidades marginales  $g(m; \mu(\mathbf{x}, \theta), \sigma, \xi)$  y  $f(c)$ , respectivamente, (4.2) se puede escribir como

$$\begin{aligned}
 &= \frac{d}{dy} \int_0^y \int_v^\infty g(u; \mu(\mathbf{x}, \theta), \sigma, \xi) f(v) dudv \\
 &= \frac{d}{dy} \int_0^y f(v) G^*(v; \mu(\mathbf{x}, \theta), \sigma, \xi) dv \\
 &= f(y) G^*(y; \mu(\mathbf{x}, \theta), \sigma, \xi)
 \end{aligned} \tag{4.3}$$

y, de manera análoga,

$$\begin{aligned}
 P[Y = y, \delta = 1] &= P[M = y, C > m] \\
 &= \frac{d}{dy} \int_0^y \int_v^\infty g_{M,C}(v, u) dudv \\
 &= \frac{d}{dy} \int_0^y \int_v^\infty g(v; \mu(\mathbf{x}, \theta), \sigma, \xi) f(u) dudv \\
 &= \frac{d}{dy} \int_0^y g(v; \mu(\mathbf{x}, \theta), \sigma, \xi) F(v) dv \\
 &= g(y; \mu(\mathbf{x}, \theta), \sigma, \xi) F(y).
 \end{aligned} \tag{4.4}$$

Considere los datos de una muestra de  $n$  máximos de bloque, la cual consiste de las tripletas  $(Y_i, \delta_i, \mathbf{x}_i)$ ,  $i = 1, \dots, n$ . Entonces, la función de verosimilitud se construye como sigue

$$L(\mu(\mathbf{x}, \theta), \sigma, \xi) = \prod_{i=1}^n [g_{Y,\delta}(y_i, \delta_i)], \tag{4.5}$$

de (4.3) y (4.4), (4.5) se puede reescribir como

$$\begin{aligned}
 &= \prod_{i=1}^n [G^*(y_i; \mu(\mathbf{x}, \theta), \sigma, \xi) f(y_i)]^{1-\delta_i} [F(y_i) g(y_i; \mu(\mathbf{x}, \theta), \sigma, \xi)]^{\delta_i} \\
 &= \left\{ \prod_{i=1}^n f(y_i)^{1-\delta_i} F(y_i)^{\delta_i} \right\} \left\{ \prod_{i=1}^n G^*(y_i; \mu(\mathbf{x}, \theta), \sigma, \xi)^{1-\delta_i} g(y_i; \mu(\mathbf{x}, \theta), \sigma, \xi)^{\delta_i} \right\} \tag{4.6}
 \end{aligned}$$

## 4.1. Incorporando un patrón estacional

---

Si la distribución de los tiempos de censura, como se mencionó previamente, no depende de los parámetros de interés, entonces, el primer término es una constante con respecto a los parámetros de interés y la función de verosimilitud toma la forma

$$L(\mu(\mathbf{x}, \theta), \sigma, \xi) \propto \prod_{i=1}^n G^*(y_i; \mu(\mathbf{x}, \theta), \sigma, \xi)^{1-\delta_i} g(y_i; \mu(\mathbf{x}, \theta), \sigma, \xi)^{\delta_i}. \quad (4.7)$$

Kalbfleisch y Prentice (2002) proporcionan un enfoque similar para construir la función de verosimilitud bajo el esquema de censura aleatoria. Smith (1985) muestra que las condiciones de regularidad para los EMV se mantienen si  $\xi > -1/2$  y además se aseguran las propiedades de consistencia, y de normalidad y eficiencia asintótica. Sin embargo, Zhang *et al.* (2006) advierte que niveles de censura altos afectan las propiedades de los EMV.

## 4.1. Incorporando un patrón estacional

Davison y Smith (1990) remueve el componente estacional de la serie de datos filtrando las observaciones por medio de una senoide antes de realizar el análisis de los valores extremos, en este trabajo la senoide se incorpora en el parámetro de localidad, reemplazando  $\mu(\mathbf{x}, \theta)$  por  $\mu_1$  en (4.7), donde  $\mu_1$  tiene la siguiente forma

$$\mu_1 = M + A \cos(\omega t_i + \eta), \quad (4.8)$$

$t_i$  es el número de bloque asociado a  $y_i, i = 1, \dots, n$  y  $\omega, M, A$  y  $\eta$  son constantes. Esto es apropiado cuando la influencia de la estacionalidad tiene la forma de una senoide. Se supone que  $\omega$  es conocido, por otro lado  $M, A$  y  $\eta$  se pueden estimar del conjunto de datos. Si el número de bloques por año es  $\tau$  entonces  $\omega = 2\pi/\tau$ . Así, la función de verosimilitud se puede reescribir como

$$L(\mu_1, \sigma, \xi) \propto \prod_{i=1}^n G^*(y_i; \mu_1, \sigma, \xi)^{1-\delta_i} g(y_i; \mu_1, \sigma, \xi)^{\delta_i}. \quad (4.9)$$

Adicionalmente, se incluye un parámetro extra en  $\mu_1$  para estudiar la tendencia en el tiempo, entonces el parámetro de localidad en (4.8) se reemplaza por

$$\mu_2 = M + A \cos(\omega t_i + \eta) + \beta_1 \text{year}_i, \quad (4.10)$$

## 4.2. Elección del modelo

---

donde  $year_i$  es el año asociado a  $y_i, i = 1, \dots, n$ . De este modo, la función de verosimilitud se puede definir de la siguiente manera

$$L(\mu_2, \sigma, \xi) \propto \prod_{i=1}^n G^*(y_i; \mu_2, \sigma, \xi)^{1-\delta_i} g(y_i; \mu_2, \sigma, \xi)^{\delta_i}. \quad (4.11)$$

Para programar las expresiones (4.9) y (4.11) se utiliza una rutina computacional basada en el paquete [VGAM](#) de [R](#). El cálculo de los EMV de (4.9) y (4.11) se realiza con la función *optim* de [R](#).

La función de los cuantiles extremos de la distribución VEG en (4.1) es

$$y_{(p,\mathbf{x})} = \begin{cases} \mu(\mathbf{x}) + \frac{\sigma}{\xi} [(-\log(1-p))^{-\xi} - 1] & , \xi \neq 0, \\ \mu(\mathbf{x}) - \sigma \log(-\log(1-p)) & , \xi = 0, \end{cases} \quad (4.12)$$

donde  $\mu(\mathbf{x})$  se puede reemplazar por  $\mu_1$  or  $\mu_2$  dependiendo de las covariables en  $\mathbf{x}$ ,  $G(y_{(p,\mathbf{x})}) = 1 - p$  con  $0 < p < 1$  y  $y_{(p,\mathbf{x})}$  es el nivel de retorno asociado con el periodo de retorno  $1/p$  condicionado en  $\mathbf{x}$  ([Beirlant et al., 2004](#)).

## 4.2. Elección del modelo

Considere la hipótesis nula  $H_0 : \Theta_0 = (M, A, \eta, \sigma, \xi)$  que corresponde al modelo en (4.9) y una hipótesis más general  $H_1 : \Theta_1 = (M, A, \eta, \beta, \sigma, \xi)$  la cual corresponde al modelo en (4.11). Entonces, se puede probar  $H_0$  contra  $H_1$  por medio de la diferencia de devianzas

$$\Delta D = 2 \left[ \log(L(\hat{\theta}_1)) - \log(L(\hat{\theta}_0)) \right], \quad (4.13)$$

donde  $\hat{\theta}_1$  y  $\hat{\theta}_0$  son los EMV de  $\Theta_1$  y  $\Theta_0$ , respectivamente. En este caso  $\Delta D \sim \chi^2(1)$ , entonces el *p-value* es  $P(\chi^2(1) > \Delta d)$ . De esta forma se rechaza  $H_0$  en favor de  $H_1$  cuando el *p-value* es menor que cierto nivel de significancia (ver [Dobson y Barnett \(2008\)](#) para mayores detalles).

# Capítulo 5

## Pruebas de bondad de ajuste para la distribución VEG bajo censura aleatoria

La distribución VEG surge como la distribución límite para modelar los extremos (máximo o mínimo) de una muestra. Sin embargo, existen otras distribuciones con colas pesadas que se emplean para modelar valores extremos como la distribución Pareto generalizada (PG) que se utiliza para modelar observaciones que están por encima de un valor denominado umbral o la distribución Dagum que surgió en la década de los 70 para modelar el ingreso [Kleiber y Kotz \(2003\)](#).

La variedad de distribuciones que se utilizan para analizar valores extremos ha motivado que en varios artículos se realicen investigaciones sobre pruebas de bondad de ajuste, pero solo unos cuantos consideran el caso de muestras censuradas. En la mayoría de los casos abordan la censura por la derecha y por la izquierda, aunque también existen trabajos sobre censura progresiva [Montfort y Gomes \(1985\)](#), [Gibson y Higgins \(2000\)](#), [Balakrishnan \*et al.\* \(2004\)](#), [Lim y Park \(2007\)](#), [Wang \(2008\)](#), [Rad \*et al.\* \(2011\)](#), [Bispo \*et al.\* \(2012\)](#), [Dey y Kundu \(2012\)](#), [Salinas \*et al.\* \(2012\)](#), [Bogdonavicius \*et al.\* \(2013\)](#), [Denecke y Müller \(2014\)](#), [Economou y Tzavelas \(2014\)](#).

Respecto a la censura aleatoria [Turnbull y Weiss \(1978\)](#) proponen un estadístico de razón de verosimilitud para realizar una prueba de bondad de ajuste de datos agrupados con censura aleatoria por la derecha, el estadístico de la prueba tiene distribución asintótica ji-cuadrada no-central bajo alternativas contiguas y lo emplea para analizar el ajuste de un conjunto de datos a las distribuciones Weibull, Gompertz y exponential power. [Koziol \(1980\)](#) introduce una nueva versión de los estadísticos de Kolmogorov-Smirnov, Kuiper y Cramer-von Mises para el problema de evaluar la bondad de ajuste en datos con censura aleatoria, en uno de los juegos de hipótesis que utilizó para ejemplificar las pruebas propuestas empleó la distribución exponencial en

## 5.1. Juego de hipótesis

---

la hipótesis nula y la distribución Weibull en la alternativa los resultados revelaron que la prueba basada en el estadístico de Cramer-von Mises tiene mayor potencia. [Habib y Thomas \(1986\)](#) proponen una prueba más general en comparación a las derivadas por [Koziol \(1980\)](#) ya que en la hipótesis nula consideran una familia paramétrica de distribuciones de sobrevivencia; es decir, una hipótesis compuesta. [Kim \(1993\)](#) considera estadísticos de prueba de bondad de ajuste ji-cuadrado generalizados a los que llama estadísticos de Pearson generalizados los cuales son formas cuadráticas definidas no negativas en las frecuencias de las celdas obtenidas del estimador producto-límite, realiza un estudio de eficiencia de Pitman que muestra la superioridad del estadístico Akritas sobre el estadístico de Pearson generalizado en situaciones de muestras con fuerte censura.

En este trabajo se considera una muestra aleatoria de la distribución VEG con censura aleatoria, cada uno de los valores censurados es transformado en un dato observado para posteriormente emplear la prueba modificada de Anderson-Darling derivada por [Heo et al. \(2013\)](#) y las generalizaciones de las pruebas de Kolmogorov-Smirnov y Cramer-von Mises derivadas por [Zhang \(2002\)](#). Se compara el tamaño y la potencia de las pruebas por medio de un estudio Monte Carlo para distintos niveles de censura y tamaños de muestra.

## 5.1. Juego de hipótesis

Sea  $Y$  una variable aleatoria continua con función de distribución  $F(y)$ , y sea  $Y_1, Y_2, \dots, Y_n$  una muestra aleatoria de  $Y$  con estadísticas de orden  $Y_{(1)}, Y_{(2)}, \dots, Y_{(n)}$ . Se desea probar la hipótesis nula

$$H_0 : F(y) = F_0(y), \forall y \in (-\infty, \infty) \quad (5.1)$$

contra la hipótesis alternativa

$$H_1 : F(y) \neq F_0(y), \text{ para alguna } y \in (-\infty, \infty) \quad (5.2)$$

donde  $F_0(y)$  es la función de distribución hipotética completamente especificada a excepción de algunos parámetros desconocidos que se pueden estimar usando la muestra. En nuestro caso, la función de distribución hipotética es la distribución VEG. Enseguida se presentan los estadísticos de prueba que se utilizan para contrastar el juego de hipótesis que se ha presentado.

## 5.2. Estadísticos de prueba

### 5.2.1. Prueba de Anderson-Darling modificada

La prueba de Anderson-Darling es un método que asigna igual peso a ambas colas de la distribución, pero en el estudio de valores extremos el principal interés es estimar los cuantiles para periodos de retorno altos. En tales casos, el interés se centra en la cola superior o inferior de la distribución. Así, [Heo et al. \(2013\)](#) proponen un estadístico de prueba de Anderson-Darling modificado empleando una función de pesos que enfatiza las desviaciones en la cola superior o inferior. Para hacer énfasis en la cola superior o inferior, la prueba utiliza la función de pesos  $\psi(y) = [1 - F(y)]^{-1}$  o  $\psi(y) = [F(y)]^{-1}$ , respectivamente. El estadístico de prueba para la cola superior ( $A_U$ ) y para la cola inferior ( $A_L$ ) se definen en las siguientes ecuaciones:

$$A_U = n \int_{-\infty}^{\infty} \frac{[F_n(y) - F_0(y)]^2}{1 - F_0(y)} dF_0(y) \quad (5.3)$$

$$A_L = n \int_{-\infty}^{\infty} \frac{[F_n(y) - F_0(y)]^2}{F_0(y)} dF_0(y) \quad (5.4)$$

donde  $F_n$  y  $F_0$  representan la f.d.a. empírica y la f.d.a. bajo la hipótesis nula, respectivamente.

Entonces, de acuerdo a [Ahmad et al. \(1988\)](#), los estadísticos de prueba (5.3) y (5.4) se pueden aproximar por las siguientes expresiones

$$A_U = \frac{n}{2} - 2 \sum_{i=1}^n F_0(y_i) - \sum_{i=1}^n \left( 2 - \frac{2i-1}{n} \right) \log(1 - F_0(y_i)) \quad (5.5)$$

$$A_L = -\frac{3n}{2} + 2 \sum_{i=1}^n F_0(y_i) - \sum_{i=1}^n \left( \frac{2i-1}{n} \right) \log(F_0(y_i)) \quad (5.6)$$

Note que la suma de (5.5) y (5.6) es el estadístico de prueba original de Anderson-Darling ya que la suma de función de pesos de estos dos estadísticos es la función de pesos del estadístico original de Anderson-Darling. Valores grandes de (5.5) o (5.6) dan evidencia para rechazar  $H_0$  en favor de  $H_1$ . En este trabajo se desea analizar la cola derecha de la distribución VEG, por lo que se utiliza el estadístico de prueba (5.5) para contrastar el juego de hipótesis mencionado previamente.

## 5.2. Estadísticos de prueba

---

### 5.2.2. Generalización de la prueba de Cramér-von Mises

Zhang (2002) propone el estadístico

$$Z = \int_{-\infty}^{\infty} Z_y dw(y), \quad (5.7)$$

para contrastar  $H_0$  y  $H_1$ , donde  $dw(y)$  es la derivada de la función de pesos  $w(y)$  y para valores grandes de  $Z$  se rechaza  $H_0$  en favor de  $H_1$ . Nuevamente, un candidato para  $Z_y$  es el estadístico de prueba de razón de verosimilitud (5.11), con  $F_n$  la función de distribución empírica y  $F_0$  la función de distribución bajo  $H_0$ .

Utilizando  $dw(y) = F_0(y)^{-1}\{1 - F_0(y)\}^{-1} dF_0(y)$  y reemplazando  $Z_y$  por  $G_y^2$  en (5.7) se obtiene

$$\sum_{i=1}^n [\log\{F_0(Y_{(i)})^{-1} - 1\} - b_{i-1} + b_i]^2 + C_n, \quad (5.8)$$

donde  $C_n$  es una constante y  $b_i = i \log(i/n) + (n - i) \log(1 - i/n)$ .

Como  $b_{i-1} - b_i \approx \log\{(n - 1/2)/(i - 3/4) - 1\}$ , el estadístico (5.8) es bien aproximado por

$$Z_C = \sum_{i=1}^n \left[ \log \left\{ \frac{F_0(X_{(i)})^{-1} - 1}{(n - \frac{1}{2}) / (i - \frac{3}{4}) - 1} \right\} \right]^2. \quad (5.9)$$

Zhang (2002) muestra nuevamente por medio de estudios Monte Carlo que el estadístico  $Z_C$  es generalmente más potente que el estadístico original de Cramér-von Mises.

### 5.2.3. Generalización de la prueba de Kolmogorov-Smirnov

Zhang (2002) también propone el estadístico

$$Z_{\max} = \sup_{y \in (-\infty, \infty)} \{Z_y w(y)\}, \quad (5.10)$$

### 5.3. Algoritmo para generar una muestra artificial

---

para contrastar  $H_0$  y  $H_1$ , donde  $w(y)$  es una función de pesos y para valores grandes de  $Z_{\max}$  se rechaza  $H_0$  en favor de  $H_1$ . Un candidato para  $Z_y$  es el estadístico de prueba de razón de verosimilitud,

$$G_y^2 = 2n \left[ F_n(y) \log \left\{ \frac{F_n(y)}{F_0(y)} \right\} + \{1 - F_n(y)\} \log \left\{ \frac{1 - F_n(y)}{1 - F_0(y)} \right\} \right], \quad (5.11)$$

donde  $F_n$  es la función de distribución empírica y  $F_0$  es la función de distribución bajo  $H_0$ .

Utilizando la función de pesos  $w(y) = 1$  y reemplazando  $Z_y$  por  $G_y^2$  en (5.10) se obtiene

$$\sup_{y \in (-\infty, \infty)} (G_y^2) = \max_{0 \leq i \leq n} \left\{ \sup_{Y_{(i)} \leq y \leq Y_{(i+1)}} (G_y^2) \right\} = \max_{1 \leq i \leq n} (G_{Y_{(i)}}^2), \quad (5.12)$$

lo cual es equivalente a

$$Z_K = \max_{1 \leq i \leq n} \left( \left( i - \frac{1}{2} \right) \log \left\{ \frac{i - \frac{1}{2}}{n F_0(X_{(i)})} \right\} + \left( n - i + \frac{1}{2} \right) \log \left\{ \frac{n - i + \frac{1}{2}}{n \{1 - F_0(X_{(i)})\}} \right\} \right). \quad (5.13)$$

Zhang (2002) ha mostrado por medio de estudios Monte Carlo que el estadístico  $Z_K$  es más potente que el estadístico original de Kolmogorov-Smirnov.

### 5.3. Algoritmo para generar una muestra artificial

Note que para poder utilizar las pruebas  $A_U$ ,  $Z_C$  y  $Z_K$  es necesario conocer los estadísticos de orden de la muestra, sin embargo bajo el esquema de censura aleatoria no es posible obtener estos estadísticos de la muestra. Una solución a este problema es completar las observaciones censuradas utilizando los estimadores de la distribución de la que proceden los datos.

Sea  $Y$  una variable aleatoria continua con f.d.c  $G(y; \Theta)$  y suponga que se tiene una muestra de tamaño  $n$  la cual contiene  $r$  datos censurados bajo el esquema de censura aleatoria, los cuales se denotarán por  $C_i, i = 1, \dots, r$ . Entonces, se desea completar la muestra simulando los valores de los  $\hat{Y}_i, i = 1, \dots, r$  que no se observaron puesto que  $C_i, i = 1, \dots, r$  resulto ser más pequeño.

## 5.4. Aproximación de los valores críticos de los estadísticos de prueba

---

La idea es, con los estimadores que se han encontrado bajo el esquema de censura aleatoria ( $\hat{\Theta}$ ), utilizar la f.d.c. de  $Y$ ,  $G(y; \hat{\Theta})$ , y como se conoce  $C_i, i = 1, \dots, r$  entonces se simula una realización de la condicional para  $\hat{Y}_i | \hat{Y}_i > c_i, i = 1, \dots, r$ ; esto es si  $U$  es una realización de una distribución uniforme en  $(0, 1)$ , simplemente se debe obtener la solución a la siguiente expresión

$$\frac{G(\hat{y}_i; \hat{\Theta}) - G(c_i; \hat{\Theta})}{1 - G(c_i; \hat{\Theta})} = u_i, i = 1, \dots, r. \quad (5.14)$$

La solución a la expresión anterior es

$$\hat{y}_i = G^{-1} \left\{ [u_i \{1 - G(c_i; \hat{\Theta})\}] + G(c_i; \hat{\Theta}) \right\}, \quad (5.15)$$

donde  $G^{-1}$  es la función cuantil.

El algoritmo se resume de la siguiente manera:

Dada una muestra  $Y_1, \dots, Y_n$  que tiene f.d.a  $G(y; \Theta)$ , la cual contiene  $r$  observaciones que presentan censura aleatoria

1. Se estiman los parámetros de f.d.a  $G(y; \Theta)$  bajo el esquema de censura aleatoria.
2. Se identifican las  $r$  observaciones censuradas  $c_1, \dots, c_r$ .
3. Se evalúa la expresión  $\hat{y}_i = G^{-1} \left\{ [u_i \{1 - G(c_i; \hat{\Theta})\}] + G(c_i; \hat{\Theta}) \right\}$  en  $c_i, i = 1, \dots, r$ .

## 5.4. Aproximación de los valores críticos de los estadísticos de prueba

En las pruebas de bondad de ajuste, el valor crítico es el punto de corte que indica el límite bajo el cual el estadístico de prueba puede rechazar o no  $H_0$  dado un nivel de significancia. Dados  $A_U, Z_C$  y  $Z_K$ , la regla de decisión es rechazar  $H_0$  en (5.1), donde  $F_0$  es la distribución VEG, con un nivel de significancia  $\alpha$  si  $A_U \geq (C_{n,lc}^{A_U}(\alpha))$  o  $Z_C \geq (C_{n,lc}^{Z_C}(\alpha))$  o  $Z_K \geq (C_{n,lc}^{Z_K}(\alpha))$ , respectivamente. Donde  $n$  es el tamaño de muestra,  $lc$  es el nivel de censura en la muestra. Además,  $C_{n,lc}^{A_U}(\alpha)$ ,  $C_{n,lc}^{Z_C}(\alpha)$  y  $C_{n,lc}^{Z_K}(\alpha)$  son tales que

$$\alpha = P(\text{Rechazar } H_0 | H_0) = P(A_U \geq C_{n,lc}^{A_U}(\alpha)), \quad (5.16)$$

#### 5.4. Aproximación de los valores críticos de los estadísticos de prueba

---

$$\alpha = P(\text{Rechazar } H_0 | H_0) = P(Z_C \geq C_{n,lc}^{Z_C}(\alpha)), \quad (5.17)$$

$$\alpha = P(\text{Rechazar } H_0 | H_0) = P(Z_K \geq C_{n,lc}^{Z_K}(\alpha)) \text{ y} \quad (5.18)$$

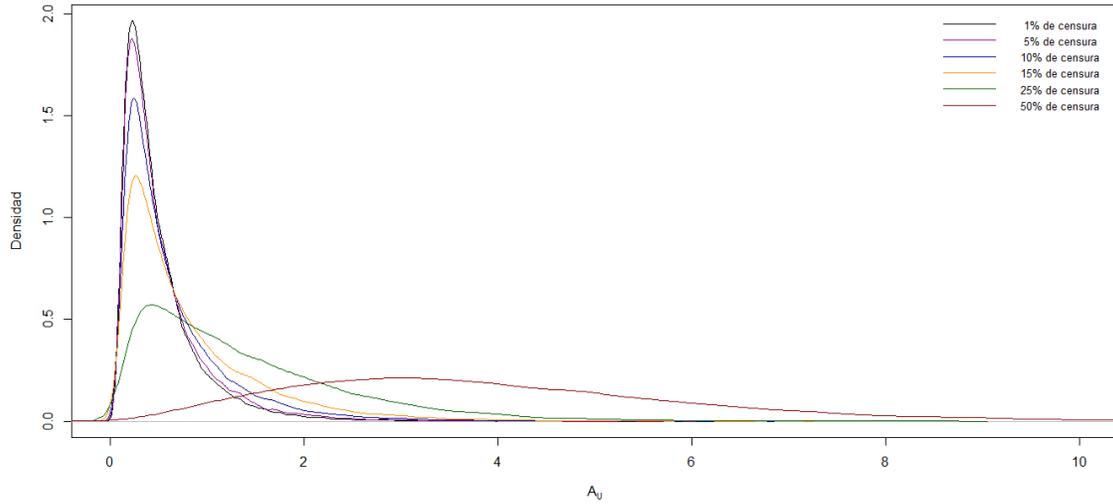
respectivamente.

Las distribuciones de  $A_U$ ,  $Z_C$  y  $Z_K$ , bajo  $H_0$  para valores fijos de  $n$ ,  $lc$ ,  $\mu$ ,  $\sigma$  y  $\xi$  se pueden obtener por medio de simulación Monte Carlo, el procedimiento se describe a continuación

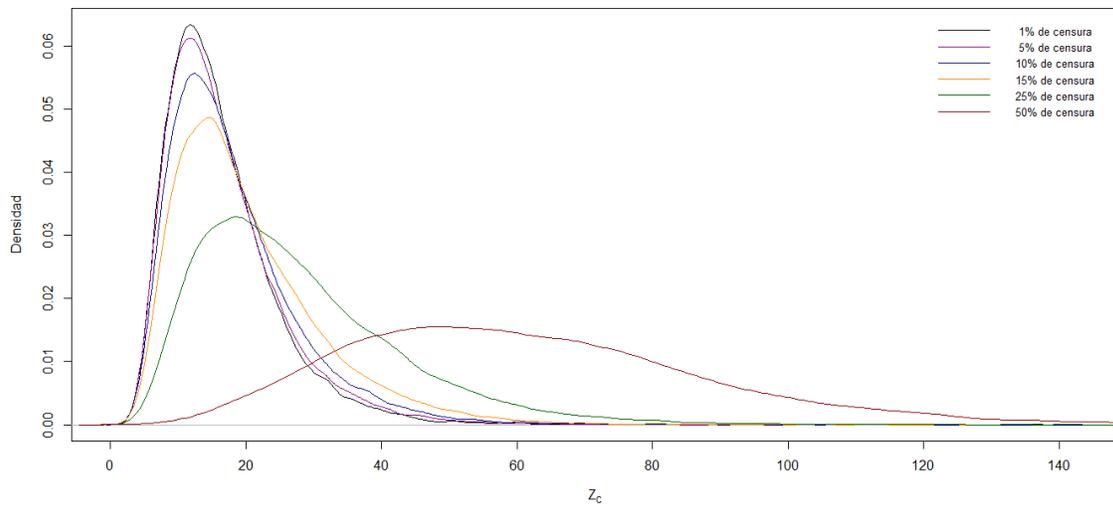
1. Fijar  $n$ ,  $lc$ ,  $\mu$ ,  $\sigma$  y  $\xi$ .
2. Simular  $n$  observaciones de la distribución VEG,  $G(\mu, \sigma, \xi)$ .
3. Censurar aleatoriamente  $n(lc)$  observaciones.
4. Completar las observaciones censuradas utilizando el algoritmo descrito en la Sección 5.3.
5. Calcular el estadístico de prueba ( $A_U$ ,  $Z_C$  o  $Z_K$ )
6. Repetir los pasos 2 a 5,  $B$  veces.

Se ejecuta el algoritmo anterior con diferentes tamaños de muestra ( $n = 100, 200, 250, 350$ ), diferentes niveles de censura ( $lc = 0.01, 0.05, 0.10, 0.15, 0.25, 0.50$ ),  $\mu = 15$ ,  $\sigma = 1$ ,  $\xi = 0.10$  y  $B = 25000$ . Los valores de los estadísticos se ordenan en orden ascendente y se calcula el cuantil 95, este es una aproximación del valor crítico para el nivel de significancia 0.05. Los valores críticos para los tamaños de muestra y niveles de censura analizados se encuentran en el Apéndice. En las Figuras 5.1, 5.2 y 5.3 se observa que la dispersión de los estadísticos de prueba  $A_U$ ,  $Z_C$  y  $Z_K$  respectivamente se incrementa conforme aumenta el nivel de censura.

## 5.4. Aproximación de los valores críticos de los estadísticos de prueba



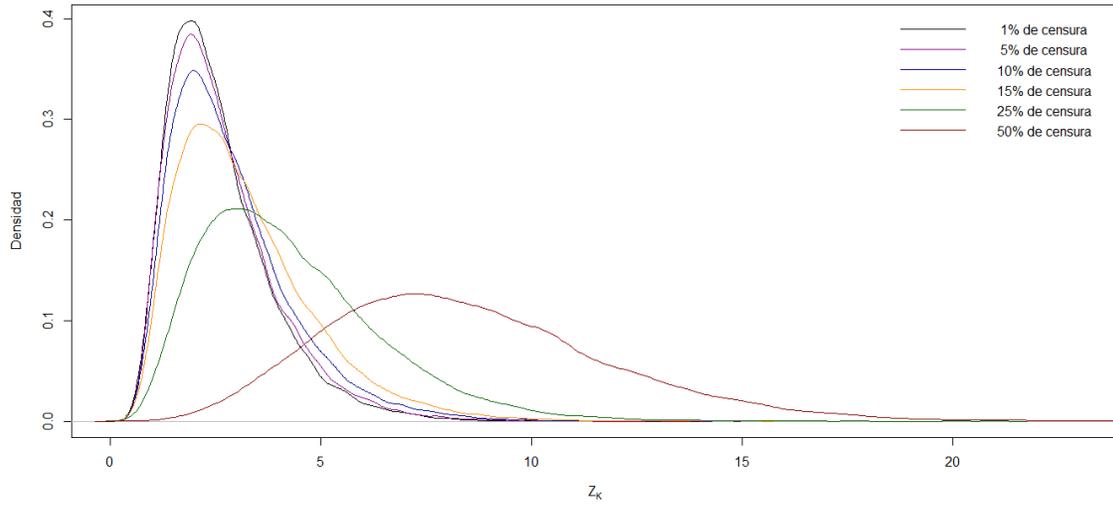
**Figura 5.1:** Distribución de la estadística de prueba  $A_U$  con  $n = 350$  y diferentes niveles de censura



**Figura 5.2:** Distribución de la estadística de prueba  $Z_C$  con  $n = 350$  y diferentes niveles de censura

#### 5.4. Aproximación de los valores críticos de los estadísticos de prueba

---



**Figura 5.3:** Distribución de la estadística de prueba  $Z_K$  con  $n = 350$  y diferentes niveles de censura

## 5.5. Comparación de la potencia de las pruebas $A_U$ , $Z_C$ y $Z_K$ por simulación

---

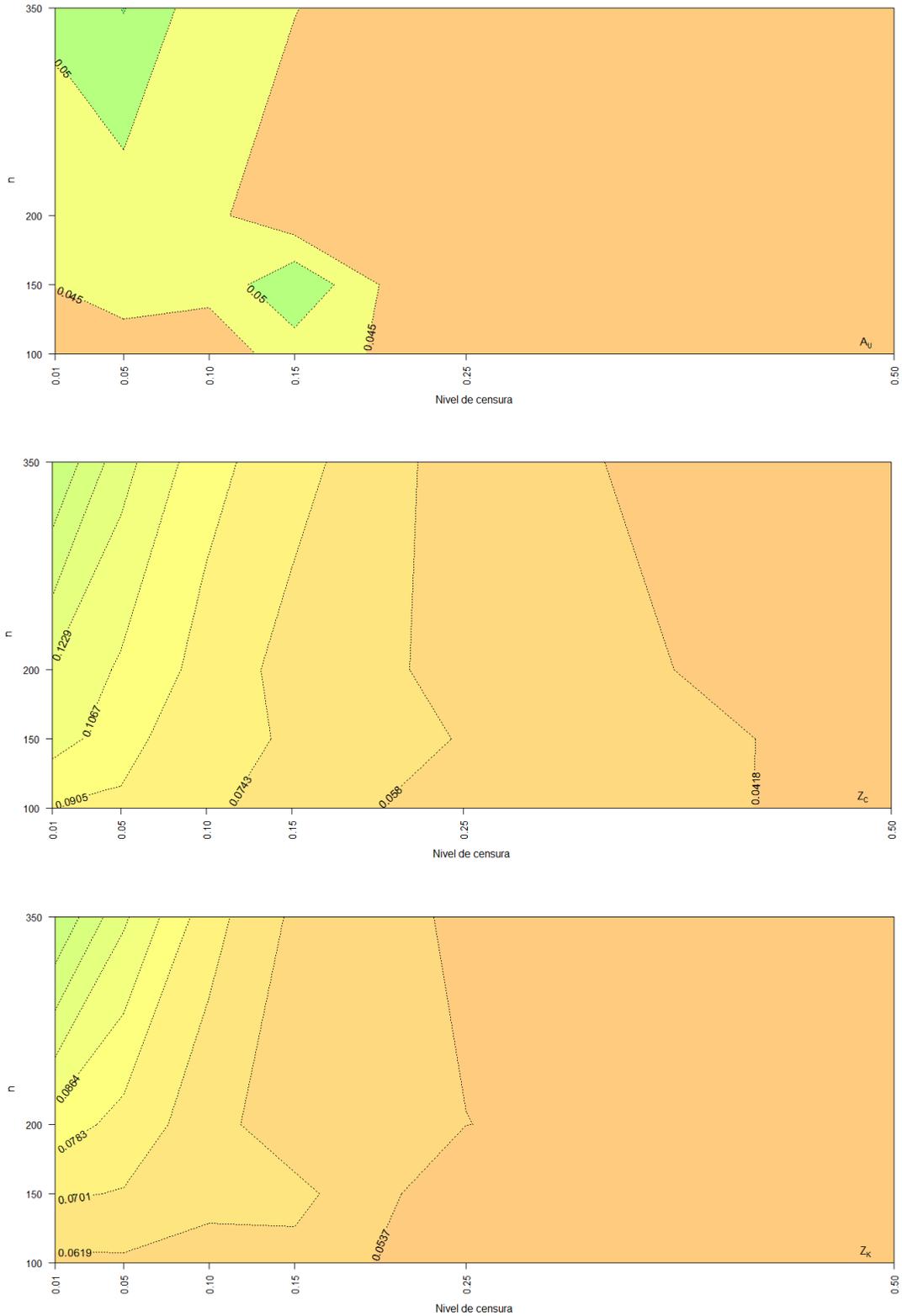
### 5.5. Comparación de la potencia de las pruebas $A_U$ , $Z_C$ y $Z_K$ por simulación

Se estudió la potencia de las pruebas  $A_U$ ,  $Z_C$  y  $Z_K$ . La potencia de la prueba se define como la probabilidad de rechazar  $H_0$  dado que  $H_0$  es falsa. Para este propósito se emplearon cuatro hipótesis alternativas: los datos siguen la distribución *Dagum*(15.26, 12.41, 18.98), *Gamma*(121.99, 7.78), *log – normal*(2.74, 0.08) y *Weibull*(7.90, 16.39). La elección de los parámetros de las distribuciones bajo la hipótesis alternativa se realizó de tal manera que la distribución en  $H_1$  fuera parecida a  $H_0$ , es decir, se generó una muestra de tamaño 10000 de la distribución bajo  $H_0$  y a partir de esta muestra se calcularon los parámetros correspondientes a cada una de las distribuciones bajo  $H_1$ . Posteriormente se empleó el siguiente algoritmo para calcular los estadísticos de prueba  $A_U$ ,  $Z_C$  y  $Z_K$  bajo cada una de las hipótesis alternativas definidas previamente.

1. Fijar  $n$ ,  $lc$ , y los parámetros de la distribución bajo  $H_1$ .
2. Simular  $n$  observaciones de la distribución bajo  $H_1$ .
3. Censurar aleatoriamente  $n(lc)$  observaciones.
4. Completar las observaciones censuradas utilizando los pasos 2 y 3 del algoritmo descrito en la Sección 5.3.
5. Calcular el estadístico de prueba ( $A_U$ ,  $Z_C$  o  $Z_K$ )
6. Repetir los pasos 2 a 5,  $B$  veces.

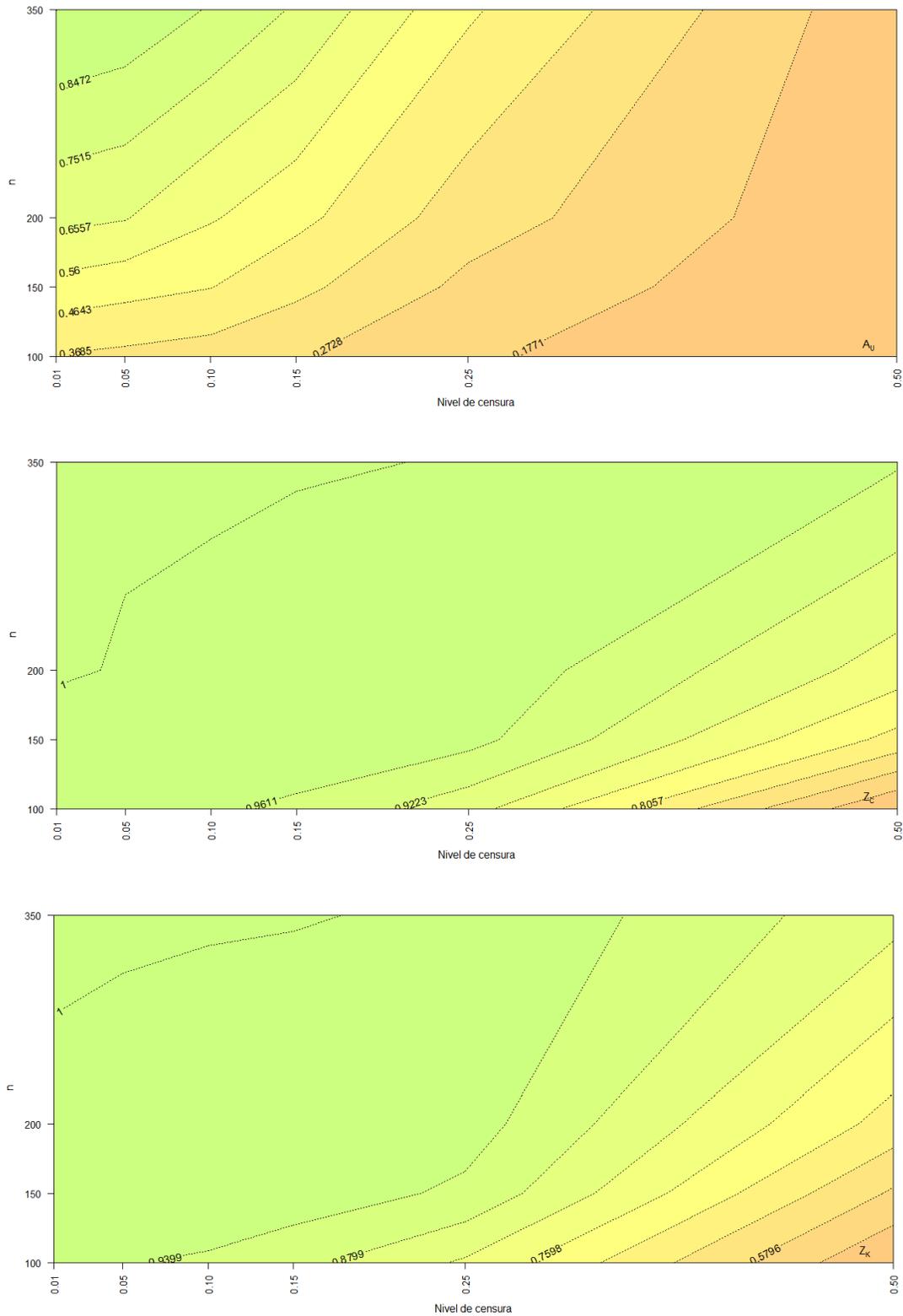
Se calcularon  $B = 5000$  simulaciones de los estadísticos  $A_U$ ,  $Z_C$  y  $Z_K$  con diferentes tamaños de muestra ( $n = 100, 150, 200, 350$ ) y diferentes niveles de censura ( $lc = 0.01, 0.05, 0.10, 0.15, 0.25, 0.50$ ). Posteriormente se calculó la proporción de rechazos empleando los valores críticos  $C_{n,lc}^{A_U}(\alpha)$ ,  $C_{n,lc}^{Z_C}(\alpha)$  y  $C_{n,lc}^{Z_K}(\alpha)$  para estimar vía Monte Carlo la potencia de las pruebas  $A_U$ ,  $Z_C$  y  $Z_K$  respectivamente. Enseguida se presentan gráficos de contorno de los resultados obtenidos

## 5.5. Comparación de la potencia de las pruebas $A_U$ , $Z_C$ y $Z_K$ por simulación



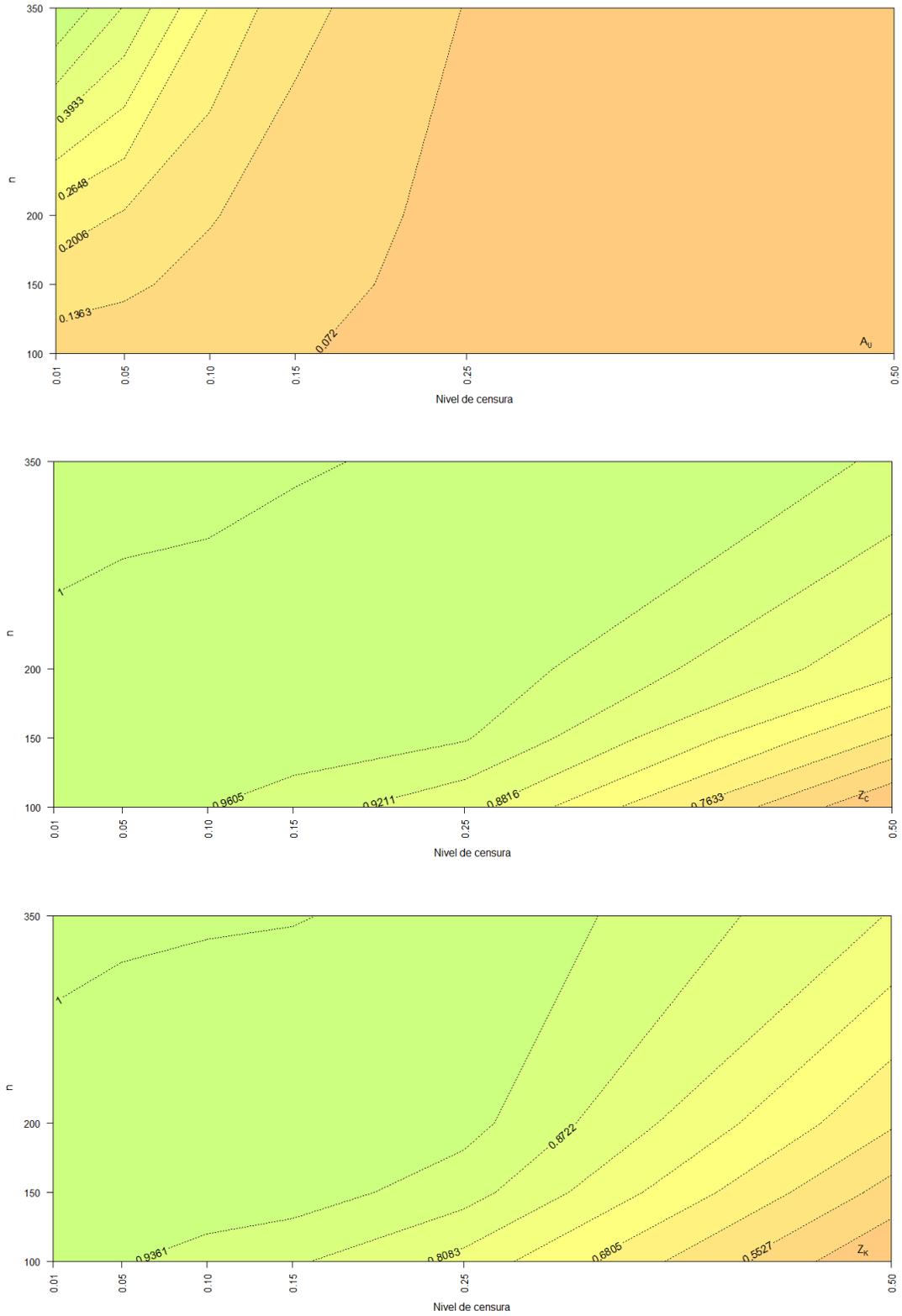
**Figura 5.4:** Potencia de la prueba  $A_U$ ,  $Z_C$  y  $Z_K$  con  $H_1 : Dagum(15.26, 12.41, 18.98)$

## 5.5. Comparación de la potencia de las pruebas $A_U$ , $Z_C$ y $Z_K$ por simulación



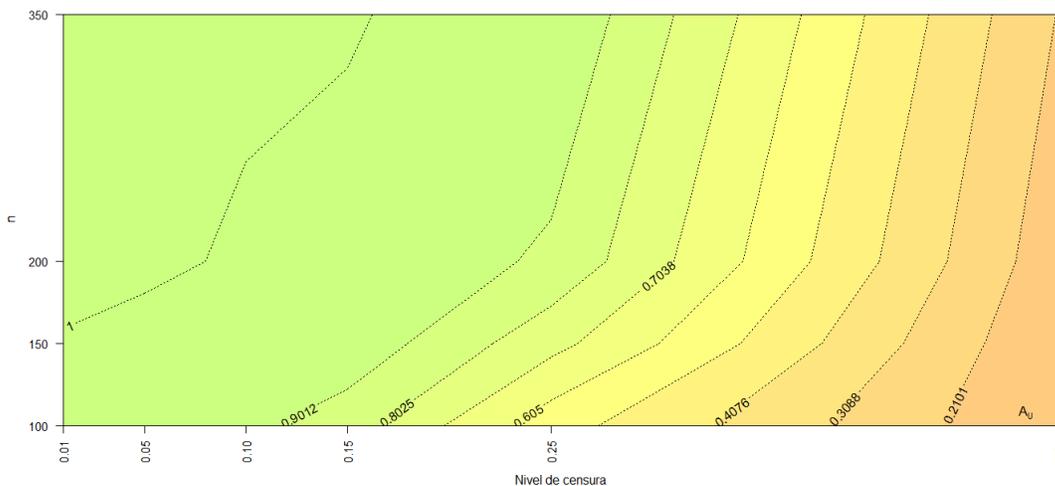
**Figura 5.5:** Potencia de la prueba  $A_U$ ,  $Z_C$  y  $Z_K$  con  $H_1 : Gamma(121.99, 7.78)$

### 5.5. Comparación de la potencia de las pruebas $A_U$ , $Z_C$ y $Z_K$ por simulación



**Figura 5.6:** Potencia de la prueba  $A_U$ ,  $Z_C$  y  $Z_K$  con  $H1 : \log - normal(2.74, 0.08)$

## 5.6. Comparación del tamaño de las pruebas $A_U$ , $Z_C$ y $Z_K$ por simulación



**Figura 5.7:** Potencia de la prueba  $A_U$  con  $H_1 : Weibull(7.90, 16.39)$

Las potencias de las pruebas  $Z_C$  y  $Z_K$  no se graficaron debido a que estas valen 1 para los diferentes niveles de censura y tamaños de muestra analizados.

De las tres pruebas, la prueba  $Z_C$  es la que tiene mayor potencia en todos los casos y la prueba  $A_U$  es la que muestra los valores de potencia más bajos en relación al resto de las pruebas. Por otro lado, la prueba  $Z_K$  tiene potencias muy parecidas a las mostradas por la prueba  $Z_C$ . Los valores de las potencias son más grandes cuando se tiene en  $H_1$  a la distribución *Weibull*; sin embargo, cuando en  $H_1$  se encuentra la distribución *Dagum* las potencias apenas y alcanzan el valor 0.1716 en el mejor de los casos. Cuando se tienen como hipótesis alternativas a la distribución *Gamma* y *Log-normal* los valores de las potencias son aceptables en el caso de las tres pruebas, sin embargo sobresalen los valores de las potencias correspondientes a las pruebas  $Z_C$  y  $Z_K$  que mantienen valores por arriba de 0.50 en escenarios donde el porcentaje de censura es de 50% y el tamaño de muestra 100. Cabe señalar, que los niveles altos de censura reducen considerablemente la potencia de la prueba  $A_U$  mientras que las potencias de las pruebas  $Z_C$  y  $Z_K$  no disminuyen de manera tan drástica como en el caso de la prueba  $A_U$ .

## 5.6. Comparación del tamaño de las pruebas $A_U$ , $Z_C$ y $Z_K$ por simulación

Se estudió el tamaño de las pruebas utilizando el algoritmo de la Sección 5.4, se realizaron  $B = 5000$  simulaciones de los estadísticos  $A_U$ ,  $Z_C$  y  $Z_K$  con diferentes

## 5.6. Comparación del tamaño de las pruebas $A_U$ , $Z_C$ y $Z_K$ por simulación

tamaños de muestra ( $n = 100, 150, 200, 350$ ) y diferentes niveles de censura ( $lc = 0.01, 0.05, 0.10, 0.15, 0.25, 0.50$ ). Posteriormente se calculo la proporción de rechazos empleando los valores críticos respectivos  $C_{n,lc}^{A_U}(\alpha)$ ,  $C_{n,lc}^{Z_C}(\alpha)$  y  $C_{n,lc}^{Z_K}(\alpha)$  para estimar el tamaño de cada una de las pruebas respectivas vía Monte Carlo. En las Figuras 5.8, 5.9 y 5.10 se presentan el tamaño de las pruebas  $A_U$ ,  $Z_C$  y  $Z_K$  respectivamente.

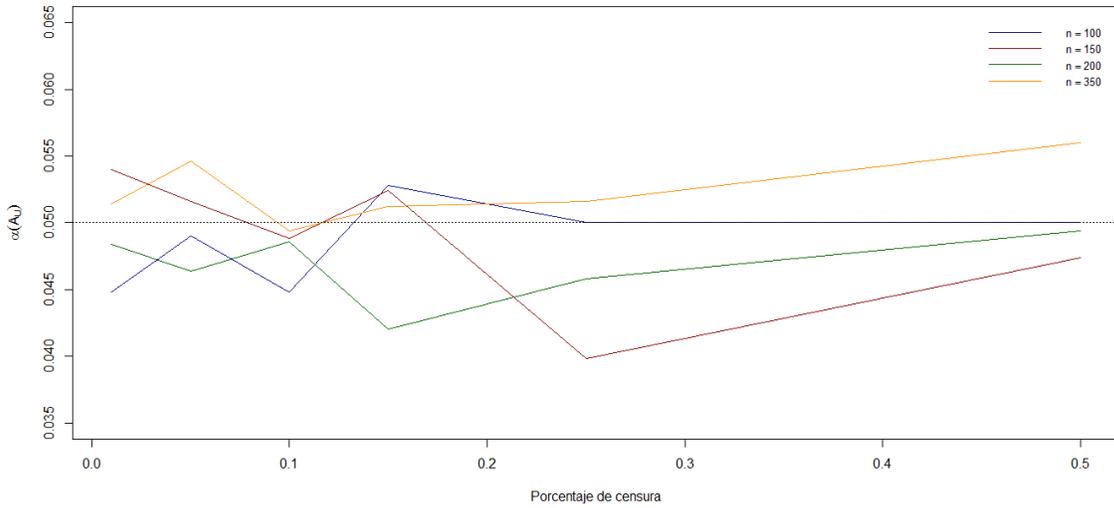


Figura 5.8: Tamaño de la prueba  $A_U$

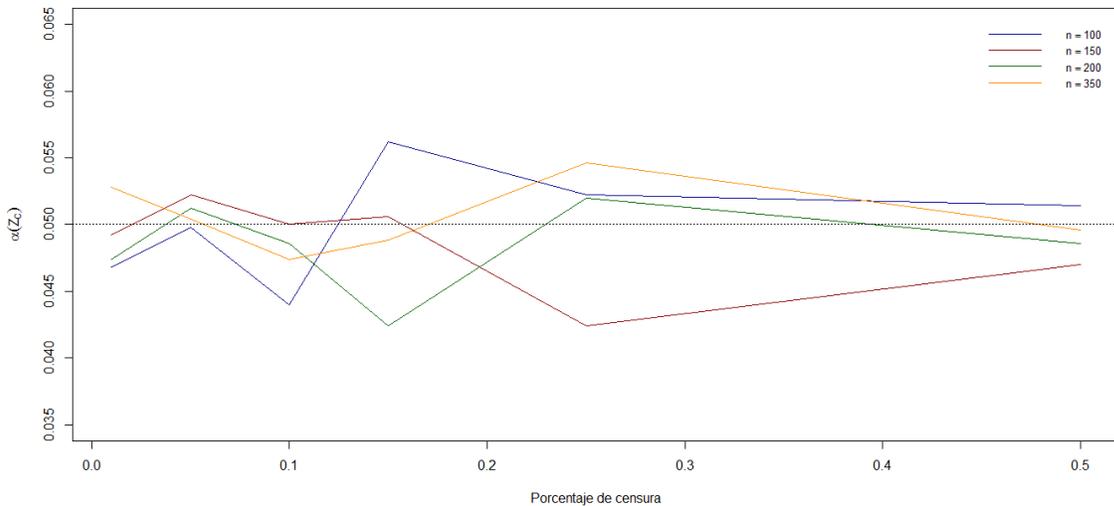
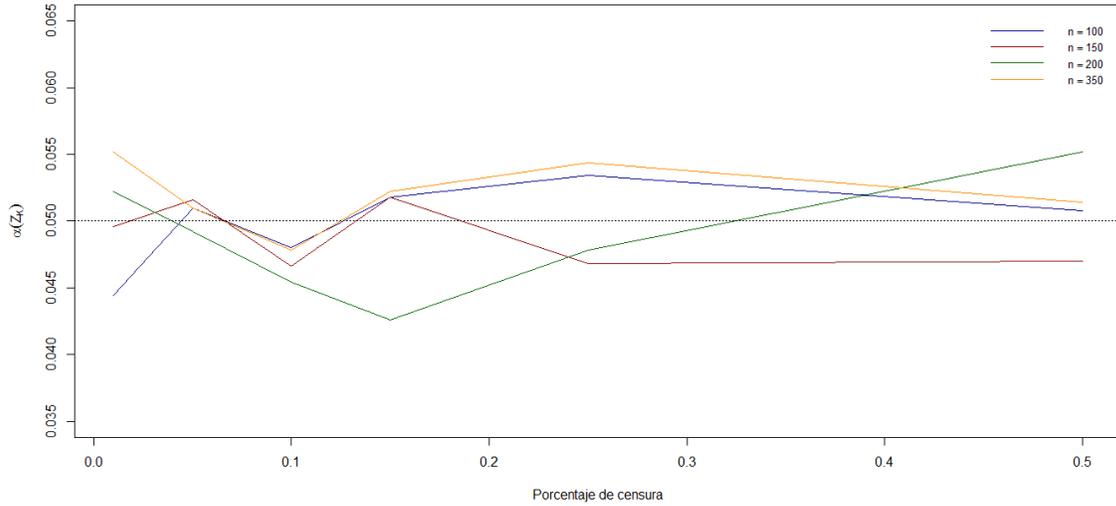


Figura 5.9: Tamaño de la prueba  $Z_C$

## 5.6. Comparación del tamaño de las pruebas $A_U$ , $Z_C$ y $Z_K$ por simulación



**Figura 5.10:** Tamaño de la prueba  $Z_K$

El tamaño de las pruebas  $A_U$ ,  $Z_C$  y  $Z_K$  se mantienen alrededor del valor ( $\alpha = 0.05$ ) fijado al calcular los valores críticos de cada una de las pruebas. En las pruebas analizadas, los niveles de censura no parecen afectar el tamaño de las pruebas aún en situaciones donde se presentan niveles altos de censura.

# Capítulo 6

## Modelación de máximos por bloque de $PM_{10}$ en ZMCM usando la distribución VEG

En esta sección, se analizan los registros de concentraciones máximas de  $PM_{10}$  en 11 estaciones de monitoreo empleando la metodología propuesta en el Capítulo 4. Las estaciones se encuentran en ZMCM. Específicamente, estas son Tlalnepantla (TLA), Xalostoc (XAL), Merced (MER), Pedregal (PED), Tultitlán (TLI), Villa de las Flores (VIF), Tláhuac (TAH), Santa Úrsula (SUR), FES Acatlán (FAC), San Agustín (SAG) e Iztacalco (IZT). [García \(2004\)](#) reporta graves problemas de altas concentraciones de  $PM_{10}$  en las estaciones XAL y TLA. La ubicación de las estaciones se muestra en el Cuadro 6.1,

**Tabla 6.1:** Ubicación en de las estaciones de monitoreo analizadas

Estación	Longitud	Latitud
TLA	-99.20423139	19.52839694
XAL	-99.07644472	19.52774806
MER	-99.11927694	19.42438667
PED	-99.20371583	19.32473472
TLI	-99.17683472	19.60197083
VIF	-99.09630667	19.65767139
TAH	-99.02688528	19.24572917
SUR	-99.14966500	19.31368861
FAC	-99.24327028	19.48192278
SAG	-99.02993917	19.53224722
IZT	-99.11763889	19.38441667

## 6.1. Partículas suspendidas ( $PM_{10}$ )

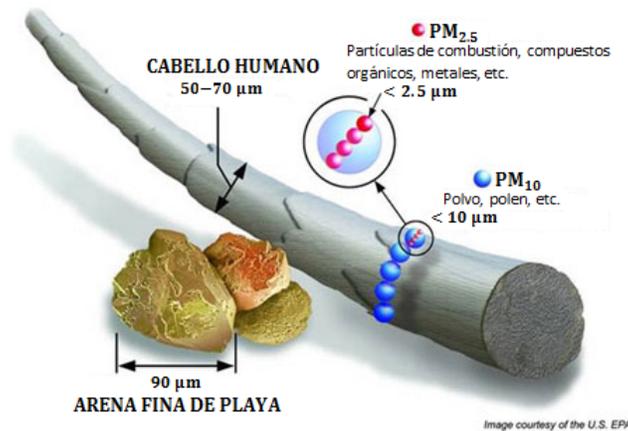


Figura 6.1: Tamaño de las fracciones de material particulado

## 6.1. Partículas suspendidas ( $PM_{10}$ )

Las partículas suspendidas (PM, por sus siglas en inglés) forman una mezcla compleja de materiales sólidos y líquidos suspendidos en el aire, que pueden variar en tamaño, forma y composición, dependiendo fundamentalmente de su origen. El tamaño de las partículas suspendidas varía desde 0.005 hasta 100  $\mu m$  de diámetro.

Las partículas pueden tener un origen natural (como el viento, reacciones químicas o fotoquímicas en la atmósfera, emisiones de volcanes, la polinización de las plantas, procesos geológicos e incendios forestales), y también a causa de actividades humanas (que puede incluir desde la fertilización de campos agrícolas y la circulación de automóviles, hasta la quema de combustibles fósiles por vehículos de transporte y por la industria así como la fundición de metales). Las partículas pueden ser directamente emitidas de la fuente, las llamadas partículas primarias, o bien formarse en la atmósfera cuando en ésta reaccionan algunas sustancias (óxidos de nitrógeno, óxidos de azufre, amoníaco, compuestos orgánicos, etc.), siendo consideradas partículas secundarias.

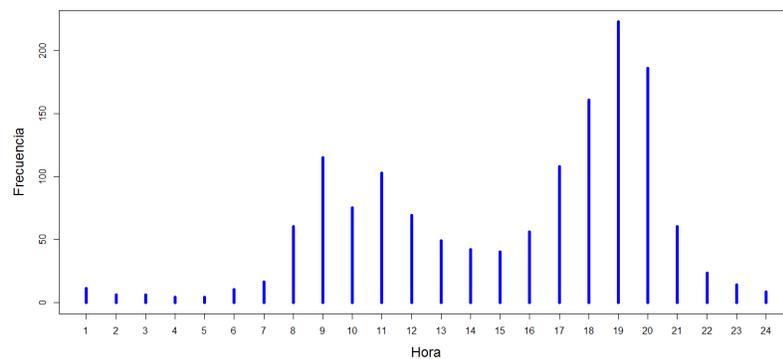
El estudio y la regulación ambiental de las partículas empezó concentrándose en las partículas suspendidas totales (PST) las cuales son menores de 100  $\mu m$ . Sin embargo, con el paso del tiempo se demostró que las partículas relativamente grandes se sedimentan fácilmente y al ingresar al sistema respiratorio se depositan en la región superior de éste en donde son fácilmente eliminadas, mientras que las menores a 10  $\mu m$  se depositan a lo largo del sistema respiratorio y las más pequeñas, menores a 2.5  $\mu m$ , logran alcanzar la región más baja del sistema respiratorio en donde se realiza el intercambio gaseoso.

La [SEDEMA \(2012\)](#) reporta que el tránsito vehicular es la principal fuente de contaminación de la ZMCM por lo que determina, en gran medida, la cantidad de  $PM_{10}$

## 6.1. Partículas suspendidas (PM<sub>10</sub>)

emitidas a la atmósfera. Las condiciones de estabilidad de la atmósfera regulan la concentración de PM<sub>10</sub> en el aire a lo largo del día, una atmósfera inestable propicia una mejor dilución de PM<sub>10</sub> a través del mezclado, difusión o dispersión; mientras que una atmósfera estable favorece el estancamiento de la contaminación.

En la Figura 6.2 se muestra el comportamiento de las frecuencias por hora de las concentraciones máximas de PM<sub>10</sub>. El patrón corresponde a un comportamiento típico de la contaminación en un ambiente urbano dominado por el tránsito vehicular, se observa una distribución bimodal con máximos en la mañana y en la tarde, durante las horas en las que el tránsito reporta la mayor intensidad.



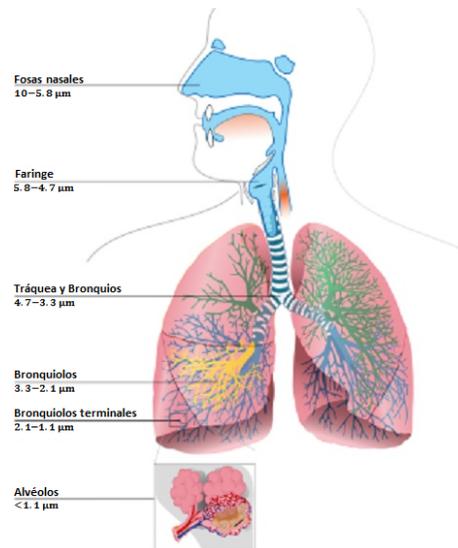
**Figura 6.2:** Frecuencias por hora de los niveles máximos de PM<sub>10</sub>

Para efectos de protección a la salud de la población más susceptible, la Norma Oficial Mexicana NOM-025-SSA1-1993 establece los valores de concentración máxima permisible para PM<sub>10</sub> en  $120 \mu\text{g}/\text{m}^3$  promedio 24 horas y  $50 \mu\text{g}/\text{m}^3$  promedio anual; mientras que la Organización Mundial de la Salud (OMS) fija estos valores en  $50 \mu\text{g}/\text{m}^3$  y  $20 \mu\text{g}/\text{m}^3$ , respectivamente. En el último informe anual de calidad del aire para la ZMCM, la SEDEMA (2012) reporta que tanto los límites de la Norma Oficial Mexicana (NOM) como los de la OMS no se cumplieron.

### 6.1.1. Efectos adversos de las partículas suspendidas (PM<sub>10</sub>) en la salud

Durante las últimas décadas, la calidad del aire en las principales ciudades del país y sus zonas conurbadas ha mostrado una clara tendencia al deterioro. Asimismo, la capacidad de renovación y recuperación del medio ambiente y de los recursos naturales también se ha visto afectada. Consecuentemente, la salud de la población está en riesgo o ya ha sido afectada debido a la presencia de contaminantes del aire ambiente. Entre éstos, las partículas suspendidas son de importancia ya que rebasan los límites de la norma vigente.

## 6.1. Partículas suspendidas ( $PM_{10}$ )



**Figura 6.3:** Material particulado en el sistema respiratorio

La toxicidad de las partículas está determinada por sus características físicas y químicas. El tamaño es un parámetro importante para caracterizar su comportamiento, ya que de él depende la capacidad de penetración y retención en diversas regiones de las vías aéreas respiratorias además de que también determina su tiempo de residencia en la atmósfera. En la Figura 6.3 se aprecia el lugar del sistema respiratorio donde se deposita el material particulado de acuerdo a su tamaño.

La composición química de  $PM_{10}$  es también importante con relación a los daños específicos a la salud. La mayoría de los estudios sobre efectos a la salud descritos en la literatura analizan las asociaciones encontradas entre las concentraciones de partículas en el aire y los daños a la salud. Sin embargo, actualmente se están haciendo esfuerzos importantes para conocer el papel que la composición química y biológica de las partículas tiene en la salud.

Los efectos nocivos de las partículas suspendidas no se limitan al aparato respiratorio, sino que pueden dañar otros aparatos y sistemas como el sistema cardiovascular. Los efectos pueden ser inmediatos o presentarse después de varios días a la exposición. Los daños a la salud inducidos por las partículas han sido estudiados en muchos países y los resultados obtenidos en todos ellos son consistentes y coherentes entre sí. [Mallone \*et al.\* \(2009\)](#), [Samet \*et al.\* \(2000\)](#) y [Zhang \*et al.\* \(2011\)](#) reportan que la exposición a altas concentraciones de  $PM_{10}$  está asociada con un incremento en la mortalidad, además [Anderson \*et al.\* \(2001\)](#), [Utell y Frampton \(2000\)](#) y [Romieu \*et al.\* \(2002a\)](#) muestran que las exposiciones prolongadas a las  $PM_{10}$  producen efectos adversos como tos, síntomas de asma, bronquitis, dificultades para respirar y otros problemas de salud.

La Organización Mundial de la Salud, [OMS \(2006\)](#), estima que más de la mitad

## 6.2. Registro de niveles de $PM_{10}$ en ZMCM

---

de la mortalidad global debida a las  $PM_{10}$  ocurre en países en desarrollo donde la concentración media anual de estas partículas excede  $70 \mu g/m^3$  y estima que una reducción a  $20 \mu g/m^3$  (valor recomendado para protección a la salud), podría reducir la tasa de mortalidad relacionada a estas partículas hasta el 15 %.

Molina y Molina (2004) encuentran que el material particulado suspendido y el ozono son los contaminantes del aire más problemáticos en ZMCM. Los efectos a la salud relacionados con las partículas suspendidas son estudiados por Loomis *et al.* (1999), Romieu *et al.* (2002b) y Holguin *et al.* (2003). Alfaro-Moreno *et al.* (2002) desarrollan estudios experimentales in vitro en diversos tipos de células e indican que las  $PM_{10}$  de la ZMCM tienen efectos citotóxicos y genotóxicos. Los experimentos realizados comparan partículas de tres diferentes regiones de ZMCM, los resultados muestran que tanto las partículas del norte, centro y sur de la ZMCM provocan muerte celular; sin embargo, existen diferencias tóxicas entre las partículas que podrían estar relacionadas con la composición de las mismas. Se ha demostrado que los extractos orgánicos de las  $PM_{10}$  obtenidos en el Centro de la Ciudad de México tienen un mayor potencial mutagénico al encontrado con partículas del Norte de la Ciudad de México y que posiblemente está relacionado con el contenido de hidrocarburos policíclicos. Además, las partículas del Norte y Centro de la ciudad tienen un mayor potencial para inducir rompimientos del ADN.

## 6.2. Registro de niveles de $PM_{10}$ en ZMCM

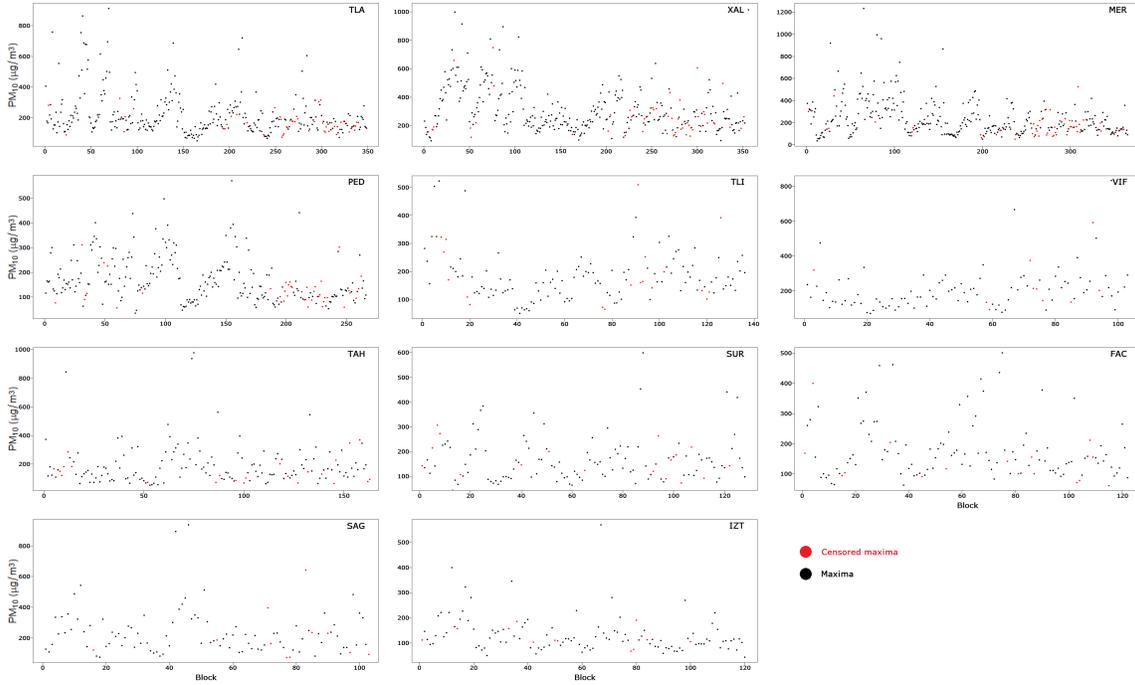
El conjunto de datos se obtuvo de la base de datos de la Red Automática de Monitoreo Atmosférico (RAMA), estos se encuentran disponibles en el siguiente enlace de internet <http://www.aire.df.gob.mx/default.php?opc='27aKBh'>. Las mediciones de las concentraciones de  $PM_{10}$  se realizan cada hora por equipos que operan automáticamente, se analizan los registros de los años 1995 a 2013.

La función de autocorrelación parcial (FAP) se calcula para los datos de  $PM_{10}$  en cada una de las estaciones de monitoreo y se percibe una fuerte dependencia entre las observaciones, sin embargo, al utilizar un tamaño de bloque de 6 días se elimina casi por completo este problema.

La Figura 6.4 muestra las concentraciones máximas de bloque de  $PM_{10}$  para cada una de las estaciones de monitoreo. En algunos bloques faltan lecturas de  $PM_{10}$  que son producto de fallas en los equipos de monitoreo, esto conduce a que se presenten máximos censurados en todas las estaciones. Además, en la Figura 6.4 se puede apreciar un patrón estacional. Los niveles de censura en las estaciones analizadas se encuentra entre el 8 y 13 %.

Para resolver el problema de la estacionalidad en las series de máximos de bloque

### 6.3. Modelación de las concentraciones máximas de $PM_{10}$ en ZMCM



**Figura 6.4:** Máximos de bloque de  $PM_{10}$  registrados en las estaciones de monitoreo localizadas en ZMCM

se incorpora una componente sinusoidal en el modelo VEG así como una covariable que corresponde al año, tal y como se propone en la expresión (4.11). También, para probar el ajuste del modelo como se describe en la Sección 4.2, se calcularon los parámetros del modelo sin la covariable año (4.9).

### 6.3. Modelación de las concentraciones máximas de $PM_{10}$ en ZMCM

En los Cuadros 6.2 y 6.3 se muestran los parámetros estimados del modelo con y sin la covariable *year*, respectivamente. También, el Cuadro 6.2 muestra los *p-values* de la prueba de hipótesis utilizando (4.13), estos indican que el modelo con la covariable *year* no proporciona de manera significativa un mejor ajuste de los datos para las estaciones PED y SUR que el modelo sin la covariable *year*. Así, solo para estas dos estaciones no hay una tendencia significativa en el tiempo.

El Cuadro 6.2 muestra que el valor estimado del parámetro,  $\beta_1$ , asociado con la covariable *year* es positivo en todas las estaciones, lo cual sugiere una tendencia positiva en el tiempo. Además, las estaciones XAL, MER, TLI y SUR presentan las mayores tendencias con valores entre 0.09 y 0.14. Para analizar la tendencia en

### 6.3. Modelación de las concentraciones máximas de PM<sub>10</sub> en ZMCM

**Tabla 6.2:** Parámetros estimados del modelo VEG con covariable *year* para cada una de las estaciones de monitoreo

Parámetros								
Estación	$\xi$	$M$	$A$	$\eta$	$\beta_1$	$\sigma$	Log-verosimilitud	p-value
TLA	0.3941	-10.4250	-39.8180	-4.0683	0.0867	65.2630	-1769.8650	0.0000
XAL	0.1557	-19.1458	-35.1983	2.1498	0.1364	104.7676	-1899.9940	0.0000
MER	0.3789	-15.8241	-46.1026	2.1004	0.0911	77.8272	-1928.4710	0.0000
PED	0.2066	-36.3697	-36.5432	2.1513	0.0829	57.0900	-1299.9897	1.0000
TLI	0.2375	-44.2847	-30.5037	2.0342	0.0935	57.7579	-676.2681	0.0000
VIF	0.2935	-22.1870	-39.8157	2.2972	0.0875	60.2102	-542.9994	0.0060
TAH	0.4501	-27.7053	-25.4159	2.1411	0.0788	62.2717	-834.2520	0.0275
SUR	0.3503	-63.2152	-31.0007	2.3949	0.0953	47.4792	-620.4319	0.1018
FAC	0.2459	-14.8430	-32.0162	2.4735	0.0754	51.8344	-603.3773	0.0000
SAG	0.3564	-1.1869	-51.1598	-4.0310	0.0880	67.5648	-549.4619	0.0000
IZT	0.3259	-41.3459	-22.9772	2.4297	0.0721	33.7645	-568.1034	0.0000

el tiempo también se utilizó un modelo con un término cuadrático adicional en la covariable *year*, sin embargo los *p - value* que se obtuvieron de la prueba en (4.13) muestran que un término cuadrático no es necesario.

Utilizando los parámetros estimados del Cuadro 6.2 se calcularon niveles de retorno a 20 años ( $p = 0.05$ ). Los resultados se muestran en el Cuadro 6.4. En el Cuadro 6.4, se puede apreciar una ligera tendencia positiva para todas las estaciones.

**Tabla 6.3:** Parámetros estimados del modelo VEG sin la covariable *year* para cada una de las estaciones de monitoreo

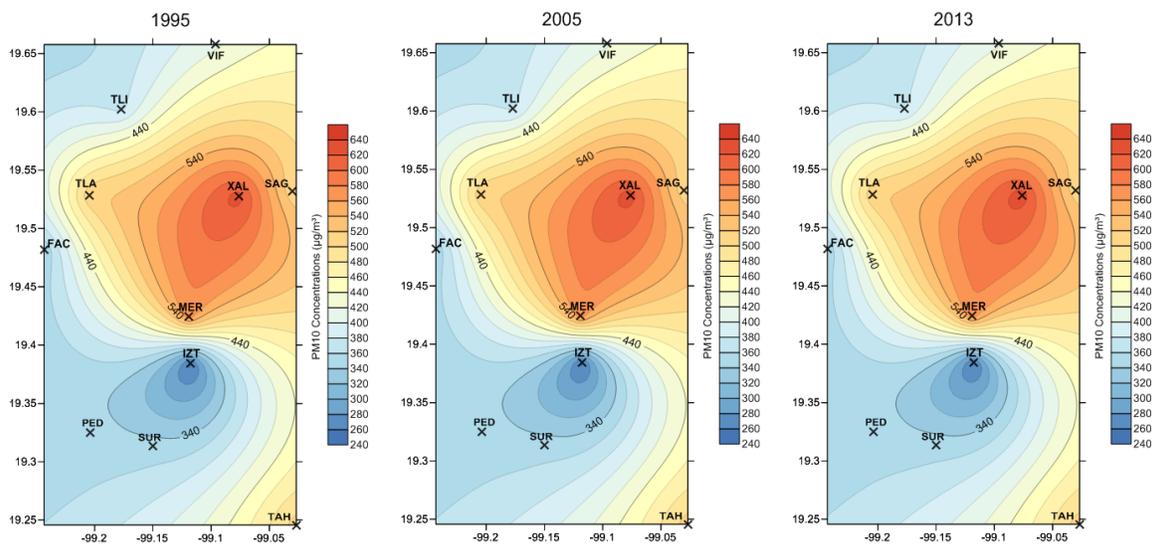
Parámetros						
Estación	$\xi$	$M$	$A$	$\eta$	$\sigma$	Log-verosimilitud
TLA	0.1345	74.6005	-185.4460	-22.9393	102.7542	-1908.6010
XAL	0.1691	263.4701	46.4026	11.8391	135.3086	-1913.5420
MER	0.1317	175.3138	-81.0655	127.5931	91.5347	-1945.7670
PED	0.1818	129.5653	-37.3690	2.1275	57.1429	-1299.5212
TLI	2.0098	60.4485	-95.6402	3.8012	92.9557	-773.6761
VIF	0.1716	147.9971	-61.4420	2.2012	57.6846	-546.7712
TAH	0.5696	122.7669	-22.7832	2.0707	62.2643	-836.6794
SUR	0.5214	123.0414	23.3499	-120.0252	45.0237	-621.7701
FAC	0.1545	122.1792	-76.6372	2.4471	60.8051	-617.8156
SAG	0.2035	159.4148	-157.6094	-129.4608	126.2587	-581.0241
IZT	0.1031	92.7969	-35.4349	-22.8102	27.6726	-583.1341

Para tener una idea del comportamiento espacio-temporal de las concentraciones máximas de PM<sub>10</sub> en ZMCM, se dibujaron mapas de contorno utilizando interpolación *kriging* para los años 1995, 2005 y 2013 para esto se utilizaron los cuantiles del Cuadro 6.4.

### 6.3. Modelación de las concentraciones máximas de $PM_{10}$ en ZMCM

**Tabla 6.4:** Estimación del cuantil 0.95 del modelo VEG con la covariable *year* para cada una de las estaciones de monitoreo

AÑO	TLA	XAL	MER	PED	TLI	VIF	TAH	SUR	FAC	SAG	IZT
1995	515.90	629.28	572.38	348.00	375.61	415.99	509.05	348.30	343.11	503.73	254.37
1996	515.99	629.42	572.48	348.09	375.70	416.08	509.13	348.39	343.18	503.82	254.44
1997	516.08	629.55	572.57	348.17	375.80	416.17	509.21	348.49	343.26	503.91	254.51
1998	516.17	629.69	572.66	348.25	375.89	416.25	509.29	348.58	343.33	503.99	254.59
1999	516.25	629.83	572.75	348.34	375.98	416.34	509.37	348.68	343.41	504.08	254.66
2000	516.34	629.96	572.84	348.42	376.08	416.43	509.44	348.77	343.48	504.17	254.73
2001	516.43	630.10	572.93	348.50	376.17	416.52	509.52	348.87	343.56	504.26	254.80
2002	516.51	630.23	573.02	348.58	376.26	416.60	509.60	348.96	343.63	504.35	254.87
2003	516.60	630.37	573.11	348.67	376.36	416.69	509.68	349.06	343.71	504.43	254.95
2004	516.69	630.51	573.20	348.75	376.45	416.78	509.76	349.15	343.79	504.52	255.02
2005	516.77	630.64	573.30	348.83	376.54	416.87	509.84	349.25	343.86	504.61	255.09
2006	516.86	630.78	573.39	348.92	376.64	416.95	509.92	349.34	343.94	504.70	255.16
2007	516.95	630.92	573.48	349.00	376.73	417.04	510.00	349.44	344.01	504.79	255.23
2008	517.03	631.05	573.57	349.08	376.82	417.13	510.08	349.54	344.09	504.87	255.31
2009	517.12	631.19	573.66	349.16	376.92	417.22	510.15	349.63	344.16	504.96	255.38
2010	517.21	631.33	573.75	349.25	377.01	417.31	510.23	349.73	344.24	505.05	255.45
2011	517.29	631.46	573.84	349.33	377.10	417.39	510.31	349.82	344.31	505.14	255.52
2012	517.38	631.60	573.93	349.41	377.20	417.48	510.39	349.92	344.39	505.23	255.59
2013	517.47	631.74	574.03	349.50	377.29	417.57	510.47	350.01	344.46	505.31	255.67



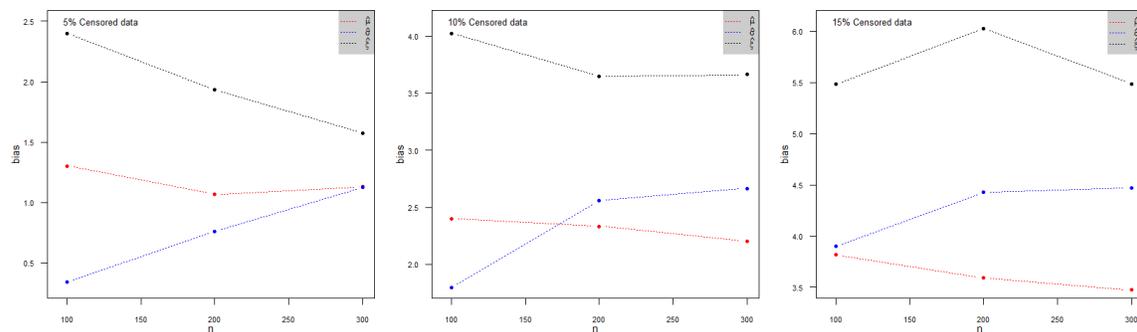
**Figura 6.5:** Estimación del cuantil 0.95 del modelo VEG con covariable *year* para cada una de las estaciones de monitoreo para los años 1995, 2005 y 2013

En la Figura 6.5 se observa que las concentraciones máximas de  $PM_{10}$  se agravan en las estaciones TAH, SAG, XAL, TLA y MER; estas corresponden al Norte y Sureste de la ZMCM. El patrón espacial de las concentraciones máximas de  $PM_{10}$  aparentemente permanece constante durante el periodo de estudio (1995-2013).

## 6.4. Estudio Monte Carlo

Se realiza un estudio Monte Carlo para investigar el efecto del nivel de censura y el tamaño de muestra en el sesgo de los parámetros estimados. El tamaño de la simulación es  $B = 10,000$ , se muestrea de una distribución VEG con parámetros:  $\mu = 150$ ,  $\sigma = 70$  y  $\xi = 0.25$ . Los tamaños de muestra ( $n$ ) que se utilizan son 100, 200, 300 y los niveles de censura 5, 10 y 15%. Se estudian los cuantiles 0.90, 0.95 y 0.99 los cuales corresponden a 361.46, 458.35 y 754.32, respectivamente.

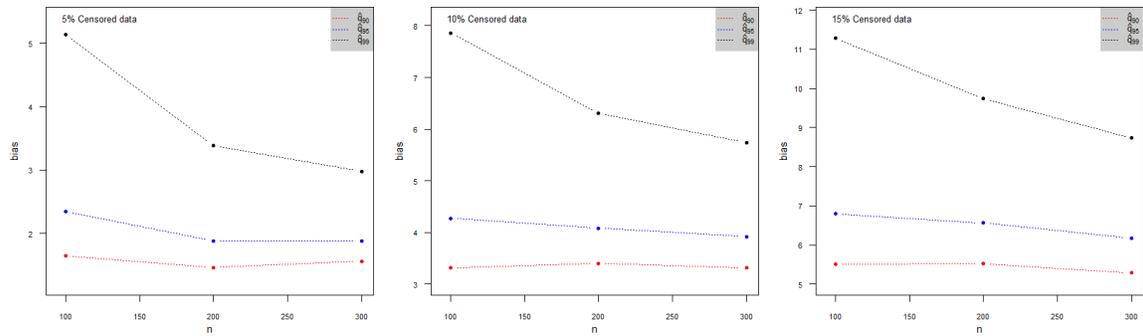
Los resultados del estudio Monte Carlo sobre el sesgo de los parámetros y cuantiles estimados se muestran en las Figuras 6.6 y 6.7, respectivamente.



**Figura 6.6:** Porcentaje de sesgo de EMV en relación a sus valores verdaderos  $\mu = 150$ ,  $\sigma = 70$  y  $\xi = 0.25$  bajo el esquema de censura aleatoria

En la Figura 6.6 se observa que el sesgo en los parámetros de localidad y forma tiende a disminuir conforme aumenta el tamaño de muestra, sin embargo en el caso del parámetro de escala este se incrementa conforme aumenta el tamaño de muestra. Adicionalmente, el sesgo de EMV disminuye cuando el nivel de censura se reduce. El sesgo de EMV de la distribución VEG es menor que 6%, 4% y 2.5% en situaciones donde el nivel de censura es menor que 15%, 10% y 5%, respectivamente. También se observa que el parámetro de forma es el que presenta los niveles de sesgo más altos en relación a los parámetros de localidad y escala.

## 6.4. Estudio Monte Carlo



**Figura 6.7:** Porcentaje de sesgo de los cuantiles estimados en relación a sus valores verdaderos  $q_{90} = 361.46$ ,  $q_{95} = 458.35$  y  $q_{99} = 754.32$  bajo el esquema de censura aleatoria

En la Figura 6.7 se observa que los cuantiles se sobreestiman y que el sesgo disminuye conforme aumenta el tamaño de muestra. Sin embargo, el valor del sesgo aumenta conforme el nivel de censura y el valor del cuantil se incrementan. Se recomienda no utilizar la estimación del cuantil 0.99 cuando el tamaño de muestra sea menor a 200 y el nivel de censura sea mayor que 10 % debido a que en estos escenarios se obtendrán sesgos superiores que 6 %. El sesgo de la estimación de los cuantiles 0.90 y 0.95 es menor que 4.5 % aún en situaciones donde el nivel de censura es 10 %.

# Capítulo 7

## Conclusiones

- El sesgo de EMV de los parámetros de localidad ( $\mu$ ) y forma ( $\xi$ ) de la distribución VEG bajo el esquema de censura aleatoria tiende a disminuir conforme aumenta el tamaño de muestra, sin embargo en el caso del parámetro de escala ( $\sigma$ ) el sesgo se incrementa conforme aumenta el tamaño de muestra.
- El sesgo de los parámetros de la distribución VEG es menor que 6 %, 4 % y 2.5 % en casos donde el nivel de censura es menor que 15 %, 10 % y 5 %, respectivamente. Además, el parámetro de forma es el que presenta los niveles de sesgo más altos en relación a los parámetros de localidad y escala.
- El estudio Monte Carlo muestra que los estimadores bajo el esquema de censura aleatoria sobreestiman el valor de los cuantiles. Se recomienda no utilizar la estimación del cuantil 0.99 en escenarios donde el tamaño de muestra sea menor a 200 y el nivel de censura sea mayor que 10 % debido a que el sesgo es mayor que 6 %. El sesgo de los cuantiles 0.90 y 0.95 es menor que 4.5 % aún en escenarios donde el nivel de censura es de 10 % y el tamaño de muestra es mayor o igual que 100.
- Un término lineal es suficiente para modelar la tendencia de las concentraciones máximas de  $PM_{10}$  y se observa un incremento de la tendencia en todas las estaciones durante el periodo de estudio (1995-2013). El modelo con la covariable *year* proporciona una significativa mejor descripción de los datos para las estaciones TLA, XAL, MER, TLI, VIF, TAH, FAC, SAG e IZT.
- Las concentraciones máximas de  $PM_{10}$  se acentúan en las estaciones TAH, SAG, XAL, TLA y MER; estas se localizan al Norte y Sureste de la ZMCM.
- Las pruebas basadas en los estadísticos  $Z_C$ ,  $Z_K$ , y  $A_U$  no detectan diferencias entre la distribución VEG y la distribución Dagum. Se observa que la potencia se incrementa conforme aumenta el tamaño de muestra y disminuye si el nivel de censura aumenta.

## 7. Conclusiones

---

- Al comparar la distribución VEG y las distribuciones *Weibull*, *Dagum* y *Log-normal* se observan potencias aceptables en las tres pruebas, sin embargo, los niveles de censura altos reducen de manera considerable la potencia de la prueba  $A_U$  mientras que las pruebas  $Z_C$  y  $Z_K$  mantienen valores por arriba de 0.50 aún en casos donde el porcentaje de censura es mayor de 50 % y el tamaño de muestra 100. La prueba  $Z_C$  es la de mayor potencia y la prueba  $A_U$  es la que tiene menor potencia, además los valores de la potencia de la prueba  $Z_K$  son cercanos a los de la prueba  $Z_C$  en todos los casos de alternativas que se estudiaron.
- En las tres pruebas analizadas, el tamaño de la prueba se mantiene alrededor del valor  $\alpha = 0.05$  fijado al calcular los valores críticos de las pruebas,  $A_U$ ,  $Z_C$  y  $Z_K$ . Entonces, se recomienda utilizar la prueba en  $Z_C$  debido a que es la que tiene mayor potencia en relación al resto de las pruebas y además conserva de buena manera el tamaño de la prueba aún en escenarios donde el nivel de censura es mayor de 50 % y el tamaño de muestra 100.

# Referencias

- Ahmad, M. I., Sinclair, C. D. y Spurr, B. D. (1988). Assessment of flood frequency models using empirical distribution function statistics. *Water Resources Research*, 24, 8, 1323–1328. Cited By (since 1996):26.
- Alfaro-Moreno, E., Martínez, L., García-Cuellar, C., Bonner, J. C., Clifford Murray, J., Rosas, I., De Rosales, S. P. L. y Osornio-Vargas, A. R. (2002). Biologic effects induced in vitro by PM 10 from three different zones of Mexico City. *Environmental health perspectives*, 110, 7, 715–720. Cited By (since 1996):90.
- Anderson, H., Bremner, S., Atkinson, R., Harrison, R. y Walters, S. (2001). Particulate matter and daily mortality and hospital admissions in the west midlands conurbation of the United Kingdom: associations with fine and coarse particles, black smoke and sulphate. *Occupational and Environmental Medicine*, 58, 504–510. ISSN 1351-0711.
- Balakrishnan, N., Ng, H. y Kannan, N. (2004). Goodness-of-fit tests based on spacings for progressively Type-II censored data from a general location-scale distribution. *IEEE Transactions on Reliability*, 53, 3, 349–356. Cited By (since 1996)14.
- Beirlant, J., Goegebeur, Y., Segers, J. y Teugels, J. (2004). *Statistics of Extremes Theory and Applications*. John Wiley & Sons, Ltd.
- Bispo, R., Marques, T. A. y Pestana, D. (2012). Statistical power of goodness-of-fit tests based on the empirical distribution function for type-I right-censored data. *Journal of Statistical Computation and Simulation*, 82, 2, 173–181.
- Bogdonavicius, V. B., Levulienė, R. J. y Nikulin, M. S. (2013). Exact goodness-of-fit tests for shape-scale families and type II censoring. *Lifetime Data Analysis*, 19, 3, 413–435.
- Castillo, E. y Hadi, A. (1994). Parameter and quantile estimation for the generalized extreme-value distribution. *Environmetrics*, 5, 4, 417–432. ISSN 1180-4009.
- Chaplin, W. S. (1880). The relation between the tensile strengths of long and short bars. *Van Nostrand's Engineering Magazine*, 23, 441–444.
- Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Verlag, London.
- Davison, A. C. y Smith, R. L. (1990). Models for Exceedances over High Thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52, 393–442.

## Referencias

---

- Denecke, L. y Müller, C. H. (2014). New robust tests for the parameters of the Weibull distribution for complete and censored data. *Metrika*, 77, 5, 585–607.
- Dey, A. K. y Kundu, D. (2012). Discriminating between the Weibull and log-normal distributions for Type-II censored data. *Statistics*, 46, 2, 197–214. Cited By (since 1996):1.
- Dobson, A. J. y Barnett, A. G. (2008). *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC.
- Economou, P. y Tzavelas, G. (2014). Kullback-Leibler divergence measure based tests concerning the biasness in a sample. *Statistical Methodology*, 21, 88–108.
- Finetti, B. D. (1932). Sulla legge di probabilità degli estremi. *Metron*, 9, 127–138.
- Fisher, R. y Tippett, L. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 24, 180–190. ISSN 0008-1981.
- Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. *Ann. Soc. Math. Polon.*, 6, 93–116.
- Gaines, S. D. y Denny, M. W. (1993). The largest, smallest, highest, lowest, longest, and shortest: extremes in ecology. *Ecology*, 74, 1677–1692.
- García, A. (2004). *Contaminantes Atmosféricos en la Zona Metropolitana de la Ciudad de México*. Secretaría de Educación Pública.
- Gibson, E. y Higgins, J. (2000). Gap-ratio goodness of fit tests for Weibull or extreme value distribution assumptions with left or right censored data. *Communications in Statistics-Simulation and Computation*, 29, 2, 541–557. ISSN 0361-0918.
- Gnedenko, B. (1943). Sur La Distribution Limite Du Terme Maximum D'Une Serie Aleatoire. *Annals of Mathematics*, 44, 3, pp. 423–453. ISSN 0003486X.
- Gumbel, E. J. (1934). Les moments des distributions finales de la première et de la dernière valeur. *Comptes Rendus de l'Académie des Sciences*, 198, 1 41–143.
- Gumbel, E. J. (1935a). Les valeurs extremes des distributions statistiques. *Annales de l'Institut 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. (1954). Statistical theory of extreme values and some practical applications. *Nat. Bur. Standards Appl. Math.*.
- Habib, M. G. y Thomas, D. R. (1986). Chi-Square Goodness-of-Fit Tests for Randomly Censored Data. *The Annals of Statistics*, 14, 2, pp. 759–765. ISSN 00905364.
- Heo, J.-H., Shin, H., Nam, W., Om, J. y Jeong, C. (2013). Approximation of modified Anderson-Darling test statistics for extreme value distributions with unknown shape parameter. *Journal of Hydrology*, 499, 41–49. Cited By (since 1996)0.

## Referencias

---

- Holguin, F., Tellez-Rojo, M., Hernandez, M., Cortez, M., Chow, J., Watsow, J., Mannino, D. y Romieu, I. (2003). Air pollution and heart rate variability among the elderly in Mexico City. *Epidemiology*, 14, 521–527. ISSN 1044-3983.
- Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, 81, 348, 158–171. ISSN 1477-870X.
- Kalbfleisch, J. D. y Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, Inc.
- Kim, J. H. (1993). Chi-Square Goodness-of-Fit Tests for Randomly Censored Data. *The Annals of Statistics*, 21, 3, pp. 1621–1639. ISSN 00905364.
- Kleiber, C. y Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. John Wiley, Hoboken, New Jersey.
- Klein, J. P. y Moeschberger, M. L. (2003). *Survival Analysis Techniques for Censored and Truncated Data*. Springer.
- Kotz, S. y Nadarajah, S. (2000). *Extreme Value Distributions: Theory and Applications*. Imperial College Press, London.
- Koziol, J. A. (1980). Goodness-of-fit tests for randomly censored data. *Biometrika*, 67, 3, 693–696. Cited By (since 1996):14.
- Laurens de Hann, A. F. (2006). *Extreme Value Theory An Introduction*. Springer.
- Leadbetter, M. R., Lindgren, G. y Rootzen, H. (1983). *Extremes and related properties of random sequences and processes*. Springer-Verlag, New York.
- Lim, J. y Park, S. (2007). Censored Kullback-Leibler information and goodness-of-fit test with type II censored data. *Journal of Applied Statistics*, 34, 9, 1051–1064. Cited By (since 1996):4.
- Loomis, D., Castillejos, M., Gold, D., McDonnell, W. y Borja-Aburto, V. (1999). Air pollution and infant mortality in Mexico City. *Epidemiology*, 10, 118–123. ISSN 1044-3983.
- Mallone, S., Stafoggia, M., Faustini, A., Gobbi, S., Forastiere, F. y Perucci, C. A. (2009). Effect of Saharan Dust on the Association Between Particulate Matter and Daily Mortality in Rome, Italy. *Epidemiology*, 20, S66–S67. ISSN 1044-3983.
- Molina, L. T. y Molina, M. J. (2004). Improving air quality in megacities - mexico city case study. En *Urban Biosphere and Society: Partnership of Cities*, tomo 1023 de *Annals of the New York Academy of Sciences*, 142–158. ISBN 1-57331-553-2. ISSN 0077-8923.
- Monroy, B. S. (2010). *Modelación de eventos extremos usando la distribución Dagum*. Tesis Doctoral, Colegio de Postgraduados.
- Montfort, M. V. y Gomes, M. (1985). Statistical choice of extremal models for complete and censored data. *Journal of Hydrology*, 77, 77 – 87. ISSN 0022-1694.

## Referencias

---

- Nakajima, J., Kuniyama, T., Omori, Y. y Fruehwirth-Schnatter, S. (2012). Generalized extreme value distribution with time-dependence using the AR and MA models in state space form. *Computational Statistics & Data Analysis*, 56, 11, SI, 3241–3259. ISSN 0167-9473.
- R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Rad, A. H., Yousefzadeh, F. y Balakrishnan, N. (2011). Goodness-of-fit test based on Kullback-Leibler information for progressively type-II censored data. *IEEE Transactions on Reliability*, 60, 3, 570–579. Cited By (since 1996):5.
- Rice, S. O. (1939). The distribution of the maxima of a random curve. *Amer. J. Math.*, 61, 409–416.
- Romieu, I., Samet, J., Smith, K. y Bruce, N. (2002a). Outdoor air pollution and acute respiratory infections among children in developing countries. *Journal of Occupational and Environmental Medicine*, 44, 640–649. ISSN 1076-2752.
- Romieu, I., Sienna-Monge, J., Ramirez-Aguilar, M., Tellez-Rojo, M., Moreno-Macias, H., Reyes-Ruiz, N., del Rio-Navarro, B., Ruiz-Navarro, M., Hatch, G., Slade, R. y Hernandez-Avila, M. (2002b). Antioxidant supplementation and lung functions among children with asthma exposed to high levels of air pollutants. *American Journal of Respiratory and Critical Care Medicine*, 166, 703–709. ISSN 1073-449X.
- Salinas, V., Pérez, P., González, E. y Vaquera, H. (2012). Goodness of fit tests for the gumbel distribution with type II right censored data. *Revista Colombiana de Estadística*, 35, 3, 407–422.
- Samet, J., Dominici, F., Curriero, F., Coursac, I. y Zeger, S. (2000). Fine particulate air pollution and mortality in 20 US Cities, 1987-1994. *New England Journal of Medicine*, 343, 1742–1749. ISSN 0028-4793.
- SEDEMA (2012). *Calidad del Aire en la Ciudad de México Informe 2011*.
- Smith, R. L. (1985). Maximun-likelihood estimation in a class of nonregular cases. *Biometrika*, 72, 1, 67–90. ISSN 0006-3444.
- Smith, T. E. (1984). A choice probability characterization of generalized extreme value models. *Applied Mathematics and Computation*, 14, 1, 35 – 62. ISSN 0096-3003.
- Turnbull, B. W. y Weiss, L. (1978). A Likelihood Ratio Statistic for Testing Goodness of Fit with Randomly Censored Data. *Biometrics*, 34, 3, pp. 367–375. ISSN 0006341X.
- Utell, M. y Frampton, M. (2000). Acute health effects of ambient air pollution: The ultrafine particle hypothesis. *Journal of Aerosol Medicine-Deposition Clearance and Effects in the Lung*, 13, 355–359. ISSN 0894-2684.
- von Mises, R. (1954). "La distribution de la plus grande de n valeurs." *American Mathematical Society*, Selected Papers Volumen II, 271–294.

## Referencias

---

- Wang, B. (2008). Goodness-of-fit test for the exponential distribution based on progressively Type-II censored sample. *Journal of Statistical Computation and Simulation*, 78, 2, 125–132. Cited By (since 1996):3.
- World Health Organization (2006). *Air Quality Guidelines Global Update 2005: Particulate matter, ozone, nitrogen dioxide and sulfur dioxide*.
- Yee, T. W. (2014). *VGAM: Vector Generalized Linear and Additive Models*. R package version 0.9-4.
- Zhang, J. (2002). Powerful goodness-of-fit tests based on the likelihood ratio. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64-2, 281–294.
- Zhang, L., M, X. y Tang, L. (2006). Bias correction for the least squares estimator of Weibull shape parameter with complete and censored data. *Reliability Engineering & System Safety*, 91, 930–939. ISSN 0951-8320.
- Zhang, Y., Zhou, M., Pan, X. y Zhang, J. (2011). Time-series Analysis of Association Between Inhalable Particulate Matter and Daily Mortality in Urban Residents in Tianjin. *Epidemiology*, 22, S228. ISSN 1044-3983.

# Apéndice

## Apéndice A: Valores críticos de las pruebas de bondad de ajuste

**Tabla .1:** Valores críticos ( $C_{n,lc}^{A_U}(\alpha)$ ) de la prueba  $A_U$  con  $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ )

n	Nivel de censura					
	0.01	0.05	0.10	0.15	0.25	0.50
100	1.327610208	1.355173074	1.507691473	1.720585689	2.319743492	4.116480213
150	1.303939157	1.422537312	1.570491215	1.821638807	2.612347381	4.972319670
200	1.304510425	1.410349764	1.646302527	1.980468699	2.826688775	5.703527464
350	1.306733931	1.431339722	1.791642180	2.253859853	3.493946286	7.688366729

**Tabla .2:** Valores críticos ( $C_{n,lc}^{Z_K}(\alpha)$ ) de la prueba  $Z_K$  con  $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ )

n	Nivel de censura					
	0.01	0.05	0.10	0.15	0.25	0.50
100	4.598406167	4.595472444	4.801950558	5.119717575	5.982450578	8.758553495
150	4.724501224	4.845341916	5.086254681	5.485974690	6.612386854	10.27401528
200	4.790359028	4.939532240	5.273132584	5.754912016	7.077946769	11.38522843
350	4.985138496	5.183549866	5.695777573	6.338794227	8.247625986	14.58709233

**Tabla .3:** Valores críticos ( $C_{n,lc}^{Z_C}(\alpha)$ ) de la prueba  $Z_C$  con  $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ )

n	Nivel de censura					
	0.01	0.05	0.10	0.15	0.25	0.50
100	28.59014550	28.22058494	29.91118956	31.76635307	38.38817526	60.09494927
150	29.19575978	30.23436188	31.74023969	34.43346538	43.20372882	72.47897749
200	30.25064642	30.96227664	33.47674594	36.57950882	46.24906453	82.92699896
350	31.83272629	33.54537564	36.88952486	41.79595818	56.41507702	110.4284473

## Apéndice B: Potencia de las pruebas de bondad de ajuste

### Alternativa $Dagum(15.26, 12.41, 18.98)$

**Tabla .4:** Potencia de la prueba  $A_U$  para alternativa  $Dagum(15.26, 12.41, 18.98)$  con  $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ )

n	Nivel de censura					
	0.01	0.05	0.10	0.15	0.25	0.50
100	0.0406	0.0436	0.0422	0.0474	0.0416	0.0366
150	0.0452	0.0464	0.0464	0.0544	0.0354	0.0404
200	0.0468	0.0476	0.0462	0.0412	0.0420	0.0370
350	0.0510	0.0552	0.0466	0.0452	0.0376	0.0330

**Tabla .5:** Potencia de la prueba  $Z_K$  para alternativa  $Dagum(15.26, 12.41, 18.98)$  con  $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ )

n	Nivel de censura					
	0.01	0.05	0.10	0.15	0.25	0.50
100	0.0602	0.0606	0.0588	0.0592	0.0470	0.0458
150	0.0712	0.0696	0.0642	0.0644	0.0472	0.0462
200	0.0828	0.0754	0.0652	0.0562	0.0538	0.0474
350	0.1192	0.0960	0.0732	0.0602	0.0522	0.0456

**Tabla .6:** Potencia de la prueba  $Z_C$  para alternativa  $Dagum(15.26, 12.41, 18.98)$  con  $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ )

n	Nivel de censura					
	0.01	0.05	0.10	0.15	0.25	0.50
100	0.0896	0.0872	0.0772	0.0662	0.0498	0.0380
150	0.1138	0.0976	0.0758	0.0738	0.0568	0.0348
200	0.1212	0.1044	0.0846	0.0684	0.0532	0.0300
350	0.1716	0.1292	0.0960	0.0804	0.0498	0.0256

**Alternativa  $\Gamma(121.99, 7.78)$**

**Tabla .7:** Potencia de la prueba  $A_U$  para alternativa  $\Gamma(121.99, 7.78)$  con  $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ )

n	Nivel de censura					
	0.01	0.05	0.10	0.15	0.25	0.50
100	0.3626	0.3474	0.3254	0.2802	0.1876	0.0814
150	0.5294	0.4980	0.4668	0.3928	0.2490	0.0822
200	0.6848	0.6618	0.5696	0.4906	0.3178	0.0908
350	0.9430	0.9174	0.8396	0.7392	0.4786	0.1036

**Tabla .8:** Potencia de la prueba  $Z_K$  para alternativa  $\Gamma(121.99, 7.78)$  con  $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ )

n	Nivel de censura					
	0.01	0.05	0.10	0.15	0.25	0.50
100	0.9494	0.9434	0.9304	0.8980	0.8108	0.4596
150	0.9946	0.9908	0.9868	0.9760	0.9274	0.5706
200	0.9994	0.9986	0.9968	0.9934	0.9672	0.6760
350	1.0000	1.0000	1.0000	1.0000	0.9986	0.8396

**Tabla .9:** Potencia de la prueba  $Z_C$  para alternativa  $\Gamma(121.99, 7.78)$  con  $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ )

n	Nivel de censura					
	0.01	0.05	0.10	0.15	0.25	0.50
100	0.9772	0.9754	0.9664	0.9520	0.8980	0.6504
150	0.9988	0.9966	0.9960	0.9926	0.9740	0.7932
200	1.0000	0.9998	0.9994	0.9978	0.9892	0.8654
350	1.0000	1.0000	1.0000	1.0000	0.9998	0.9650

**Alternativa  $Log - normal(2.74, 0.08)$**

**Tabla .10:** Potencia de la prueba  $A_U$  para alternativa  $Log - normal(2.74, 0.08)$  con  $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ )

n	Nivel de censura					
	0.01	0.05	0.10	0.15	0.25	0.50
100	0.1064	0.1058	0.0920	0.0756	0.0438	0.0248
150	0.1672	0.1462	0.1178	0.0920	0.0486	0.0142
200	0.2362	0.1942	0.1408	0.1020	0.0544	0.0126
350	0.5862	0.4522	0.2606	0.1546	0.0694	0.0078

**Tabla .11:** Potencia de la prueba  $Z_K$  para alternativa  $Log - normal(2.74, 0.08)$  con  $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ )

n	Nivel de censura					
	0.01	0.05	0.10	0.15	0.25	0.50
100	0.9444	0.9404	0.9070	0.8820	0.7868	0.4250
150	0.9918	0.9864	0.9804	0.9694	0.8994	0.5290
200	0.9992	0.9980	0.9954	0.9894	0.9600	0.6252
350	1.0000	1.0000	1.0000	1.0000	0.9962	0.8046

**Tabla .12:** Potencia de la prueba  $Z_C$  para alternativa  $Log - normal(2.74, 0.08)$  con  $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ )

n	Nivel de censura					
	0.01	0.05	0.10	0.15	0.25	0.50
100	0.9798	0.9722	0.9608	0.9386	0.8934	0.6450
150	0.9984	0.9974	0.9926	0.9874	0.9640	0.7586
200	0.9998	0.9996	0.9994	0.9976	0.9882	0.8540
350	1.0000	1.0000	1.0000	1.0000	0.9992	0.9570

**Alternativa *Weibull*(7.90, 16.39)**

**Tabla .13:** Potencia de la prueba  $A_U$  para alternativa *Weibull*(7.90, 16.39) con  $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ )

n	Nivel de censura					
	0.01	0.05	0.10	0.15	0.25	0.50
100	0.9800	0.9708	0.9256	0.8458	0.5464	0.1114
150	0.9998	0.9986	0.9918	0.9700	0.7356	0.1204
200	1.0000	1.0000	0.9994	0.9968	0.8824	0.1470
350	1.0000	1.0000	1.0000	1.0000	0.9934	0.2050

**Tabla .14:** Potencia de la prueba  $Z_K$  para alternativa *Weibull*(7.90, 16.39) con  $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ )

n	Nivel de censura					
	0.01	0.05	0.10	0.15	0.25	0.50
100	1	1	1	1	1	0.9990
150	1	1	1	1	1	1.0000
200	1	1	1	1	1	1.0000
350	1	1	1	1	1	1.0000

**Tabla .15:** Potencia de la prueba  $Z_C$  para alternativa *Weibull*(7.90, 16.39) con  $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ )

n	Nivel de censura					
	0.01	0.05	0.10	0.15	0.25	0.50
100	1	1	1	1	1	1
150	1	1	1	1	1	1
200	1	1	1	1	1	1
350	1	1	1	1	1	1

## Apéndice C: Tamaño de las pruebas de bondad de ajuste

**Tabla .16:** Tamaño de la prueba  $A_U$  con  $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ )

n	Nivel de censura					
	0.01	0.05	0.10	0.15	0.25	0.50
100	0.0448	0.0490	0.0448	0.0528	0.0500	0.0500
150	0.0540	0.0516	0.0488	0.0524	0.0398	0.0474
200	0.0484	0.0464	0.0486	0.0420	0.0458	0.0494
350	0.0514	0.0546	0.0494	0.0512	0.0516	0.0560

**Tabla .17:** Tamaño de la prueba  $Z_K$  con  $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ )

n	Nivel de censura					
	0.01	0.05	0.10	0.15	0.25	0.50
100	0.0444	0.0510	0.0480	0.0518	0.0534	0.0508
150	0.0496	0.0516	0.0466	0.0518	0.0468	0.0470
200	0.0522	0.0492	0.0454	0.0426	0.0478	0.0552
350	0.0552	0.0510	0.0478	0.0522	0.0544	0.0514

**Tabla .18:** Tamaño de la prueba  $Z_C$  con  $\alpha = 0.05$ , diferentes tamaños de muestra ( $n$ ) y diferentes niveles de censura ( $lc$ )

n	Nivel de censura					
	0.01	0.05	0.10	0.15	0.25	0.50
100	0.0468	0.0498	0.0440	0.0562	0.0522	0.0514
150	0.0492	0.0522	0.0500	0.0506	0.0424	0.0470
200	0.0474	0.0512	0.0486	0.0424	0.0520	0.0486
350	0.0528	0.0504	0.0474	0.0488	0.0546	0.0496

## Apéndice D: Programas de R usados en el trabajo

### Ajuste de la distribución VEG bajo el esquema de censura aleatoria y estacionalidad

Para realizar el ajuste de la distribución VEG es necesario instalar el paquete [VGAM](#) en el programa [R](#).

```
require(VGAM)

###1.Lectura de los datos (registros por hora)
Y=as.matrix(read.csv(file="/Users/BED/Desktop/PM10/PM10_t.csv",header=TRUE))
###-----

###2. Selección de datos para la estación bajo estudio
###Carga (PM10 de Estación de estudio,Hora,t,Año,Día)
###En est, seleccionar la estación que se desea analizar
est=1
Y2=Y[,c((est+1),1,13,14,15)]

###3. Máximos por bloque
### En bloq, seleccionar la longitud del bloque en días
vare=1
cov1=4
bloqd=6
bloq=24*bloqd
delta=round(365/bloqd)
b2=bloq-1
nr=trunc(nrow(Y2)/bloq)
nc=ncol(Y2)

####3.1 Observaciones completas por bloque####
Mb_2=matrix(rep(0,nr*(nc+1)),nr,nc+1)
for (i in 1:nr){
temp=matrix(Y2[((bloq*i)-b2):(bloq*i),],bloq,nc)
pos=min(which(temp[,vare]==max(temp[,vare])))
Mb_2[i,1:nc]=temp[pos,]
Mb_2[i,nc+1]=round((max(temp[,nc]))/bloqd,0)
}

####3.2 Observaciones censuradas por bloque####
Mb_2c=matrix(rep(0,nr*(nc+1)),nr,nc+1)
for (i in 1:nr){
temp=matrix(Y2[((bloq*i)-b2):(bloq*i),],bloq,nc)
pos=min(which(temp[,vare]==max(temp[,vare],na.rm=T)))
```

## Apéndice

---

```
Mb_2c[i,1:nc]=temp[pos,]
Mb_2c[i,nc+1]=round((max(temp[,nc]))/bloqd,0)
}

####3.3 Variable censura####
censure=matrix(is.na(Mb_2[,vare]),nr,1)
censure=replace(censure, censure == FALSE, 0)
####3.4 Eliminando datos perdidos####
Mb_2cd=cbind(Mb_2c,censure)
lost_ve=is.na(Mb_2cd[,vare])
lve_Mb_2cd=which(lost_ve)
Mb_2cd=Mb_2cd[-lve_Mb_2cd,]
nlost=length(lve_Mb_2cd)

####3.4 Datos perdidos y censurados####
d_2cen=which(Mb_2cd[,ncol(Mb_2cd)]==1)
d_2obs=which(Mb_2cd[,ncol(Mb_2cd)]==0)
data_2obs=Mb_2cd[-d_2cen,]
data_2cen=Mb_2cd[-d_2obs,]
t_obs=round(data_2obs[,3]/bloq)
t_cen=round(data_2cen[,3]/bloq)

###3.5 Porcentaje de censura###
obs=nrow(data_2obs)
cen_o=nrow(data_2cen)
p_cen_t=round(runif(1,0.10,0.15),digits=2)
cen=round(p_cen_t*obs/(1-p_cen_t))
delet=sample.int(cen_o,cen_o-cen)
data_2cen=data_2cen[-delet,]
cen=nrow(data_2cen)
lev_censure=round(cen*100/(obs+cen),digits=2)
DATA_=rbind(data_2obs,data_2cen)
DATA=DATA_[order(DATA_[,3]),]
DATAgraf=cbind(matrix(seq(1,nrow(DATA)),by=1),nrow(DATA),1),DATA)
DATAgraf_cen=DATAgraf[-(which(DATAgraf[,8]==0)),]
DATAgraf_obs=DATAgraf[-(which(DATAgraf[,8]==1)),]
#-----#

###3.6 Gráfico FACP y de los datos###
x11()
pacf(DATA[,1],main="")
legend("topright","FAC",bty="n",cex=0.8,pt.cex=1,text.font=2)
x11()
acf(DATA[,1],main="")
legend("topright","FAC",bty="n",cex=0.8,pt.cex=1,text.font=2)
x11()
plot(DATAgraf_obs[,1],DATAgraf_obs[,2],type="p",pch=20,col="black",cex=1,
```

## Apéndice

---

```
xlab="Block",ylab=expression(paste(PM[10]," (" ,mu,"g"/m[3],")")),cex.axis=0.7,cex.lab=0.9,1
lines(DATAgraf_cen[,1],DATAgraf_cen[,2],type="p",pch=20,col="red",cex=1)
legend("topright","FAC",bty="n",cex=0.8,pt.cex=1,text.font=2)
#-----#
```

```
####4. Ajuste del modelo DVEG####
```

```
w=2*pi/delta
```

```
#####4.1 (Tendencia y Estacionalidad)#####
```

```
vero1=function(a){
temp1=matrix(NA,obs,1)
temp1=log(dgev(data_2obs[,1],a[1]+a[2]*cos(w*data_2obs[,6]+a[3])+a[4]*
data_2obs[,4],a[5],a[6], log = FALSE))
temp2=matrix(NA,cen,1)
temp2=log(1-pgev(data_2cen[,1],a[1]+a[2]*cos(w*data_2cen[,6]+a[3])+a[4]*
data_2cen[,4],a[5],a[6]))
vero=-1*(sum(temp1)+sum(temp2))
return(vero)
}
par.ajus1=optim(c(-34,-28,2.31,0.09,77,0.38),vero1)
par.opt1=as.data.frame(par.ajus1[1])
          veropt1=as.data.frame(par.ajus1[2])
          conver1=as.data.frame(par.ajus1[4])
```

```
#####4.2 (Estacionalidad)#####
```

```
vero2=function(a){
temp1=matrix(NA,obs,1)
temp1=log(dgev(data_2obs[,1],a[1]+a[2]*cos(w*data_2obs[,6]+a[3]),a[4],a[5],
log = FALSE))
temp2=matrix(NA,cen,1)
temp2=log(1-pgev(data_2cen[,1],a[1]+a[2]*cos(w*data_2cen[,6]+a[3]),a[4],a[5]))
vero=-1*(sum(temp1)+sum(temp2))
return(vero)
}
par.ajus2=optim(c(-34,-28,2.31,77,0.38),vero2)
par.opt2=as.data.frame(par.ajus2[1])
          veropt2=as.data.frame(par.ajus2[2])
          conver2=as.data.frame(par.ajus2[4])
```

```
#####5. Prueba de hipótesis para tendencia#####
```

```
DEVIANCE=2*(-veropt1[1,1]+veropt2[1,1])
p_value=pchisq(DEVIANCE, df=1, ncp = 0, lower.tail = FALSE, log.p = FALSE)
```

```
#####6. Estimación de Cuantiles por año y guardando resultados#####
```

```
year=seq(1995,2013,by=1)
Q_year=matrix(round(qgev(0.95,par.opt1[1,1]+par.opt1[2,1]*
cos(w*mean(data_2obs[,6])+par.opt1[3,1])+par.opt1[4,1]*
```

```
(year),par.opt1[5,1],par.opt1[6,1]),digits=2),1,19)
colnames(Q_year)=c(year)
MEyT=matrix(c(par.opt1[1,1],par.opt1[2,1],par.opt1[3,1],par.opt1[4,1],
par.opt1[5,1],par.opt1[6,1],veropt1[1,1],p_value),1,8)
colnames(MEyT)=c("M","A","eta","beta[1]","sigma","xi","log-likelihood","p-value")
ME=matrix(c(par.opt2[1,1],par.opt2[2,1],par.opt2[3,1],par.opt2[4,1],
par.opt2[5,1],veropt2[1,1]),1,6)
colnames(ME)=c("M","A","eta","sigma","xi","log-likelihood")

write.csv(Q_year,"01_Q95_FAC.csv")
write.csv(MEyT,"01_MEyT_FAC.csv")
write.csv(ME,"01_ME_FAC.csv")
write.csv(DATAgraf,"01_DATA_FAC.csv")
```

## Cálculo de los valores críticos

```
require(VGAM)

EST_Akac_D=function(N=50,lcen=0.5,location=15,scale=1,shape=0.1)
{
data=matrix(NA,N,2)
n_cen=round(N*lcen)
n_com=N-n_cen
if (n_cen==0){n_cen==1}
if (n_com==0){n_com==1}

cen_0=rep(0,n_cen)
cen_1=rep(1,n_com)
censure=c(cen_0,cen_1)
rm(cen_0);rm(cen_1)
data[,2]=sample(censure)
rm(censure)

#En data[,1] se introducen las observaciones muestreadas de la
#distribución bajo Ho
data[,1]=matrix(rgev(N,15,1,0.1),N,1)
min_d=min(data[,1])

for (i in 1:N){
  if (data[i,2]==0){
    data[i,1]=runif(1,min_d,data[i,1])}
}

cen=which(data[,2]==0)
d_cen=matrix(data[cen,],n_cen,1)
```

## Apéndice

---

```
com=which(data[,2]==1)
d_com=matrix(data[com,],n_com,1)
rm(com)

par.opt=matrix(c(location,scale,shape),3,1)

#####Muestra artificial#####
art_sam=function(c_i,par.opt){
U=runif(1,0,1)
GEV=pgev(c_i,par.opt[1,1],par.opt[2,1],par.opt[3,1])
S_GEV=1-GEV
GEV.Y_i_hat=(U*S_GEV)+GEV
Y_i_hat=qgev(GEV.Y_i_hat,par.opt[1,1],par.opt[2,1],par.opt[3,1])
return(Y_i_hat)
}

#Aplicando la función para completar la muestra
d_cen_aux=matrix(d_cen[,1],n_cen,1)
d_cen_est=matrix(apply(d_cen_aux,1, art_sam, par.opt),n_cen,1)
data=data[-cen,-2]
data=matrix(c(data,d_cen_est),N,1)
par.opt2=matrix(c(location,scale,shape),3,1)

#Cálculo de los estadísticos de prueba bajo Ho
data=matrix(sort(data),N,1)
c1=N/2
agr1=-2*sum(pgev(c(data),par.opt2[1,1],par.opt2[2,1],par.opt2[3,1]))

Z_Akac=matrix(NA,N,4)
for (i in 1:N){
p_i=pgev(data[i,1],par.opt2[1,1],par.opt2[2,1],par.opt2[3,1])
Z_Akac[i,1]=(i-1/2)*log((i-1/2)/(N*p_i))+(N-i+1/2)*log((N-i+1/2)/(N*(1-p_i)))
Z_Akac[i,2]=((log(p_i))/(N-i+1/2))+((log(1-p_i))/(i-1/2))
Z_Akac[i,3]=(log(((p_i^-1)-1)/((N-1/2)/(i-3/4)-1)))^2
Z_Akac[i,4]=(2-((2*i-1)/N))*log(1-p_i)
}

Z_k=max(Z_Akac[,1])
Z_c=sum(Z_Akac[,3])
Au=c1+agr1-sum(Z_Akac[,4])

out=matrix(c(Z_k,Z_c,Au),1,3)
return(out)
}

#Implementación Monte Carlo
START=Sys.time()
```

## Apéndice

---

```
B=25000
KACs=matrix(NA,B,3)
colnames(KACs)=c("Z_k","Z_c","Au")
for (i in 1:B){
print(i)
KACs[i,]=EST_Akac_D(N=100,lcen=0.01,location=15,scale=1,shape=0.10)
}
write.csv(KACs,"EST_Akac_100_0.01.csv")
END=Sys.time()
START
END

#Valor de alfa
p=0.95

#Estimación vía Monte Carlo del valor crítico
CRITICAL_Akac_100_0.01=as.matrix(read.csv("CRITICAL_Akac_100_0.01.csv", header = TRUE))
Cvalue_k_100_0.01=quantile(CRITICAL_Akac_100_0.01[,2],P)
Cvalue_c_100_0.01=quantile(CRITICAL_Akac_100_0.01[,3],P)
Cvalue_Au_100_0.01=quantile(CRITICAL_Akac_100_0.01[,4],P)
rm(CRITICAL_Akac_100_0.01)
```

## Cálculo de las potencias

```
require(VGAM)

#Sys.time()
#####INICIO DE LA PRUEBA#####

EST_Akac_L=function(N=50,lcen=0.5,location=15,scale=1,shape=0.1)
{
data=matrix(NA,N,2)
n_cen=round(N*lcen)
n_com=N-n_cen
if (n_cen==0){n_cen==1}
if (n_com==0){n_com==1}

cen_0=rep(0,n_cen)
cen_1=rep(1,n_com)
censure=c(cen_0,cen_1)
rm(cen_0);rm(cen_1)
data[,2]=sample(censure)
rm(censure)

###Selección de la hipótesis H1
```

## Apéndice

---

```
data[,1]=matrix(rlnorm(N, 2.74, 0.08),N,1)
#data[,1]=rdagum(N,15.26,12.41,18.98)
#data[,1]=rgamma(N, 121.99, 7.78)
#data[,1]=rweibull(N, 7.90 ,16.39)

min_d=min(data[,1])

for (i in 1:N){
  if (data[i,2]==0){
    data[i,1]=runif(1,min_d,data[i,1])}
}

cen=which(data[,2]==0)
d_cen=matrix(data[cen,],n_cen,1)
com=which(data[,2]==1)
d_com=matrix(data[com,],n_com,1)
rm(com)
par.opt=matrix(c(location,scale,shape),3,1)

#####Muestra artificial#####
art_sam=function(c_i,par.opt){
U=runif(1,0,1)
GEV=pgev(c_i,par.opt[1,1],par.opt[2,1],par.opt[3,1])
S_GEV=1-GEV
GEV.Y_i_hat=(U*S_GEV)+GEV
Y_i_hat=qgev(GEV.Y_i_hat,par.opt[1,1],par.opt[2,1],par.opt[3,1])
return(Y_i_hat)
}

##Completando la muestra
d_cen_aux=matrix(d_cen[,1],n_cen,1)
d_cen_est=matrix(apply(d_cen_aux,1, art_sam, par.opt),n_cen,1)
data=data[-cen,-2]
data=matrix(c(data,d_cen_est),N,1)
par.opt2=matrix(c(location,scale,shape),3,1)

#Cálculo de los estadísticos de prueba
data=matrix(sort(data),N,1)
c1=N/2
agr1=-2*sum(pgev(c(data),par.opt2[1,1],par.opt2[2,1],par.opt2[3,1]))

Z_Akac=matrix(NA,N,4)
for (i in 1:N){
p_i=pgev(data[i,1],par.opt2[1,1],par.opt2[2,1],par.opt2[3,1])
Z_Akac[i,1]=(i-1/2)*log((i-1/2)/(N*p_i))+(N-i+1/2)*log((N-i+1/2)/(N*(1-p_i)))
Z_Akac[i,2]=((log(p_i))/(N-i+1/2))+((log(1-p_i))/(i-1/2))
Z_Akac[i,3]=(log(((p_i^-1)-1)/((N-1/2)/(i-3/4)-1)))^2
```

## Apéndice

---

```
Z_Akac[i,4]=(2-((2*i-1)/N))*log(1-p_i)
}

Z_k=max(Z_Akac[,1])
Z_c=sum(Z_Akac[,3])
Au=c1+agr1-sum(Z_Akac[,4])

out=matrix(c(Z_k,Z_c,Au),1,3)
return(out)
}

#Implementación Monte Carlo
START=Sys.time()
B=5000
KACs=matrix(NA,B,3)
colnames(KACs)=c("Z_k","Z_c","Au")
for (i in 1:B){
print(i)
KACs[i,]=EST_Akac_L(N=100,lcen=0.01,location=15,scale=1,shape=0.10)
}
write.csv(KACs,"EST_Akac_L_100_0.01.csv")
END=Sys.time()
START
END

#Valores críticos de A_U(Ver Tabla .1 del Apéndice)
est_0.05a=as.matrix(read.csv("CRITICALa_0.05.csv", header = T))
#Valores críticos de Z_K(Ver Tabla .2 del Apéndice)
est_0.05k=as.matrix(read.csv("CRITICALk_0.05.csv", header = T))
#Valores críticos de Z_C(Ver Tabla .3 del Apéndice)
est_0.05c=as.matrix(read.csv("CRITICALc_0.05.csv", header = T))

#Estimación de la potencia de las pruebas vía Monte Carlo
D_100_0.01_a=as.matrix(read.csv("EST_Akac_L_100_0.01.csv", header = T))
rej=which(D_100_0.01[,4] >= est_0.05a[1,2])
P_100_0.01_a=length(rej)/B

D_100_0.01_c=as.matrix(read.csv("EST_Akac_L_100_0.01.csv", header = T))
rej=which(D_100_0.01[,3] >= est_0.05c[1,2])
P_100_0.01_c=length(rej)/B

D_100_0.01_k=as.matrix(read.csv("EST_Akac_L_100_0.01.csv", header = T))
rej=which(D_100_0.01[,2] >= est_0.05k[1,2])
P_100_0.01_k=length(rej)/B

#Estimación del tamaño de las pruebas vía Monte Carlo
EST_100_0.01_a=as.matrix(read.csv("EST_Akac_100_0.01.csv", header = T))
```

## Apéndice

---

```
rej=which(EST_100_0.01[,4] >= est_0.05_a[1,2])  
S_100_0.01_a=length(rej)/B
```

```
EST_100_0.01_c=as.matrix(read.csv("EST_Akac_100_0.01.csv", header = T))  
rej=which(EST_100_0.01_c[,3] >= est_0.05_c[1,2])  
S_100_0.01_c=length(rej)/B
```

```
EST_100_0.01_k=as.matrix(read.csv("EST_Akac_100_0.01.csv", header = T))  
rej=which(EST_100_0.01_k[,2] >= est_0.05_k[1,2])  
S_100_0.01_k=length(rej)/B
```