UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

PROYECTO FIN DE CARRERA

CARACTERIZACIÓN DE TRÁFICO A PARTIR DE TRAZAS DE TRÁFICO REALES: APLICACIÓN A RedIRIS

Gonzalo Polo Vera

Noviembre 2010

CARACTERIZACIÓN DE FLUJOS A PARTIR DE TRAZAS DE TRÁFICO REALES: APLICACIÓN A RedIRIS

Autor: Gonzalo Polo Vera

Tutor: Javier Aracil Rico

High Performance Computing and Networking

Dpto. de Ingeniería Informática

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Noviembre 2010

PALABRAS CLAVE

Flujo de tráfico; Distribución de cola pesada; Índice de cola; Distribución Generalizada de Pareto; Distribución Pura de Pareto; Distribución log-Normal; Sentido de tráfico ascendente y descendente; Diversidad temporal y espacial.

RESUMEN

Tanto los proveedores de servicio de Internet, como la comunidad investigadora, han comprendido la importancia de las medidas de tráfico real de redes de datos. Por un lado, la comunidad científica utiliza estas medidas para caracterizar Internet contribuyendo al mejor conocimiento de su dinámica. Por otro lado, los proveedores las utilizan para asegurar la calidad de servicio, detectar problemas, facturar y para dimensionar correctamente sus redes, entre otras aplicaciones.

El presente PFC pertenece al primer grupo. Su objetivo es la caracterización de flujos a partir de trazas de tráfico provenientes de RedIRIS. Concretamente analiza la distribución del tamaño en Bytes que poseen los flujos Web. Se apoya en distintos estudios que han analizado distintas métricas del tráfico Web. Unos, mediante la utilización de distribuciones de cola pesada, para obtener una mayor comprensión de fenómenos observados como la autosimilaridad o la dependencia a larga escala del tráfico y otros, que por el contrario utilizan otro tipo de distribuciones como la log-Normal.

Mediante esta clase de distribuciones se analiza el tamaño de los flujos Web para determinar cuál es la más apropiada y a su vez, caracterizarlos respecto a sus propiedades tanto temporales como espaciales.

ABSTRACT

Both Internet service providers and the research community have understood the importance of real traffic measurements. On the one hand, the research community uses these measurements to characterize the network traffic in order to improve the understanding of the Internet dynamics. On the other hand, ISP's use them to ensure the quality of service, detect problems, invoice clients and perform capacity planning, among other applications.

This project is accounted into the second group. Its aim is the description of Web flow from traffic traces retrieved from RedIRIS. Specifically it analyzes the Web flows' size distribution in Bytes. To this end, several studies of a variety of Web traffic measures have been taken into consideration. Some of them use heavy-tail distributions to achieve a better understanding of the observe traffic phenomena, such as self-similarity and long range dependence. Others use other kind of distribution such as log-Normal.

By means of these distributions, Web traffic flows sizes are analyzed to test which is the most appropriate and at the same time, it defines them regarding their temporary and spatial features.

Agradecimientos

Quiero agradecer a todo el grupo HCPN-UAM, es especial a mi tutor, Javier Arazil, el trabajo y paciencia que han tenido, fundamentales para que este proyecto se haya podido realizar. A mis padres por su incondicional apoyo y cariño. Y finalmente, a todos quienes han hecho posible que estos años hayan sido inolvidables.

Gonzalo Polo Vera

Noviembre 2010.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

1. INTRODUCCIÓN

La recolección de medidas de tráfico de red dentro de Internet supone una información muy valiosa para los investigadores, proveedores de servicio (ISPs[1]) y otros miembros de la comunidad de Internet.

Por un lado, los operadores de red se benefician de esta información para cumplir su objetivo de asegurar una apropiada calidad de servicio (QoS[2]) a sus clientes. El constante crecimiento de la demanda por parte de los usuarios y las exigencias que requieren las nuevas aplicaciones están obligando a los ISPs a desarrollar sus planes de capacidad de red con extrema atención, no sólo para mantener la calidad de servicio que deben proveer, sino también para reducir las necesidades de inversión. Los ISPs no han infravalorado los beneficios que aportan las medidas de tráfico y tradicionalmente han aplicado su potencial para otras funciones relacionadas como puede ser la evaluación del comportamiento de las redes, la detección de anomalías, rechazo de ataques e incluso la facturación de los clientes.

Por otro lado, la comunidad investigadora se ha dado cuenta que es fundamental la utilización de medidas reales de red para profundizar en el conocimiento de la dinámica de Internet y posteriormente aplicarlo en el desarrollo de los modelos de red, con aplicación directa para las necesidades de los operadores de red mencionadas anteriormente.

Sin embargo, la recolección de medidas de tráfico representativas no es proceso sencillo. En este sentido, los autores de [1] proporcionan una

[1] **ISP** Del inglés, Internet Service Provider

[2] **QoS** Del inglés, Quality of Service

explicación detallada de los problemas que se pueden encontrar en las simulaciones de Internet, algunos de las cuales surgen en el proceso de realizar medidas de red. Ejemplos de estos problemas incluyen el enorme tamaño y naturaleza heterogénea de Internet, el constante incremento del número de aplicaciones nuevas que se introducen en la red, la forma rápida e impredecible con la cual Internet cambia y el tamaño de las muestras recolectada en el proceso de medida. Esto provoca que existan una amplísima variedad de estudios que se realizan sobre medidas de tráfico de red cuyos análisis son muy diferentes.

1.1 MOTIVACIÓN Y OBJETIVOS

La comprensión de la naturaleza del tráfico de red es crítica para diseñar e implementar correctamente redes de computadoras y servicios como la Web[3]. Estudios sobre tráfico de redes WAN[4] [2] han demostrado que los modelos comúnmente aceptados para el tráfico de red, es decir, el proceso de Poisson, no son del todo apropiados. Él tráfico donde el proceso de llegadas fuera de tipo Poisson o Markov, presenta una longitud de ráfaga característica, la cual tiende a disminuir al promediar en una escala temporal cada vez mayor. En lugar de esto, las medidas de tráfico real presentan tráfico a ráfagas en un amplio rango de escalas temporales.

El tráfico que presenta ráfagas en muchas o todas las escalas temporales puede ser descrito estadísticamente usando el concepto de autosimilaridad (SS[5]). La autosimilaridad es una propiedad asociada a los fractales (objeto semigeométrico cuya estructura básica, fragmentada o irregular, se repite a diferentes escalas). En el caso de fenómenos

[3] **WWW :** Del inglés, World Wide Web

[4] **WAN:** Del ingles, Wide Area Network

[5] **SS:** Del inglés, Self-similarity

estocásticos como las series temporales, la autosimilaridad produce que ciertas medidas estadísticas como la correlación, permanezcan invariables ante el cambio de escala. Como resultado, estas series temporales presentan ráfagas (bursts), largos periodos por encima de la media, en un rango amplio de escalas temporales.

Debido a que un proceso autosimilar posee ráfagas significativas en un rango amplio de escalas temporales, puede presentar lo que se llama dependencia a larga escala (LRD[6]): los valores en cualquier instante tienen, generalmente, una correlación positiva no despreciable con todos los valores futuros. La importancia del LRD en el tráfico de redes se ha observado en varios numerosos estudios como [3]. Éstos muestran que el comportamiento respecto a la pérdida de paquetes y el retardo es completamente distinto cuando en las simulaciones se utilizan datos de tráfico real o sintético con LRD, respecto al tradicional (Poisson sources) que carece de él.

A pesar de que estos fenómenos (SS y LRD) no son exactamente lo mismo, dentro del contexto de tráfico de red, están íntimamente ligados. Se ha comprobado que ambos están presentes, es decir, el tráfico se puede modelar con un proceso autosimilar que presenta dependencia a larga escala.

Matemáticamente ambos fenómenos se pueden expresar de forma sencilla desde el punto de vista de las series temporales. Se define la serie temporal $X = (X_t;\ t = 1,2,3 \dots)$ como el número de flujos activos en un punto en el instante t. A partir de ella se crean las series m-agregadas $X^{(m)} = (X_k^{(m)}; k = 1,2,3 \dots)$ sumando la serie original en bloques no solapados de longitud m. Esto se puede entender como un cambio de escala temporal. Entonces se dice que X es H-autosimilar si para todo m positivo, $X^{(m)}$ tiene

[6] **LRD:** Del inglés, Long-Range Dependece.

la misma distribución que X escalado por m^H. H se conoce como parámetro de Hust y es el que mide el grado de autosimilaridad. Esto es:

$$X_t \triangleq m^{-H} \sum_{i=(t-1)m+1}^{tm} X_i \;\; \forall \; m \in \mathbb{N}$$

El símbolo $\triangleq$ significa iguales en distribución. Esta es la expresión matemática que refleja la observación de comportamientos similares a distintas escalas temporales. El concepto de LRD, se expresa mediante la función de autocorrelación $r(k) = E[X_t X_{t+k}]$ de la siguiente forma:

$$r(k) \sim k^{-\beta}\,;\; k \to \infty \;,\;\; 0 < \beta < 1$$

Las distribuciones de cola pesada han sido sugeridas para explicar ambos fenómenos. Los autores en [4] demuestran que si el tráfico es construido como la suma de muchos procesos ON/OFF, en los cuáles los periodos ON (u OFF) son independientes entre sí y siguen una distribución de cola pesada, entonces las series temporales que lo representan serán asintóticamente autosimilares. Si la distribución de los tiempos ON u OFF son de cola pesada con parámetro α (serán explicadas en detalle más adelante), entonces las series resultantes serán autosimilares con parámetro $\mathrm{H} = (3 - \alpha)/2$ y presentarán LRD siempre que $H > \frac{1}{2}$. Si ambos, ON y OFF son de cola pesada el parámetro H resultante vendrá determinado por la distribución "más" pesada (un α menor). Por ejemplo, en el contexto de la Web, se puede considerar a los procesos ON/OFF como sesiones de navegación. Cada sesión puede estar en silencio o recibiendo datos transmitidos a una tasa regular. Es una simplificación del entorno real de la Web, pero indica que si la duración de las transmisiones es de cola pesada, entonces es probable que el tráfico resultante sea autosimilar.

En lo referido a LRD, en un sencillo modelo del tráfico, como es un sistema M/G/∞, que representa flujos activos en un punto concreto de una red, si la distribución del tiempo de duración de los flujos es de cola pesada con parámetro α, la función de autocorrelación de X_t presentará un decaimiento asintótico polinómico de exponente $\beta = (\alpha - 1)$, lo que se traduce en que presenta LRD si $1 < \alpha < 2$.

La figura 1-1, muestra una visión conceptual de este comportamiento. Cada línea horizontal representa un flujo que atraviesa un enlace (el tiempo comienza con el primer paquete y finaliza con el último). La posición vertical es aleatoria para separarlos visualmente. La agregación vertical representa el tráfico total que atraviesa el enlace en ese momento. La duración de los flujos se podría decir que sigue una distribución de cola pesada, hay unos pocos flujos muy largos (denominados “elefantes”), y muchos muy cortos (denominados “ratones”). Si las duraciones estuvieran distribuidas exponencialmente con el mismo valor medio, habría muchos más de tamaño medio (figura 1-2). Estos elefantes son los que provocan que el tráfico posea la propiedad LRD. Concretamente, puntos que están ampliamente separados en el tiempo, podrán tener en común alguno de ellos, lo que supone que el tráfico total en esos puntos esté correlado.

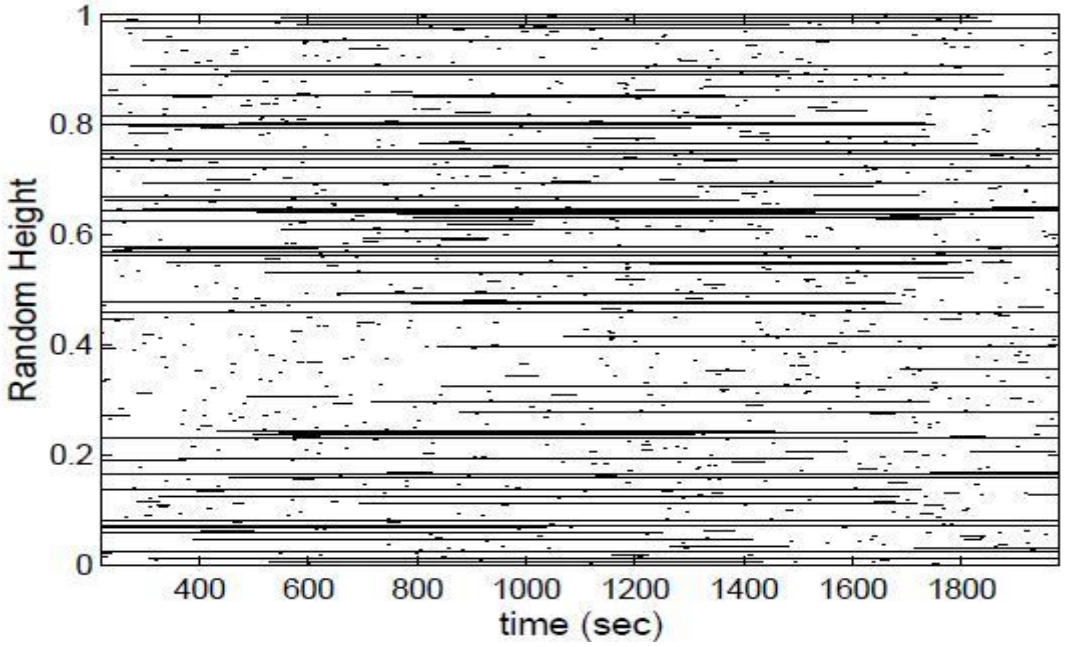

Figura 1-1: Longitud de flujos de cola pesada

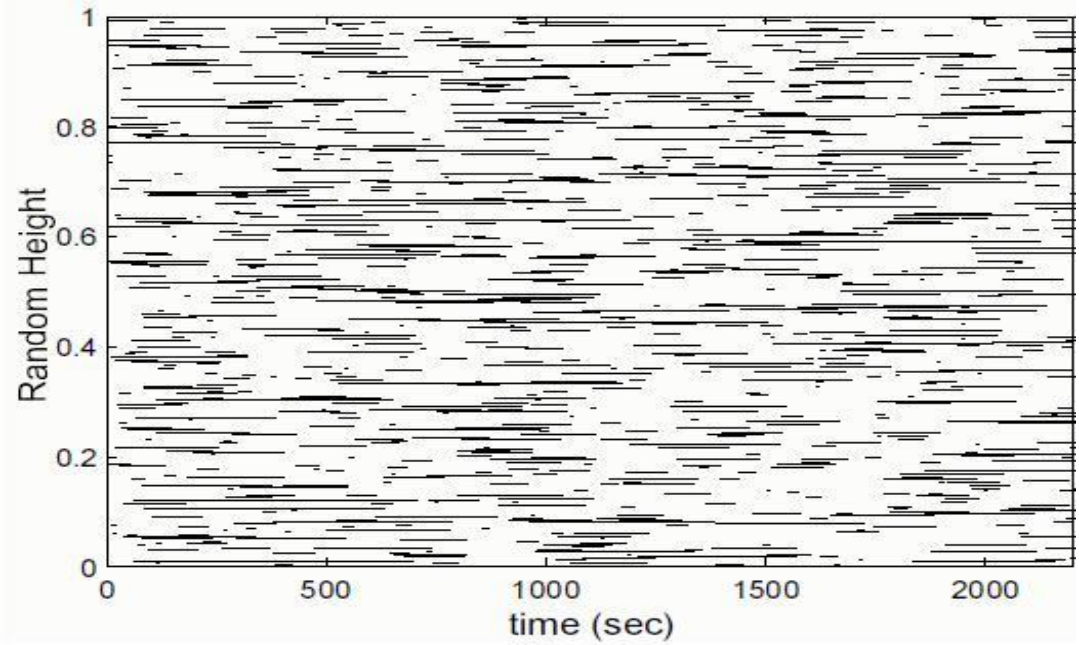

Figura 1-2: Longitud de flujos exponenciales

Para comprender lo importante de esto, hay que ver que cuando el tráfico es fuertemente autosimilar/LRD (si H=1, el proceso se denomina perfectamente autosimilar), las ráfagas pueden ocurrir en un rango muy

amplio de escalas temporales. Cuando se producen ráfagas muy largas, muchos paquetes requerirán ser almacenados temporalmente en cola (buffering). Esto puede producir dos efectos negativos:

-Primero, los paquetes almacenados en largas colas tendrán que esperar largos periodos de tiempo antes de que puedan ser transmitidos. Este es el problema del retardo de transmisión[7].

-Segundo, debido a que las colas tienen un espacio finito, una ráfaga muy larga podría exceder su capacidad. En este caso, las redes descartan estos paquetes y genera el problema de disminución de rendimiento[8]. Esto es debido a que el ancho de banda debe utilizarse para la retransmisión de estos paquetes[9].

En la práctica, las colas tienen la capacidad suficiente para evitar el problema de la pérdida de paquetes y así mantener un buen rendimiento. Sin embargo, el retardo de los paquetes sigue produciéndose, lo que desde el punto de vista del usuario se percibe como un mal funcionamiento (video que no se ve fluido etc.) Entonces, debido a las grandes ráfagas en el tráfico de red, los usuarios experimentan largos retardos en las transmisiones, y la red parece que no responde a su demanda.

Una muestra de la gravedad de este efecto es observada por los autores de [5] tal y como se puede ver en la figura 1-3. En esta figura se puede observar la relación que existe entre el retardo de paquetes y la tasa de transmisión de una red simulada en la cual se envían archivos cuya distribución es de cola pesada. Las cuatro curvas corresponden a diferentes

[7] En inglés denominado packet delay, ya que se refiere al retardo que sufre un paquete desde su punto de origen al punto de destino

[8] En inglés denominado decreased throughput

[9] En este caso, se denomina rendimiento a la tasa neta de transmisión.

valores del parámetro α de la distribución de los archivos; el eje *x* es la medida del rendimiento de la red como un porcentaje de la máxima tasa de transmisión posible, el eje *y* mide número medio de paquetes en cola, el cual es proporcional al retardo medio por paquete.

Se observa que cuando α tiene un valor cercano a 2, es posible conseguir un alto rendimiento sin un retardo en los paquetes muy significativo. Sin embargo cuando α está próximo a 1, es prácticamente imposible conseguir un nivel de rendimiento comparable debido al incremento que es produce en el retardo. Entonces vemos que a medida que la distribución del tamaño de los archivos se hace más "pesada" (α menor), es más complicado obtener un alto rendimiento en la red sin sufrir un retardo en los paquetes importante.

Viendo esto, se ve la importancia de comprender la naturaleza del tráfico para así poder provisionar a las redes basándose en suposiciones precisas del tráfico que transportan.

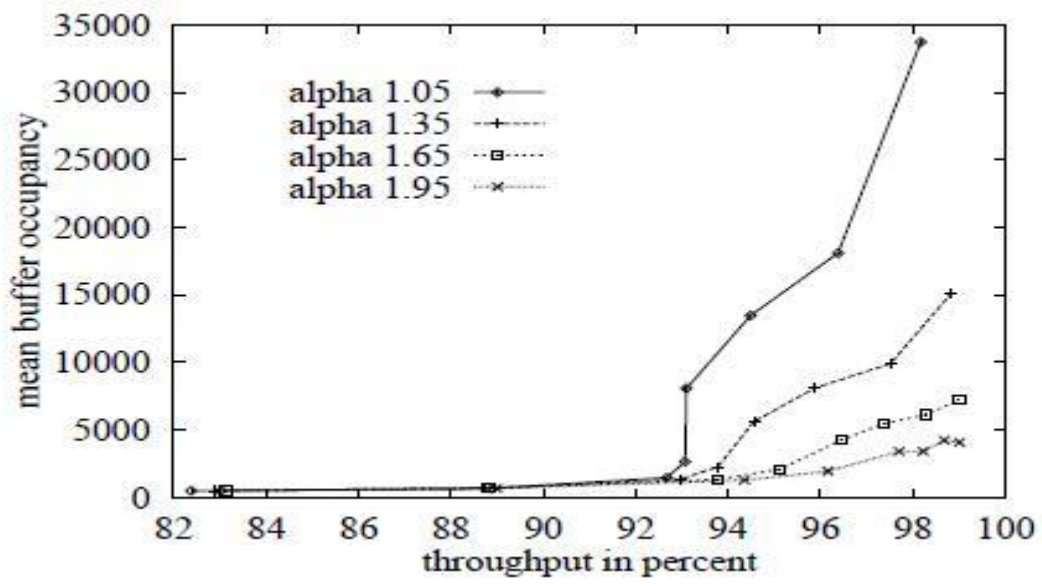

Figura 1-3: Numero medio de paquetes en cola en función del índice de cola α

Este proyecto tiene como objetivo arrojar un poco de luz en lo referido a este tema. Para su realización, se ha tenido acceso a las medidas de tráfico realizadas en RedIRIS [6] (figura 1-4), la red nacional española de investigación y educación que está compuesta por más de 300 instituciones. Se han recopilado las trazas producidas durante un período de tiempo de un mes, concretamente Junio de 2007. La medida a analizar es la longitud, en bytes, de los flujos que atraviesan un nodo concreto agrupándolos según el puerto TCP/UDP. Un flujo se define como el conjunto de paquetes que comparten la quíntupla de dirección IP origen y destino, puerto origen y destino y número de aplicación/protocolo de forma continuada en el tiempo. Los datos de los que se dispone provienen de los nodos que realizan un muestreo del tráfico total y los exportan al repositorio de datos (figura 1-4). En [7], sus autores explican en detalle la validación y obtención de las medidas.

Se ha trabajado principalmente con los provenientes del puerto 80, que es el puerto que transporta el tráfico Web. Este puerto supone una gran parte del total del tráfico dentro del conjunto de los puertos llamados bien-conocidos (de 1 a 1023). Se han analizado las distribuciones de la longitud de estos flujos. Concretamente, se utilizan la distribución Generalizada de Pareto (GP) y la distribución Log-Normal (LN). En capítulos posteriores se justificará su utilización, ya que ambas son muy diferentes entre sí, considerándose una de cola pesada (GP si $\alpha < 2$ posee varianza infinita) y otra de cola “ligera” (LN tiene todos los momentos finitos).

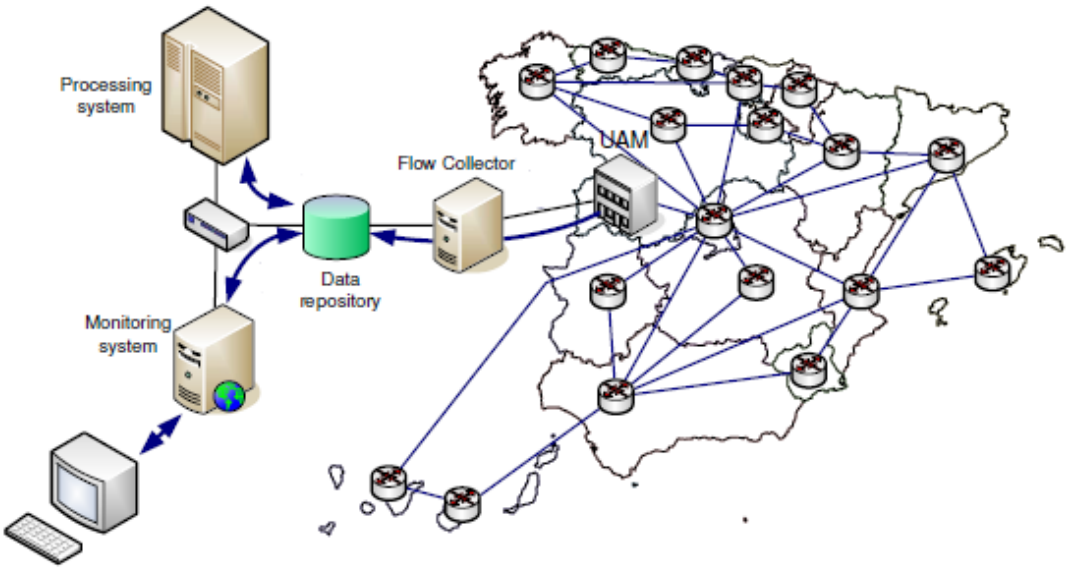

Figura 1-4: Arquitectura del sistema de medida y topología de RedIRIS

1.2 ORGANIZACIÓN DE LA MEMORIA

La memoria consta de los siguientes capítulos:

CAPÍTULO 1 INTRODUCCIÓN

- Objetivos y motivación del proyecto.
- Organización de la memoria.

CAPÍTULO 2 ESTADO DEL ARTE

- Visión general de estudios similares que analizan trazas de tráfico reales, centrándose especialmente en los referidos al análisis de distribuciones de cola pesada y log-Normal.

CAPÍTULO 3 ASPECTOS TEÓRICOS DEL ANÁLISIS

- Descripción matemática de los aspectos analíticos utilizados para la caracterización del tráfico mediante la distribución de Pareto.
- Descripción matemática de los aspectos analíticos utilizados para la caracterización del tráfico mediante la distribución de log-Normal.

CAPÍTULO 4 ASPECTOS PRÁCTICOS DEL ANÁLISIS

- Descripción de los métodos y procedimientos realizados para llevar a cabo la caracterización del tráfico propuesta en el capítulo 3.
- Descripción de las representaciones gráficas que se muestran en el capítulo 5.

CAPÍTULO 5 RESULTADOS

- Presentación y explicación de los resultados obtenidos tras el análisis del tráfico mediante las técnicas descritas en el capítulo 4.

CAPÍTULO 6 CONCLUSIONES Y TRABAJO FUTURO

- Conclusiones obtenidas tras el análisis completo del tráfico.
- Relación de posibles líneas futuras de mejora para la caracterización del tráfico.

2. ESTADO DEL ARTE

Este capítulo presenta y explica varios estudios sobre tráfico de redes que me han aportado numerosas ideas y me han ayudado a situarme en el contexto del análisis de medidas de red, siendo útil para justificar y comprender los métodos que se han llevado a cabo para realizar este proyecto.

Varios son los métodos y conclusiones extraídas de los análisis. Algunas de ellas parecen totalmente contradictorias pero como se verá a lo largo del proyecto no son tan lejanas. Se han separado en dos grupos, los que se centran principalmente en el estudio mediante distribuciones de cola pesada, en el sentido de que poseen un decaimiento hiperbólico y los que lo hacen mediante la log-Normal.

2.1 ANÁLISIS BASADOS EN DISTRIBUCIONES DE COLA PESADA

Los autores de [8] llevan a cabo un estudio de la presencia de distribuciones de cola pesada en la Web. Señalan que la Web genera, dentro de Internet, más tráfico que ninguna otra aplicación y que, por esta razón, sus características son fundamentales para la ingeniería y planificación de red, y evaluación del comportamiento de la misma.

Las mediciones de la actividad Web que analizan, se realizan tanto en el lado de los clientes como en los servidores. En la parte de los primeros, recopilan toda la actividad Web llevada a cabo en una LAN durante el período aproximado de un mes. Esta información consta básicamente del nombre, tiempo de transmisión y tamaño (número de Bytes incluyendo la sobrecarga producida por el protocolo HTTP, flujos HTTP) de todos los archivos solicitados vía Web por los usuarios de la red. Por parte de los servidores, recopilan el tamaño de los archivos disponibles en 32 de ellos durante el mismo periodo de tiempo.

A partir de la recopilación de los datos tratan de estimar el valor del índice de las colas, α, del conjunto de distribuciones empíricas muestreadas realizando dos métodos diferentes.

El primero que utilizan es la representación de la función de distribución acumulada complementaria CCDF[10] con ambos ejes en escala logarítmica[11]. Una de las propiedades que poseen las distribuciones de cola pesada, es que su CCDF $\bar{\mathrm{F}}(\mathrm{x}) = 1 - \mathrm{F}(\mathrm{x}) = \mathrm{P}[\mathrm{X} > x]$ cumple que:

$$\frac{\mathrm{d}\log \bar{F}(x)}{d\log x} \sim -\alpha$$

para valores de x suficientemente grandes (esto es lo mismo que decir que $\bar{\mathrm{F}}(\mathrm{x})$ sigue una ley potencial[12]). Para calcular el valor del índice, seleccionan un valor mínimo de x a partir del cual la gráfica parece ser lineal. Después, de este conjunto de puntos que son mayores, seleccionan un conjunto de ellos que sean equidistantes logarítmicamente, y estiman sobre ellos la pendiente mediante regresión por mínimos cuadrados. Esto lo hacen así, ya que la densidad de puntos para valores pequeños es muy superior a la existente en la cola y, de otra manera, predominaría mucho en la regresión.

El segundo método utilizado es el llamado estimador de Hill. Éste aproxima el valor de α, como una función de los k elementos mayores de las muestras de sus medidas empíricas. Representan el estimador respecto a los valores de k (desde ser sólo el máximo hasta ser la muestra completa) y, si se estabiliza a un valor aproximadamente constante, representará un valor de α (indicando también que la distribución es de cola pesada).

Además de calcular el índice con ambos métodos, realizan un test que denominan Distribución Limite (LD test[13]). Este test que sirve para

[10] **CCDF:** Del ingles, Complementary Cumulative Distribution Function.

[11] Estas representaciones se denominan log-log CCDF plots

[12] Más conocido con su denominación inglesa, power-law

[13] **LD test:** Del inglés, Limit Distribution test.

comprobar si las muestras realmente provienen de una distribución subyacente que posee varianza infinita. Es el caso de las de cola pesada. Para ello, a partir de la muestra original (X_i), generan otra agregándola en bloques de longitud m, tal que:

$$X_t^{(m)} = \sum_{i=(t-1)m+1}^{tm} X_i$$

Este proceso lo repiten para distintos valores de m, cada vez mayores ($m = 10,100,500$ en su caso). Se basa en la teoría de distribuciones estables. Esta teoría demuestra que, si la muestra original pertenece al dominio de atracción de una distribución estable con $\alpha < 2$ (varianza infinita), entonces las colas de las muestras $m-$agregadas tenderán a seguir una ley potencial con el mismo α. En caso contrario, si la distribución subyacente posee una varianza finita, las muestras agregadas tenderán a la Normal y sus colas decrecerán de forma exponencial.

A partir de esto, representan las log-log CCDF de las muestras m-agregadas, para comprobar si a medida que m se incrementa, la pendiente aumenta, reflejando que la distribución resultante se aproxima a una Normal o si se mantiene constante. En la figura 2-1 se puede observar la diferencia existente entre la distribución de Pareto con varianza infinita (izquierda), y la log-normal con varianza finita (derecha). En la primera las líneas son aproximadamente paralelas mientras que en la segunda parecen convergentes.

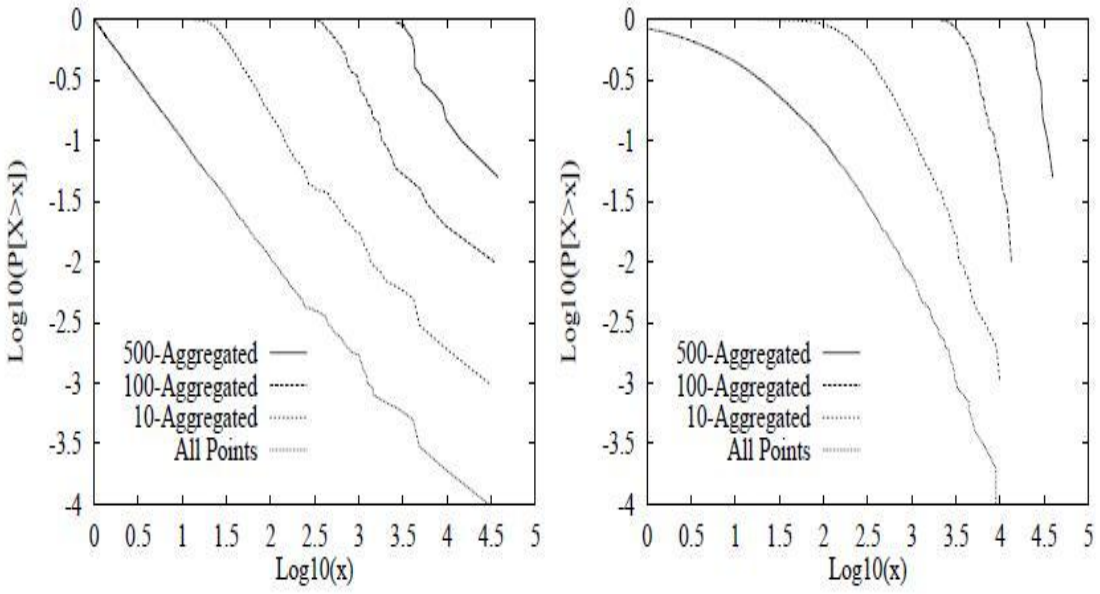

Figura 2-1 CCDF de muestras m-agregadas

Primero muestran el análisis sobre la distribución que siguen los tiempos de transmisión de los archivos solicitados en la parte de los clientes. Esto tiene bastante importancia ya que, la conclusión de que sean de cola pesada, es una posible explicación a la autosimilaridad observada en el tráfico de red, como ya he comentado en el apartado introductorio. Señalan que la muestra presenta las características atribuibles a las distribuciones de cola pesada. En el log-log CCDF plot toman como punto inicial $x \approx 0.3$ segundos, tal que, $\log_{10} x = -0.5$. A partir de la regresión por mínimos cuadrados estiman una pendiente de -1.21, lo que implica $\hat{\alpha} = 1.21$. Ellos mismos comentan que la función parece presentar cierta curvatura. Por ello realizan el LD test, donde concluyen que la distribución pertenece al dominio de atracción de de distribuciones estables con varianza infinita, ya que como se ha explicado arriba, las pendientes de las colas no cambian de forma significativa (figura2-2 y figura 2-3).

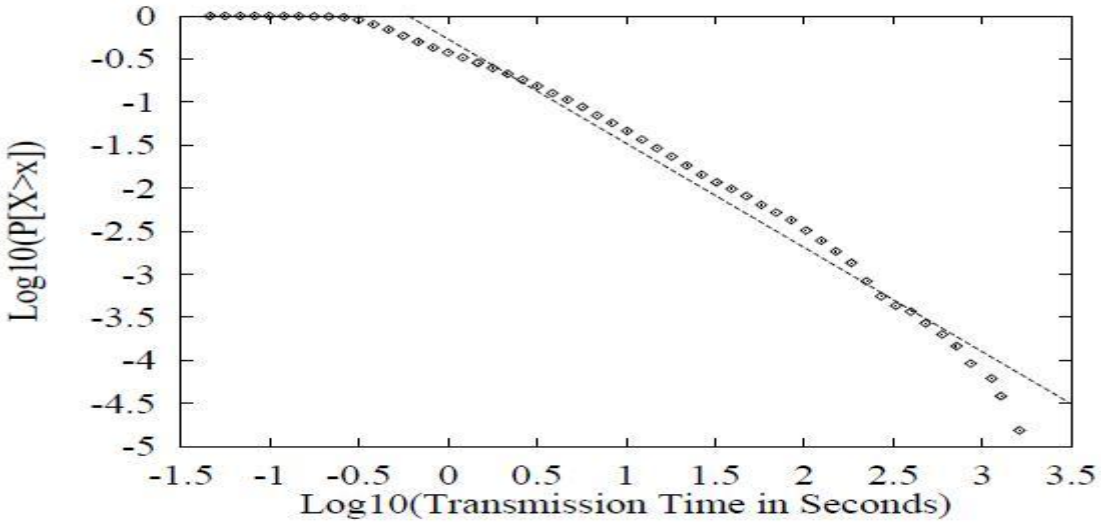

Figura 2-2: log-log CCDF de la duración de los flujos

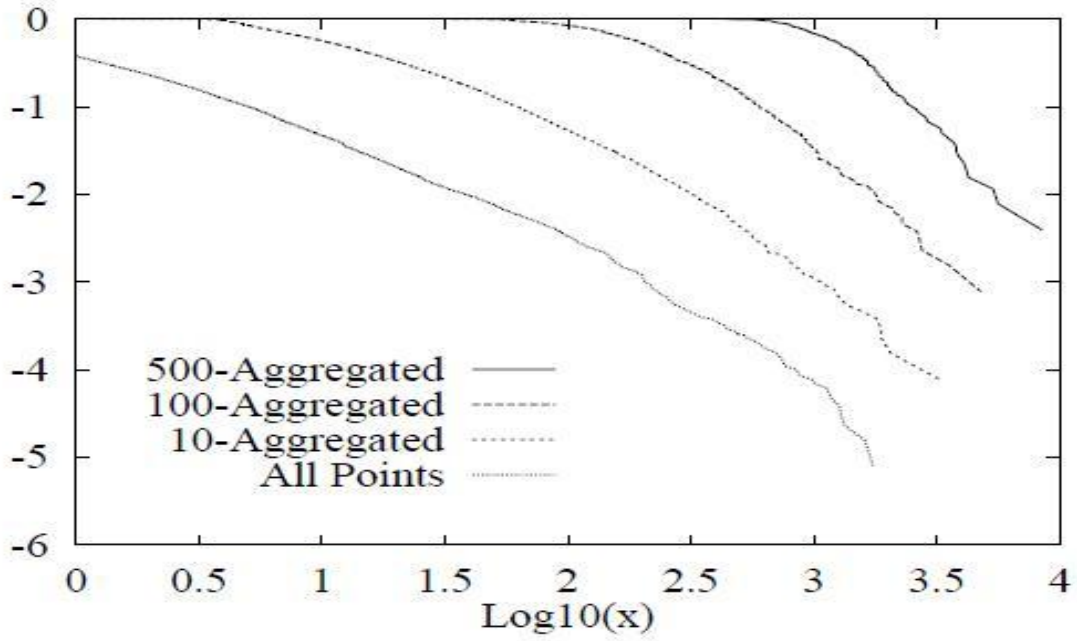

Figura 2-3: log-log CCDF de muestras m-agregadas de la duración de los flujos

Finalmente, muestran la representación del estimador de Hill. Consideran que, el estimador, permanece relativamente constante a partir de alcanzar la mediana (percentil 0.5). El valor estimado es de $\hat{\alpha} = 0.83$.

Con estos resultados concluyen que la distribución de los tiempos de transmisión posee una varianza infinita con un índice de cola α en un rango aproximado de 0.8 a 1.2.

A continuación, para comprender porque los tiempos de transmisión son de cola pesada, analizan las distribuciones de los tamaños de los archivos Web. Realizan un análisis y razonamiento muy interesante, que no tiene especial relevancia aquí, con el cual explican que, gracias al "caching" cada vez más efectivo, el conjunto de archivos transmitidos por la red (que implican fallo de caché), tiende a ser el de los disponibles en los servidores cuando el tiempo tiende a infinito, independientemente del comportamiento de los usuarios. El análisis que realizan para el tamaño de éstos es idéntico al comentado anteriormente, concluyendo que son de cola pesada con un índice de cola α en un rango aproximado de 0.6 a 1.1. Con esto, señalan que parece probable que la naturaleza de cola pesada de los tiempos de transmisión, esté directamente relacionada con que los archivos disponibles posean la misma al igual que el tamaño de los flujos transmitidos.

Terminan relacionando estos resultados con sus implicaciones en el tráfico de red. Señalan que esto puede ser la causa de que el tráfico Web sea autosimilar y, como representa la gran mayoría del tráfico presente en Internet, es un tema de estudio que hay que tener muy en cuenta para poder comprender sus efectos negativos que ya han sido explicados en la introducción.

En [9], sus autores tratan arrojar un poco de luz acerca del comportamiento de las distribuciones de cola pesada. Plantean una solución al problema de obtener, de una muestra, toda la información posible discriminando entre que aspectos se deben a las propiedades de la distribución subyacente y cuáles a la "variabilidad del muestreo".

Resaltan que el efecto de tener muestras mayores, es obtener una mayor comprensión en las regiones “ricas” en información pero que siempre habrá una región de incertidumbre con información escasa en las colas. Uno de los resultados principales de su análisis es que, a lo largo de la cola, las distribuciones empíricas que analizan poseen varias oscilaciones irregulares bastante pronunciadas que llaman “wobbles” que no están presentes en las clásicas como puede ser la de Pareto o la log-Normal (aunque no sea de cola pesada la tienen en cuenta para su estudio). Explican que estos wobbles no pueden ser atribuidos a la variabilidad debida al muestreo ya que después de analizar varios conjuntos de muestras, éstas no sólo presentan la misma cantidad de wobbles sino que además se encuentran en los mismos lugares y con la misma forma.

Así, concluyen que estas oscilaciones son fenómenos producidos por las características de la distribución subyacente y no son artificios debidos al muestreo.

Para poder llegar a esa conclusión, profundizan en la compresión del comportamiento de las colas. Las dividen en tres regiones, según la cantidad de información/variabilidad disponible en las muestras bajo estudio. Está lo que denominan como el extremo de la cola (“extreme tail”). Esta zona se sitúa más allá del máximo valor de la muestra, es decir, no hay ningún tipo de información en la muestra. Luego viene la zona lejana de la cola (“far tail”). Aquí existe cierta cantidad de información pero no la suficiente para poder comprender de forma fiable las propiedades de la distribución a partir de la muestra. Por último, la región moderada de la cola (“moderate tail”). En ésta, la información disponible en los datos es “rica”, y es la que menos se ve afectada por la variabilidad del muestreo.

Estos conceptos son completamente heurísticos, pero son fundamentales en el análisis que realizan y están basados en el concepto de lo que denominan envoltura muestral.

Estudian la duración de flujos HTTP al igual que en [8]. La muestra, recopilada a lo largo de una semana, la dividen en 21 bloques de cuatro horas, tres periodos distintos a lo largo de cada día para poder analizar también las diferencias que existen según la situación horaria. La duración de los flujos la analizan de forma individual para cada bloque.

Primero utilizan un método estadístico gráfico denominado QQ-plot[14]. Así les permite realizar una comparación gráfica entre los cuantiles de la distribución teórica de Pareto con los de la muestra.

Para seleccionar los dos parámetros de la distribución, hacen que coincidan los cuantiles 0.8 y 0.99 con los de la muestra (figura 2-4). Ahora para comprobar si los wobbles, antes mencionados, son producidos por la variabilidad del muestreo, superponen cien simulaciones (generaciones de números aleatorios) con el mismo número de muestras de la misma distribución de Pareto. Esto gráficamente aparece como "nube" de puntos alrededor de la línea recta de los cuantiles teóricos, la antes mencionada envoltura muestral. Señalan que, si la muestra fuera realmente de Pareto, entonces la mayoría de los wobbles caerían dentro de la envoltura. Esto ocurre para la mayoría de los (wobbles) situados en la región lejana de la cola, pero no para el resto.

Con esto concluyen que no son producidos por la variabilidad del muestreo, sino que es una propiedad intrínseca de la distribución real subyacente que no es exactamente de Pareto.

[14] **QQ-plot:** Son representaciones de los cuantiles de una variable aleatoria teórica o empírica frente a los de otra.

Con la representación de la envoltura, explican lo que consideran las regiones de la cola. Apuntan que el cuerpo y la zona moderada de la cola llegan aproximadamente hasta el cuantil 0.9999. En esta región, la nube es inapreciable, cae completamente debajo de la línea teórica y alcanza flujos de un tamaño de hasta 1.2 megabytes. Comentan que, dentro del tráfico HTTP, los tamaños superiores son los denominados "elefantes" (comentados en la introducción) que aparecen en las distribuciones de cola pesada. La zona lejana de la cola llegaría hasta el cuantil 0.99999, ya que a pesar de que la cantidad de datos es escasa, la variabilidad del muestreo aún es aceptable. El extremo de la cola en este caso sería para valores superiores a 980 megabytes, que es el valor superior de la muestra que analizan. Llevan a cabo el mismo análisis pero utilizando como distribución teórica la log-normal y los resultados obtenidos son ligeramente peores.

Como segundo método, también utilizan la representación el log-log CCDF plot. La razón argumentada, es que, a pesar de que el método QQ plot permite una comparación precisa con cualquier distribución, unido a la visualización de la variabilidad del muestreo (con el concepto de la envoltura de puntos), tiene el problema de que debe ser realizado sobre una distribución teórica particular (requiere el cálculo previo de los parámetros pertinentes), mientras que éste no. Comprueban si los CCDF plot son aproximadamente lineales, y vuelven a observar que aparecen wobbles a lo largo de las curvas. En este caso, no pueden visualizar la variabilidad del muestreo para diferenciar las distintas regiones, debido a que no lo comparan con ninguna distribución teórica, pero en cambio, presenta la ventaja de poder superponer todos los CCDF plot obtenidos de todos los conjuntos de muestras y así, vuelven a confirmar lo que concluyeron con el método anterior: los wobbles en las colas son sistemáticos y no provocados por la variabilidad del muestreo debido a la asombrosa similaridad entre los aparecidos de todas las muestras.

Por estas razones, concluyen que es necesario intentar comprenderlos y modelarlos. Para ello buscan distribuciones más complejas. Utilizan la llamada doble Pareto log-normal. Esta distribución es el producto de una doble Pareto (consta de dos pendientes, es decir, su densidad es proporcional a $x^{-\alpha-1}$ para $x > 1$ y a $x^{-\beta-1}$ para $x < 1$) y una log-normal independientes. Se puede ver como una ampliación de la idea de Downey (referencias [11], [12] y [13] que comentaré más adelante en el capítulo) pero en la cual el número de factores es aleatorio con distribución exponencial. Vuelven aplicar el concepto de la envoltura, ahora en la representación log-log CCDF. Llegan a la conclusión de que, a pesar de que esta distribución mejora los resultados, está lejos de proporcionar un buen ajuste. Los problemas vuelven aparecer por los wobbles que están en la zona moderada de la cola.

Para conseguir la flexibilidad necesaria para estas oscilaciones, recurren al método de las mixturas. Con la mixtura de 3 doble Pareto log-normal consiguen un ajuste casi perfecto, teniendo únicamente un "pequeño" escape de la envoltura en la zona lejana de la cola, donde la variabilidad el muestreo es bastante alta (figura 2-5). El ajuste de los parámetros lo realizan mediante prueba y error comprobándolo visualmente, lo cual es una ardua y trabajosa tarea. Resaltan que, a pesar de que el ajuste es muy bueno, no se puede concluir que la muestra provenga de este modelo. Esto se debe a que, en lo referido a mixturas, si se incluyen un número suficiente de componentes cualquier distribución puede ser ajustada satisfactoriamente pero hay que evitar caer en la inclusión de demasiadas componentes, denominado "sobreajuste[15]". En la figura 2-6, se puede ver el ajuste que obtienen con la mixtura de 3

[15] Más comúnmente conocido por su denominación inglesa, overfitting. Se produce cuando un modelo estadístico es demasiado complejo y modela los errores aleatorios, ruido, en lugar de centrarse únicamente en la relación subyacente.

log-normal calculando los parámetros de nuevo mediante prueba y error. Es bastante bueno y podría ser perfecto si añadieran un nuevo componente que captara el comportamiento al final de la cola, pero sería "overfitting".

Concluyen diciendo que a pesar de que estas dos distribuciones son muy diferentes en el sentido clásico del comportamiento asintótico de la cola, son muy parecidas en la región moderada, con lo que no pueden ser diferenciadas utilizando únicamente muestras. Por ello es necesario no descartar ninguno de los modelos para posteriores análisis y simulaciones admitiendo que se obtienen mejores resultados con la mixtura de 3 doble Pareto log-normal.

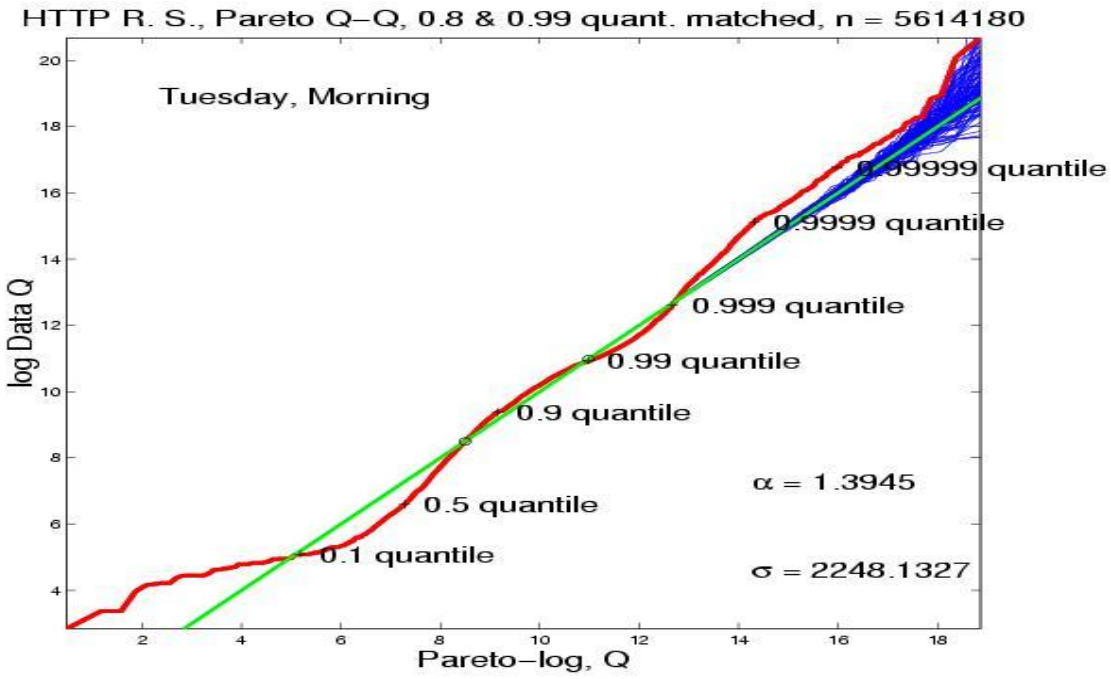

Figura 2-4: QQ-plot Duracion de los flujos vs Pareto

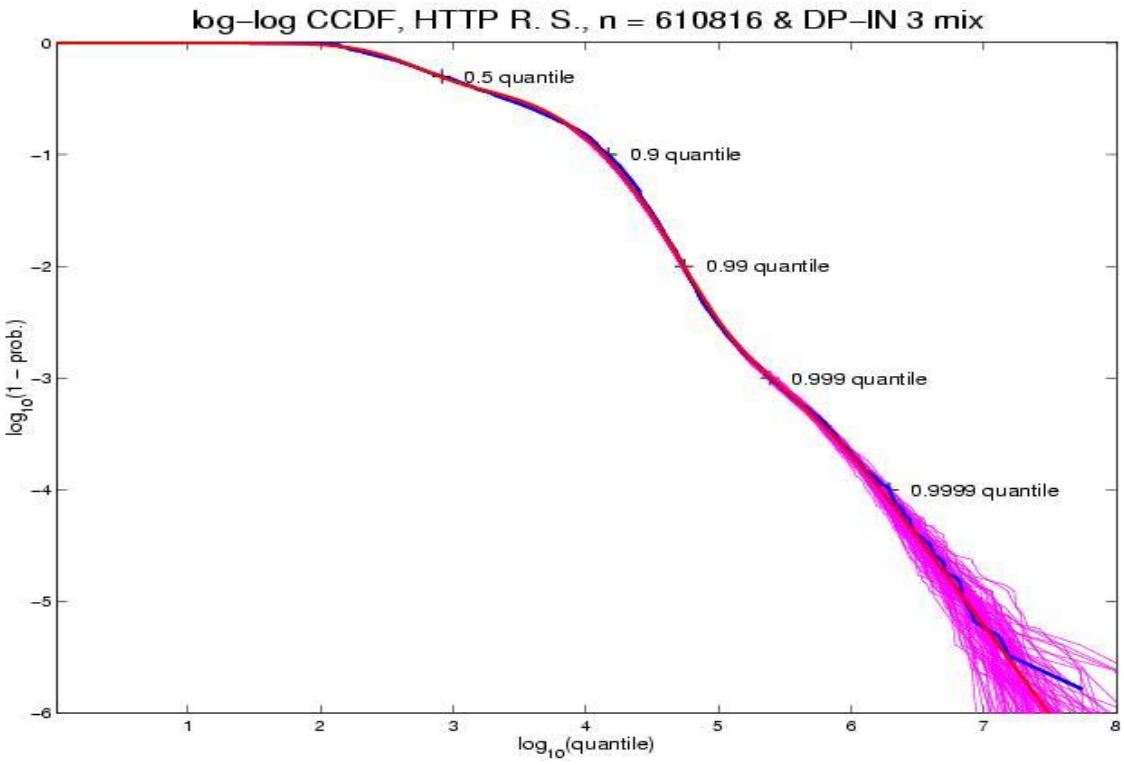

Figura 2-5: log-log CCDF Duración de flujos con envoltura muestral de mixtura de 3 doble-Pareto-LN

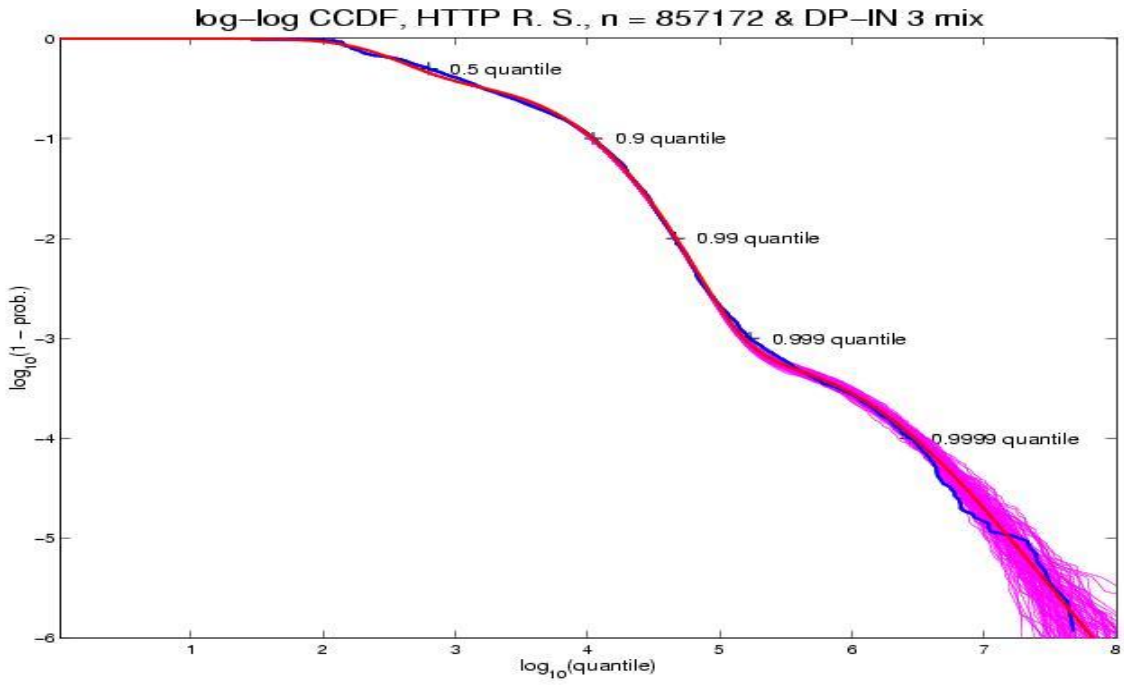

Figura 2-6: log-log CCDF Duración de flujos con envoltura muestral de mixtura de 3 log-Normal

Estos mismos autores en [10] profundizan de forma teórica en lo referido a los wobbles. Generalizan la teoría "clásica" de las duraciones que siguen distribución de cola pesada que producen LRD. Esta teoría

asume que el índice de la cola es constante, como posee la ya mencionada distribución de Pareto, o que varía de una forma regular permitiendo wobbles pero estabilizándose de forma asintótica a medida que avanzas en la cola. Éste último caso parece consistente con los datos observados pero no les resulta del todo satisfactorio. Señalan que la estabilización debe producirse en las regiones lejana y extrema de la cola, donde la cantidad de información disponible es tan escasa que no se puede comprobar que ocurre con seguridad. Por ello, amplían la noción de índice variable, en el sentido de que ya no es necesario que se estabilice a medida que se avanza a lo largo de la cola, pero permanece dentro de unos límites que, dentro de la teoría clásica, producen LRD. Demuestran que, con estas restricciones, este tipo de distribuciones también producen LRD en el sentido de decaimiento polinómico de la función de autocorrelación del tráfico agregado.

En estos estudios, donde la distribución utilizada es la de Pareto de cola pesada, aparece varias veces la log-normal. Ésta es una distribución que, como se comenta a continuación, ha sido utilizada en varios estudios para el análisis de tráfico como prioritaria. Esto se debe a que, a pesar de no ser de cola pesada y tener todos sus momentos finitos, en ciertos rangos de valores pueden ser indistinguibles.

2.2 ANÁLISIS BASADOS EN LA DISTRIBUCIÓN LOG-NORMAL

Los autores de [11] proponen un modelo que, basándose en el comportamiento de los usuarios, explica cómo deben ser las distribuciones de los tamaños de los archivos tanto en los sistemas locales como en la Web. Sostienen que esta distribución es la log-Normal.

Su modelo se basa en que la mayoría de las operaciones de generación de archivos (copia, filtrado, cambio de formato, edición etc.) se puede caracterizar como una transformación lineal de un archivo ya existente, es decir, si el archivo original tiene tamaño s, el archivo final tendrá tamaño γs, donde γ es una variable aleatoria. Para comprobar si su modelo es preciso y realista, realizan simulaciones de sistemas reales cuyos archivos son generados de esta forma. Apuntan que equivalen a la resolución numérica de una ecuación en derivadas parciales en la que se establecen una serie de condiciones iniciales. Concretamente la ecuación del calor, por lo que llaman a este modelo, modelo de difusión. Obtienen resultados bastante satisfactorios y la solución analítica de la ecuación sugiere que la distribución del tamaño de los archivos será aproximadamente log-Normal. Los parámetros que definen la distribución los obtienen a partir de las condiciones iniciales del sistema que quieren simular, las cuales definen basándose en muestras de datos.

Es más sencillo ver que, si un sistema comienza con un archivo inicial de tamaño s_0, el tamaño del archivo n, s_n vendrá determinado por:

$$s_n = s_0\, \gamma_1 \gamma_{2\ldots\ldots} \gamma_m$$

donde m es el número de predecesores. Tomando logaritmos:

$$\log(s_n) = \log(s_0) + \log(\ \gamma_1) + \log(\ \gamma_2\ ) + \cdots + \log\ (\ \gamma_m\ )$$

Por el Teorema Central del Límite, a medida que $m \to \infty$, esta suma converge a la distribución Normal asumiendo que las variables γ_i son independientes entre sí y poseen varianza finita. Esto implica que entonces s_n sigue una distribución log-Normal, como será explicado en al capítulo siguiente.

Debido a los buenos resultados que obtienen con las simulaciones, estos mismos autores en [12] y [13] revisan varias observaciones publicadas por otros pero ahora utilizando la distribución log-Normal como referencia. En todos los casos basan su análisis principalmente en el ajuste visual que consiguen mediante la log-log CCDF. Realizan también lo que llaman test de curvatura, basado en la derivada de la log-log CCDF y comparándola con la de generaciones aleatorias que siguen una distribución de Pareto.

Por ejemplo, analizan de nuevo los datos con los que los autores de [8] (comentado en el apartado anterior) concluían que los archivos disponibles tanto en la parte de los servidos como de los clientes seguían una distribución de cola pesada. Concluyen que su modelo proporciona un mejor ajuste en la mayoría del rango de valores, incluyendo la zona extrema de la cola. También que refleja de forma precisa el comportamiento aparente de la cola que no se mantiene como una línea recta sino que va incrementando su pendiente, lo que significa que estos datos soportan mejor un modelo de distribución log-Normal que un modelo de Pareto.

No sólo estudian distribuciones referidas al tamaño de archivos. También analizan otros parámetros como pueden ser los tiempos de transmisión HTTP. Contradicen a los autores de [14] concluyendo que el modelo de la log-Normal produce un mejor ajuste para la CCDF y que ese conjunto de datos no soporta la hipótesis de que los tiempos de transmisión sean de cola pesada.

Tras concluir que hay varias magnitudes que tienen una pequeña evidencia de ser de cola pesada y que parece más probable que sigan una distribución log-Normal, argumentan que cabe la posibilidad de que el tráfico de red no sea realmente autosimilar. Explican que en el modelo M/G/∞, si la distribución es log-normal, el proceso de conteo resultante no presenta ni autosimilaridad ni LRD, pero que en un rango amplio de escalas

temporales puede ser estadísticamente indistinguible de un proceso realmente autosimilar.

3. ASPECTOS TEÓRICOS DEL ANÁLISIS

En este capítulo se presenta, desde el punto de vista teórico, todos los procesos que se han llevado a cabo para el análisis del tráfico en la realización de este proyecto.

Como se ha expuesto en el apartado anterior, la mayoría de los autores que han realizado análisis de tráfico similares, centran su estudio utilizando las ya mencionadas distribuciones de cola pesada[16] , mediante el uso de la distribución de Pareto[17] o utilizan la distribución log-Normal, que dentro de este contexto se considera de cola ligera al converger todos sus momentos.

En este proyecto, se ha optado por la utilización de ambas. La razón es que a pesar de poseer naturalezas muy diferentes, se ha observado que varios autores han llegado a conclusiones contradictorias analizando las mismas medidas de tráfico. Se debe a que, al trabajar con este tipo de distribuciones, las conclusiones muchas veces se determinan a partir de aspectos subjetivos como pueden ser el decidir si algo sigue un comportamiento "aproximadamente" lineal o no, o cualificar la calidad de los ajustes, ambos mediante inspección visual.

Para no cometer el error de descartar ninguna, se han analizado todos los datos disponibles con ambas distribuciones. Además de obtener

[16] El término cola pesada aquí se refiere a que la distribución posea un comportamiento hiperbólico asintótico, es decir que la cola se comporte de la forma $x^{-\alpha}$, $1 < \alpha < 2$, conjunto en el cual la distribución log-Normal no está incluida, pero que si lo está en otras definiciones del mismo término.

[17] En adelante a la distribución utilizada por los autores de los estudios del apartado 2 se la denominará pura, para diferenciarla de la aplicada en este proyecto, la distribución generalizada que se explicará a lo largo de este capítulo.

conclusiones a partir de inspección visual, se han buscado otras formas de comparación no tan subjetivas para poder concluir cuál de las dos se aproxima más a la realidad que muestran los datos. Se debe a que, tal y como se puede observar en la figura 3-1, pueden presentar formas muy similares en la generación de muestras aleatorias que las poseen como distribuciones subyacentes. Además hay que apuntar que, en cuanto a las implicaciones de la distribución log-Normal sobre la naturaleza del tráfico comentado en el capítulo 1, los autores de [15] demuestran de forma teórica que ésta, bajo ciertas condiciones, también puede producir el fenómeno LRD observado empíricamente en el tráfico de red.

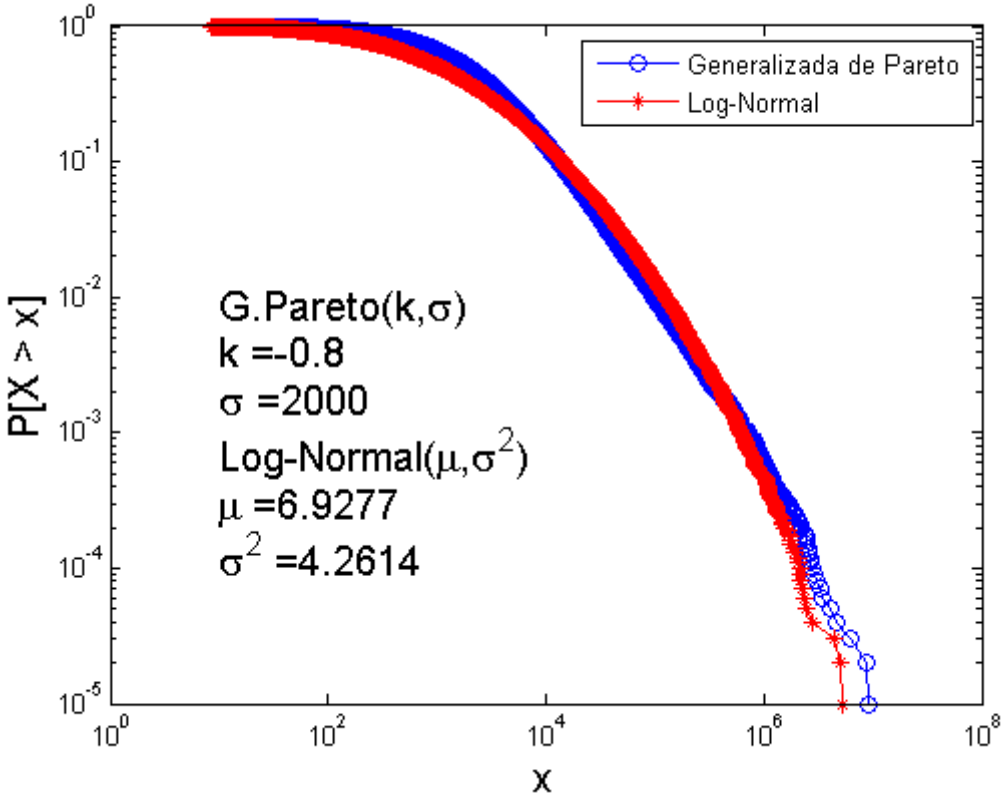

Figura 3-1: log-log CCDF G.Pareto vs Log-Normal

A pesar de que los parámetros, explicados a continuación en este mismo capítulo, han sido determinados con el propósito de mostrar la similitud existente entre ambas, se observa que muestras distintas obtenidas de estas distribuciones pueden llegar a ser prácticamente indistinguibles.

En este sentido, los autores de [16], exponen las similitudes existentes entre ellas. Exponen que, respecto a los modelos generativos, ambas distribuciones conectan de forma muy natural y no es sorprendente que las distribuciones log-Normal aparezcan como alternativa a las "power law" en numerosos campos de investigación. Esto mismo es tratado por los autores de [17], lo denominan como "fragilidad distribucional" y alegan que, normalmente, la información disponible para clasificar una distribución, es insuficiente para clasificar la cola. Como ilustración, se presenta un ejemplo de modelo generativo (de varios que proponen), en el cual, a partir de una pequeña variación se converge a una distribución de Pareto en lugar de a una log-Normal:

-Sea $\{X_i\}_{i=1}^{\infty}$ una secuencia de variables aleatorias positivas i.i.d.[18] con los dos primeros momentos finitos. Entonces sean:

$$Y_n = Y_{n-1}X_n = \prod_{i=1}^{n} X_i$$

$$Y_n = \max(Y_{n-1}X_n\,, \varepsilon)$$

En el primer caso, Y_n converge a la distribución log-Normal a medida que $n \to \infty$. En el segundo caso, se puede demostrar que si $\varepsilon > 0$ entonces Y_n converge a una distribución de Pareto. Es fácil ver que si $\varepsilon = 0$, equivale al primer caso. Así se observa la denominada fragilidad distribucional, si a Y_n se le añade una restricción de no permitir que tome un valor menor que un umbral $\varepsilon > 0$, cambia completamente la distribución a la que converge.

Tras justificar porque se utilizan ambas distribuciones, a lo largo de este capítulo se exponen las propiedades que han sido utilizadas para el análisis de tráfico, comenzando primero con la distribución de Pareto y terminando con la log-Normal.

[18] **i.i.d** : Independientes e idénticamente distribuidas.

3.1 DISTRIBUCIÓN DE PARETO PURA Y GENERALIZADA

3.1.1 DEFINICIÓN Y PROPIEDADES BÁSICAS: DISTRIBUCIÓN DE PARETO

En este apartado se presenta la denominada distribución de Pareto ya mencionada en apartados anteriores. Primero señalar que lo que aquí se denomina como pura, es la distribución empleada en todos los estudios ya comentados en apartados anteriores, llamada así por el economista italiano Vilfredo Pareto. La generalizada, como su propio nombre indica, consiste en una extensión de la pura, siendo más versátil al contar con un parámetro más. Todas las diferencias se verán a continuación.

Se dice que una variable aleatoria X sigue una distribución pura de Pareto, $X \in PPD(\alpha, x_m)$, si:

$$F(x;\alpha,x_m) = P\{X \leq x\} = \begin{cases} 1-(\frac{x}{x_m})^{-\alpha} & \text{para} \quad \alpha > 0, x \geq x_m \\ 0 & \text{para} \quad \alpha > 0, x < x_m \end{cases} \qquad (3.1)$$

Donde $F(x;\alpha,x_m)$ es la función de distribución acumulada, x_m es un parámetro de escaldo[19] que representa el valor mínimo que puede tomar la variable y α es un parámetro de forma[20] también llamado índice de cola.

Dentro del contexto de la autosimilaridad, el rango de valores de interés es $0 < \alpha < 2$ ya que, como se ha comentado anteriormente, para generar procesos autosimilares con tiempos de servicio de cola pesada, es necesario que posean una varianza infinita. Como se verá a continuación,

[19] En inglés, scale parameter.

[20] En inglés, shape parameter.

esto se produce si y sólo si $\alpha < 2$. Realmente, el valor de este parámetro suele ser también mayor que 1, lo que implica que la media sea finita.

A partir de la función de distribución es inmediato ver que:

$$\overline{\mathrm{F}}(\mathrm{x};\alpha,\mathrm{x_m}) = \mathrm{P}\{\mathrm{X} > x\} = \begin{cases} (\frac{\mathrm{x}}{\mathrm{x_m}})^{-\alpha} & \text{para} \quad \alpha > 0, x \geq \mathrm{x_m} \\ 1 & \text{para} \quad \alpha > 0, x < \mathrm{x_m} \end{cases} \qquad (3.2)$$

Con esto se ve que esta distribución es de cola pesada en el sentido de que posee un comportamiento hiperbólico, no sólo asintótico (en la cola) si no en todo el rango de valores. Es la distribución que sigue una ley potencial por excelencia, al ser la más sencilla. Esto se traduce en que:

$$\frac{\partial \log \bar{\mathrm{F}}(\mathrm{x})}{\partial \log \mathrm{x}} = \frac{\partial(-\alpha(\log \mathrm{x} - \log \mathrm{x_m}))}{\partial \log \mathrm{x}} = -\alpha \qquad (3.3)$$

Así en las representaciones CCDF, estas distribuciones aparecerán como una línea recta con pendiente $-\alpha$.

La función de densidad se obtiene a partir de la de distribución acumulada anterior:

$$\mathrm{f}(\mathrm{x};\alpha,\mathrm{x_m}) = \frac{\partial \mathrm{F}(\mathrm{x};\alpha,\mathrm{x_m})}{\partial \mathrm{x}} = \begin{cases} \frac{\alpha}{\mathrm{x_m}} (\frac{\mathrm{x}}{\mathrm{x_m}})^{-(\alpha+1)} & \text{para} \quad \alpha > 0, x \geq \mathrm{x_m} \\ 0 & \text{para} \quad \alpha > 0, x < \mathrm{x_m} \end{cases} \qquad (3.4)$$

De forma similar a la función de distribución complementaria, en una representación con ambos ejes logarítmicos, la función de densidad también aparecerá como una línea recta, con la misma pendiente -1:

$$\frac{\partial \log f(x)}{\partial \log x} = \frac{\partial((-\alpha - 1)\log x + \alpha \log x_m + \log \alpha)}{\partial \log x} = -\alpha - 1 \qquad (3.5)$$

La función inversa de distribución $x(F)$, se obtiene calculando la inversa de la acumulada y es fácil ver que:

$$x(F) = \frac{x_m}{(1-F)^{1/\alpha}}, 0 \leq F \leq 1 \qquad (3.6)$$

Los momentos m_n de la distribución son:

$$m_n = E[X^n] = \int_{-\infty}^{\infty} x^n f(x)dx = \frac{\alpha x_m^n}{\alpha - n} \qquad (3.7)$$

El momento m_n existe sólo si $\alpha > n$, lo que implica que dentro del rango de valores importante en este contexto, la distribución posea todos los momentos infinitos salvo la media $m_1 = E[X] = \frac{\alpha x_m}{\alpha - 1}$ que existirá en el caso de que $\alpha > 1$. Es importante resaltar que, el hecho de que la varianza sea infinita, provoca que con este tipo de distribuciones no se cumpla el Teorema Central del Límite clásico, sino el generalizado[21].

A continuación, se define la distribución generalizada y se verá la relación existente entre ambas.

Se dice que una variable aleatoria X sigue una distribución generalizada de Pareto, $X \in GPD(k, \sigma, \mu)$, si:

[21] Este teorema no posee la restricción de que las variables aleatorias posean varianza finita. En el caso de las variables pertenezca al conjunto de las de cola pesada con índice $\alpha < 2$, la suma de las variables tiende a una distribución tipo estable cuyo parámetro de forma es concretamente α. De hecho, en el caso de que α=2, esa distribución es simplemente la Gaussiana.

$$F(x;k,\sigma,\mu)=\begin{cases}1-\left(1-\dfrac{k(x-\mu)}{\sigma}\right)^{\frac{1}{k}} & \text{si } k\neq 0\,,\sigma>0,x>\mu\\ 1-e^{-\frac{(x-\mu)}{\sigma}} & \text{si } k=0\,,\sigma>0,x>\mu\end{cases} \qquad (3.8)$$

Donde $F(x;k,\sigma,\mu)=P\{X\leq x\}$ es la función de distribución acumulada, σ es un parámetro de escaldo, k es un parámetro de forma que determina lo que antes se denominó índice de cola y μ es un parámetro de posición[22] que, como x_m en el caso de la pura, determina el valor mínimo que puede tomar la variable.

Para el caso de $k>0$, los valores permitidos para la variable x son $\mu<x<\mu+\frac{\sigma}{k}$. Las distribuciones resultantes para este caso no son de interés dentro de este contexto. Por ejemplo, si $k=1$, la distribución se convierte en la uniforme $U(\mu\,,\,\mu+\sigma)$. El rango de valores de interés es $-\infty<k<-\frac{1}{2}$, que es el equivalente al comentado anteriormente para el parámetro α de la distribución pura.

A partir de ahora se omitirán los casos de $k\geq 0$. Con la función de distribución se obtiene:

$$\overline{F}(x;\alpha,x_m)=P\{X>x\}=\begin{cases}\left(1-\dfrac{k(x-\mu)}{\sigma}\right)^{\frac{1}{k}} & \text{si } \sigma>0,x>\mu\\ 1 & \text{si } \sigma>0,x<\mu\end{cases} \qquad (3.9)$$

Se puede ver que es de cola pesada ya que posee un comportamiento hiperbólico que en este caso es asintótico, es decir, en la cola, para valores grandes de x. Por abreviar $\overline{F}(x;k,\sigma,\mu)=\overline{F}(x)$, y sin perder generalidad $\mu=0$:

[22] En inglés, location parameter.

$$\lim_{x\to\infty}\frac{\partial \log \bar{F}(x)}{\partial \log x}=\lim_{y\to\infty}\frac{\partial \log \bar{F}(e^{y})}{\partial y}=\lim_{y\to\infty}\frac{\partial(\log((1-\frac{ke^{y}}{\sigma})^{\frac{1}{k}}))}{\partial y}$$

$$=\lim_{y\to\infty}\frac{1}{k}\frac{1}{(1-\frac{ke^{y}}{\sigma})}\frac{(-ke^{y})}{\sigma}=\lim_{y\to\infty}\frac{-e^{y}}{\sigma-ke^{y}}=\frac{1}{k}\qquad(3.10)$$

Aquí también se puede ver que la pendiente de la distribución acumulada complementaria en escalas logarítmicas es:

$$\frac{\partial\, log\, \bar{F}(x)}{\partial\, log\, x}=\frac{-x}{\sigma-k(x-\mu)}\qquad(3.11)$$

Se ve que no es constante como ocurre en el caso de la pura (3.3), sino que va aumentando a medida que x va creciendo. Cuando $x \gg \sigma$, se puede considerar constante con valor $\frac{1}{k}$. Esta propiedad es la que provoca que, a diferencia de como ocurre con la pura, no sea necesario determinar un valor a partir del cual se considere que la distribución sigue un comportamiento lineal ya que, al poseer curvatura, permite el ajuste en todo el rango de valores. La curvatura será mayor cuanto mayor sea el valor de σ.

La función de densidad será por tanto:

$$f(x;k,\sigma,\mu)=\frac{\partial F(x;k,\sigma,\mu)}{\partial x}=\begin{cases}\frac{1}{\sigma}(1-\frac{k(x-\mu)}{\sigma})^{\frac{1}{k}-1} & si\ \ \sigma>0, x>\mu\\ 0 & si\ \ \sigma>0, x>\mu\end{cases}\qquad(3.12)$$

Al igual que en el caso de la $\overline{F}(x)$, la $f(x)$ también presenta un comportamiento hiperbólico asintótico, con el mismo valor -1:

$$\lim_{x\to\infty}\frac{\partial\, log\, f(x)}{\partial\, log\, x}=\lim_{x\to\infty}(1-k)\frac{-x}{\sigma-k(x-\mu)}=\frac{1}{k}-1\qquad(3.13)$$

La función inversa de distribución $x(F)$, se obtiene calculando la inversa de la acumulada (3.8) y es fácil ver que:

$$x(F) = \mu + \frac{\sigma}{k}(1 - (1 - F)^k)\,, 0 \leq F \leq 1 \qquad (3.14)$$

Los momentos $\mathrm{m_n}$ de la distribución es muy complejo obtenerlos de forma general. La media existirá si $k > -1$ y tendrá un valor:

$$m_1 = E[X] = \mu + \frac{\sigma}{1 + k} \qquad (3.15)$$

Como ya se ha explicado con la distribución pura, los resultados obtenidos suelen generar un rango de valores para k que posea el primer momento finito y el resto infinitos, es decir, $-1 < k < \frac{-1}{2}$. Esto provoca que para el ajuste de la distribución a las muestras, métodos de estimación de parámetros como MOM[23] o PWM[24] no se puedan utilizar. En el próximo capítulo se verá el método EPM[25] que es el utilizado para llevar a cabo la tarea del ajuste ya que el de máxima verosimilitud MLE no ha proporcionado tampoco resultados satisfactorios.

Por último, se va a ver cómo están relacionadas entre ellas, es decir, que valor deben tener los parámetros de la generalizada para convertirse en la pura:

$$F(x; k, \sigma, \mu) = 1 - \left(1 - \frac{k(x - \mu)}{\sigma}\right)^{\frac{1}{k}}$$

[23] **MOM** Del inglés, Method of Moments

[24] **PWM** Del inglés, Probability-Weighted-Moments

[25] **EPM** Del inglés, Elemental Percentil Method. Este método será explicado en el capítulo siguiente, dentro del apartado de algoritmo EPM.

$$\Rightarrow Si\ k = \frac{-1}{\alpha}, \sigma = \frac{x_m}{\alpha}, \mu = x_m$$

$$\Rightarrow F(x; \tfrac{-1}{\alpha}, \tfrac{x_m}{\alpha}, x_m) = 1 - (1 + \tfrac{(x - x_m)}{x_m})^{-\alpha} = 1 - (\tfrac{x}{x_m})^{-\alpha}$$

$$\Rightarrow GPD(\frac{-1}{\alpha}, \frac{x_m}{\alpha}, x_m) \equiv PPD(\alpha, x_m)$$

Dentro de este contexto, como normalmente $1 < \alpha < 2$, se podría decir que la distribución generalizada se aproxima mucho a la pura cuando, además de poseer un índice de cola negativo, posee un valor de σ relativamente pequeño. Éste parámetro por lo tanto, se puede entender como un indicador de la curvatura, siendo curvatura cero para el valor de σ= $\frac{x_m}{\alpha} = -k\mu$.

3.1.2 UMBRALIZACIÓN: DISTRIBUCIÓN DE PARETO

En este apartado se va a ver, cómo se comportan las distribuciones del capítulo anterior ante umbralizaciones, es decir, ver como se ven afectadas las variables aleatorias al truncarlas inferiormente.

Por su simplicidad se comenzará por la distribución pura de Pareto seguida de la generalizada.

Sea $X \in \mathrm{PPD}(\alpha, x_m)$, si se restringe a que toma valores mayores que un $x'_m > x_m$, utilizando la ecuación (3.1) se obtiene que:

$$P\{X \leq x | X > x'_m\} = \frac{P\{x'_m \leq X < x\}}{P\{X > x'_m\}} = \frac{F(x;\alpha,x_m) - F(x'_m;\alpha,x_m)}{1 - F(x'_m;\alpha,x_m)} =$$

$$= \frac{(\frac{x'_m}{x_m})^{-\alpha} - (\frac{x}{x_m})^{-\alpha}}{(\frac{x'_m}{x_m})^{-\alpha}} = 1 - (\frac{x}{x'_m})^{-\alpha} = F(x; \alpha, x'_m).$$

Entonces se ve que para $x'_m > x_m$:

$$X \in PPD(\alpha, x_m) \Leftrightarrow (X|X > x'_m) \in PPD(\alpha, x'_m) \qquad (3.16)$$

Esto significa que, si truncamos una muestra a partir de un valor, el índice de la cola α, debería mantenerse invariable ante esta alteración si la distribución subyacente es realmente una PPD.

La distribución generalizada posee una característica similar, pero más interesante a la hora de trabajar con los datos. Se debe a que el parámetro x_m de la pura suele no tener mucha importancia y su cálculo ser heurístico. Como ahora se verá, la generalizada se ve alterada en su parámetro σ, el cual es calculado algorítmicamente. Permite que, a partir de una muestra que aparenta poseer una GPD subyacente, se generen submuestras mediante umbralización y se recalculen los parámetros para comprobar si su alteración es consistente con la supuesta distribución que siguen.

Sea $X \in GPD(k, \sigma, \mu)$, utilizando la ecuación (3.8), para un valor de $u > \mu$ cumple que:

$$P\{X \leq x|X > u\} = \frac{P\{u \leq X \leq x\}}{P\{X > u\}} = \frac{F(x; k, \sigma, \mu) - F(u; k, \sigma, \mu)}{1 - F(u; k, \sigma, \mu)}$$

$$= \frac{\left(1 - \frac{k(u-\mu)}{\sigma}\right)^{\frac{1}{k}} - \left(1 - \frac{k(x-\mu)}{\sigma}\right)^{\frac{1}{k}}}{\left(1 - \frac{k(u-\mu)}{\sigma}\right)^{\frac{1}{k}}} = 1 - \left(\frac{1 - \frac{k(x-\mu)}{\sigma}}{1 - \frac{k(u-\mu)}{\sigma}}\right)^{\frac{1}{k}}$$

$$= 1 - \left(1 - \frac{\frac{k(x-u)}{\sigma}}{1 - \frac{k(u-\mu)}{\sigma}}\right)^{\frac{1}{k}} = 1 - \left(1 - \frac{k(x-u)}{\sigma - k(u-\mu)}\right)^{\frac{1}{k}} = F(x; k, \sigma - k(u-\mu), u).$$

Entonces se ve que, para $u > \mu$:

$$X \in GPD(k,\sigma,\mu) \Leftrightarrow (X|X > u) \in GPD(k,\sigma - k(u-\mu),u) \qquad (3.17)$$

Esta propiedad será de utilidad para poder comprobar si las muestras que se están ajustando a ésta distribución aumentan su parámetro σ al ser umbralizadas de esta forma ya que al ser $\mathrm{k} < 0$, $\sigma(u) = \sigma_0 - \mathrm{k}(\mathrm{u} - \mu) > \sigma_0$, donde $\sigma_0 = \sigma(\mu)$. Esto significa que teóricamente, σ varía de forma lineal respecto a la umbralización con una pendiente de valor $-\mathrm{k}$.

De forma análoga se puede ver que, para $\mathrm{u} > \mu$:

$$X \in GPD(k,\sigma,\mu) \Leftrightarrow (X - u|X > u) \in GPD(k,\sigma - k(u-\mu),0) \qquad (3.18)$$

$$k(u) = k_0 \; \forall \, u \qquad (3.19)$$

$$\sigma(u) = \sigma(\mu) - \; k(u-\mu) = \sigma_0 - k(u-\mu) \qquad (3.20)$$

La única diferencia es el parámetro de posición μ, que en lugar de pasar a valer u se convierte en 0, así para todo umbral las representaciones comenzarán en el mismo punto.

Esta difiere mucho de la archiconocida propiedad de no memoria de la distribución exponencial, la cual no se vería alterada por esta umbralización, es decir, $\mathrm{X} \in \mathrm{Exp}(\lambda) \Leftrightarrow (\mathrm{X} - \mathrm{u}|\mathrm{X} > u) \in \mathrm{Exp}(\lambda)$

Debido a estas umbralizaciones, lógicamente el valor de la esperanza también se verá afectado. Con los desarrollos anteriores unido a las ecuaciones (3.7) y (3.15):

$$X \in PPD(\alpha, x_m) \Rightarrow E[X - u|X > u] = E[X|X > u] - u = \frac{\alpha u}{\alpha - 1} - u = \frac{u}{\alpha - 1} \propto u \qquad (3.21)$$

$$X \in GPD(k,\sigma,\mu) \Rightarrow E[X - u|X > u] = \frac{\sigma - k(u-\mu)}{1+k} = \frac{\sigma + k\mu}{1+k} - \frac{ku}{1+k} \propto u \qquad (3.22)$$

Se puede observar, que a medida que aumenta el valor de u, aumenta proporcionalmente la media. A pesar de que las muestras son finitas, se podrá comprobar si en cierta manera siguen también esta tendencia al truncarlas inferiormente. Esta propiedad es paradójica. Por ejemplo, si se está midiendo el tamaño de un flujo concreto que sigue una distribución de este tipo, cuanto más bytes se hayan contabilizado, más se debe suponer que queden por contabilizar. Esto suele ocurrir al contrario en muchas otras distribuciones. Por ejemplo, la distribución exponencial por la propiedad de no memoria, $E[X-u|X>u]=E[X]$ es decir, no depende de u.

Hay que añadir que todas las distribuciones con soporte acotado, como por ejemplo la uniforme, su media disminuye a medida que aumenta u en lugar de incrementarse.

3.1.3 CURVA DE LORENZ Y COEFICIENTE DE GINNI: DISTRIBUCIÓN DE PARETO

En este apartado se va a presentar la llamada Curva de Lorenz, que será utilizada en apartados posteriores. Conceptualmente, esta curva es una representación gráfica que plasma la distribución relativa de una variable (aleatoria) dentro de un dominio determinado. Dentro del contexto de este proyecto, el dominio son los flujos que atraviesan un nodo de red y la variable es su tamaño. Con esta curva se puede observar la influencia de los llamados “elefantes” con respecto a los “ratones” comentados en el apartado introductorio. Mide la desigualdad que hay entre ellos en el porcentaje del total del tráfico que transportan, representándose en orden creciente, desde los más pequeños (ratones) hasta los más grandes (elefantes). Esta curva es otra forma de realizar comparaciones entre los resultados empíricos y teóricos, diferente a todas las vistas en análisis similares. Además de caracterizar el tráfico desde un

punto de vista que aporta mucha información de cómo los flujos lo "transportan", mediante el llamado coeficiente de Ginni, se obtiene un resultado cuantitativo que permite las comparaciones sin ningún tipo de subjetividad.

3.1.3.1 CURVA DE LORENZ.

Matemáticamente, la curva de Lorenz de una variable aleatoria X con función de densidad f(x) y función inversa de distribución x(F) es:

$$L(F) = \frac{\int_{-\infty}^{x(F)} xf(x)dx}{\int_{-\infty}^{\infty} xf(x)dx} = \frac{\int_{-\infty}^{x(F)} xf(x)dx}{E[X]} \qquad (3.23)$$

Aquí se ve que sólo está definida para aquellas variables que posean una distribución con media finita, casos que ya han sido comentados anteriormente para la *PPD, GPD y LN.*

Para el caso de la distribución de Pareto pura, aplicando la ecuación anterior unido a (3.4), (3.6) y (3.7):

$$L(F) = \frac{\alpha - 1}{\alpha x_m} \int_{x_m}^{\frac{x_m}{(1-F)^{1/\alpha}}} \alpha \left(\frac{x}{x_m}\right)^{-\alpha} dx = -x_m^{(\alpha-1)} \left(\frac{x_m^{(1-\alpha)}}{(1-F)^{\frac{1-\alpha}{\alpha}}} - x_m^{(1-\alpha)} \right)$$

$$= 1 - (1-F)^{1-\frac{1}{\alpha}}, \quad \alpha > 1, \quad 0 \leq F \leq 1 \qquad (3.24)$$

Para el caso de la distribución generalizada, su cálculo es un poco más complicado y se ha realizado en varias etapas.

Primero ver que una primitiva de xf(x), integrando por partes es:

$$\int xf(x)dx = xF(x) - \int F(x)dx$$

Ahora para el cálculo de una primitiva de F(x) (3.8), se realiza el cambio de variable $t = 1 - \frac{k(x-\mu)}{\sigma}$ tal que $F(t) = 1 - t^{\frac{1}{k}}$ con lo que se obtiene:

$$\int F(x)dx = \int F(t)(-\frac{\sigma}{k}dt) = \frac{-\sigma}{k}t + \frac{\sigma}{k+1}t^{1+\frac{1}{k}}$$

Ahora deshaciendo el cambio de variable y obviando todos los elementos constantes (no tendrán importancia al evaluar la primitiva):

$$\int F(x)dx = x + \frac{\sigma}{k+1}(1 - \frac{k(x-\mu)}{\sigma})^{1+\frac{1}{k}}$$

Con esto, agrupando se obtiene que:

$$\int xf(x)dx = \left(1 - \frac{k(x-\mu)}{\sigma}\right)^{\frac{1}{k}}(\frac{k}{k+1}(x-\mu) - x - \frac{\sigma}{k+1})$$

Entonces, para el cálculo de la curva, se evalúa esta primitiva tal que:

$$\int_{\mu}^{x(F)} xf(x)dx = \left(1 - \frac{k(x(F)-\mu)}{\sigma}\right)^{\frac{1}{k}}\left(\frac{k}{k+1}(x(F)-\mu) - x(F) - \frac{\sigma}{k+1}\right) + \mu + \frac{\sigma}{k+1}$$

Ahora, introduciendo el valor de x(F) y E[X] correspondientes (ecuaciones (3.14) y (3.15) respectivamente) y agrupando lo máximo posible, se obtiene que:

$$L(F) = \frac{\int_{-\infty}^{x(F)} xf(x)dx}{E[X]} =$$

$$= \frac{\sigma}{k}\frac{(1-F)}{\sigma+(1+k)\mu}((1-F)^k - 1) + F,\ \ k > -1\ ,\ 0 \leq F \leq 1 \qquad (3.25)$$

Se puede observar que para el caso de $\mu = 0$, la forma de la curva depende únicamente del parámetro k tal que $L(F) = \frac{1}{k}\Big((1-F)^{1+k} - (1-F)\Big) + F$ al igual que ocurría en el caso de la distribución pura con el parámetro α. Hay que resaltar que las diferencias que se producen por definir μ como el mínimo valor de la muestra o con valor cero, son prácticamente inapreciables en todos los resultados obtenidos. Esto es así dentro de este contexto $(-1 < k < -\frac{1}{2})$. En otros casos, como puede ser $k = 1$, ya se ha comentado que la distribución se convierte en la uniforme $U(\mu, \mu + \sigma)$ donde el parámetro sí que jugaría un papel importante.

Así es fácil comprobar que los cálculos son correctos comprobando los casos extremos de igualdad y desigualdad perfecta:

$$\lim_{k\to -1} L(F) = 0$$

$$\lim_{k\to \infty} L(F) = F$$

Donde el primero es un caso extremo donde la media no converge, y el segundo la distribución es determinista.

3.1.3.2 COEFICIENTE DE GINNI

El coeficiente de Ginni se obtiene a partir de la curva de Lorenz. Este coeficiente sirve para cuantificar la desigualdad que la curva representa gráficamente. El coeficiente se define como:

$$G = 1 - 2\int_0^1 L(F)dF \quad 0 \leq G \leq 1 \qquad (3.26)$$

El coeficiente está acotado entre 0 y 1. Esto se debe a que la integral de la curva de Lorenz para variables aleatorias positivas vale un máximo de $\frac{1}{2}$. Este caso se denomina igualdad perfecta, en el cual la curva es la recta $L(F) = F,\ 0 \leq F \leq 1$ que corresponde con una distribución determinística, de valor constante, caso de $\mathrm{G} = 0$. Por el contrario, la desigualdad perfecta que se corresponde con $\mathrm{G} = 1$, se produce cuando la curva de Lorenz posee área cero, es decir $L(F) = 0,\ 0 \leq F < 1$ y $L(1) = 1$. Esto se produce únicamente con casos extremos, como puede ser la distribución pura con el índice α tendiendo a 1, donde la media deja de converger.

Para el caso de una variable aleatoria $X \in PPD(\alpha, x_m)$, observando la ecuación anterior unido a (3.24) se obtiene que:

$$G = 1 - 2\int_0^1 \left(1 - (1-F)^{1-\frac{1}{\alpha}}\right) dF = 1 - 2\left(1 - \frac{1}{2 - \frac{1}{\alpha}}\right) = \frac{1}{2\alpha - 1}$$

$$\Rightarrow G = \frac{1}{2\alpha - 1} \quad \alpha > 1 \qquad (3.27)$$

Para el caso de una variable aleatoria $X \in GPD(k, \sigma, \mu)$, si se define $\varphi_0 = \frac{\sigma}{k}\frac{1}{\sigma + (1+k)\mu}$ y se utiliza la ecuación (3.25) se obtiene que:

$$L(F) = \varphi_0\Big((1-F)^{k+1} - (1-F)\Big) + F$$

A continuación, aplicando la definición del coeficiente a esta ecuación se obtiene que:

$$G = 1 - 2\int_0^1 (\varphi_0\big((1-F)^{k+1} - (1-F)\big) + F)dF = \varphi_0 \frac{k}{2+k}$$

$$\Rightarrow G = \frac{\sigma}{(\sigma + (1+k)\mu)(2+k)} \ , k > -1 \qquad (3.28)$$

Al igual que ocurre con la curva de Lorenz, si $\mu = 0$ el coeficiente sólo depende del parámetro k tal que $G = \frac{1}{(2+k)}$ y al igual que se mostró antes si $k \to -1 \Rightarrow G = 1$ y si $k \to \infty \Rightarrow G = 0$, lo que es totalmente coherente con lo esperado.

3.2 DISTRIBUCIÓN LOG-NORMAL

3.2.1 DEFINICIÓN Y PROPIEDADES BÁSICAS DE LA DISTRIBUCIÓN LOG-NOMRAL

Una variable aleatoria X sigue una distribución log-normal, $X \in LN(\mu, \sigma)$, si su logaritmo se distribuye gaussianamente[26]. Esto se puede ver de dos formas equivalentes:

$$Si\ X \in LN(\mu, \sigma) \Rightarrow Y = \log(X) \in N(\mu, \sigma)$$

ó

$$Si\ X \in N(\mu, \sigma) \Rightarrow Y = e^X \in LN(\mu, \sigma)$$

Por esta razón, debido a que la suma de dos variables aleatorias normales Y_1 e Y_2 es normal, entonces la multiplicación de dos variables log-normales será también log-Normal.

[26] Normalmente.

Si $X \in N(\mu, \sigma)$, su función de densidad de distribución es la archiconocida función gaussiana con media μ y desviación típica σ:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty < x < \infty$$

donde $-\infty < \mu < \infty \ y \ \sigma > 0$.

Entonces $Y = e^X \in LN(\mu, \sigma)$ posee una función de densidad de distribución:

$$f_Y(y) = \begin{cases} \frac{1}{y} f_X(\log y) = \frac{1}{\sqrt{2\pi}\sigma y} e^{-(\log y-\mu)^2/2\sigma^2} & y > 0 \\ 0 & y < 0 \end{cases} \qquad (3.29)$$

Se puede ver que la distribución log-normal es estrictamente positiva y que sus parámetros μ y σ son la media y la desviación típica de la variable aleatoria normal asociada, no de ella misma. Señalar que $\log x = \log_e x$ en todos los casos en los que se utilice el logaritmo.

La función de distribución acumulada, al igual que ocurre con la distribución Normal, no se puede expresar como una función elemental. Se expresa a partir de $\Phi(x)$, la función de distribución acumulada de la $N(0,1)$:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} \, dt \qquad (3.30)$$

Entonces si $X \in LN(\mu, \sigma)$, haciendo el cambio de variable $z = \frac{\log t - \mu}{\sigma}$,es sencillo observar que:

$$F(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}} \int_0^x \frac{1}{\sigma t} e^{-(\log t-\mu)^2/2\sigma^2} dt = \Phi\left(\frac{\log x - \mu}{\sigma}\right) \qquad (3.31)$$

Con lo que la función de distribución complementaria[27] es:

$$\overline{F}(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}}\int_x^{\infty}\frac{1}{\sigma t}e^{-(\log t-\mu)^2/2\sigma^2}dt = 1-\Phi\left(\frac{log\,x-\mu}{\sigma}\right) \qquad (3.32)$$

Los momentos m_n de la distribución son:

$$m_n = E[X^n] = e^{n\mu+\frac{1}{2}n^2\sigma^2}$$

La distribución log-normal, en contraposición a la de Pareto (con los parámetros dentro del rango de este contexto), tiene todos sus momentos finitos. Por lo tanto posee una media y una varianza definidas:

$$E[X] = e^{\mu+\frac{1}{2}\sigma^2} \qquad (3.33)$$

$$V[X] = E[X^2] - E^2[X] = \left(e^{\sigma^2}-1\right)e^{2\mu+\sigma^2} \qquad (3.34)$$

Gracias a esto, a partir de la media y la varianza, es posible calcular los parámetros de la distribución simplemente despejando las ecuaciones anteriores:

$$\mu = \log E[X] - \frac{1}{2}log\left(1+\frac{V[X]}{E^2[X]}\right) \qquad (3.35)$$

$$\sigma^2 = \log\left(1+\frac{V[X]}{E^2[X]}\right) \qquad (3.36)$$

Para ajustar la distribución a las muestras, bastará con calcular la media y la varianza muestrales y obtener los parámetros con las ecuaciones que se acaban de presentar.

[27] En muchas ocasiones esta función se expresa mediante la función $Q(x) = 1-\Phi(x)$ tal que $\overline{F}(x;\mu,\sigma) = Q(\frac{\log x-\mu}{\sigma})$

A pesar de que, al contrario de la distribución de Pareto, la log-normal se puede considerar de "cola ligera" al poseer todos sus momentos finitos, se va a ver que es posible que en cierto rango de valores, su comportamiento sea muy parecido. Utilizando (3.29) es fácil ver que:

$$\log f(x) = -\log x - \log\sqrt{2\pi}\sigma - \frac{(\log x - \mu)^2}{2\sigma^2}$$
$$= -\frac{(\log x)^2}{2\sigma^2} + \left(\frac{\mu}{\sigma^2} - 1\right)\log x - \log\sqrt{2\pi}\sigma - \frac{\mu^2}{2\sigma^2}$$
$$\Rightarrow \frac{\partial \log f(x)}{\partial \log x} = \left(\frac{\mu}{\sigma^2} - 1\right) - \frac{\log x}{\sigma^2} \qquad (3.37)$$

Si σ es suficientemente grande, el segundo término de la derivada será pequeño para un rango amplio de valores de x, y por lo tanto con ambas escalas logarítmicas, la densidad parecerá casi lineal para este rango. Esto mismo le ocurre a la CCDF, como se vio en la figura 3-1 mediante generaciones aleatorias. A pesar de que en ciertos casos parecen comportarse de forma similar, hay que destacar que la log-Normal posee curvatura en todo su rango.

Observando la ecuación anterior, se puede intuir que la log-log CCDF, presentará una pendiente mayor[28] a medida que se desplaza por el eje x al estar restando el segundo término. Esta similitud de comportamientos también se produce en el caso de la distribución de Pareto como ya se pudo observar mediante las ecuaciones (3.3) y (3.5).

3.2.2 UMBRALIZACIÓN: DISTRIBUCIÓN LOG-NORMAL

Para el caso de la distribución log-normal, como se comentó en el apartado anterior, su función de distribución acumulada no se puede

[28] Aquí mayor se refiere a mayor en valor absoluto ya que la CCDF posee pendientes negativas, con lo que realmente la pendiente es cada vez menor en sentido matemático

expresar como función elemental. Por esto, no es tan sencillo determinar cómo se ve alterada de forma teórica frente a una umbralización. Para evaluar su comportamiento frente a las muestras umbralizadas, no ha sido posible determinar cómo se alteran de forma teórica sus parámetros μ, σ frente a los umbrales u. Para poder compararlo con la distribución de Pareto se recurrirá a las medias umbralizadas teóricas.

Sea X una variable aleatoria con función de densidad $f_X(x)$. Entonces, $X|X > u$ tendrá una función de densidad $f_{X|X>u}(x) = \frac{f_X(x)}{P\{X>u\}}$ $x > u$. Por lo tanto:

$$E[X|X > u] = \int_u^\infty x f_{X|X>u}(x)dx = \frac{1}{P\{X > u\}} \int_u^\infty x f_X(x)dx$$

Así se define la esperanza parcial $g(u) = E[X|X > u]P\{X > u\} = \int_u^\infty x f_X(x)dx$ que será utilizada también en apartados posteriores.

Para la variable log-normal, esta sí se puede expresar en función de $\Phi(x)$, que como ya se vio en el capítulo anterior es la función de distribución acumulada de la variable aleatoria gaussiana estándar (3.30):

$$g(u) = \int_u^\infty x f_X(x)dx = \int_u^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{-(\log x - \mu)^2/2\sigma^2} dx$$

Para su cálculo se realiza el cambio de variable siguiente:

$$\tau = \frac{\log x - \mu}{\sigma} \Rightarrow x = e^{\sigma\tau+\mu} \Rightarrow dx = \sigma e^{\sigma\tau+\mu} d\tau$$

Así:

$$g(u) = \frac{1}{\sqrt{2\pi}} \int_{\frac{log\, u-\mu}{\sigma}}^{\infty} e^{(\frac{-\tau^2}{2}+\sigma\tau+\mu)}\, d\tau$$

Completando cuadrados en el exponente:

$$g(u) = e^{\mu+\frac{\sigma^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{\frac{\log u-\mu}{\sigma}}^{\infty} e^{-(\tau-\sigma)^2/2}\, d\tau$$

Para finalizar, $z = \tau - \sigma$, con lo que:

$$g(u) = e^{\mu+\frac{\sigma^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{\frac{\log u-\mu-\sigma^2}{\sigma}}^{\infty} e^{-z^2/2}\, dz = e^{\mu+\frac{\sigma^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{-\log u+\mu+\sigma^2}{\sigma}} e^{-z^2/2}\, dz$$

Expresándolo en función de Φ(x):

$$g(u) = e^{\mu+\frac{\sigma^2}{2}}\, \Phi\left(\frac{\mu+\sigma^2-log\, u}{\sigma}\right) \qquad (3.38)$$

Así expresamos el valor de las medias umbralizadas en función del valor de los parámetros u de la siguiente manera:

$$E[X-u|X>u] = \frac{g(u)}{P\{X>u\}} - u = \frac{g(u)}{1-\Phi(u)} - u \qquad (3.39)$$

Con estas propiedades, se podrán realizar umbralizaciones sobre los datos y realizar comparaciones entre los resultados empíricos obtenidos sobre el conjunto de medidas bajo estudio y los teóricos, con el fin de determinar cuáles son más aproximados.

3.2.3 CURVA DE LORENZ Y COEFICIENTE DE GINNI: DISTRIBUCIÓN LOG-NORMAL

3.2.3.1 CURVA DE LORENZ

Para la distribución log-normal, no es posible expresar la curva mediante funciones elementales. Vuelve a aparecer la función $\Phi(\mathrm{x})$ y su inversa $\Phi^{-1}(F)$ que en ocasiones se la denomina función probit, la cual aparece en la expresión de la inversa de la log-normal de la siguiente manera:

$$F(x) = F = \Phi\left(\frac{\log x - \mu}{\sigma}\right) \Rightarrow \Phi^{-1}(F) = \frac{\log x - \mu}{\sigma}$$

Con lo que:

$$x_{LN} = x_{LN}(F) = e^{\mu + \sigma\Phi^{-1}(F)} \qquad (3.40)$$

Ahora para el cálculo de la expresión de la curva, recurriendo a la expresión de, la esperanza parcial $g(u)$ (3.38), la esperanza $E[X]$ (3.33) y la función de densidad (3.29) se obtiene que:

$$\int_{-\infty}^{x_{LN}(F)} xf(x)dx = \int_{0}^{e^{\mu+\sigma\Phi^{-1}(F)}} xf(x)dx = E[X] - g(e^{\mu+\sigma\Phi^{-1}(F)})$$

Entonces:

$$L(F) = \frac{\int_{-\infty}^{x_{LN}(F)} xf(x)dx}{E[X]} = 1 - \Phi\left(\frac{\mu + \sigma^2 - log\left(e^{\mu+\sigma\Phi^{-1}(F)}\right)}{\sigma}\right)$$

Lo que finalmente se expresa:

$$L(F) = 1 - \Phi\left(\sigma - \Phi^{-1}(F)\right) \qquad (3.41)$$

Se puede observar que los valores de la curva no dependen del parámetro μ. Dependen únicamente de la desviación típica σ de la variable aleatoria gaussiana asociada. Cuanto mayor sea su valor, mayor desigualdad existirá. Esto no sería así en el caso de haber introducido un nuevo parámetro de posición en la definición de la distribución, pero al igual que en el caso de la distribución de Pareto, las diferencias son inapreciables, complicando en exceso los desarrollos mostrados. Se comprueba, al igual que para la distribución de Pareto, si los resultados son correctos con los casos extremos de igualdad y desigualdad perfecta:

$$\lim_{\sigma \to 0} L(F) = 1 - \Phi\left(-\Phi^{-1}(F)\right) = 1 - \Phi\left(\Phi^{-1}(1-F)\right) = F$$

$$\lim_{\sigma \to \infty} L(F) = 1 - \Phi(\infty) = 0$$

Se puede observar que en ambos casos ocurre lo esperado. El primero es un caso degenerativo, en el cual la gaussiana asociada es constante con valor μ, es decir, determinista. El segundo, la distribución también degenera a poder tomar valores infinitamente grandes con la misma probabilidad que el resto, lo que provoca la desigualdad perfecta.

3.2.3.2 COEFICIENTE DE GINNI

Para el coeficiente de Ginni, no ha sido posible calcular una fórmula cerrada que dependa de los parámetros de la distribución. Para su cálculo se recurrirá a métodos numéricos que generen esta curva y luego aproximen su área para poder determinar un valor de G teórico y compararlo contra el empírico obtenido de los datos.

4. ASPECTOS PRÁCTICOS DEL ANÁLISIS

En este capítulo se van a explicar todos los aspectos relacionados con la forma en la que se ha llevado a cabo de forma práctica el análisis teórico presentado en el capítulo 3. No consiste en los aspectos propios de la programación en sí, sino más bien de explicar cómo se calcula y obtienen los resultados desde un nivel superior al del código de los programas utilizados.

4.1 CÁLCULO DE LOS PARÁMETROS DISTRIBUCIONALES

4.1.1 DISTRIBUCIÓN GENERZALIZADA DE PARETO: ALGORITMO EPM

En este apartado se va a explicar el algoritmo utilizado para el cálculo de los parámetros que ajustan la distribución $GPD(k, \sigma, \mu)$ a las muestras. En el rango de $-\infty < k < -\frac{1}{2}$, de interés en este contexto, ningún momento salvo la media convergen. Esto provoca que no tenga sentido utilizar métodos de estimación como pueden ser MOM ó PWM. La estimación mediante MLE[29] tampoco se ha utilizado porque, además de que no es posible una maximización analítica de la función log-likelihood[30] (requiriendo métodos numéricos para su resolución), se ha comprobado que en ciertos casos es imposible encontrar un máximo local debido a que no existe, lo que provoca una total inconsistencia de los parámetros obtenidos con las muestras analizadas. Él método elegido es por tanto el EPM[31]. Los autores de [18] aseveran que, aunque no existe ningún método

[29] **MLE:** Del inglés, Maximum Likelihood Estimation

[30] Función que se maximiza en el método MLE.

[31] **EPM:** Del inglés, Elemental Percentile Method

que sea el mejor para todos los valores de los parámetros, dentro del rango de interés en el que se centra el proyecto, el mejor es el EPM.

A continuación se explicará su funcionamiento resaltando las pequeñas variaciones que se han llevado a cabo para adaptarlo al proyecto. Se asume $k \neq 0$, y $\mu = 0$ y se reparametriza la función de distribución de la siguiente forma:

$$F(x; k, \sigma) = 1 - (1 - \frac{x}{\delta})^{1/k} \; ; \; \delta k = \sigma > 0 \qquad (4.1)$$

Sean i, j dos números enteros tal que $i < j$. Sean por tanto $x_{i:n} < x_{j:n}$ dos estadísticos de orden dentro de una muestra aleatoria de tamaño n proveniente de la distribución anterior. El parámetro de posición μ no se tiene en cuenta debido a que las muestras son modificadas antes de introducirlas en el algoritmo. Son desplazadas al origen restándoles su valor mínimo menos uno. Esta alteración se vuelve a deshacer tras su ejecución determinándose su valor como el mínimo de ésta, es decir, $x_{1:n}$ manteniendo la notación anterior. Ahora evaluando la función de distribución acumulada (4.1):

$$F(x_{i:n}; k, \sigma) = 1 - \left(1 - \frac{x_{i:n}}{\delta}\right)^{\frac{1}{k}} = p_{i:n} \qquad (4.2)$$

$$F\left(x_{j:n}; k, \sigma\right) = 1 - \left(1 - \frac{x_{j:n}}{\delta}\right)^{\frac{1}{k}} = p_{j:n} \qquad (4.3)$$

donde en el algoritmo definen:

$$p_{i:n} = \frac{i}{n+1}$$

Esta definición tiene sentido al ser una distribución continua ya que es muy improbable que aparezcan dos valores con el mismo valor dentro de una misma muestra. En el caso de este proyecto, el tamaño de los flujos (en número de Bytes) no es realmente una variable aleatoria continua, es discreta al tomar únicamente valores enteros. A pesar de ello, se puede modelar con esta distribución al tomar un rango de valores muy amplio y estar muy próximos entre sí. Por esta razón, en el algoritmo aplicado ha sido necesario variar ese valor de $p_{i:n}$ por el que se obtiene al realizar la ECDF[32] sobre la muestra. Esto provoca que los valores repetidos dentro de las muestras aparezcan sólo una vez y en ellos se produzca un incremento de probabilidad superior a $\frac{1}{(n+1)}$. La mejora que se obtuvo tras esta variación fue notable. Sin la modificación, muchos resultados eran bastante insatisfactorios, siendo necesaria una fase de post-ajuste prácticamente mediante prueba y error. Tras la modificación, casi todos estos casos mejoraron considerablemente proporcionando un ajuste bastante aceptable como se va a observar en el capítulo de resultados. Tomando logaritmos en las ecuaciones (4.2) y (4.3) se obtiene:

$$\begin{cases} log(1-\frac{x_{i:n}}{\delta}) = kC_i \\ log(1-\frac{x_{j:n}}{\delta}) = kC_j \end{cases} \qquad (4.4)$$

donde $C_i = \log(1-p_{i:n}) < 0$. Se puede ver que (4.4) es un sistema de dos ecuaciones con dos incógnitas $\delta\ y\ k$. Entonces eliminando k:

$$k = C_i\, log(1-\frac{x_{j:n}}{\delta}) = C_j\, log\left(1-\frac{x_{i:n}}{\delta}\right) \qquad (4.5)$$

Para hallar δ basta con resolver la ecuación:

[32] **ECDF:** Del inglés, Empirical CDF.

$$h(\delta) = C_i\, log(1 - \frac{x_{j:n}}{\delta}) - C_j\, log(1 - \frac{x_{i:n}}{\delta}) = 0 \quad (4.6)$$

Hay que ver que al ser $x_{i:n} < x_{j:n}$ y $C_i > C_j$, $h(\delta)$ está definida en el conjunto $\{(-\infty, 0) \cup (x_{j:n}, \infty)\}$. Dos soluciones triviales son $\delta = \pm\infty$. Para probar que existe una solución finita y es única hay que ver que:

$$h(\pm\infty) = 0; \qquad lim_{\delta \to 0^-}\, h(\delta) = \infty \qquad lim_{\delta \to x_j^+}\, h(\delta) = \infty \qquad (4.7)$$

Esto unido a que:

$$\frac{dh(\delta)}{d\delta} = \frac{1}{\delta}\left[\frac{C_i x_{j:n}}{\delta - x_{j:n}} - \frac{C_j x_{i:n}}{\delta - x_{i:n}}\right] = 0$$

que tiene como solución $\delta = \delta_0$, además de $\delta = \pm\infty$, tal que:

$$\delta_0 = \frac{x_{i:n} x_{j:n}(C_j - C_i)}{C_j x_{i:n} - C_i x_{j:n}} \qquad (4.8)$$

Se puede comprobar también que en δ_0, $\frac{d^2 h(\delta)}{d\delta^2} > 0$, es decir, se produce un mínimo relativo, y sólo existe ese punto crítico. Así entonces se ve que:

$$\begin{aligned} \delta_0 &= x_{j:n} \quad si \quad x_{i:n} = x_{j:n} \\ \delta_0 &> 0 \quad si \quad x_{i:n} > C_i x_{j:n}/C_j \\ \delta_0 &\to \pm\infty \quad si \quad x_{i:n} \to C_i x_{j:n}/C_j \\ \delta_0 &< 0 \quad si \quad x_{i:n} < C_i x_{j:n}/C_j \end{aligned}$$

El primer caso nunca se producirá porque si $i \neq j \Rightarrow x_{i:n} \neq x_{j:n}$. Esto se produce por lo comentado anteriormente de asignar los $p_{i:n}$ mediante la

ECDF eliminando los valores repetidos. Así, conjuntamente con obtener mejores resultados, se evita el problema de que esta situación sin solución se produzca (fijarse que en $x_{j:n}$ aparecerá el mínimo pero a su vez es una asíntota vertical). Entonces la continuidad de $h(\delta)$, unido a las propiedades (4.7) y la existencia de un único punto crítico, mínimo relativo, implican que ésta función tenga un único 0 finito si $C_j x_{i:n} \neq C_i x_{j:n}$. Este 0 se situará en el intervalo $(x_{j:n}, \delta_0)$ si $x_{i:n} > C_i\, x_{j:n}/C_j$ ó en el intervalo $(\delta_0, 0)$ si $x_{i:n} < C_i\, x_{j:n}/C_j$.

Con esto entonces, se aplica el método de la bisección para hallar la raíz de la ecuación $h(\delta) = 0$ (4.6). Después de obtener el valor estimado $\hat{\delta}(i,j)$, se obtiene el de los parámetros k, σ de la siguiente forma:

$$\hat{k}(i,j) = \frac{log\left(1 - \frac{x_{i:n}}{\hat{\delta}(i,j)}\right)}{C_i} \qquad (4.9)$$

$$\hat{\sigma}(i,j) = \hat{k}(i,j)\hat{\delta}(i,j) \qquad (4.10)$$

4.1.1.1 MÉTODOS

El algoritmo ahora continúa calculando $\hat{k}(i,j), \hat{\sigma}(i,j)$ para todos los pares de valores distintos tal que $x_{i:n} < x_{j:n}$. Después, tras obtener el conjunto de todos los resultados, el parámetro estimado final se calcula mediante la mediana de éstos. Esto puede ser válido cuando el número de muestras n, no es muy grande. Para un conjunto de n muestras, es fácil ver que el número total de pares es el binomial $\binom{n}{2} = \frac{n(n-1)}{2}$. Entonces el número de pares es proporcional a n^2, lo que supone un coste computacional enorme cuando n toma valores grandes. Éste es el caso del proyecto, donde las muestras poseen una cantidad del orden de 10^4 valores distintos. Por ello se ha recurrido a un conjunto de métodos que

seleccionan los pares de valores de tal forma que se reduce su número e intenta no alterar el resultado final.

A continuación se van han explicar los métodos implementados, donde $n = \#\text{muestras}$, $i, j = \text{posición de los estadísticos de orden } x_{i:n} \text{ y } x_{j:n}$ tal que $i < j, N = \#ejecuciones$:

1. Sistemático-mitad

   En este método se define $s = \left\lfloor \frac{n}{2} \right\rfloor$, donde $\lfloor \rfloor$ representa el entero inferior más próximo. Así para todos los pares, $j = i + s, \quad i \in [1, s]$. Cada par está compuesto por un estadístico de orden y el que se sitúa la mitad de muestras alejado de él. Entonces este procedimiento supone $N \cong \frac{n}{2}$.

2. Sistemático-cuarta

   En este método se define $s = \left\lfloor \frac{n}{4} \right\rfloor$. Entonces, $j = i + s$ y en este caso $i \in [1,3s]$. Es muy similar al anterior a diferencia que ahora los estadísticos que forman los pares están alejados una cuarta parte de las muestras totales. Entonces $N \cong \frac{3n}{4}$.

3. Todos-último

   En este método, se fija $j = n, \; i \in [1, n-1]$. Simplemente es que los pares son todas las muestras con la muestra mayor siempre como el estadístico superior. Entonces $N = n - 1$.

4. Todos-último-primero

   En éste, primero fija $j = n$, $i \in [1, \left\lfloor \frac{n}{2} \right\rfloor]$. Después fija $i = 1$ y $j \in [\left\lfloor \frac{n}{2} \right\rfloor + 1, n]$. Aquí se emparejan la primera mitad de los estadísticos de orden con el último, y la segunda con el primero. Entonces $N \cong n$.

5. Tercios-final

   En éste, se define $s = \left\lfloor \frac{n}{3} \right\rfloor$. Primero $j = i + 2s, i \in [1, s]$. Después $j = i + s, i \in (s, 2s]$. Se empareja el primer y el segundo tercio de los estadísticos de orden con el último. Entonces $N \cong \frac{2n}{3}$

6. Cuartos-final

   Es muy similar al anterior, pero en lugar de separarse en tercios, la muestra se separa en cuartos. Se define $s = \left\lfloor \frac{n}{4} \right\rfloor$. Primero $j = i + 3s, i \in [1, s]$. Después $j = i + 2s, i \in (s, 2s]$. Por último, $j = i + s, i \in (2s, 3s]$. Se emparejan los tres primeros cuartos de los estadísticos de orden con el último. Entonces $N \cong \frac{3n}{4}$.

7. Acercándose

   En este método, $j = n - i + 1, i \in [1, \left\lfloor \frac{n}{2} \right\rfloor]$. Así se empareja el primer punto con el último, el segundo con el penúltimo y así hasta llegar a los dos estadísticos centrales de la muestra. Esto supone $N \cong \frac{n}{2}$.

   La ejecución de todos estos métodos supone un total de:

$$N \cong \frac{n}{2} + \frac{3n}{4} + \frac{2n}{3} + \frac{3n}{4} + \frac{n}{2} + n + n - 1 \approx 5n \text{ ejecuciones.}$$

Además de obtener un k, σ por cada método, se realiza la mediana del conjunto de todas las ejecuciones obteniendo otro par, siendo un total de 8.

4.1.1.2 DISTANCIAS

En este apartado se va a explicar, cómo, de entre las 8 parejas de parámetros (k, σ) obtenidas mediante los métodos anteriores, se opta por una u otra. Para ello se calcula la "distancia" que existe entre la CCDF teórica de la $GPD(k, \sigma, \mu)$ y la CCDF empírica obtenida a partir de las muestras. Obviamente, se selecciona la que posea una menor distancia.

La distancia se calcula de tal forma que, a pesar de que en las colas la densidad de puntos es mucho menor, el ajuste en éstas influya sensiblemente en el valor de la distancia. Por esto, por ejemplo la distancia vectorial euclídea punto a punto no es válida, ya que todos los puntos contribuyen por igual a su valor, dando por lo tanto menor importancia a la zona lejana de la cola.

El valor de la distancia se calcula de forma teórica como la integral del valor absoluto de la diferencia entre las funciones en escalas logarítmicas. Para llevarlo a cabo, las funciones continuas se crean mediante interpolación de orden 0 de la función de distribución acumulada empírica generada a partir de las muestras, generando funciones "escalera[33]". Entonces, el área encerrada entre ellas se reduce a la suma de áreas de rectángulos, cuya altura será la diferencia logarítmica entre las probabilidades empíricas y teóricas, y base la distancia (logarítmica) entre los valores de las muestras (en orden creciente). De esta forma, se solucionan los inconvenientes de, primero, la baja densidad de puntos

[33] Término más conocido por su nombre en inglés, staircase function.

existente en la cola y segundo, la proximidad de los valores de estos puntos a la probabilidad cero. El primero mediante la utilización del concepto de área entre las curvas en lugar de puntos y el segundo mediante el uso del logaritmo de las probabilidades en lugar de su propio valor. Para no excederse en la prioridad que representa el ajuste en la cola respecto al resto, se decidió utilizar también la distancia logarítmica entre los valores de muestras consecutivas. En la cola, las distancias absolutas son muy grandes y la utilización de su valor real provocaba que los desajustes producidos en las otras zonas de la distribución no tuvieran a penas influencia en el valor final de la distancia. Matemáticamente equivale a lo siguiente:

Sea un conjunto de datos X formado por n muestras, $X_i\ i \in [1,n]$. Se ordena la muestra tal que se obtiene $x_i\ i \in [1,n]$ tal que $x_i < x_{i+1}$. Para un $k_0, \sigma_{0,}\mu_0$ estimado a partir de X_i se calcula $\overline{F_i} = \overline{F}(x_i; k_0, \sigma_{0,}\mu_0)$ (ver (3.9)) y $\overline{E_i}$ como los valores empíricos obtenidos mediante la ECDF. Si se denomina $\overline{F_0}(x)$ y $\overline{E_0}(x)$ las funciones continuas creadas mediantes interpolación de orden 0 de los pares $(\overline{F_i}, x_i)$ y $(\overline{E_i}, x_i)$, se tiene que la distancia $D(k_0, \sigma_{0,}\mu_0)$ es:

$$D\left(k_0, \sigma_{0,}\mu_0\right) = \int_{\mu_0}^{\infty} \left|\log \frac{\overline{F_0}(x)}{\overline{E_0}(x)}\right| d\log x = \sum_{i=1}^{n-1} \left|\log \frac{\overline{F_i}}{\overline{E_i}}\right| \log \frac{x_{i+1}}{x_i}$$

Esta distancia otorga una coherencia absoluta entre los ajustes "visuales" y matemáticos que se obtienen en las representaciones de las CCDF en escalas logarítmicas.

4.1.2 DISTRIBUCIÓN LOG-NORMAL: MÉTODO DE LOS MOMENTOS

Para la distribución log-Normal, $LN(\mu, \sigma)$, se optó en un principio por el método MLE para la obtención de sus parámetros. Los resultados

obtenidos no fueron del todo satisfactorios. Se observó que, principalmente en el parámetro σ, se obtenía "desplazado[34]" ya que alterándolo de forma manual los resultados mejoraban considerablemente.

Por esta razón se recurrió a las ecuaciones (3.35) y (3.36) que simplemente seleccionan los valores de μ, σ para que la media y varianza teóricas sean iguales a las muestrales. Este método es conocido como MOM, ya comentado anteriormente. Los resultados que se obtienen son sensiblemente mejores, sin ser posible una mejora mediante alteración manual como en el caso anterior. Por esta razón es el método utilizado en este proyecto.

4.2 REPRESENTACIONES Y ENVOLTURA DISTRIBUCIONAL

Tras el cálculo de los parámetros de las distribuciones ajustadas a las medidas realizadas, se procederá a su representación para poder realizar una inspección visual del ajuste. Para ello se van a realizar los CCDF-plots. Este método ha sido ya comentado en el capítulo 2, debido a que ha sido utilizado por numerosos autores.

Para tener en cuenta la fuerte variabilidad que presentan éstas distribuciones, se recurrirá a la denominada envoltura muestral utilizada por los autores de [9]. Para ello, tras obtener los parámetros que caracterizan las supuestas distribuciones subyacentes, se realizarán $N = 100$ generaciones aleatorias que sigan esa distribución. Estas generaciones poseerán el mismo número de elementos que la muestra de medidas que se está ajustando para tener en cuenta la fuerte variabilidad que presentan estas distribuciones incluso con un número de elementos

[34] En inglés se conoce como biased parameter.

elevado. Se busca obtener un resultado similar al mostrado en las figuras 2-4, 2-5, y 2-6 del capítulo 2.

4.3 DIVERSIDAD TEMPORAL Y ESPACIAL

4.3.1 ESTACIONARIDAD

En este apartado se va a explicar cómo se va a analizar la diversidad temporal, que consiste sencillamente en estudiar si el conjunto de medidas bajo estudio presentan características estacionarias, es decir, distribuciones que no cambian con el tiempo. Para llevar esto acabo, se evaluará el número de días de datos necesarios para que los parámetros de las distribuciones propuestas permanezcan estables.

Entonces para un parámetro cualquiera, sea por ejemplo σ de la $GPD(k,\sigma,\mu)$, se considerará estable si su variación porcentual es menor que un umbral dado durante un período de tiempo determinado. Matemáticamente es simplemente:

$$D(j) \equiv Medidas\ del\ día\ j.$$

$$\sigma(i) \equiv Parámetro\ obtenido\ con\ las\ medidas\ D(j), j \in [1,i]$$

$$\Delta\sigma(i) = \frac{|\sigma(i+1) - \sigma(i)|}{\sigma(i)} \equiv Incremento\ porcentual\ del\ parámetro.$$

Entonces, si $\Delta\sigma(i) < M$ durante N conjuntos de medidas, se considerará que el parámetro de esa distribución es estacionario. Valores escogidos por otros autores [7], son $M = 0.05$ (5%) y $N = 5$, que son los valores escogidos como referencia en este proyecto.

4.3.2 DIVERSIDAD ESPACIAL

El conjunto de datos proviene de RedIris [6]. Gracias a esto, se dispone de una muestra de tráfico que proviene de más de 70 centros universitarios con diferente número de usuarios, ancho de banda de los enlaces de acceso, políticas de filtrado (aplicaciones P2P), utilización de proxys o traducción de direcciones de red NAT[35]. Evidentemente este tipo de propiedades intrínsecas tienen un impacto en el tráfico que manejan.

Se han seleccionado 3 universidades del conjunto total con características similares para así poder comparar sus resultados en cuanto a la distribución que sigue el tamaño de los flujos que manejan y poder comprobar si los resultados que se obtienen en una red, son extrapolables a otras con características similares o de otra forma, eso no es correcto al existir una diversidad espacial que hay que tener en cuenta. Las seleccionadas son un subconjunto de las escogidas por los autores de [7], donde se explica en detalle las propiedades que poseen y donde realizan un análisis de este tipo de diversidad pero referido a la distribución de las direcciones IP más visitadas mediante una distribución denominada Zipf[36]. También hay que remarcar que las medidas de todas han sido tomadas en el mismo periodo de tiempo, evitando de esta forma alteraciones a la diversidad espacial provocada por factores temporales.

[35] **NAT:** Del inglés, Network Address Translation.

[36] La distribución Zipf es una distribución que se puede considerar como la versión discreta de la de Pareto Pura.

5. RESULTADOS

Como ya se ha comentado a lo largo de todo el documento, el análisis se ha realizado principalmente en el puerto 80. Este puerto es el utilizado por defecto por el protocolo HTTP[37]. Por esta razón aparece en todas las transacciones de la Web, lo que implica que sea, de los puertos bien conocidos, el más utilizado. Esto se refleja en los datos disponibles ya que es del que más muestras se disponen y mayor cantidad de tráfico transporta.

Por esta razón, los datos se estudian dividiéndolos en agrupaciones de un único día al ser suficientes para obtener resultados concluyentes. Esto supone que se dispongan de 30 bloques de muestras para cada universidad y sentido[38] del tráfico al tratarse de una recopilación de un mes de duración, así se posee un conjunto total de 180. Además se ha diferenciado entre los bloques que provenían de uno de los 21 días de diario o de los 9 de fin de semana. Parece lógico pensar que las características que presenten sean distintas, empezando porque el número de muestras disponibles en el primer caso es sensiblemente superior al segundo. Así aparecen 12 subconjuntos atendiendo a la terna compuesta por la universidad (U1, U2, U3), sentido (Up, Down) y zona semanal (Diario, fin de semana). A pesar de que cada bloque se ha analizado de forma independiente con el resto, se han extraído generalidades de cada uno de estos 12 subconjuntos para su comparación (diversidad espacial) y análisis de estacionaridad comentado en el capítulo anterior.

[37] **HTTP**: Del inglés, Hyper Text Transfer Protocol

[38] A continuación se explicará que se entiende por sentido del tráfico, tanto ascendente como descendente.

Antes de proseguir hay que aclarar lo que se considera como sentido ascendente (UP) y tráfico descendente (DOWN). En la figura 5-1 se ilustra esta definición. Se considera "tráfico ascendente" a la colección de flujos cuya fuente es algún host situado en la red de la universidad y va destinado a algún otro situado en Internet. Por otra parte, se considera "tráfico descendente" a lo contrario, cuando la fuente está situada en algún lugar de internet y el destino está situado en la universidad.

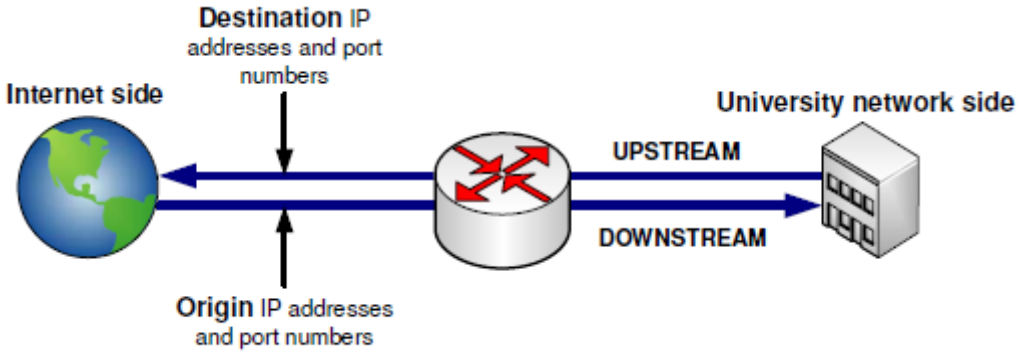

Figura 5-1: Sentido del tráfico en la red

Se muestran los resultados obtenidos tanto por la distribución generalizada de Pareto (apartado 3.1) como por la distribución log-Normal (apartado 3.1). Al producirse las mayores diferencias entre el tráfico ascendente y descendente se presentan con una clara distinción entre ambos grupos.

5.1 TRÁFICO ASCENDENTE

A continuación se va a mostrar de forma gráfica, el proceso de análisis seguido con cada uno de los bloques de muestras de forma independiente. Este proceso es idéntico para todos los casos, siendo al final de este cuando se concluye que distribución de las dos se adapta

mejor a las propiedades empíricas de la muestra. Se muestra un caso representativo del total.

5.1.1 AJUSTE VISUAL LOG-LOG CCDF

A continuación se muestra el ajuste obtenido con la distribución generalizada de Pareto mediante el algoritmo EPM (apartado 4.1.1) y el obtenido con la log-Normal mediante el método de los momentos (apartado 4.1.2) para un bloque de muestras determinado. Se comenzará con la distribución de Pareto:

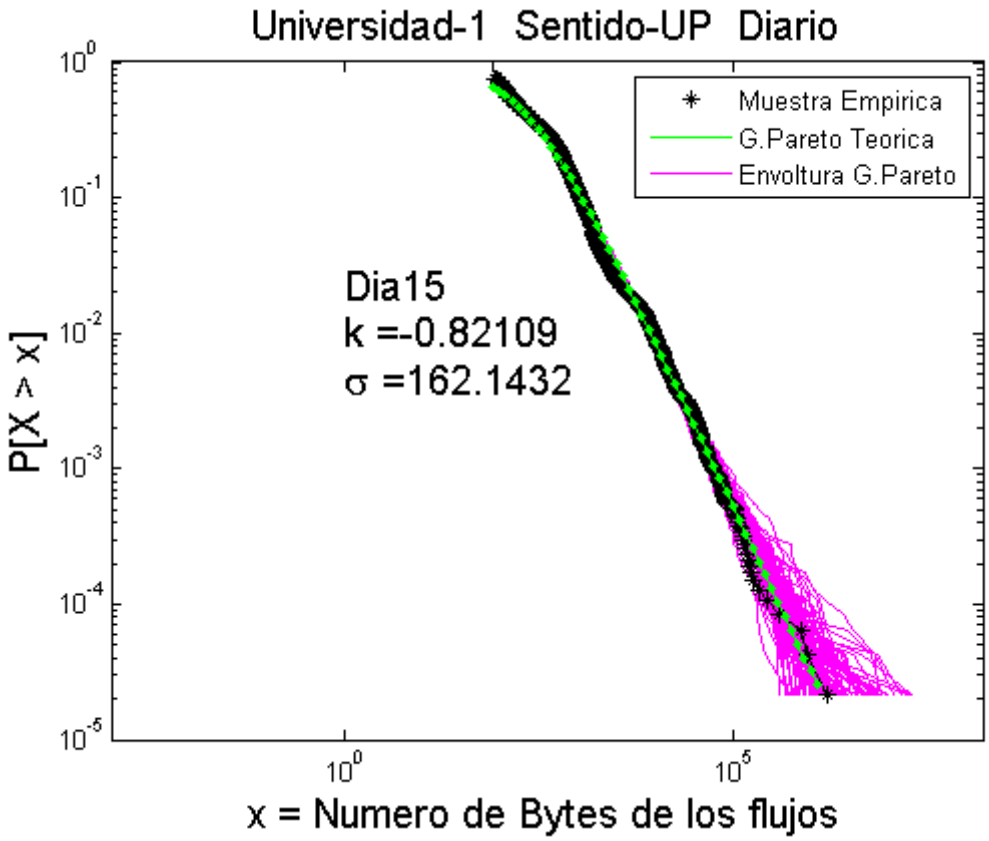

Figura 5-2: Ajuste visual CCDF de la GPD – UP

En la figura se observa el ajuste visual entre la CCDF empírica y la teórica generalizada de Pareto, con los parámetros mostrados con ambas escalas logarítmicas. Comentar que, a pesar de que en la figura no se

aprecia suficientemente, la curva empírica presenta pequeñas ondulaciones que oscilan alrededor de la curva teórica, algo similar a lo que llaman "wobbles" los autores de [9]. A su vez, se puede apreciar la llamada variabilidad de la región lejana de la cola, gracias a la envoltura muestral. La curva empírica permanece dentro de los límites marcados por ésta, con lo que se puede concluir que estos parámetros producen un ajuste satisfactorio con las muestras respecto a la CCDF. El valor de k indica que la distribución posee un índice de cola $\alpha = \frac{-1}{k} \approx 1.22$ perteneciente al rango necesario para producir el fenómeno de la autosimilaridad comentado en el apartado introductorio. El valor del parámetro σ es relativamente pequeño, reflejando la escasa curvatura que presentan esta muestra empírica, siendo próxima a una distribución pura de Pareto en prácticamente todo su rango.

En la siguiente figura se muestra el ajuste visual que se obtiene mediante la distribución log-Normal:

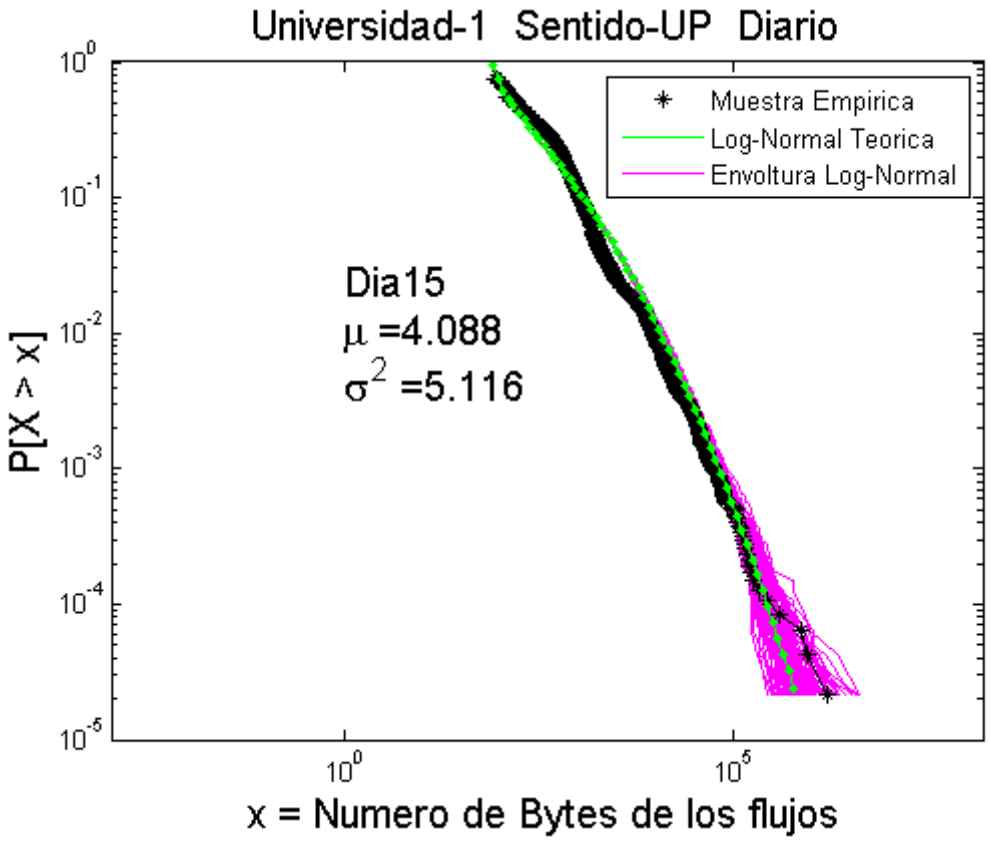

Figura 5-3: Ajuste visual CCDF de la LN - UP

Se puede observar que, debido a la curvatura que presenta este tipo de distribución, en la región media la curva teórica no se corresponde con la empírica y a pesar de que en la región lejana permanece dentro de la envoltura, también ahí el ajuste es menor que el obtenido con la generalizada de Pareto. Visualmente parece mejor la distribución GPD para esta muestra.

5.1.2 UMBRALIZACIÓN DE LA MUESTRA

En este apartado se va a mostrar los resultados gráficos del proceso de umbralización de las muestras y ver cómo afectan a las distribuciones teóricas correspondientes.

Primero se comenzará con la distribución GPD. Como ya se vio en el apartado 3.1.2, esta distribución se adapta de una forma muy sencilla a la umbralización de las muestras mediante la variación de su parámetro σ tal que, si $X \in GPD(k,\sigma,\mu) \Leftrightarrow (X-u|X>u) \in GPD(k,\sigma-k(u-\mu),0)$. Los umbrales u aplicados se han calculado a partir de percentiles de la propia muestra. Para todos los casos, se han utilizado los 10 percentiles siguientes:

$$\vec{u} = [x_{0.1}, x_{0.5}, x_{0.6}, x_{0.7}, x_{0.8}, x_{0.85}, x_{0.9}, x_{0.95}, x_{0.99}, x_{0.995}]$$

$x_F \quad 0 < F < 1$ es el percentil F.

Para el bloque de muestras mostrado antes, los resultados son los siguientes:

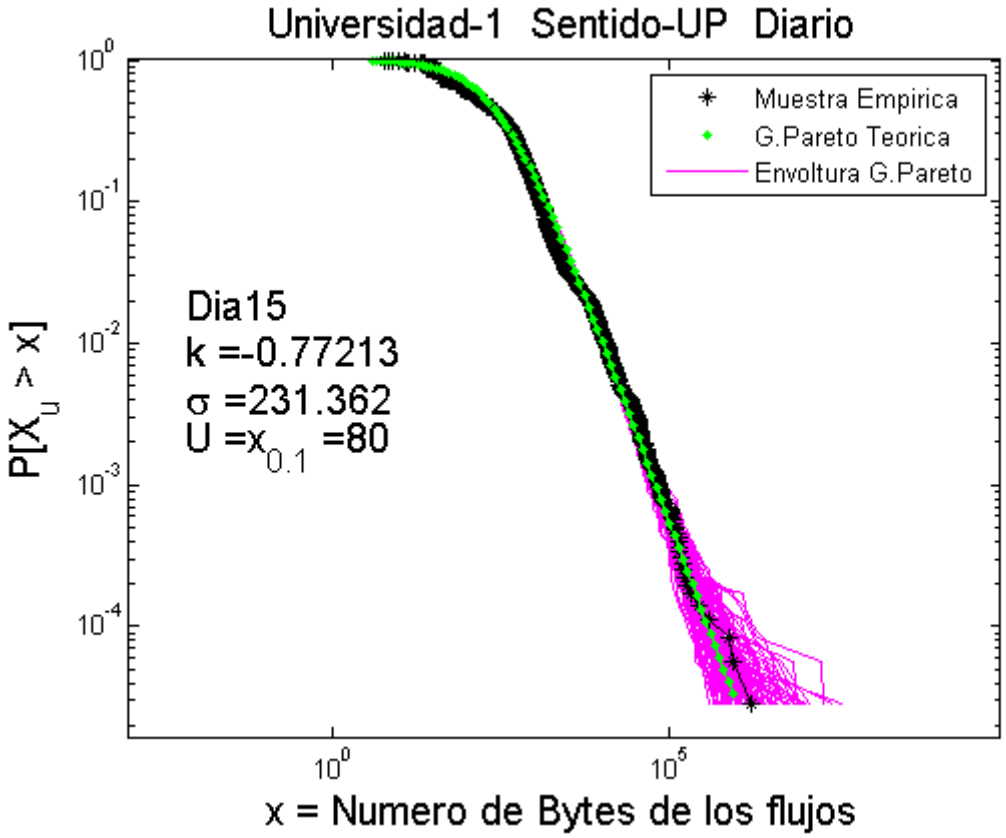

Figura 5-4: Ajuste visual CCDF GPD u1 - UP

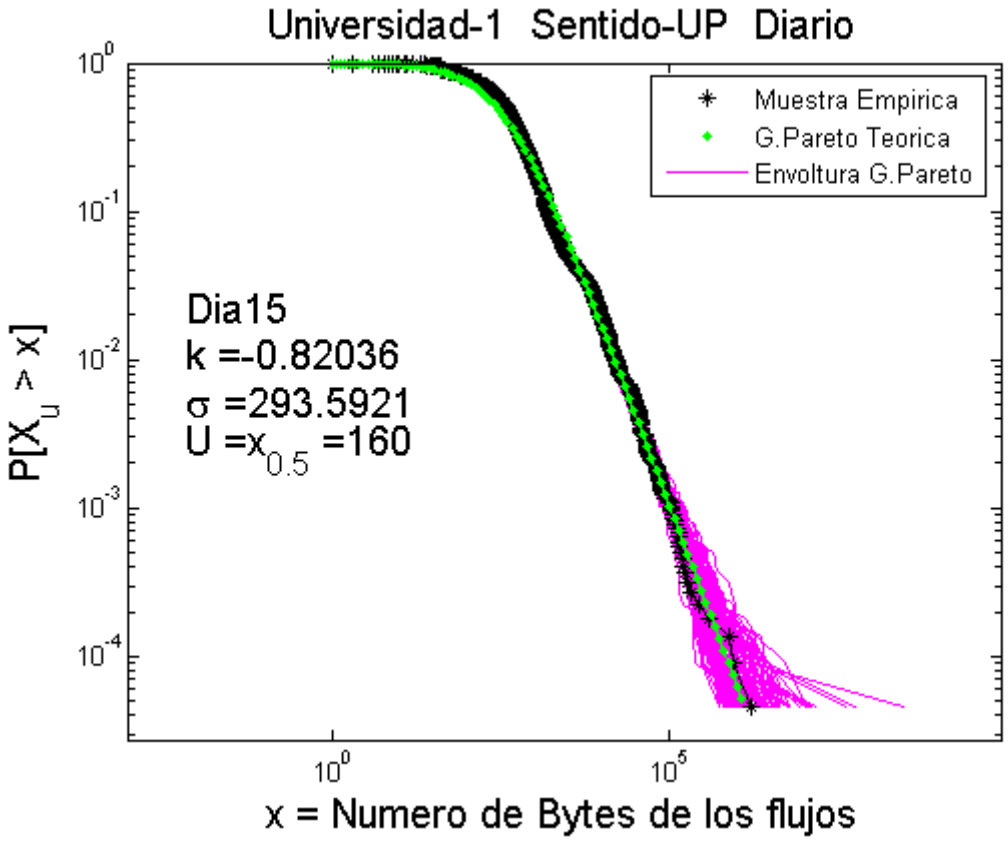

Figura 5-5: Ajuste visual CCDF GPD u2 - UP

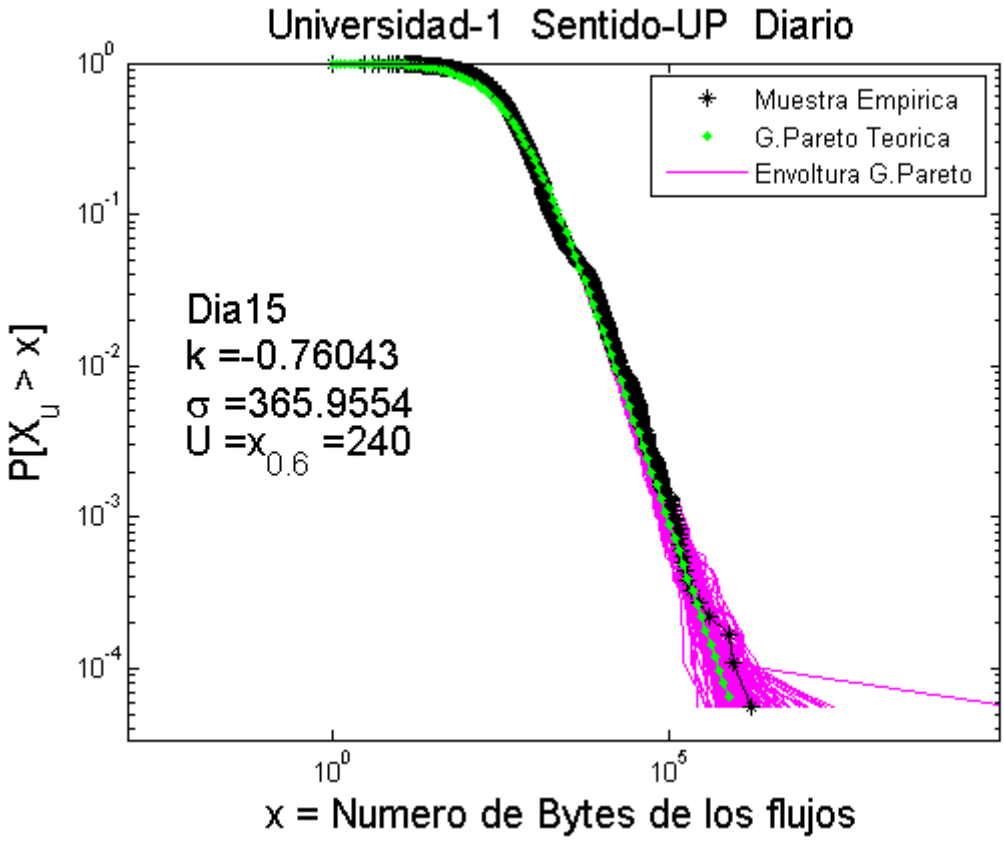

Figura 5-6: Ajuste visual CCDF GPD u3 - UP

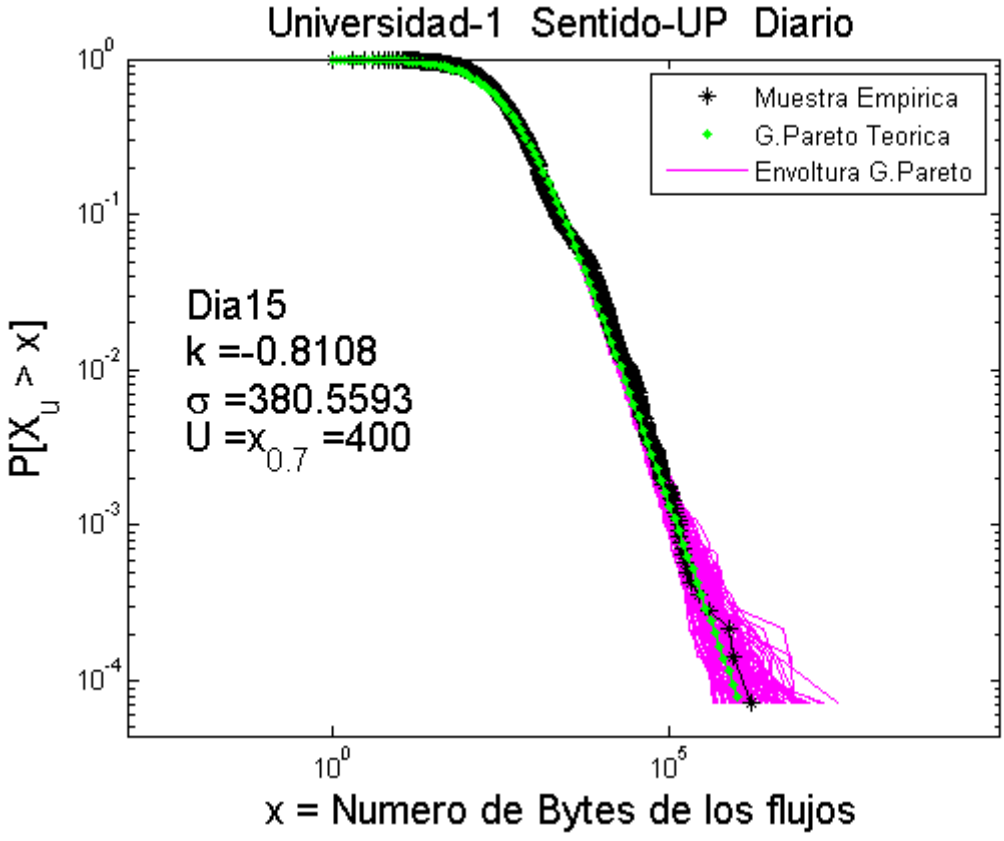

Figura 5-7: Ajuste visual CCDF GPD u4 - UP

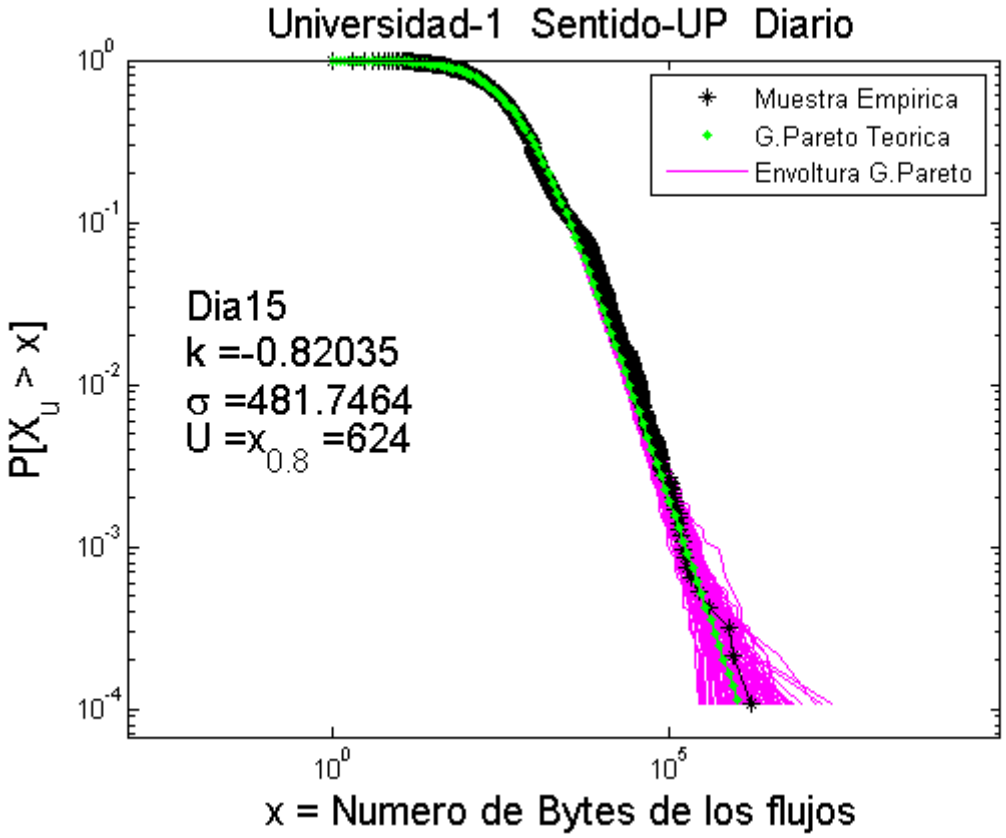

Figura 5-8: Ajuste visual CCDF GPD u5 - UP

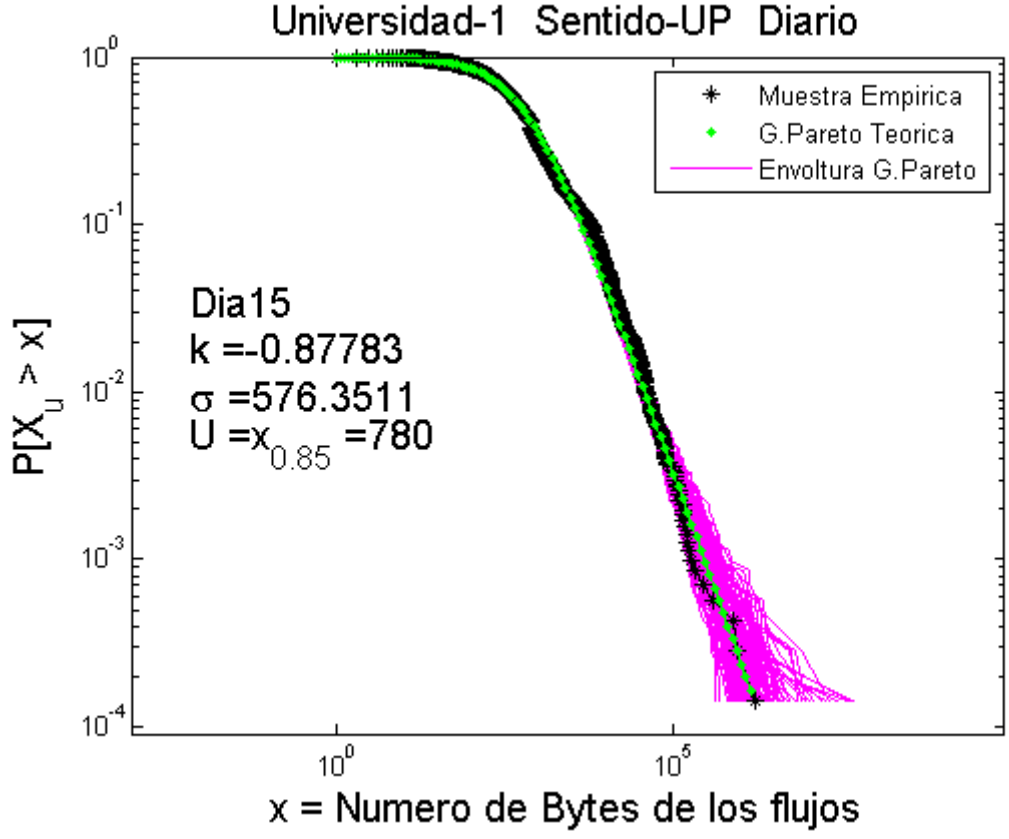

Figura 5-9: Ajuste visual CCDF GPD u6 - UP

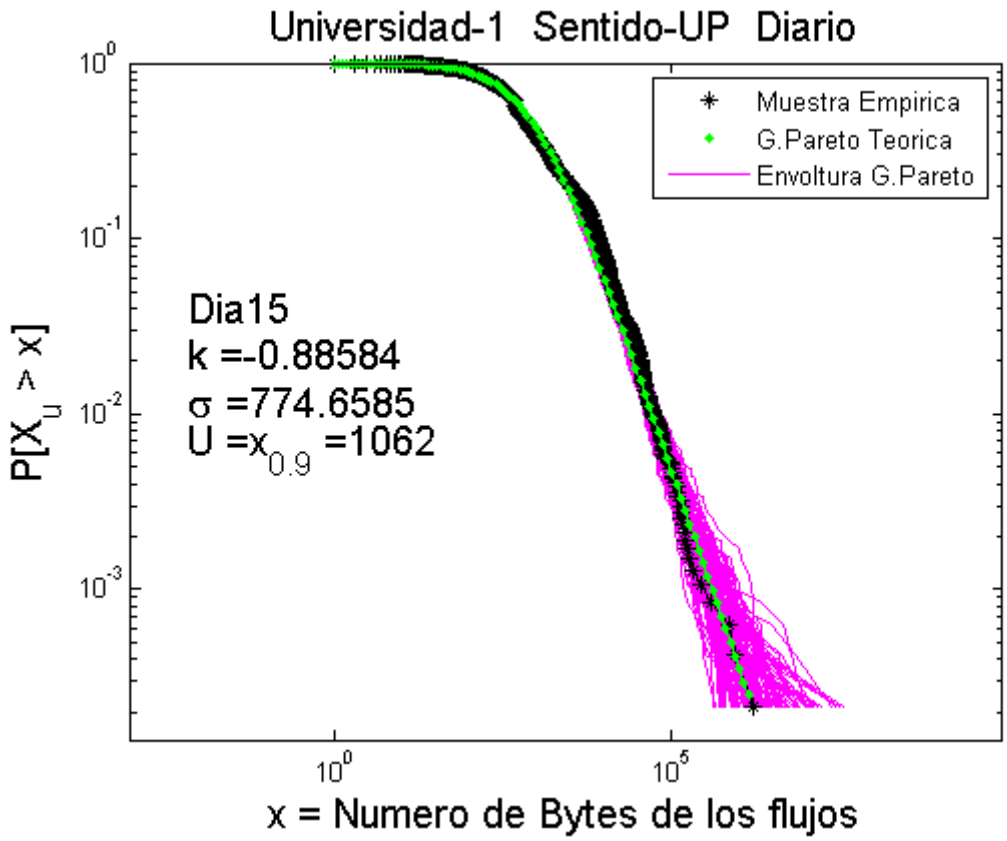

Figura 5-10: Ajuste visual CCDF GPD u7 - UP

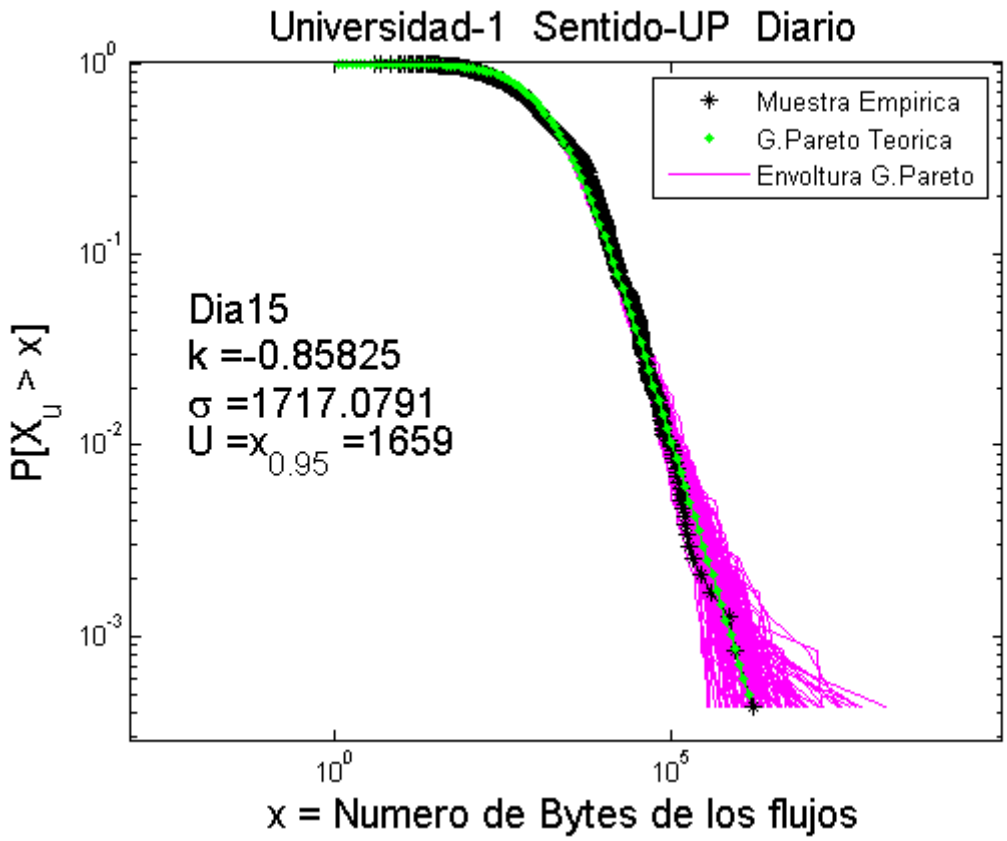

Figura 5-11: Ajuste visual CCDF GPD u8 - UP

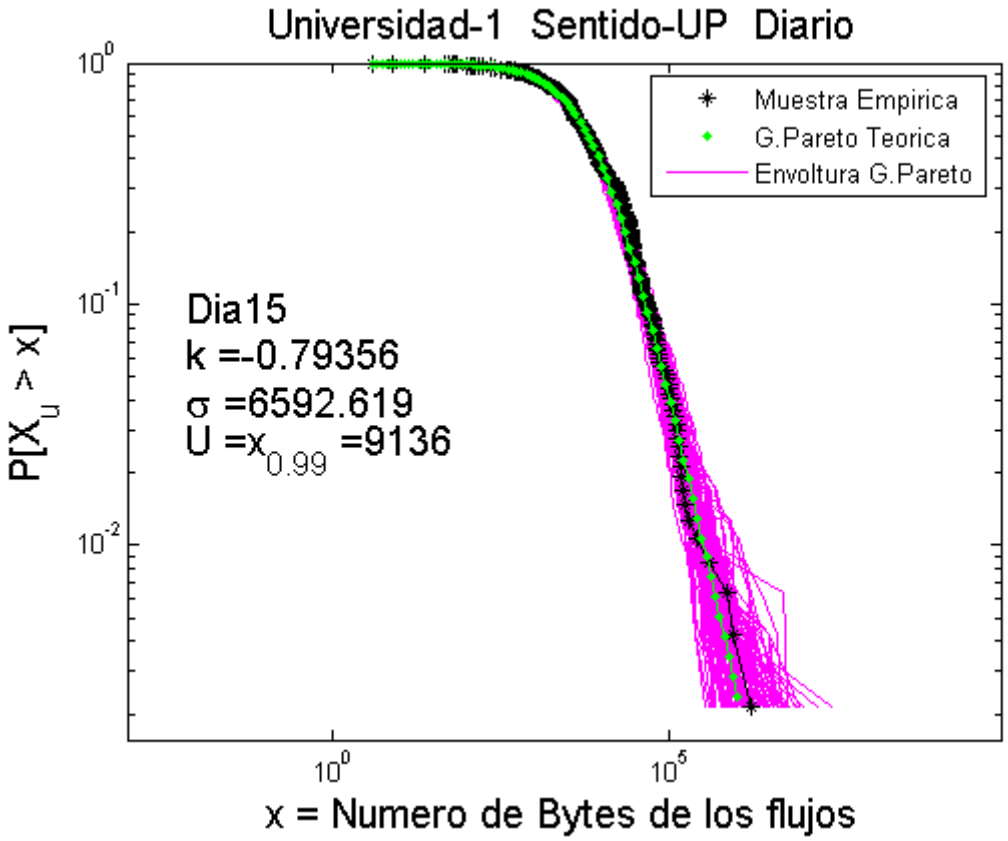

Figura 5-12: Ajuste visual CCDF GPD u9 - UP

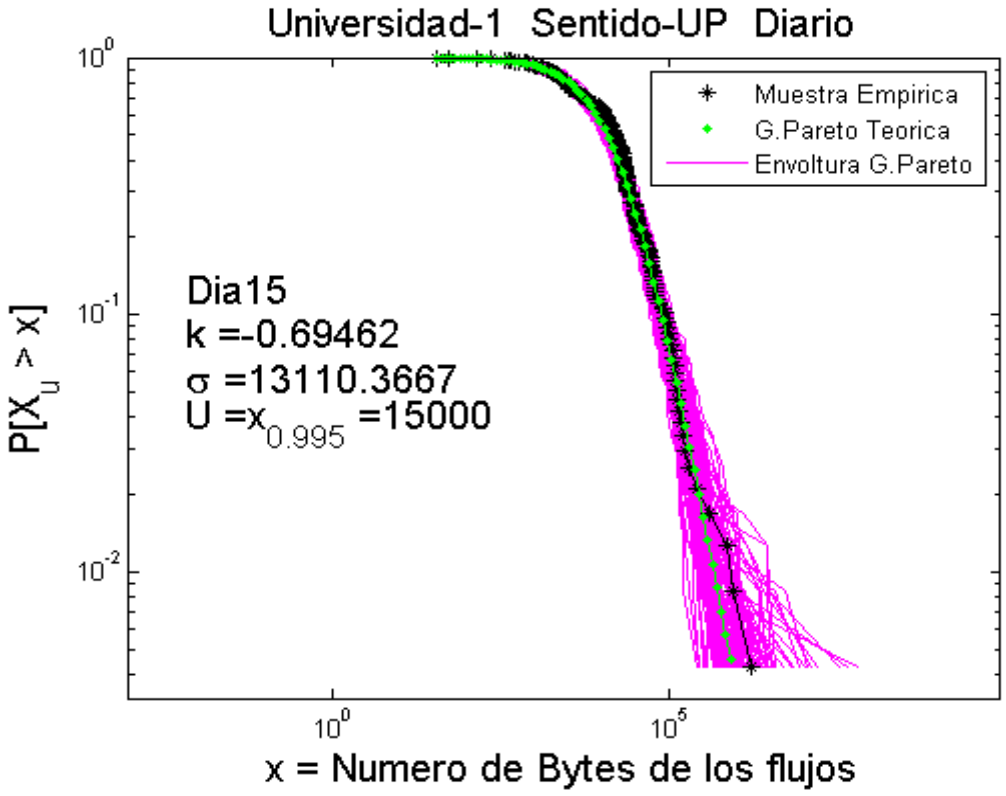

Figura 5-13: Ajuste visual CCDF GPD u10 - UP

En las figuras se denomina $X_u = \{X - u | X > u\}$. Son, al igual que en el apartado anterior representaciones log-log CCDF de X_u . Se puede observar que en general el ajuste visual es bastante bueno, ya que las muestras empíricas permanecen dentro de la envoltura. Para ver como los parámetros varían a media que aumenta el valor de los umbrales u aplicados. En las figuras siguientes se podrá observar su relación:

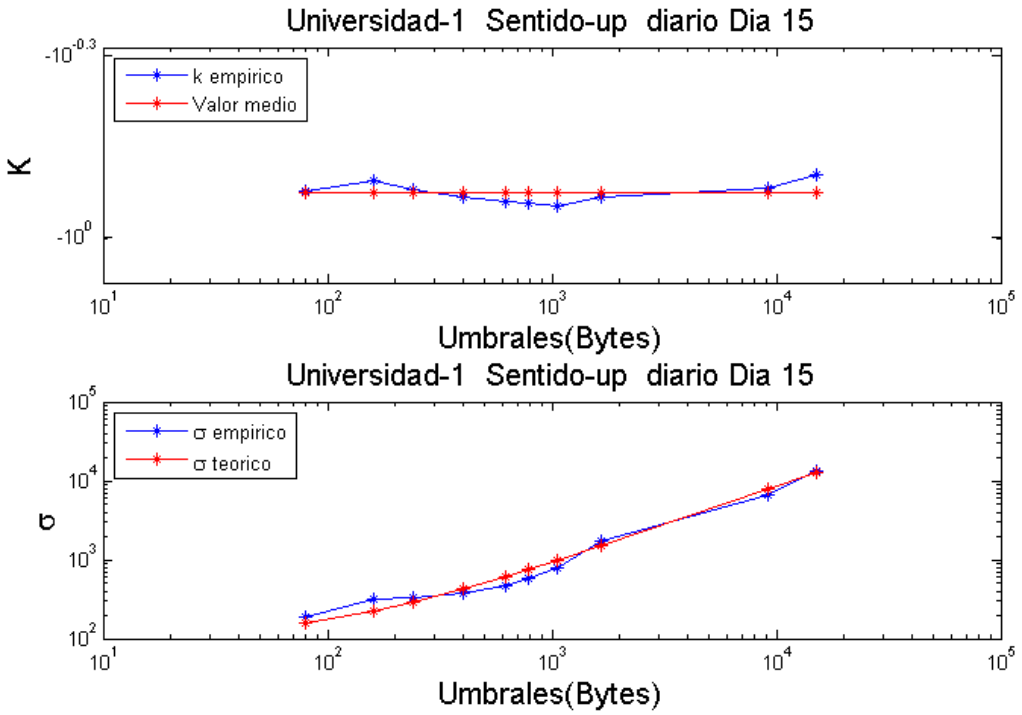

Figura 5-14: Parámetros k,sigma respecto a los umbrales u - UP

Se han representado en escalas logarítmicas para su observación con mayor claridad ya que, con ejes lineales, los umbrales u aplicados están demasiado próximos entre sí al inicio, en relación con los incrementos que se producen al final lo que provocaba que no se pudiera apreciar de forma clara sus valores. Se puede observar que los resultados empíricos son coherentes con los teóricos. En este caso, se obtienen mediante regresión lineal por mínimos cuadrados los siguientes resultados:

$$k(u) = \hat{a}u + \hat{b} \;\Rightarrow\; \hat{a} = 6.98\,10^{-7}\,, \hat{b} = -0.827$$
$$\sigma(u) = \hat{c}u + \hat{d} \Rightarrow\; \hat{c} = 0.83\,, \hat{d} = 47.12$$

Entonces se ve que $\hat{a} \approx 0 \Rightarrow \hat{b} \approx E[k(u)] \approx k(0) = k_o = -0.821$ que es lo mismo que decir que k permanece prácticamente constante frente a la umbralización. A su vez, se observa que $\hat{c} \approx -k_0$. Estos resultados son coherentes con las ecuaciones (3.19) y (3.20) explicadas en el apartado 3.1.2. También se comprueba cómo se comportan los parámetros $k(u)$ y $\sigma(u)$ obtenidos con respecto a la media de las muestras umbralizadas:

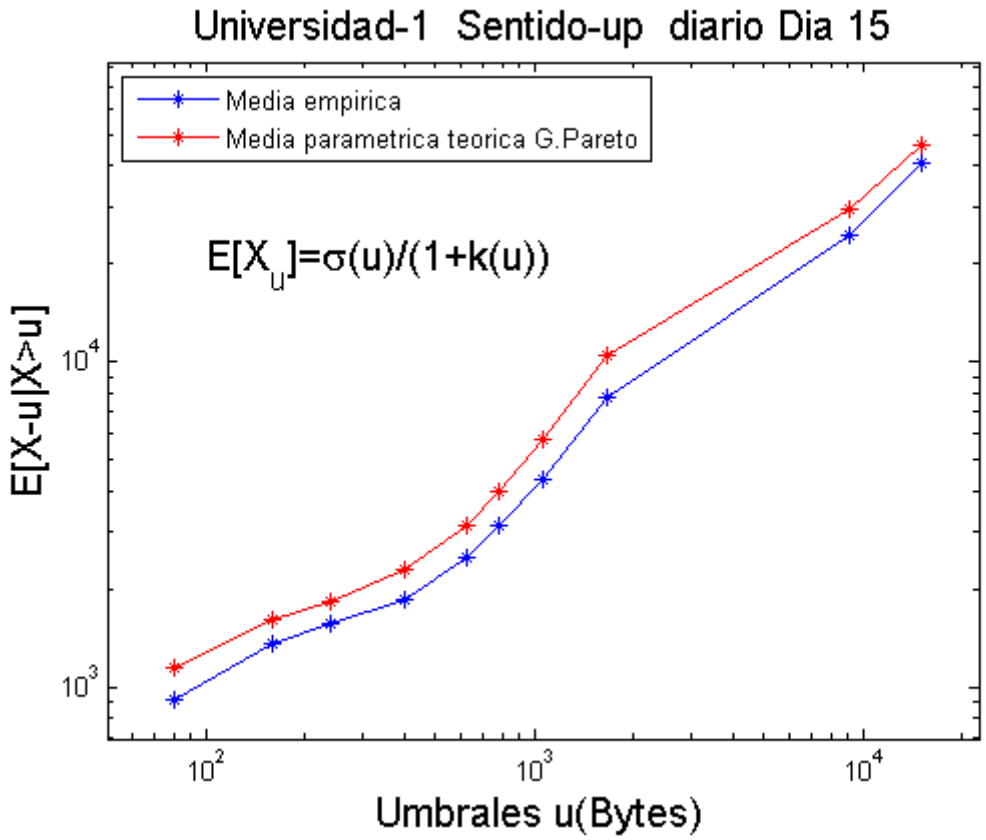

Figura 5-15: Medias umbralizadas paramétricas GPD y empíricas - UP

Se observa que se obtienen valores aproximados calculando la media a partir de los valores de $k(u)$ y $\sigma(u)$. Los empíricos son ligeramente menores, esto se debe a que el número de muestras es finito y no llegan a converger al mismo valor teórico de la distribución continua. Pero se observa que siguen un comportamiento muy similar, lo que indica que los

parámetros calculados reflejan correctamente el comportamiento de la muestra.

Los ajustes visuales obtenidos mediante la distribución log-Normal se muestran a continuación:

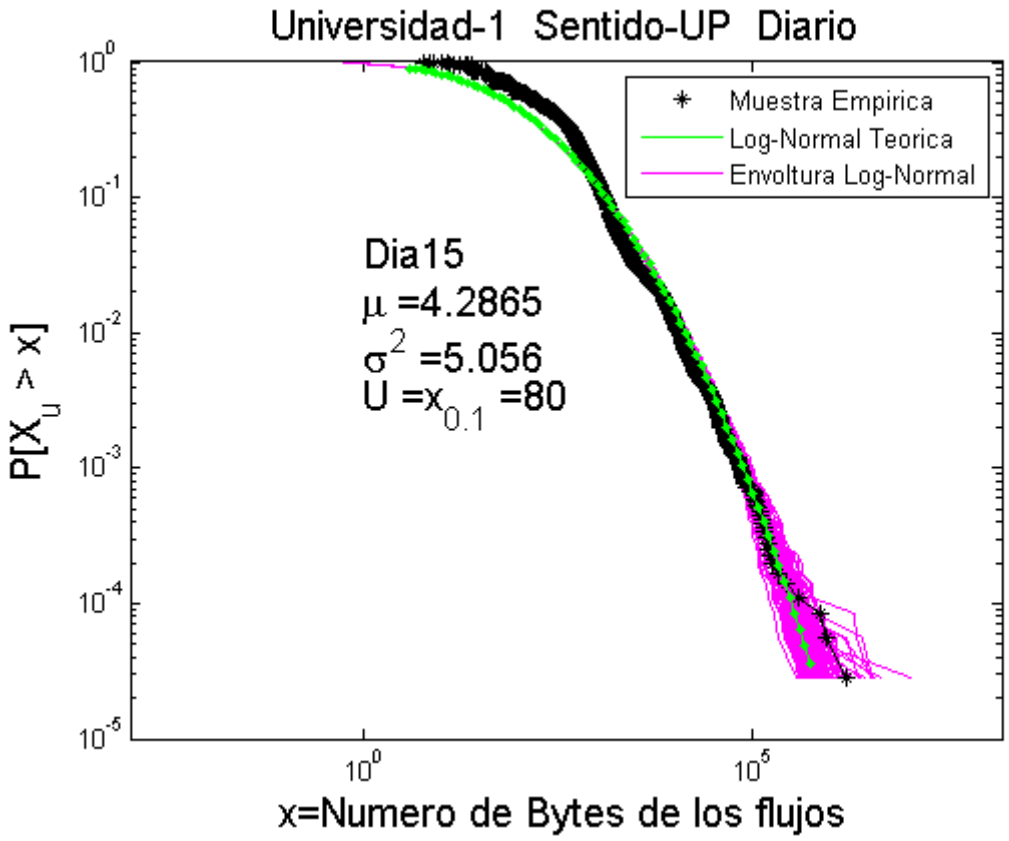

Figura 5-16: Ajuste visual CCDF LN u1 –UP

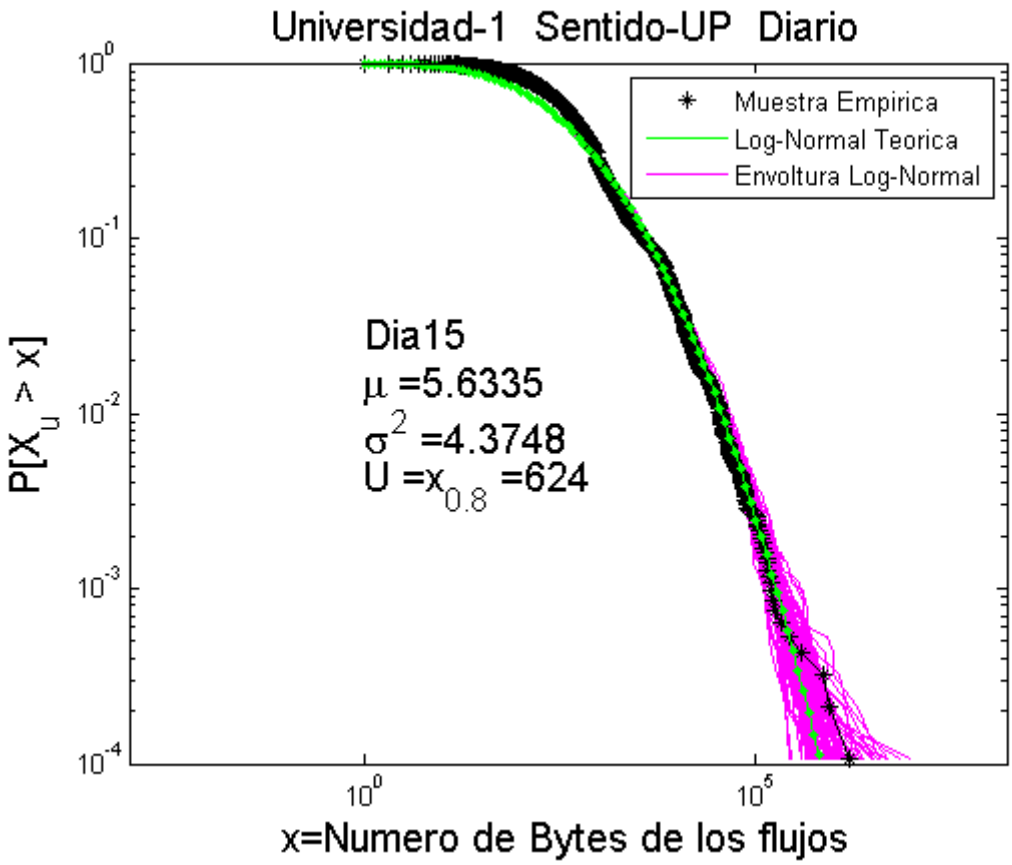

Figura 5-17: Ajuste visual CCDF LN u5 – UP

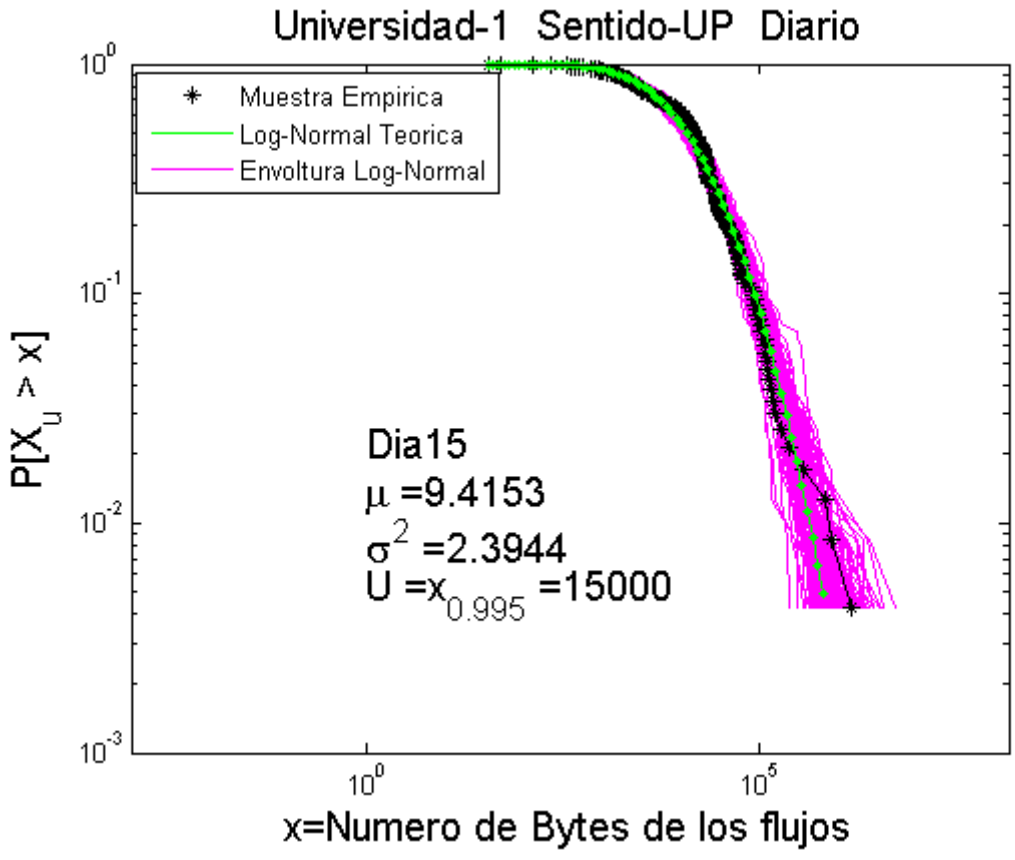

Figura 5-18: Ajuste visual CCDF LN u10 – UP

Se muestran únicamente estos casos porque son suficientes para observar que, visualmente, la distribución log-Normal se va ajustando cada vez mejor a medida que el valor del umbral aumenta. Visualmente parece que, a partir de cierto umbral, la distribución log-Normal podría también ser apropiada para caracterizar la muestra, pero lo que interesa es el bloque completo.

En este caso, no tiene sentido representar las medias de las muestras umbralizadas empíricas frente a las teóricas obtenidas mediante $\mu(u)\ y\ \sigma^2(u)$ debido a que, estos parámetros, se determinan concretamente para que sean iguales, al calcularlos mediante el método de los momentos. Como lo que interesa es considerar el conjunto total del bloque de datos, se realiza la representación de las medias umbralizadas y las teóricas que se obtienen con los parámetros determinados para $u = 0$, en la fase de ajuste a la escala temporal elegida:

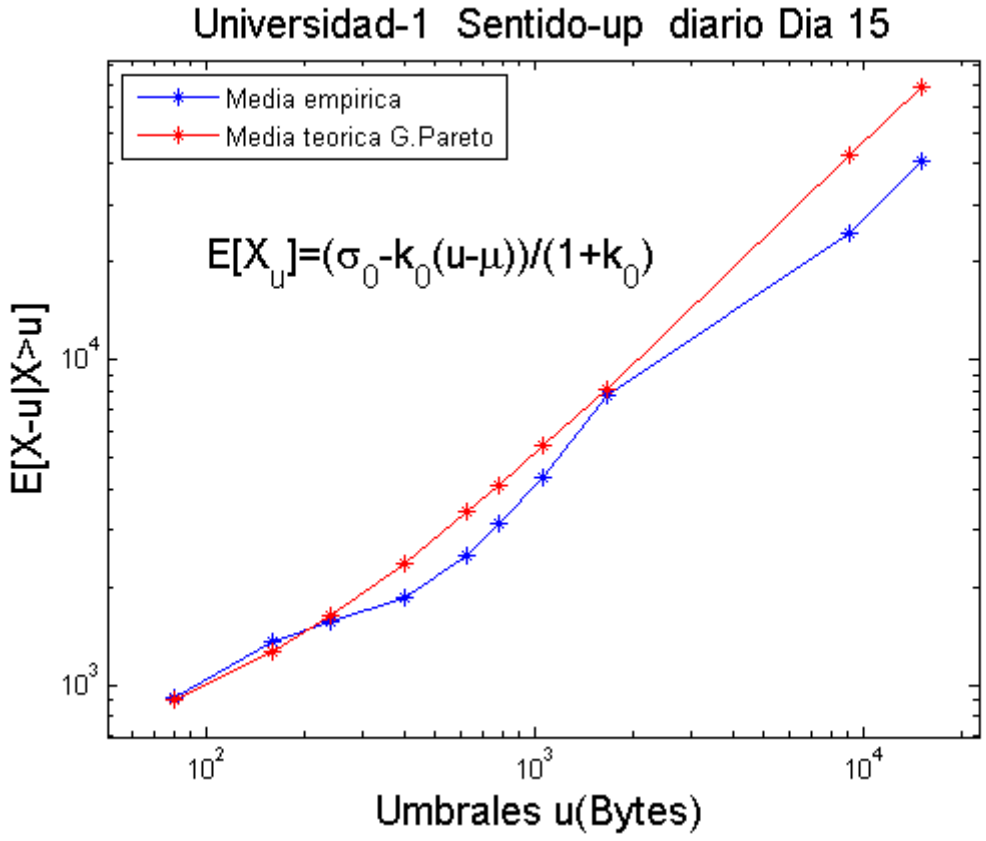

Figura 5-19: Medias umbralizadas teóricas GPD y empíricas - UP

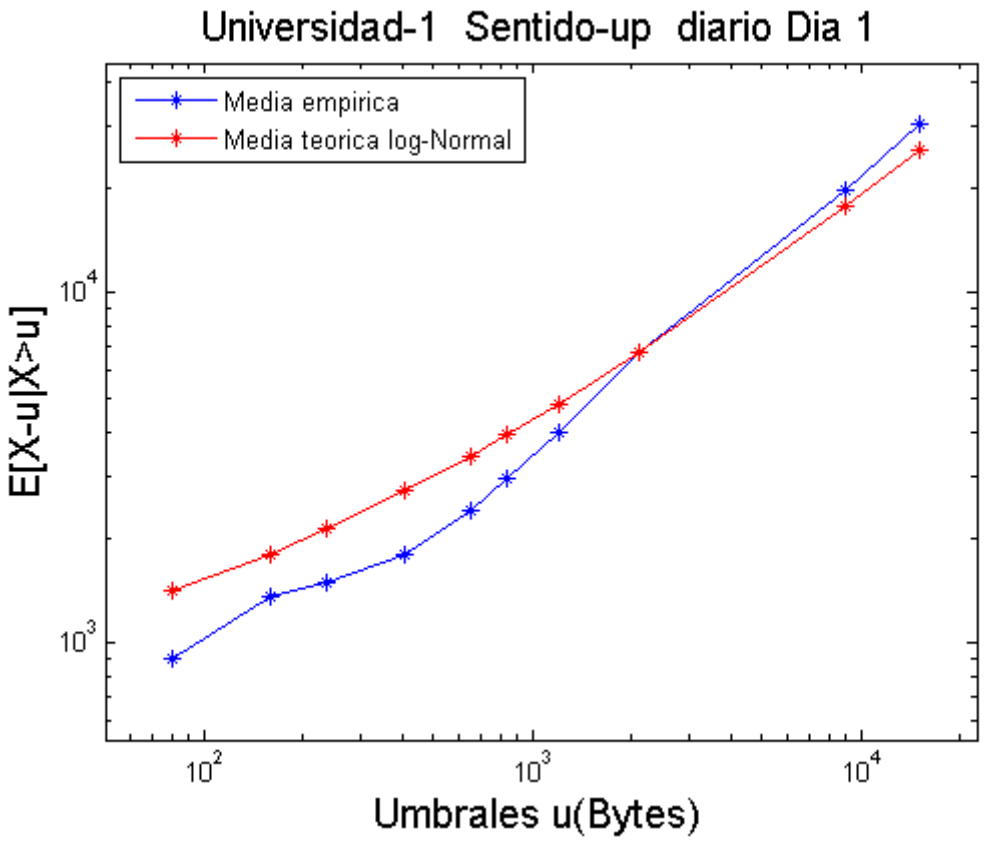

Figura 5-20: Medias umbralizadas teóricas LN y empíricas - UP

Se puede ver que el ajuste del bloque completo mediante la GPD se comporta mejor que el log-Normal también respecto al valor de las medias umbralizadas a pesar de que éste último no parece ser erróneo del todo.

5.1.3 CURVA DE LORENZ Y COEFICIENTE DE GINNI

La curva de Lorenz y el coeficiente de Ginni (apartados 3.1.3 y 3.2.3) sirven para analizar la distribución relativa del tamaño de los flujos respecto al total del tráfico que transportan. Aquí también se aprovechará para visualizar y cuantificar las diferencias, no sólo de la GPD y la log-Normal respecto de la muestra empírica, sino también la distribución pura de Pareto respecto a la GPD ya que en su caso depende únicamente del índice de cola $\alpha = \frac{-1}{k}$. Se podrá observar la mejora que se obtiene por la utilización de este modelo respecto al que utilizan todos los autores comentados en el capítulo 2.

En la siguiente figura se muestra la comparación entre la GPD y la log-Normal:

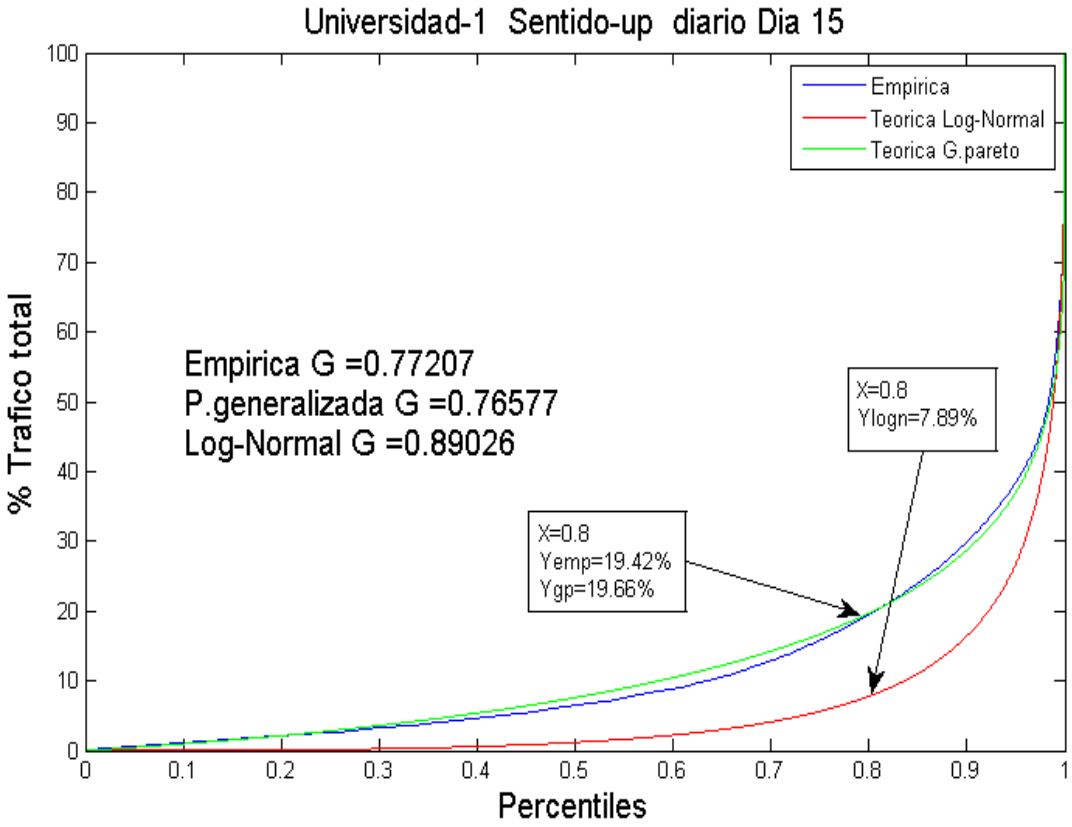

Figura 5-21: Curva de Lorenz GPD vs LN – UP

Como se puede ver la curva de la GPD se aproxima sensiblemente más que la de log-Normal. El coeficiente de Ginni, G, cuantifica esta diferencia entre ambas, observando que la generalizada está más próxima que la log-Normal. Se han indicado, los valores de las curvas para un valor de F=0.8. Se ve que los valores obtenidos tanto con la empírica como con la GPD son cercanos al 20%. Como curiosidad, esta distribución parece cumplir el llamado principio de Pareto (o regla 80-20) debida al economista del mismo nombre con la que describió la distribución de la riqueza en Italia mediante la distribución pura de Pareto. Entonces, se ve que en este sentido, también es mejor la aproximación a la muestra que se consigue con la GPD que con la log-Normal.

Gracias a esta curva, se puede observar la diferencia entre aproximar esté bloque de datos completo mediante una distribución pura de Pareto (recta log-log) o mediante la GPD:

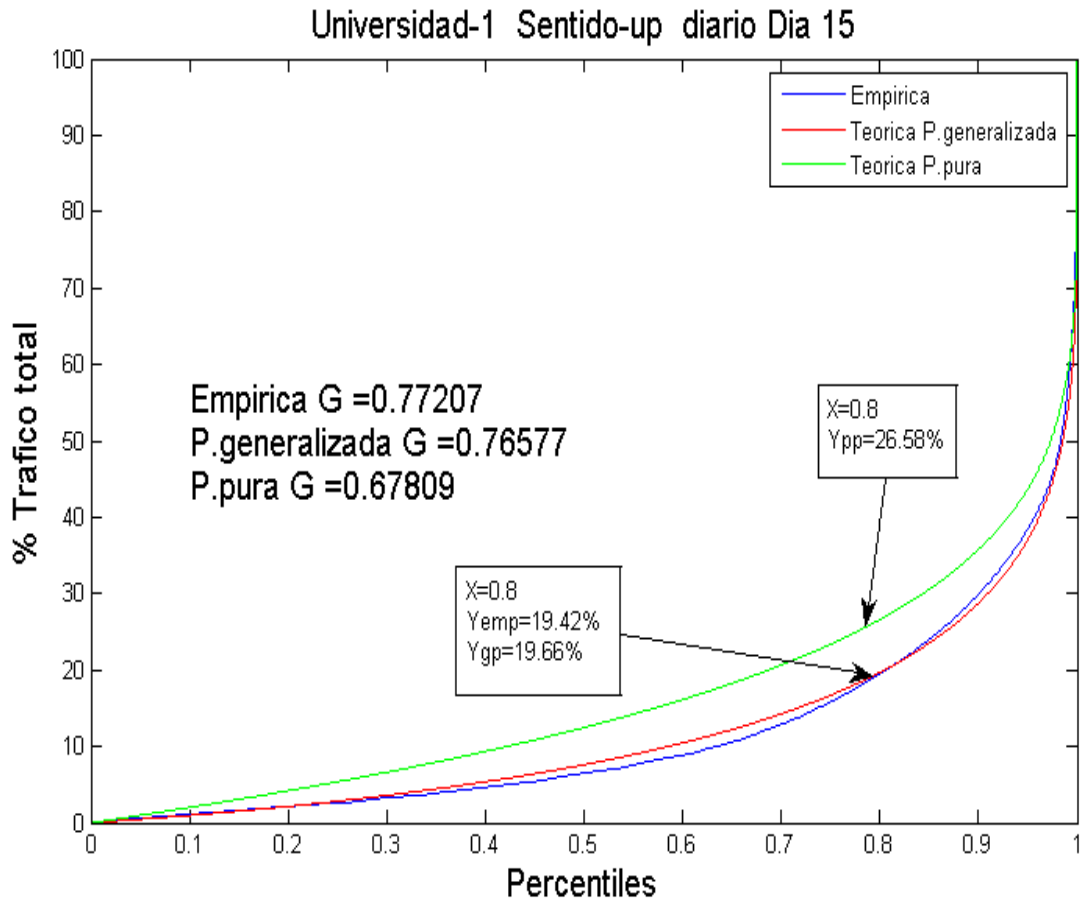

Figura 5-22: Curva de Lorenz GPD vs PPD – UP

Se observa que la generalizada es más fiel a la muestra que la pura. Esta diferencia es la que existe si se modela toda la muestra con una distribución pura, determinando x_m como el mínimo valor. Esto, como se comenta en el capítulo 2, no se hace así realmente. La distribución pura modela la muestra a partir de un umbral x_m a partir del cual parece que la muestra parece seguir un comportamiento log-log lineal. Si esto se realizará, las diferencias no serían tan grandes, pero la determinación de ese parámetro son completamente heurísticas. Con la GPD esto no es

necesario gracias a que es más versátil y puede modelar y caracterizar el rango de valores de forma completa.

Después de este análisis mostrado en los últimos apartados, se concluye que este bloque de datos que proviene de la universidad U1, en sentido ascendente (Up), día 15 del subconjunto Diario, sigue una distribución $GPD(k, \sigma, \mu)$ tal que:

$$k = -0.82 \; ; \; \sigma = 162 \; ; \; \mu = 80$$

El resto de bloques del subconjunto del sentido ascendente han sido analizados de forma idéntica. Los resultados obtenidos con el resto será, mostrados de forma conjunta en apartados posteriores aprovechando el análisis de diversidad temporal y espacial.

5.2 TRÁFICO DESCENDENTE

Al igual que en el sentido ascendente, se presentan de forma gráfica, unas muestras representativas de los resultados obtenidos en este caso.

5.2.1 AJUSTE VISUAL DE LA LOG-LOG CCDF

A continuación se muestra el ajuste obtenido con la distribución generalizada de Pareto mediante el algoritmo EPM (apartado 4.1.1) y el obtenido con la log-Normal mediante el método de los momentos (apartado 4.1.2) para un bloque de muestras determinado. Se comenzará con la distribución de Pareto:

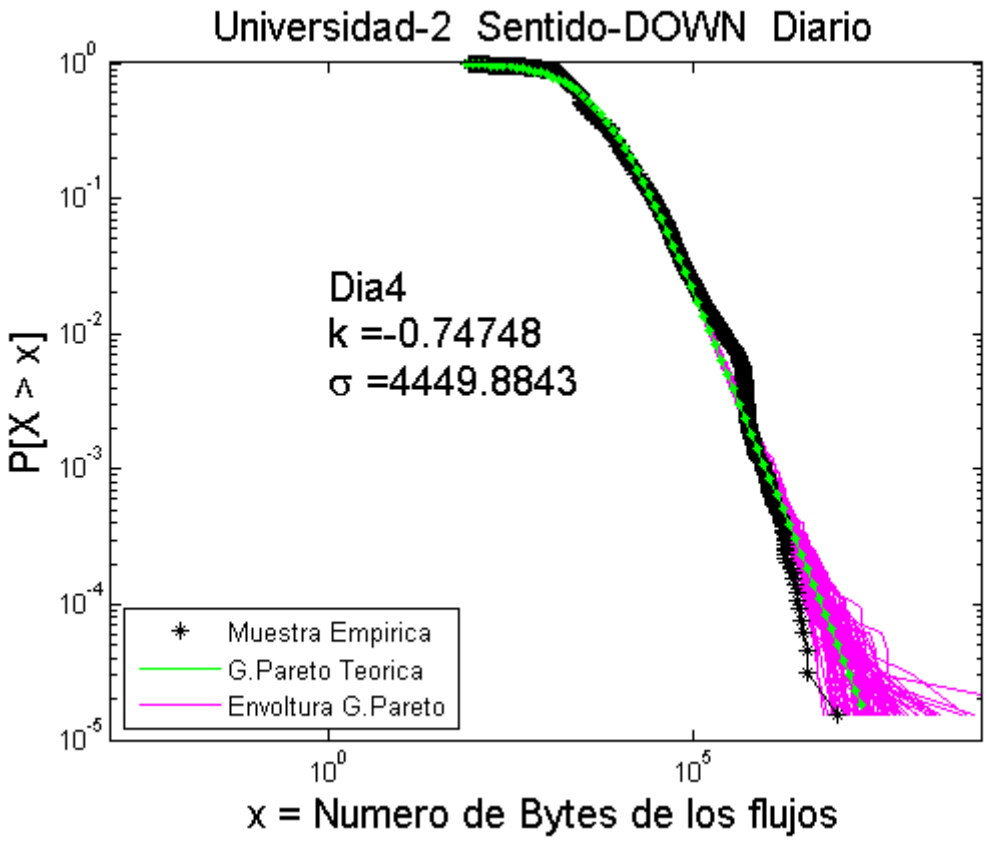

Figura 5-23: Ajuste visual CCDF de la GPD – DOWN

En la figura se observa el ajuste visual entre la CCDF empírica y la teórica generalizada de Pareto, con los parámetros mostrados con ambas escalas logarítmicas. En este caso se puede observar que la aproximación obtenida es sensiblemente menor que la mostrada para el caso ascendente. Se aprecia un incremento de la pendiente a medida que se avanza a lo largo del eje x, lo cual es contrario a la naturaleza de la distribución GPD. las medidas empíricas no se mantienen dentro de la envoltura al ir decrementándose cada vez de forma más pronunciada. Este efecto que se produce es similar al observado por el autor de [12] y [13], comentado en el capítulo 2 donde la distribución elegida para el análisis es la log-Normal.

Los resultados obtenidos mediante esta distribución se muestran en la siguiente figura:

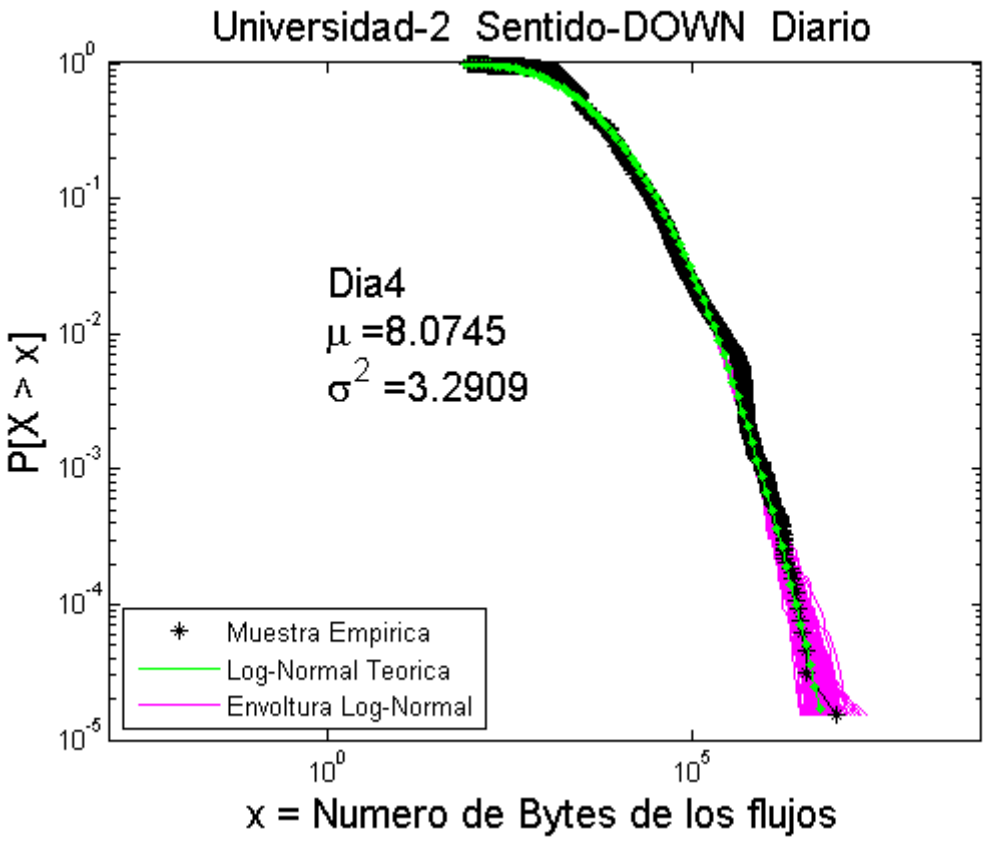

Figura 5-24: Ajuste visual CCDF de la LN - DOWN

Al contrario que en el sentido ascendente, aquí la distribución log-Normal proporciona un ajuste visual mayor que la GPD. Debido a ciertas ondulaciones que se producen en la zona media de la cola, el ajuste no es perfecto pero parece aceptable desde este punto de vista.

5.2.2 UMBRALIZACIÓN DE LA MUESTRA

Al igual que en el caso del sentido del tráfico ascendente, se van a mostrar los resultados gráficos del proceso de umbralización de las muestras y ver cómo afecta a las distribuciones teóricas correspondientes. En este caso se focalizará más en la distribución log-Normal debido a que sus resultados son más satisfactorios. Para este bloque de muestras, aplicando los umbrales de igual forma a los comentados anteriormente, los resultados son los siguientes:

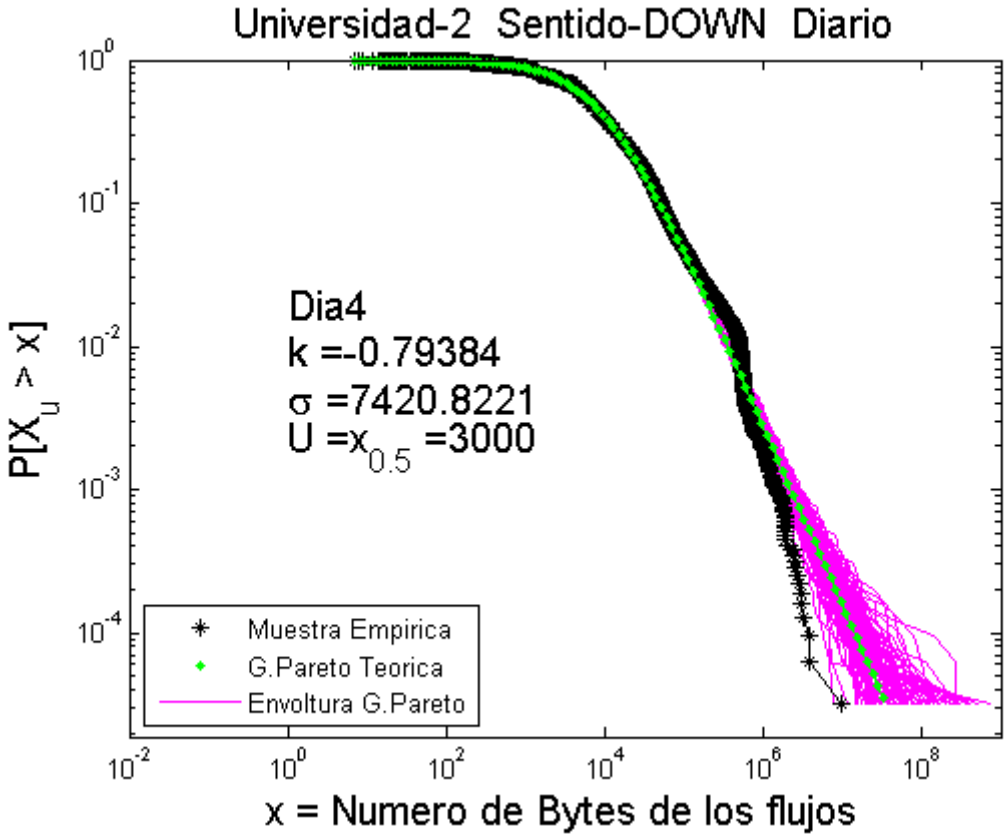

Figura 5-25: Ajuste visual CCDF GPD u2 - DOWN

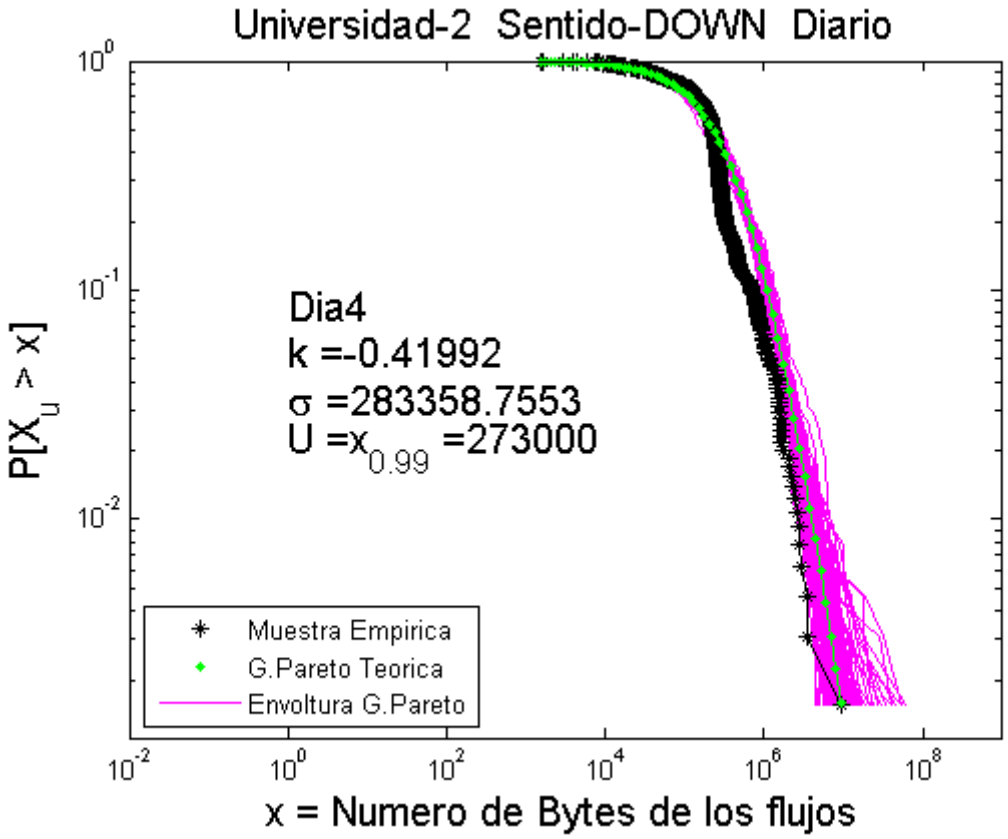

Figura 5-26: Ajuste visual CCDF GPD u9 - DOWN

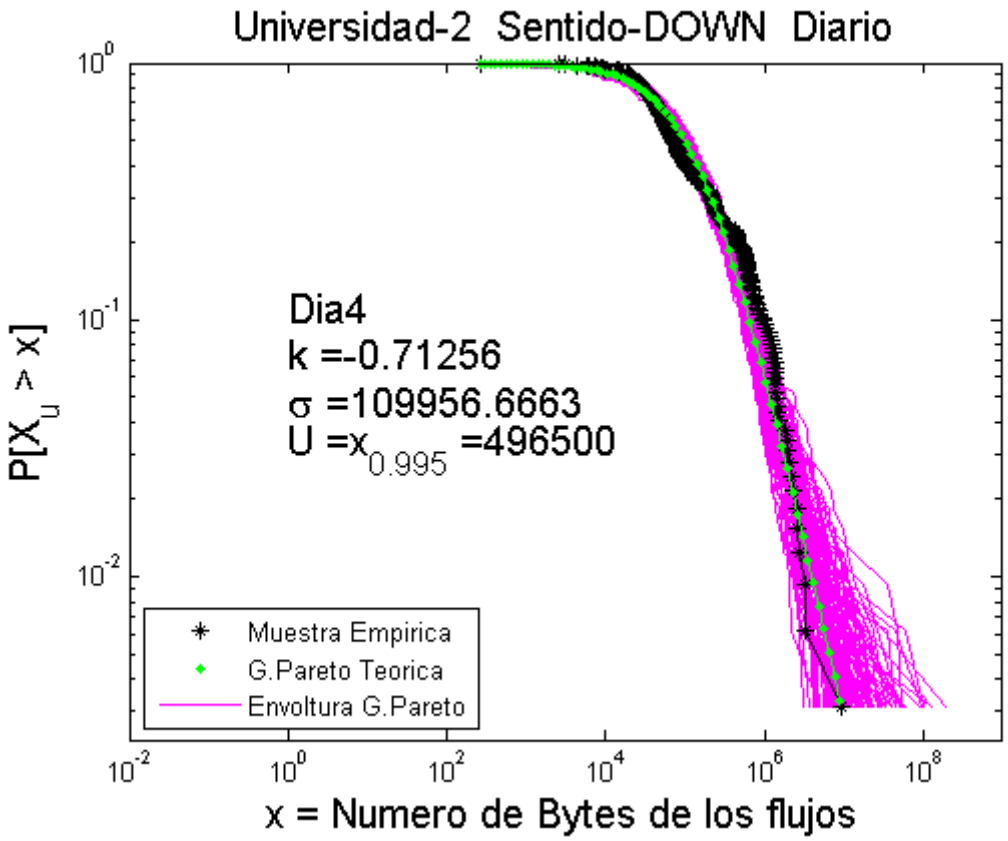

Figura 5-27: Ajuste visual CCDF GPD u10 - DOWN

En las figuras, al igual que en el apartado anterior se muestran las representaciones log-log CCDF de $X_u = \{X - u | X > u\}$. Se puede observar que el ajuste visual es bastante malo. A pesar de esto también se comprueba el comportamiento de los parámetros respecto a los umbrales para obtener mayor información de los resultados. En las figuras siguientes se podrá observar su relación:

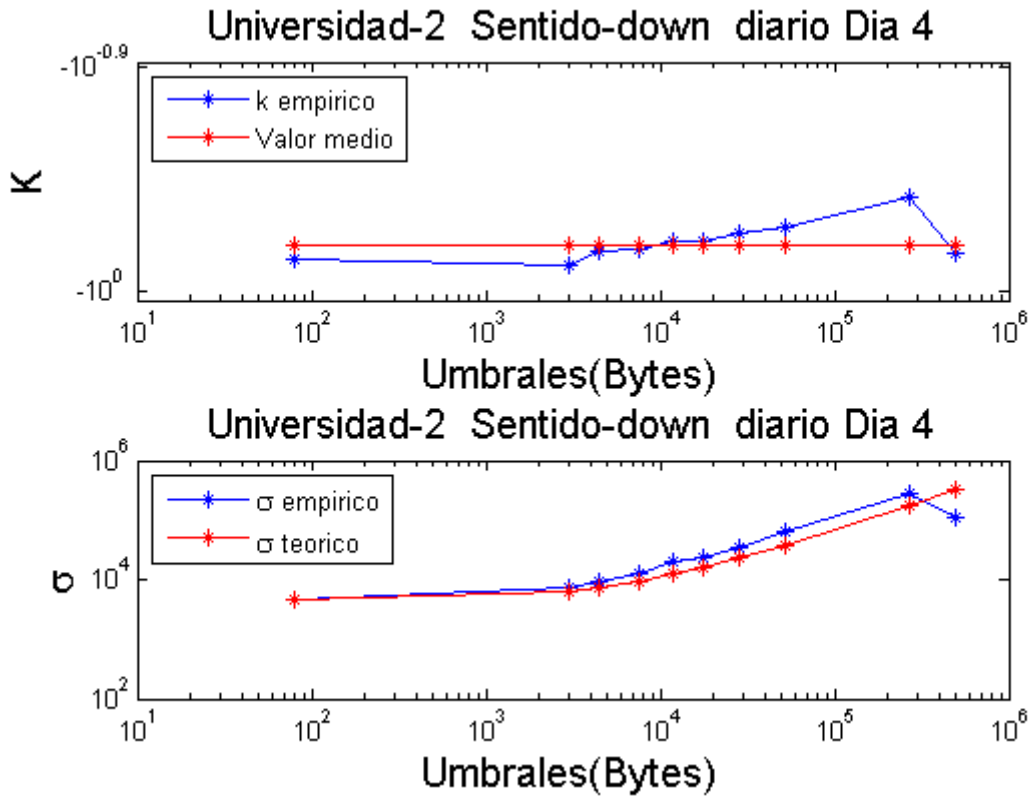

Figura 5-28: Parámetros k,sigma respecto a los umbrales u – DOWN

En este caso, desde este punto de vista podría parecer que el comportamiento podría ser el de una GPD aunque los ajustes visuales se vio que no eran del todo satisfactorios. Al igual que en el caso del sentido ascendente, mediante regresión lineal por mínimos cuadrados se cuantifica en cierta manera este comportamiento:

$$k(u) = \hat{a}u + \hat{b} \;\Rightarrow\; \hat{a} = 1.82\,10^{-6}, \hat{b} = -0.673$$
$$\sigma(u) = \hat{c}u + \hat{d} \Rightarrow\; \hat{c} = 0.359, \hat{d} = 2.27\,10^{4}$$

Entonces vemos que $\hat{a} \approx 0 \;\Rightarrow\; \hat{b} \approx E[k(u)] \approx k(0) \neq k_o = -0.74$. A pesar de presentar muy pequeñas variaciones, el valor obtenido dista sensiblemente del teórico. A su vez, se observa que $\hat{c} \neq -k_0$, lo que no es coherente con las ecuaciones (3.19) y (3.20) del apartado 3.1.2.

A continuación se muestran los ajustes visuales obtenidos mediante la distribución log-Normal:

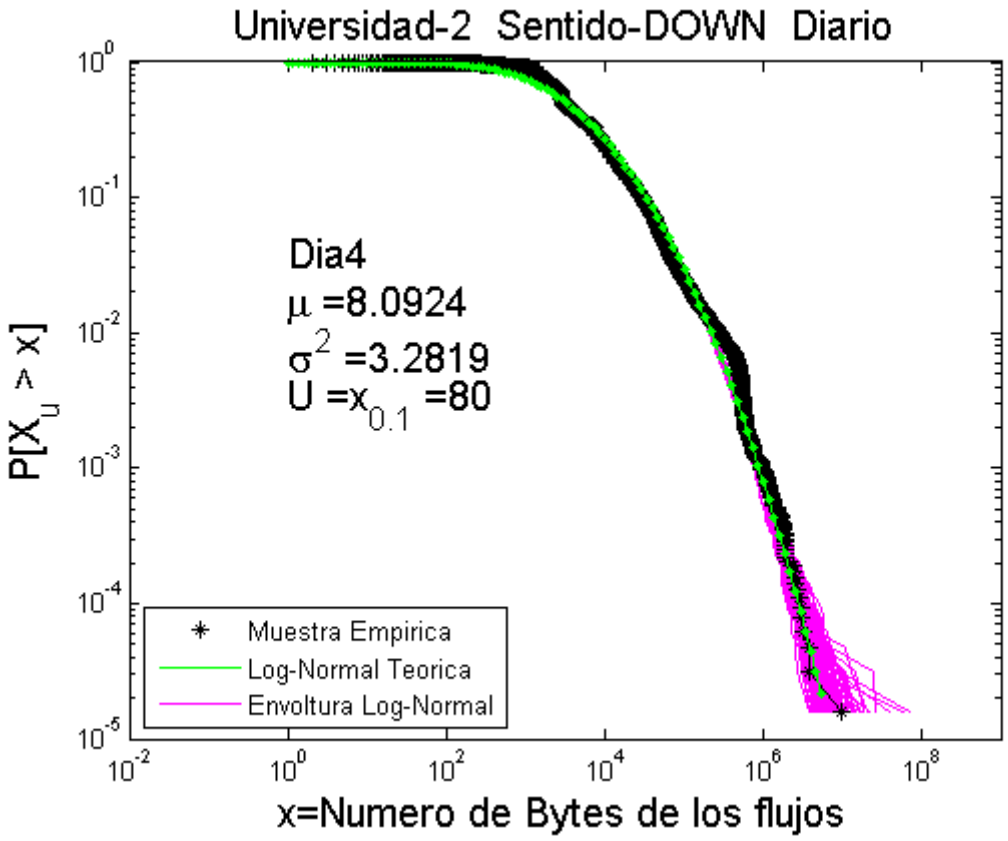

Figura 5-29: Ajuste visual CCDF LN u1 - DOWN

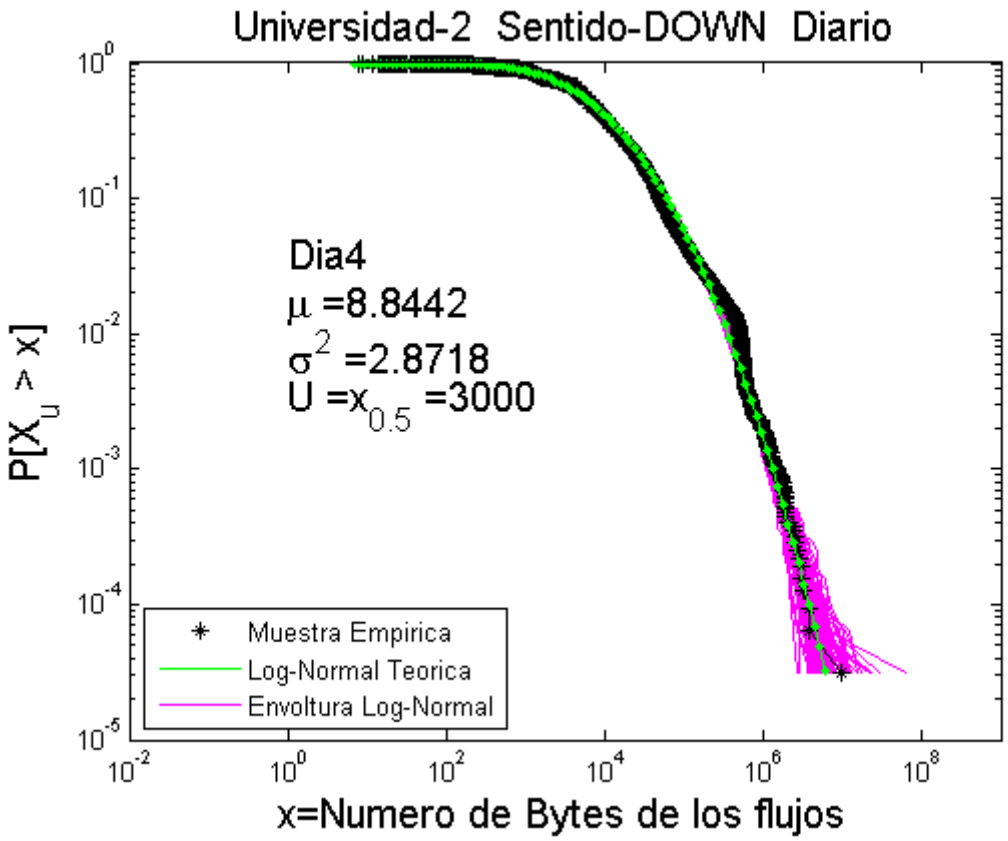

Figura 5-30Ajuste visual CCDF LN u2 - DOWN

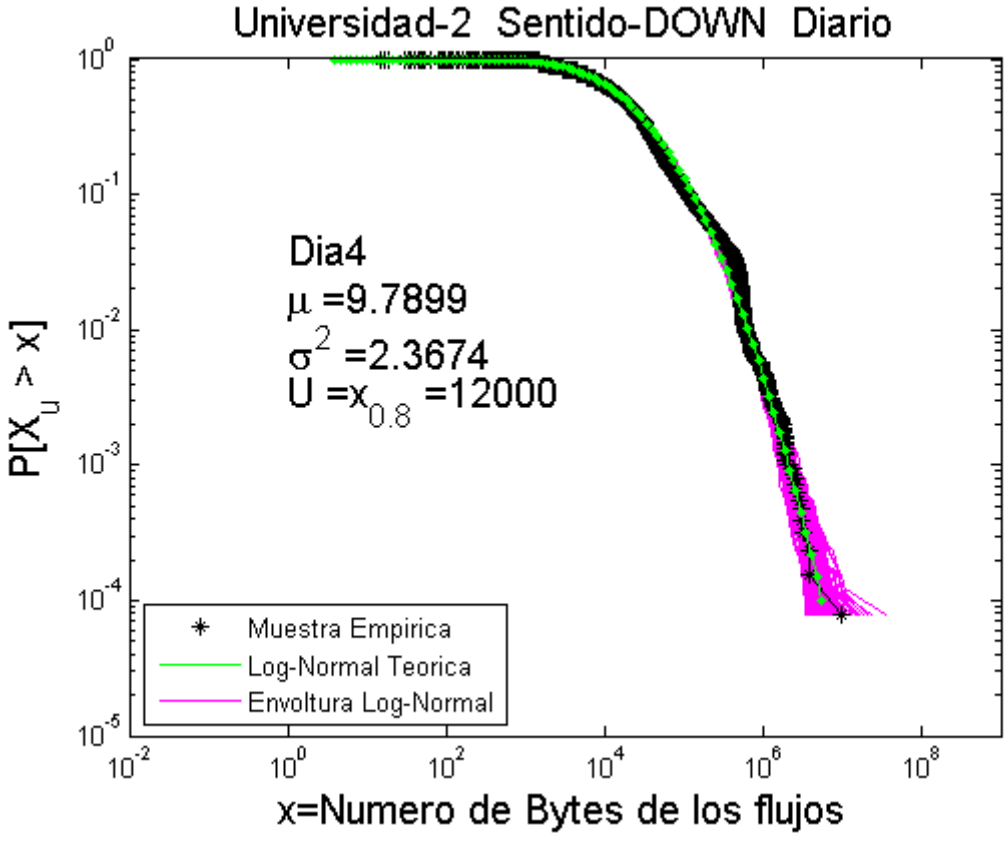

Figura 5-31: Ajuste visual CCDF LN u5 - DOWN

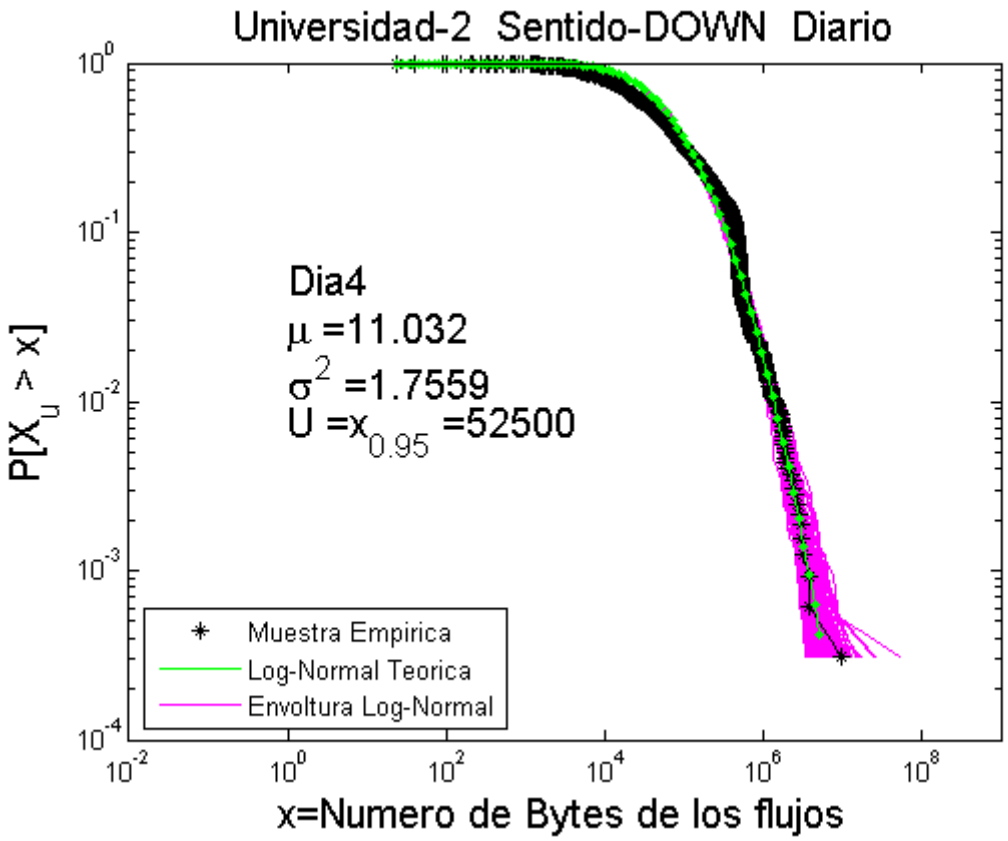

Figura 5-32: Ajuste visual CCDF LN u9 - DOWN

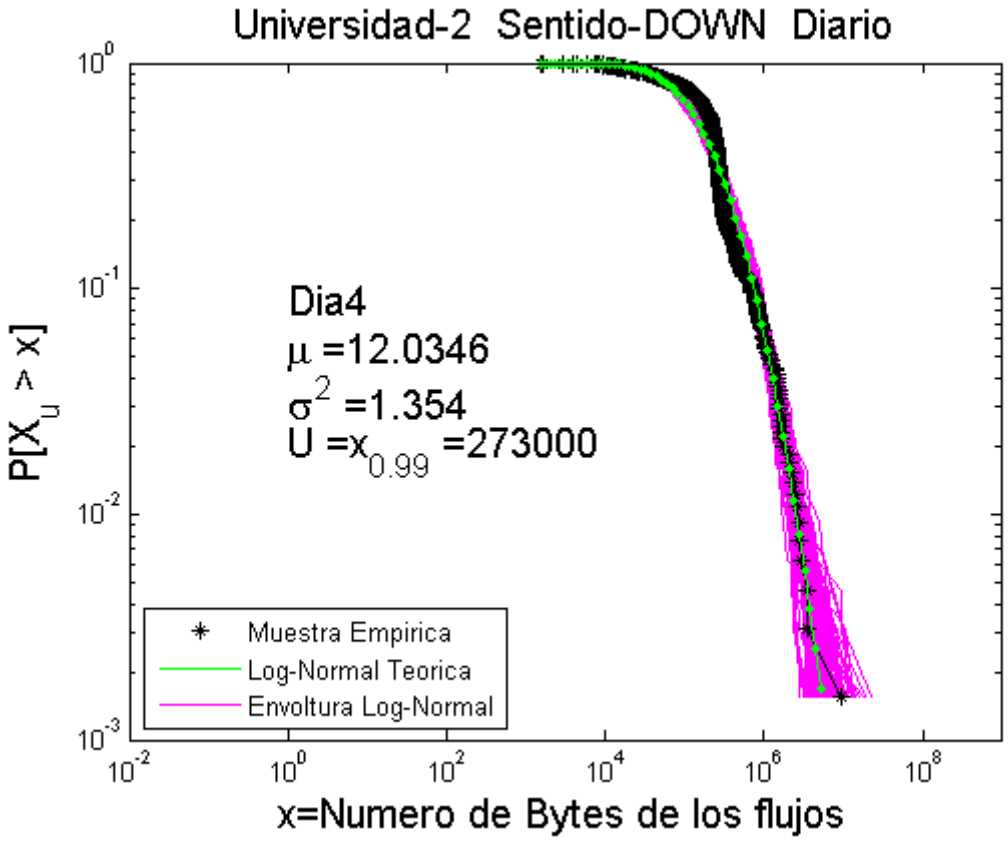

Figura 5-33: Ajuste visual CCDF LN u10 - DOWN

Se observa que los ajustes visuales son bastante buenos para todos los umbrales aplicados. Al observar que funciona mejor la log-Normal para todos los casos, se comprueba que la muestra no sigue una distribución GPD a partir de ningún valor umbral. Resaltar que, por ejemplo en [3], [8] y [14] sus autores concluyen que los datos siguen una distribución pura de Pareto a partir de cierto umbral, situación que con estos resultados queda descartada. En cambio, se puede decir que la muestra está más próxima a la distribución log-Normal en todo su conjunto, independientemente del punto que se escoja como valor inicial de la muestra. En este caso no es posible determinar cómo se ven alterados los parámetros de forma teórica a medida que aumentan los umbrales. No existe una forma cerrada de $\mu(u)$ y $\sigma^2(u)$ con la cual compararlos con los obtenidos empíricamente. A pesar de esto, se puede cualificar el ajuste y compararlo con el de la GPD mediante las medias umbralizadas teóricas determinadas por los parámetros obtenidos para el conjunto total, es decir, $u = 0$, durante la fase de ajuste a la escala temporal de un único día:

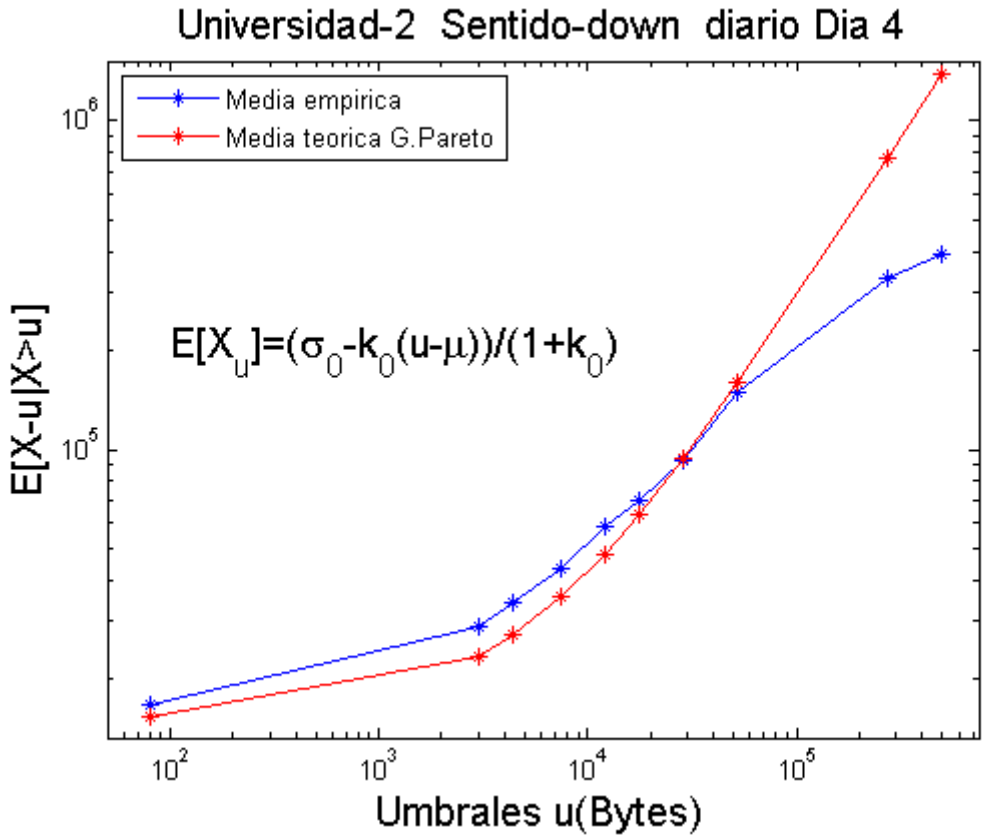

Figura 5-34: Medias umbralizadas teóricas GPD y empíricas - DOWN

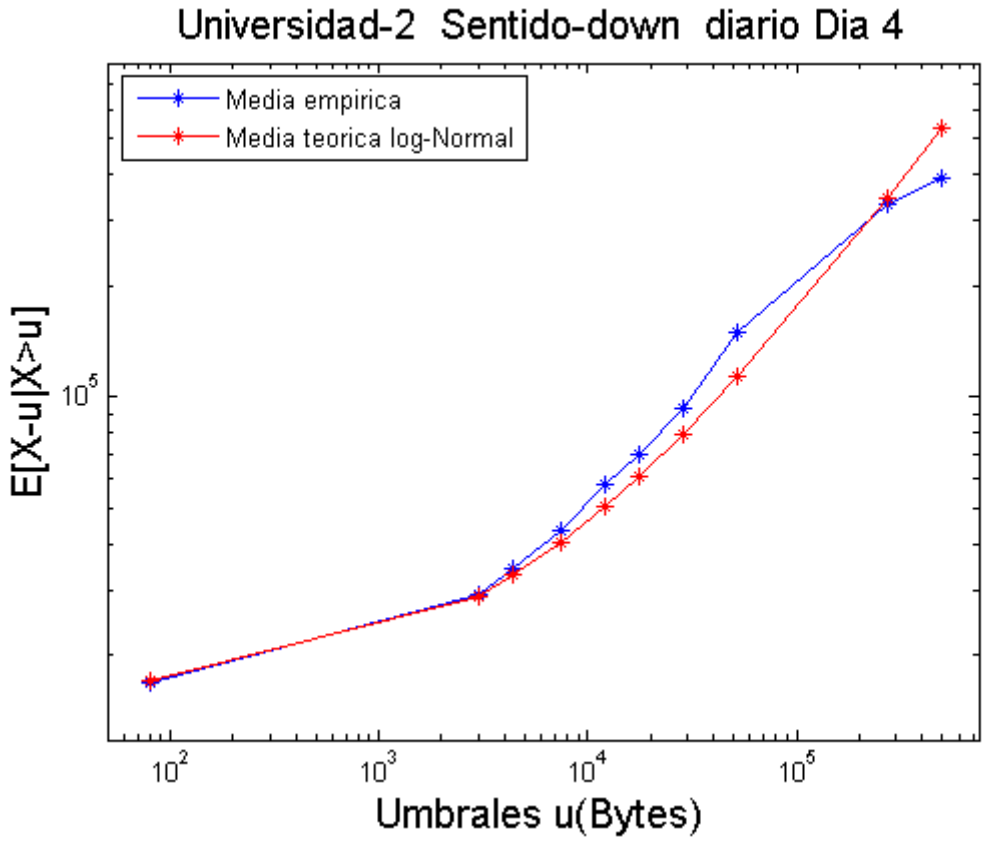

Figura 5-35: Medias umbralizadas teóricas LN y empíricas – DOWN

Así se puede ver que el ajuste del bloque completo mediante la log-Normal se comporta mejor que el de la GPD también respecto al valor de las medias umbralizadas y el resultado es satisfactorio.

5.2.3 CURVA DE LORENZ Y COEFICIENTE DE GINNI

En este apartado, de igual modo que en el sentido ascendente, se utiliza la curva de Lorenz y el coeficiente de Ginni para visualizar y cuantificar las diferencias entre la GPD y la log-Normal respecto a la muestras empíricas comprobando al mismo tiempo si los resultados son coherentes con los obtenidos hasta ahora en el sentido descendente.

A continuación se muestra la comparación entre la GPD y la log-Normal con las muestras empíricas:

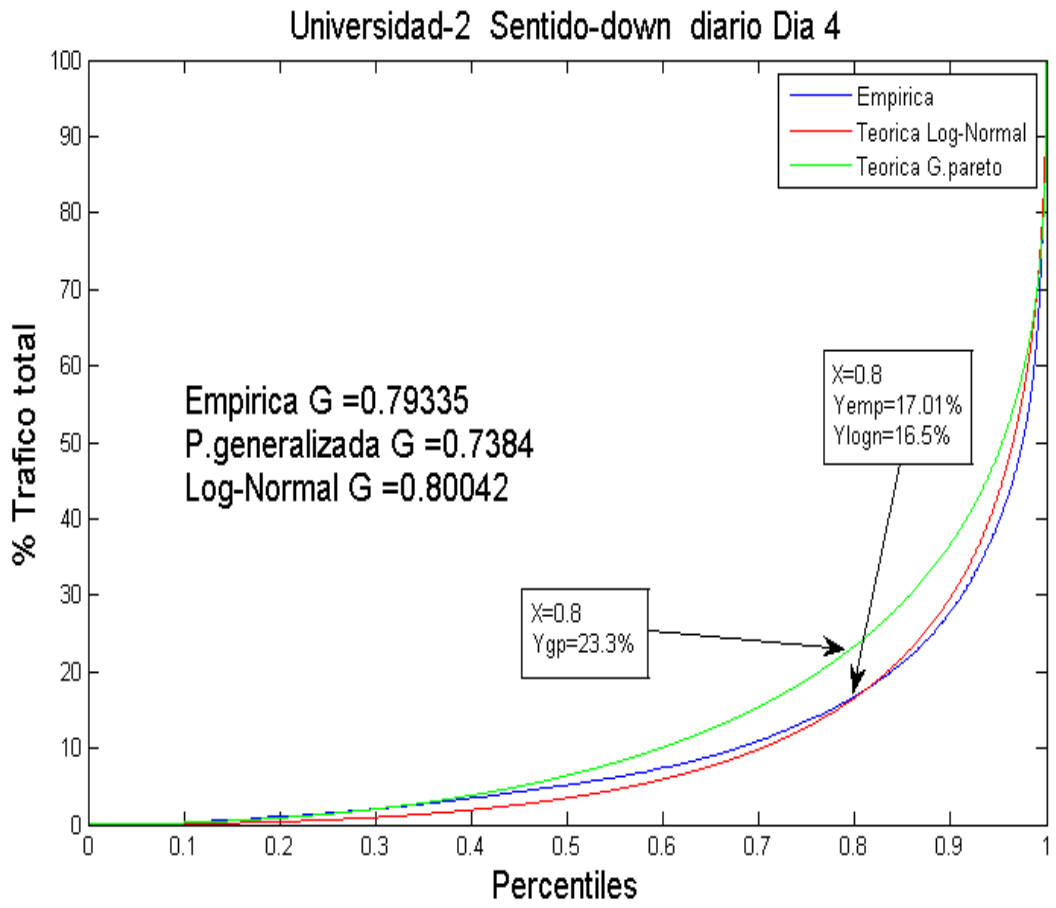

Figura 5-36: Curva de Lorenz GPD vs LN - DOWN

Como se puede ver, la curva de la log-Normal se aproxima sensiblemente más que la de GPD. El coeficiente de Ginni, G, cuantifica esta diferencia entre ambas, en coherencia con las curvas. De nuevo, se han indicado, los valores de las curvas para un valor de F=0.8.

Después de este análisis, se concluye que este bloque de datos que proviene de la universidad U2, en sentido descendente (Down), día 4 del subconjunto Diario, sigue una distribución $logN(\mu, \sigma^2)$ tal que:

$$\mu = 8.07;\ \sigma^2 = 3.29$$

El resto de bloques del subconjunto del sentido descendente han sido analizados de forma idéntica. Los resultados obtenidos con el resto

será, mostrados de forma conjunta en apartados posteriores aprovechando el análisis de diversidad temporal y espacial.

5.3 DIVERSIDAD TEMPORAL Y ESPACIAL

En este apartado se van a presentar los resultados sobre los conjuntos de datos vistos desde un punto de vista más global. Esto significa que se va a caracterizar las muestras del tráfico en todo su conjunto atendiendo a la universidad y sentido del que provengan como se ha explicado en la sección 4.3. Sólo se mostrarán los resultados obtenidos en el subconjunto de los días de Diario. En los días de fin de semana, a pesar de que se ha encontrado un comportamiento ligeramente similar, el número de muestras disponible es mucho menor y en muchas ocasiones ninguno de los dos modelos parece apropiado al mostrar formas muy irregulares. Por ello, después de analizar el subconjunto de forma global desde el punto vista de alguna de las dos distribuciones, se ha concluido que no son apropiadas para modelarlo.

Debido a que, como ya se ha mostrado en los apartados anteriores, en el estudio de las muestras en longitudes de un único día existen diferencias cualitativas entre ambos sentidos del tráfico, aquí se vuelve a realizar la misma separación.

5.3.1 SENTIDO ASCENDENTE

Los resultados referidos a la diversidad temporal (estacionaridad, apartado 4.3.1) se mostrarán de forma independiente para cada centro universitario para finalmente realizar la comparativa que supondrá el análisis sobre la diversidad espacial (apartado 4.3.2).

5.3.1.1 ESTACIONARIDAD

Universidad U1

En este caso para el 91% de los bloques ($^{19}/_{21}$), el ajuste obtenido mediante la distribución GPD es superior al obtenido mediante la LN. A pesar de no ocurre en un 100% de los casos, el análisis del subconjunto total se realiza mediante la distribución de cola pesada ya que es evidente que ésta predomina sobre la otra. A continuación se muestra, el ajuste visual log-log CCDF y la evolución paramétrica a lo largo del tiempo para este subconjunto de bloques:

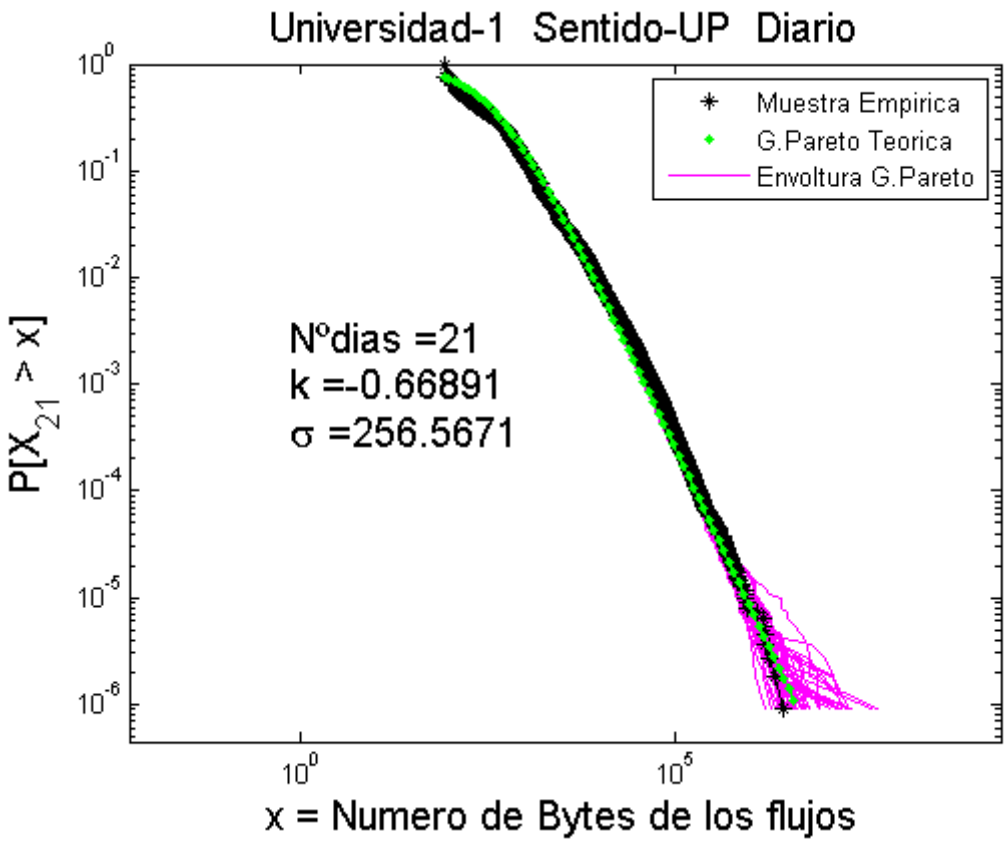

Figura 5-37: Ajuste visual log-log CCDF GPD U1 con 21 días - UP

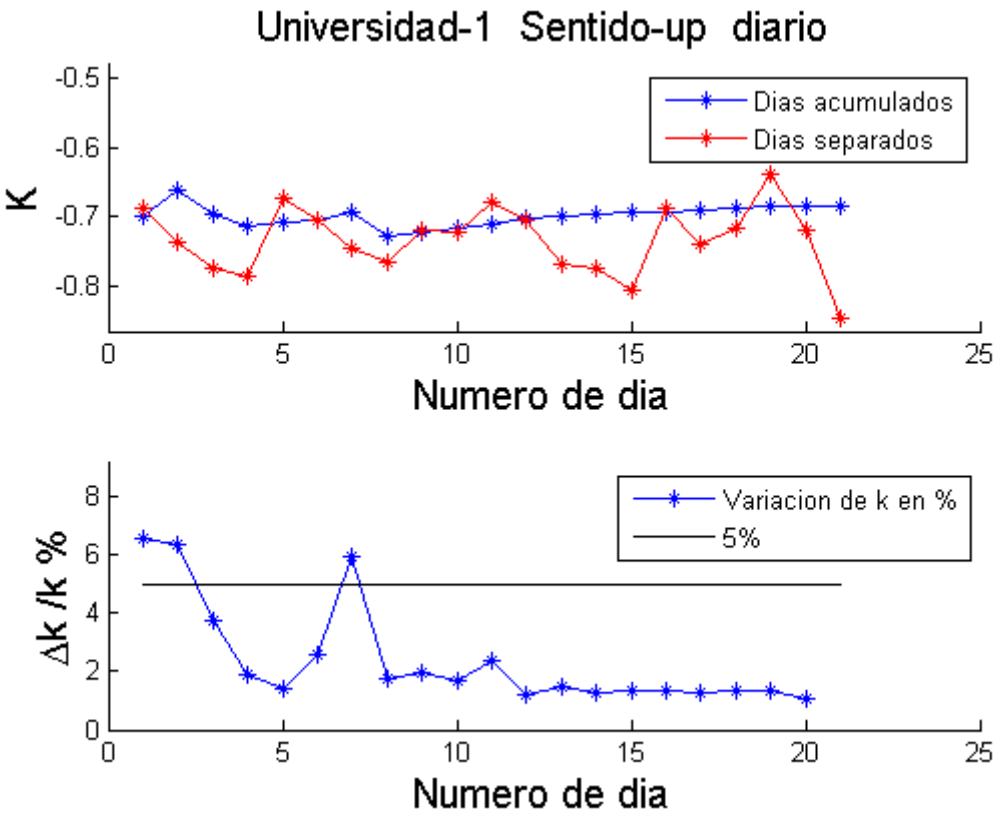

Figura 5-38: Estacionaridad parámetro k en U1 - UP

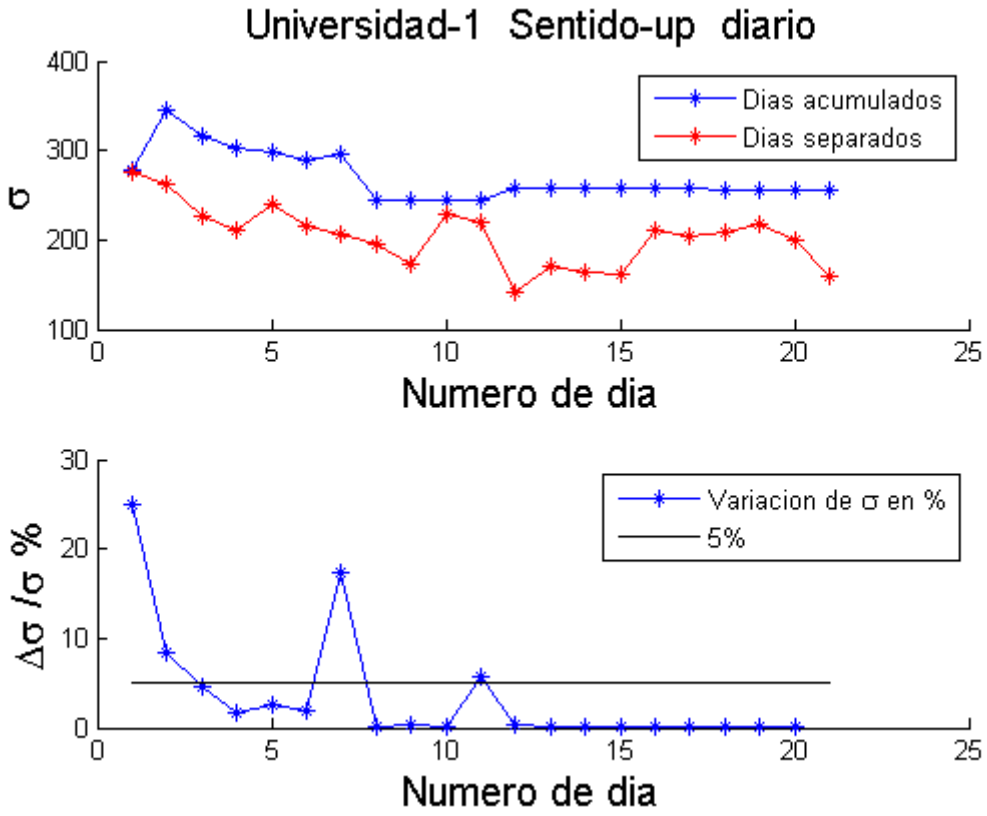

Figura 5-39: Estacionaridad parámetro sigma en U1 – UP

En la primera figura, se denomina $X_N = X_{21}$ a la variable aleatoria que abarca los 21 días de tráfico. Los resultados obtenidos para el resto de X_N, $N \in [1,20]$, son muy similares. Se observa claramente que el subconjunto total sigue una distribución GPD. A su vez, en las figuras 5-38 y 5-39 se puede ver que presenta un comportamiento estacionario. Respecto al parámetro de mayor interés, k, 9 días son suficientes para que se estabilice mientras que σ necesita más tiempo.

Con estos resultados, se puede concluir que el tráfico en sentido ascendente perteneciente a la universidad U1 posee una distribución GPD estacionaria con parámetros $\mathrm{k} \approx -0.67$ y $\sigma \approx 257$. Entonces esto supone un índice de cola $\alpha = \frac{-1}{\mathrm{k}} \approx 1.5$ que está dentro del rango necesario para producir autosimilaridad y LRD con parámetro de Hurst $\mathrm{H} = \frac{(3-\alpha)}{2} \approx 0.75$ y parámetro $\beta = \alpha - 1 \approx 0.5$.

Universidad U2

En este caso para el 100% de los bloques, el ajuste obtenido mediante la distribución GPD es superior al obtenido mediante la LN. Esta universidad presenta ciertas diferencias con la anterior. Por ello se mostrará el ajuste visual log-log CCDF para varios casos (incrementando el número de días acumulados) para poder observar la evolución que presenta y no solamente para el subconjunto total. De nuevo, también se verá comportamiento de los parámetros a lo largo del tiempo para analizar su estacionaridad:

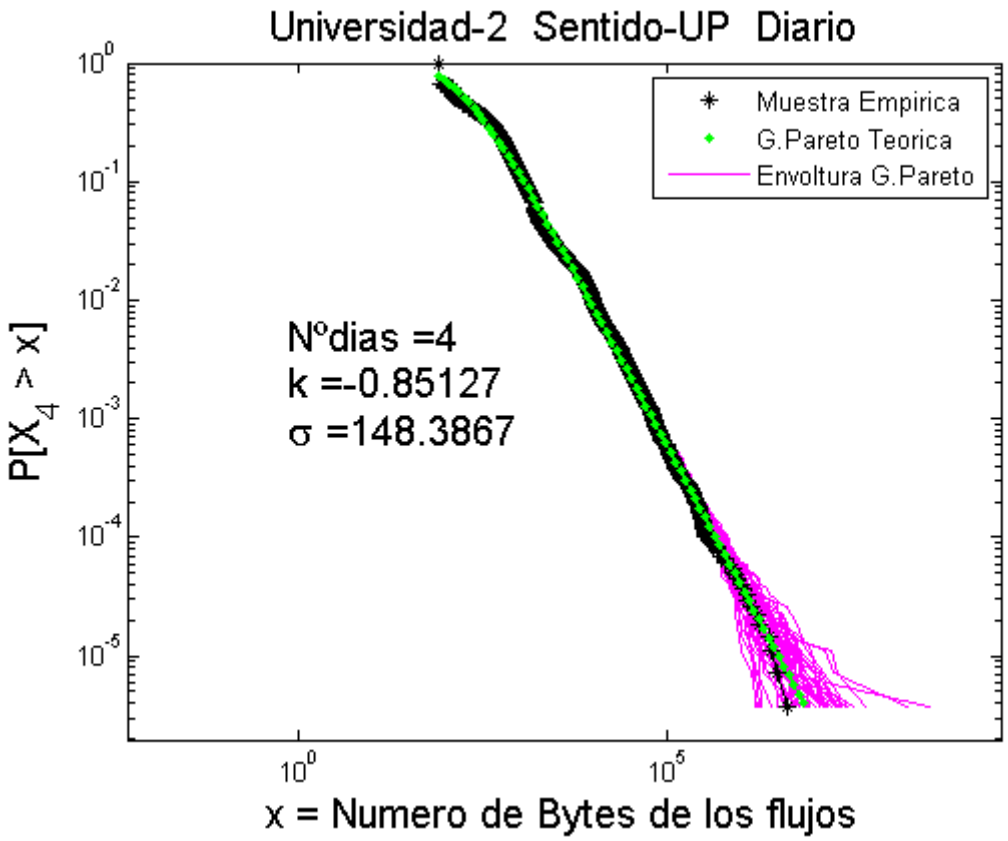

Figura 5-40: Ajuste visual log-log CCDF GPD U2 con 4 días - UP

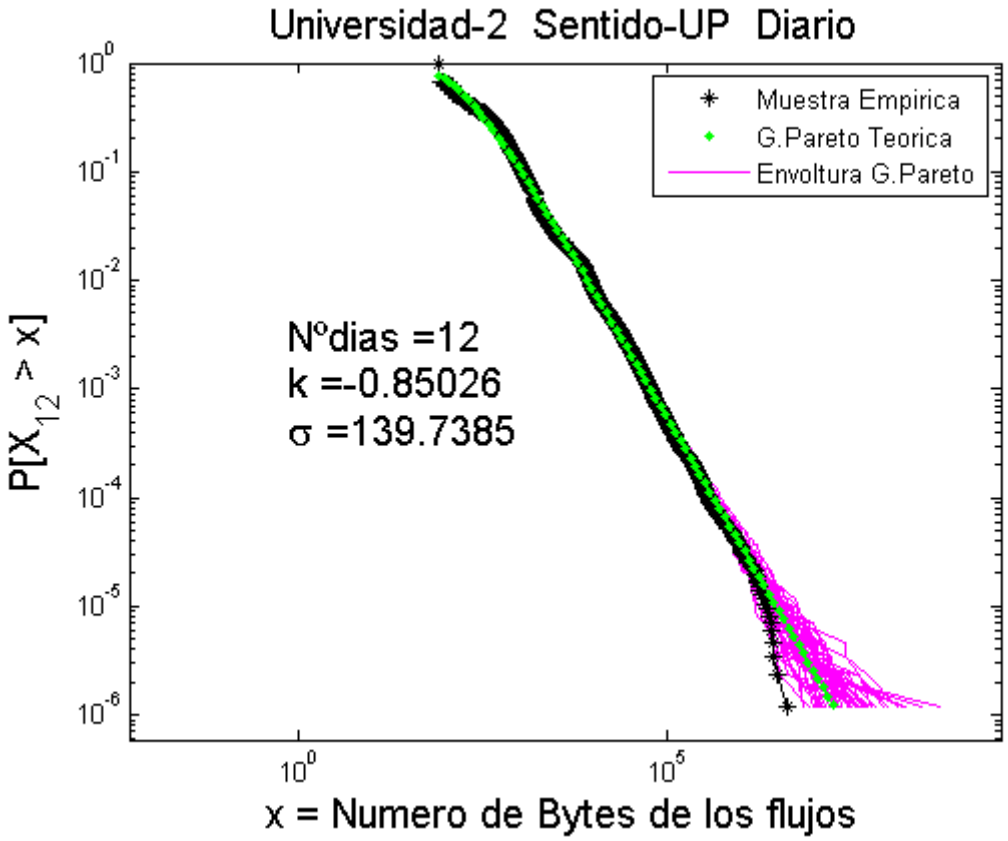

Figura 5-41: Ajuste visual log-log CCDF GPD U2 con 12 días - UP

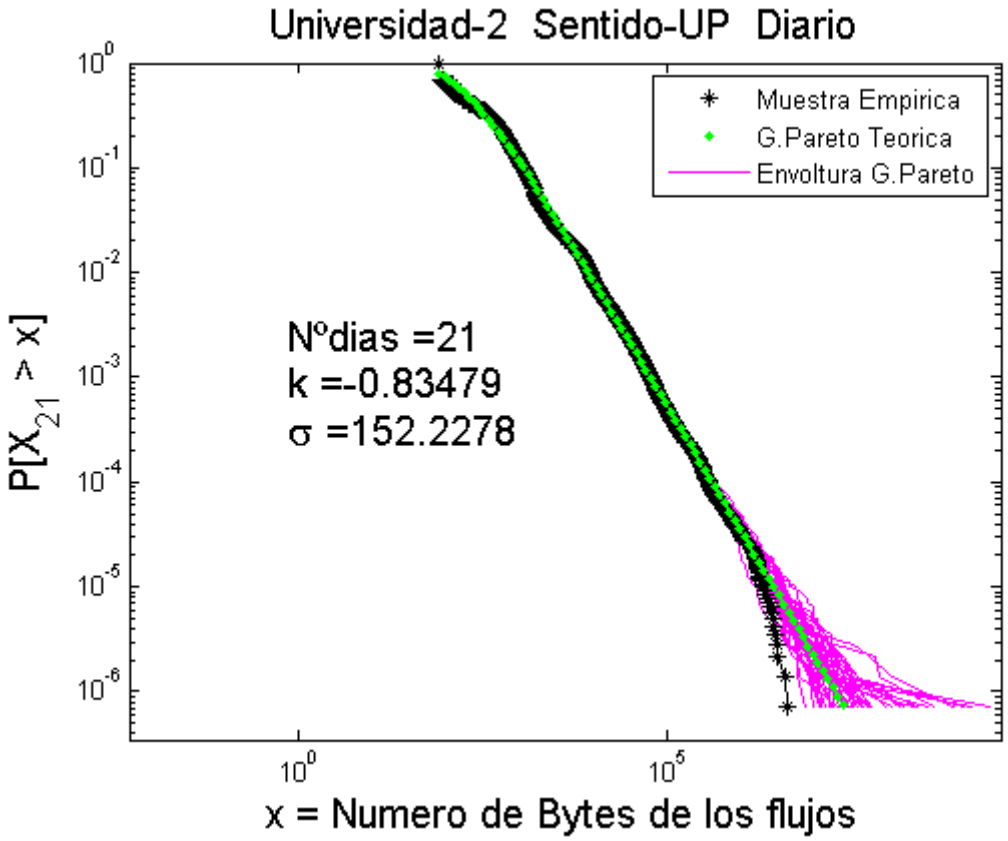

Figura 5-42: Ajuste visual log-log CCDF GPD U2 con 21 días - UP

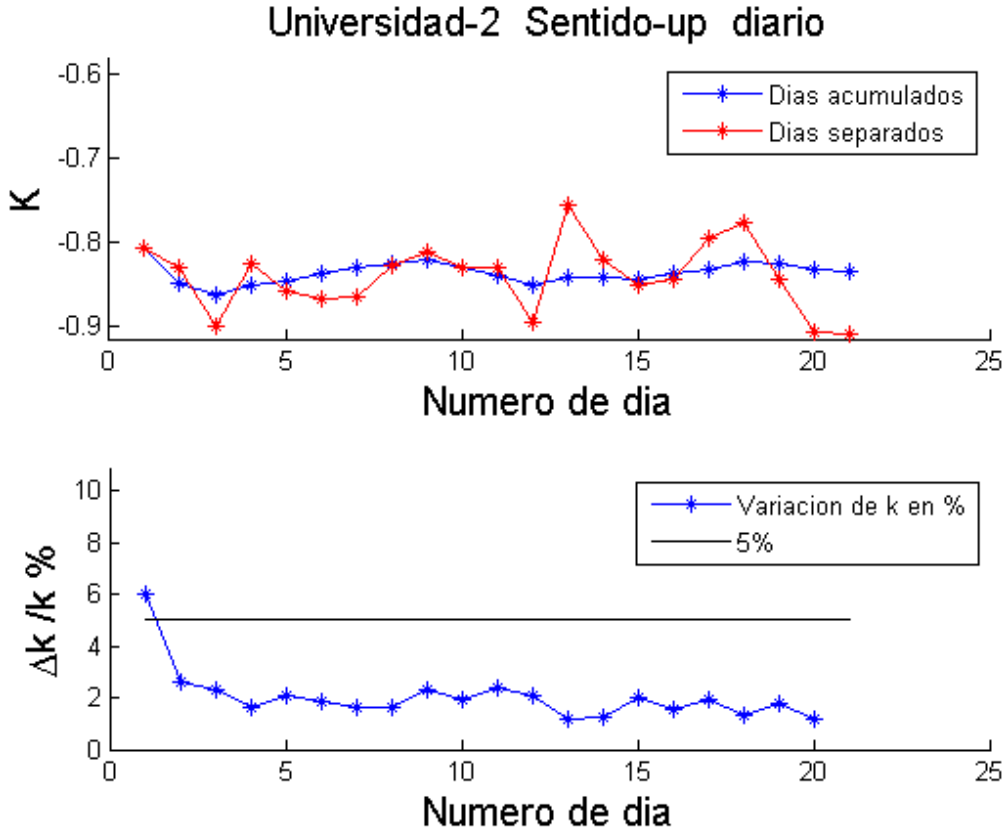

Figura 5-43: Estacionaridad parámetro k en U2 – UP

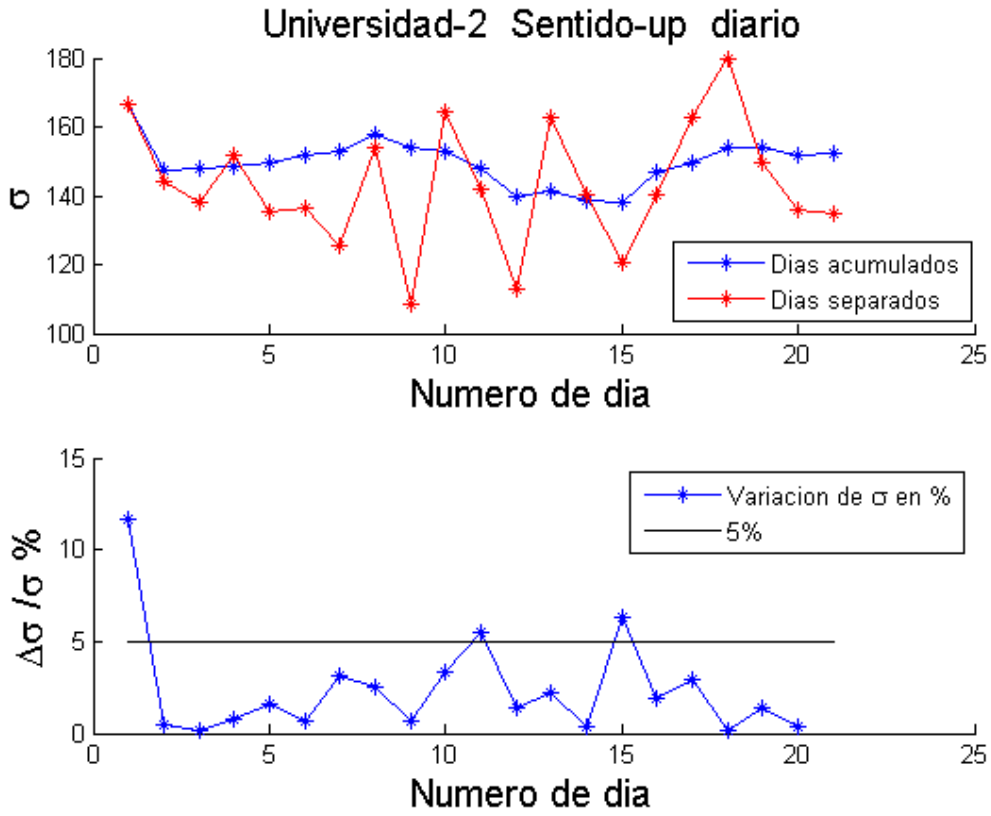

Figura 5-44: Estacionaridad parámetro sigma en U2 – UP

Respecto al ajuste visual, se han mostrado los casos X_n de X_4, X_{12} y X_{21}. Se puede observar que a medida que el número de días acumulados va incrementándose, la zona lejana de la cola va incrementando su pendiente respecto a la recta teórica de la GPD. No es que presente una curvatura que provoque que la pendiente vaya aumentando a lo largo del todo el rango como ocurre con la log-Normal. Esto significa que, a medida que N aumenta, la distribución se aleja de la GPD "ideal" y se aproxima a una versión truncada superiormente de la misma, es decir, $\{X|X < Max\}$. Su distribución CCDF se ve alterada de la siguiente forma:

$$\bar{F}_{X|X<Max}(x) = \frac{P\{x \leq X \leq Max\}}{P\{X \leq Max\}} = 1 - \frac{F_X(x)}{F_X(Max)} \quad , \qquad x \leq Max$$

siendo $F_X(x)$ la función CDF de la GPD "ideal".

Esto es una consecuencia de los límites que poseen los datos reales. Si N aumenta, el número de puntos/muestras $n \to \infty$, con lo que la probabilidad asociada al valor máximo mediante la ECCDF[39] $\frac{1}{n} \to 0$. La distribución teórica ideal alcanza ese valor únicamente en el infinito, valor que, evidentemente, no se puede alcanzar con datos provenientes de trazas de tráfico real. La diferencia que se produce con respecto al caso de la U1, es que, mostrando ambas un valor $\mathrm{Max} \approx 10^7$ Bytes, (figuras 5-37 y 5-42) el número de muestras n es ahora superior y la distribución es más pesada (α menor), lo que, para no alejarse de la distribución "ideal", debería venir acompañado de un incremento en su valor máximo Max que no se produce. A pesar de esto, en esencia, se puede considerar que la distribución subyacente es la GPD no truncada al abarcar la gran mayoría de su rango. Se puede observar que, en la representación del subconjunto completo, también están presentes los llamados wobbles comentados en el capítulo 2, lo que apoya la hipótesis de que ocurren de forma sistemática y no son debidos a la variabilidad del muestreo de los autores de [9]. En las figuras 5-43 y 5-44 se puede ver que presenta un comportamiento estacionario. Respecto al parámetro de mayor interés, k, presenta estabilidad rápidamente, siendo necesarios únicamente 3 días. El parámetro σ no es tan estable aunque las variaciones que sufre parecen estar acotadas en torno al 6%.

Con estos resultados, se puede concluir que el tráfico en sentido ascendente perteneciente a la universidad U2 posee una distribución GPD estacionaria con parámetros $k \approx -0.84$ y $\sigma \approx 152$. Entonces esto supone un índice de cola $\alpha = \frac{-1}{k} \approx 1.2$ que está dentro del rango necesario para producir autosimilaridad y LRD con parámetro de Hurst $H = \frac{(3-\alpha)}{2} \approx 0.9$ y parámetro $\beta = \alpha - 1 \approx 0.2$.

[39] **ECCDF**: Del inglés, Experimental Complemetary Cumulative Distribution Function

Universidad U3

En este caso para el 96% de los bloques ($^{20}/_{21}$) el ajuste obtenido mediante la distribución GPD es superior al obtenido mediante la LN. Como en el caso de la universidad $U1$, se analiza el total del conjunto mediante la distribución de cola pesada a pesar de no suponer el 100%:

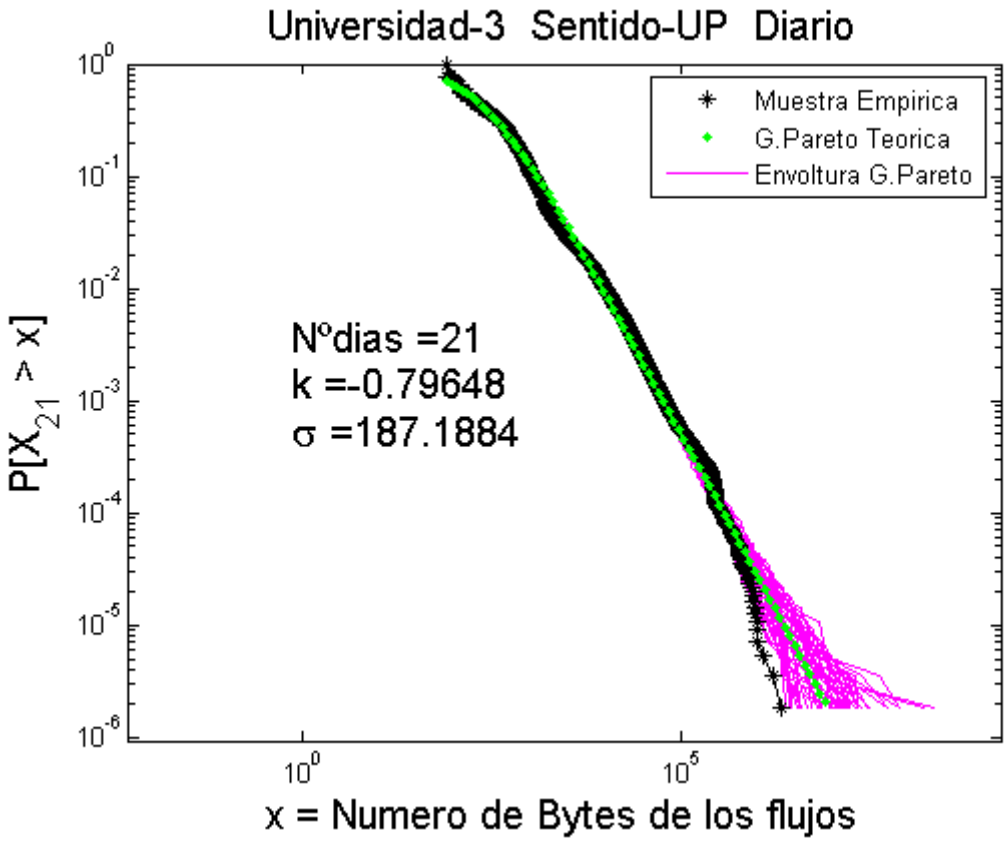

Figura 5-45: Ajuste visual log-log CCDF GPD U3 con 21 días - UP

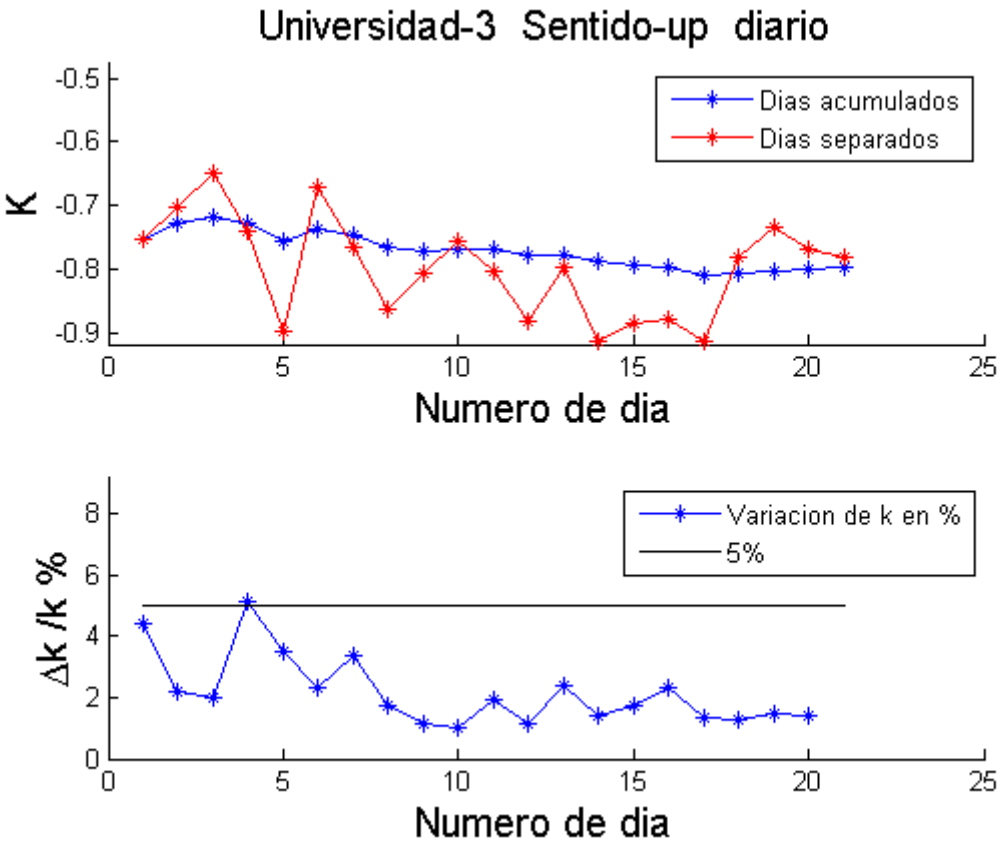

Figura 5-46: Estacionaridad parámetro k en U3 – UP

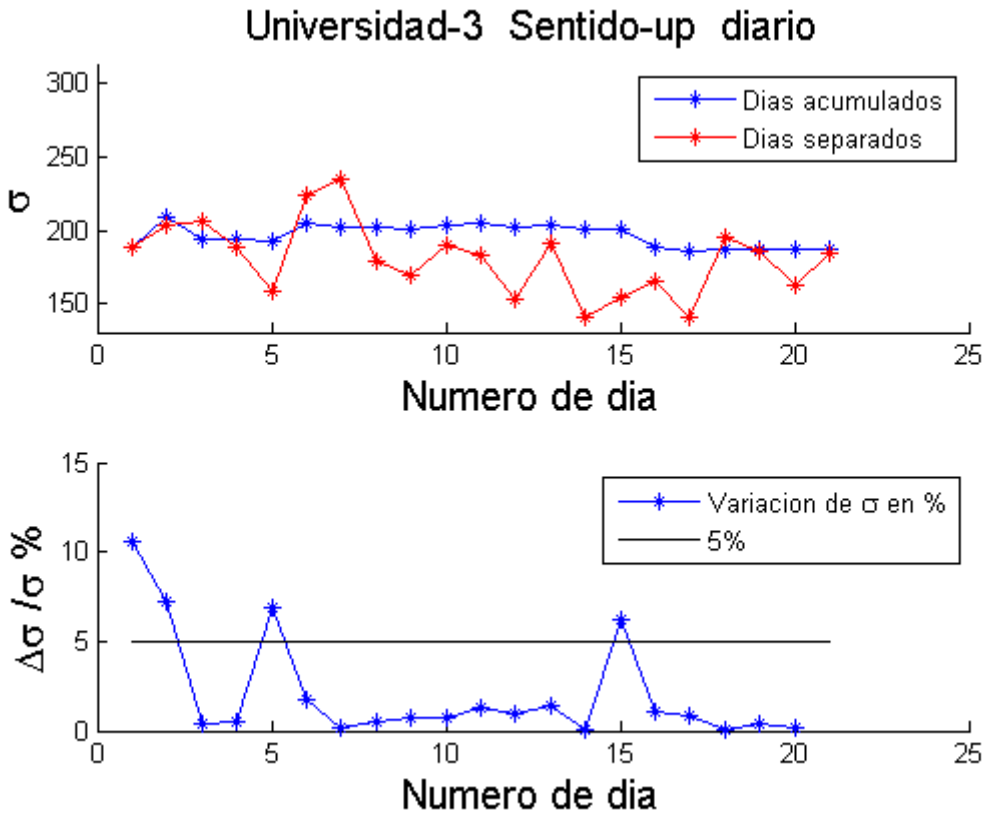

Figura 5-47: Estacionaridad parámetro sigma en U3 – UP

En lo referido al ajuste visual, se produce un efecto muy similar al comentado en la universidad U2. En este caso, solamente se muestra el caso de X_{21}. El alejamiento de la distribución "ideal" es menos acusado debido a que el número de muestras n es menor y la distribución no es tan pesada (α superior). Se puede observar que, de nuevo, $\mathrm{Max} \approx 10^7$ Bytes. También se ha observado que aquí los wobbles son más pronunciados que en los casos anteriores. Respecto a la estacionaridad, el comportamiento es prácticamente idéntico al de U2. El parámetro k se estabiliza rápidamente, necesitando únicamente 5 días para presentar variaciones menores del 5% y σ no es tan estable pero sus variaciones están acotadas en torno al 6% como se observa en las figuras 5-46 y 5-47 respectivamente.

Con estos resultados, se puede concluir que el tráfico en sentido ascendente perteneciente a la universidad $U3$ posee una distribución GPD estacionaria con parámetros $k \approx -0.8$ y $\sigma \approx 187$. Entonces esto supone un índice de cola $\alpha = \frac{-1}{k} \approx 1.25$ que está dentro del rango necesario para producir autosimilaridad y LRD con parámetro de Hurst $H = \frac{(3-\alpha)}{2} \approx 0.875$ y parámetro $\beta = \alpha - 1 \approx 0.25$.

5.3.1.2 DIVERSIDAD ESPACIAL

En el apartado 4.3.2 ya se ha comentado lo que se denomina diversidad espacial. Tras haber analizado el tráfico ascendente de las tres universidades en todo su conjunto, los perfiles de tráfico observados, determinados principalmente por los valores del índice de cola $\alpha = \frac{-1}{k}$, son diferentes de un caso a otro, a pesar de que las redes han sido escogidas con propiedades intrínsecas muy similares. Hay que señalar que, el número de usuarios de las tres universidades es suficientemente grande (más de 20.000 usuarios de internet) como para esperar que las CCDF converjan a la misma distribución. La mayor diferencia se produce entre la $U1$, que presenta un $\alpha \approx 1.5$ con las otras dos, que como ya se ha visto poseen un tráfico más pesado ($\alpha \approx 1.2$ y $\alpha \approx 1.25$). Entonces, con esto se puede concluir que, las medidas recopiladas en una

universidad generalmente no son válidas para otra aunque tengan características intrínsecas similares. En [7], sus autores explican de forma muy convincente la razón de esta diversidad tras concluir lo mismo mediante el análisis del parámetro s de la Zipf aplicado a la distribución de las direcciones IP más visitadas.

5.3.2 SENTIDO DESCENDENTE

Al igual que para el sentido ascendente, los resultados referidos a la diversidad temporal (estacionaridad, apartado 4.3.1) se mostrarán de forma independiente para cada centro universitario para finalmente realizar la comparativa que supondrá el análisis sobre la diversidad espacial (apartado 4.3.2).

5.3.2.1 ESTACIONARIDAD

Universidad U1

En este caso para el 71% de los bloques ($^{15}/_{21}$), el ajuste obtenido mediante la distribución LN es superior al obtenido mediante la GPD. A pesar de no ocurre en un 100% de los casos, el análisis del subconjunto total se realiza mediante la distribución log-Normal ya que ésta predomina sobre la otra. A continuación se muestra, el ajuste visual log-log CCDF y la evolución paramétrica a lo largo del tiempo para este subconjunto de bloques:

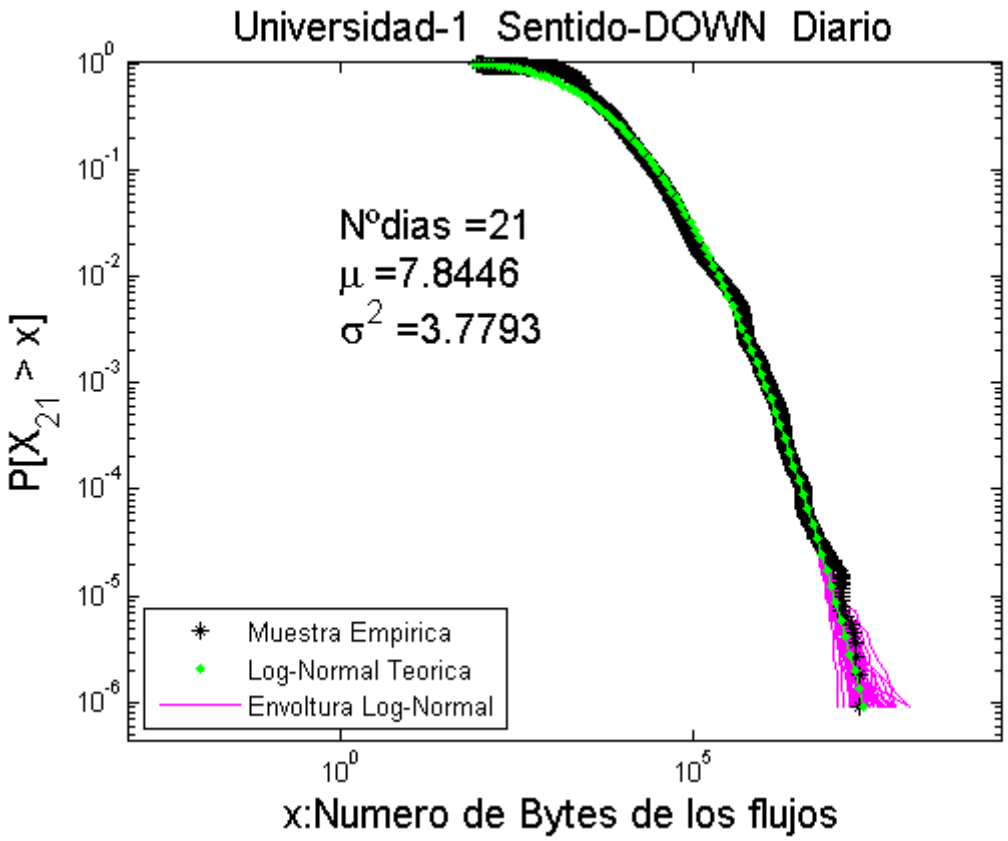

Figura 5-48: Ajuste visual log-log CCDF LN U1 con 21 días - DOWN

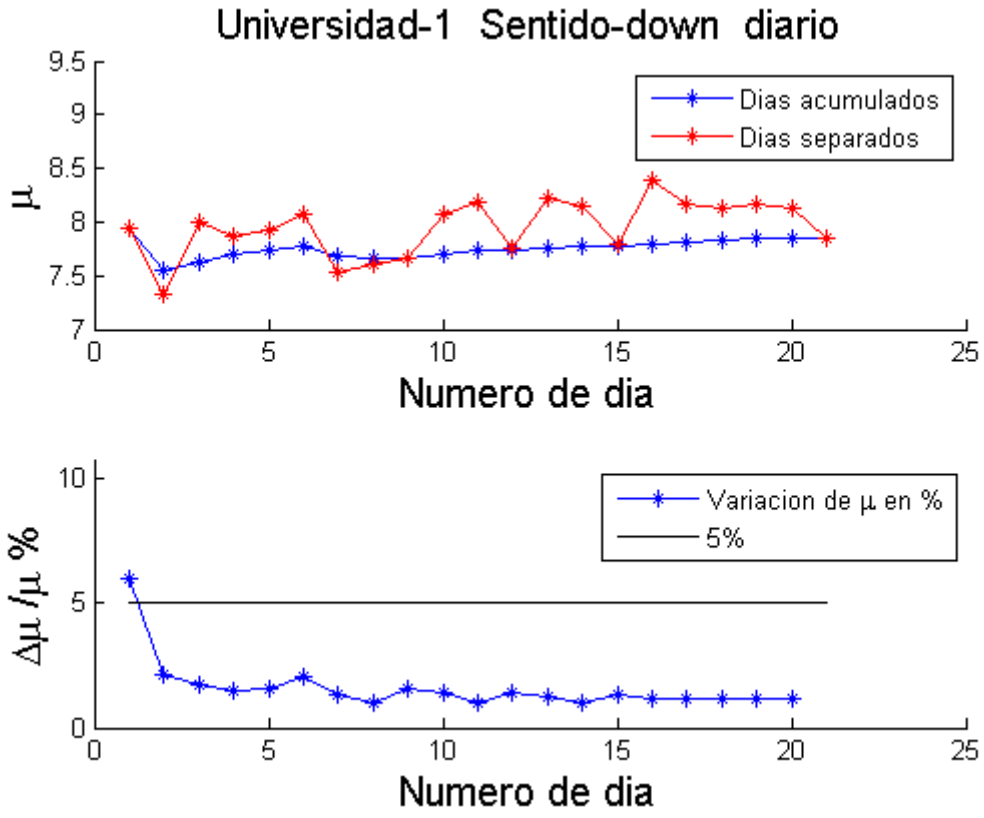

Figura 5-49: Estacionaridad parámetro mu en U1 – DOWN

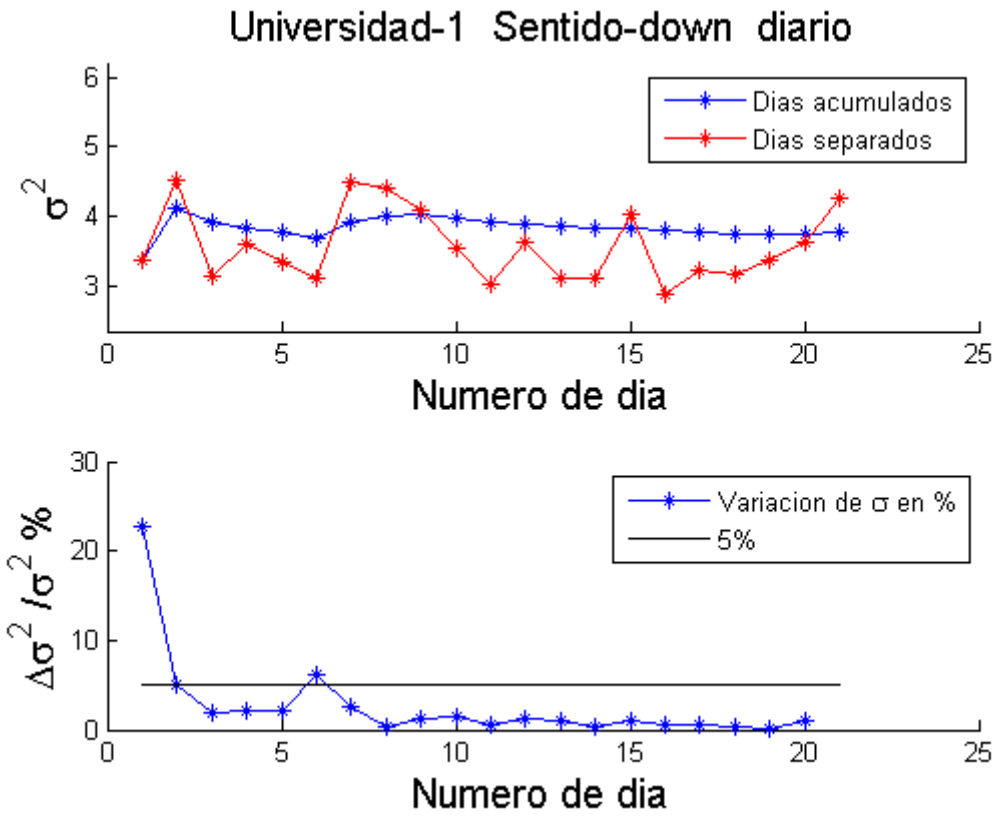

Figura 5-50: Estacionaridad parámetro sigma^2 en U1 – DOWN

En la primera figura, de nuevo, se denomina $X_N = X_{21}$ a la variable aleatoria que abarca los 21 días de tráfico. Los resultados obtenidos para el resto de X_N, $N \in [1,20]$, son muy similares. Se observa que el subconjunto total se aproxima a una distribución LN. Resaltar de nuevo la aparición de wobbles de forma más acusada que para el caso ascendente. En las figuras 5-49 y 5-50 se puede ver que presenta un comportamiento estacionario. Ambos parámetros se estabilizan rápidamente permaneciendo debajo del límite de 5% de variación.

Con estos resultados, se puede concluir que el tráfico en sentido descendente perteneciente a la universidad $U1$ se aproxima a una distribución LN estacionaria con parámetros $\mu \approx 7.8$ y $\sigma^2 \approx 3.8$. Notar que en este caso, la conclusión que se obtiene (extensible a las otras dos universidades) realmente es que la distribución del tráfico se aproxima más, al comportamiento de una distribución log-Normal, que al de una generalizada de Pareto. A la hora de ver las implicaciones que conlleva sobre la naturaleza del tráfico, comentadas en el capítulo 1, no significa que éste no pueda

presentar autosimilaridad o LRD[40]. Significa que, o hay que encontrar otras posibles causas, o hay que intentar modelar este tipo de tráfico con distribuciones más sofisticadas que se aproximen más al comportamiento que presenta y además puedan producir estos efectos o por último estudiar si es posible que la duración de los flujos pueda ser de cola pesada a pesar de que su tamaño siga una distribución log-Normal.

Universidad U2

En este caso para el 76% de los bloques ($^{16}/_{21}$), el ajuste obtenido mediante la distribución LN es superior al obtenido mediante la GPD. A continuación se muestra, el ajuste visual log-log CCDF y la evolución paramétrica a lo largo del tiempo para este subconjunto de bloques:

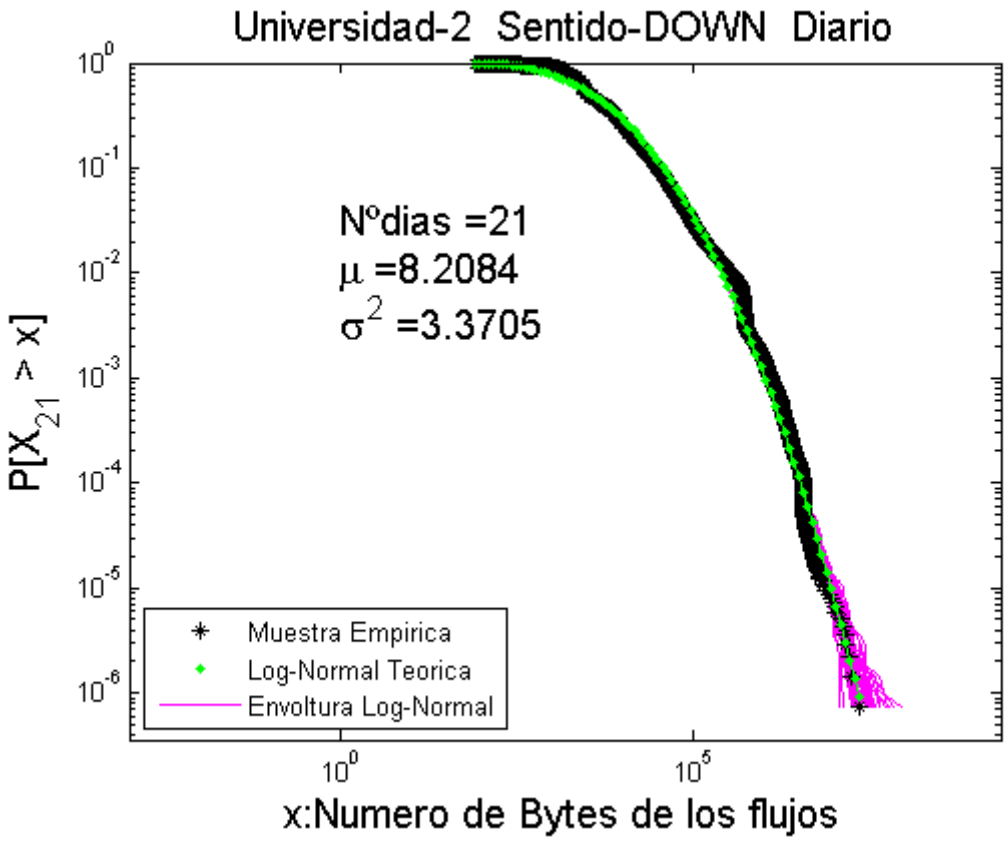

Figura 5-51: Ajuste visual log-log CCDF LN U2 con 21 días - DOWN

[40] Como ya se ha comentado, los autores de [15] demuestran que la distribución log-Normal puede implicar LRD bajo ciertas condiciones aunque se alejan bastante de los resultados obtnidos.

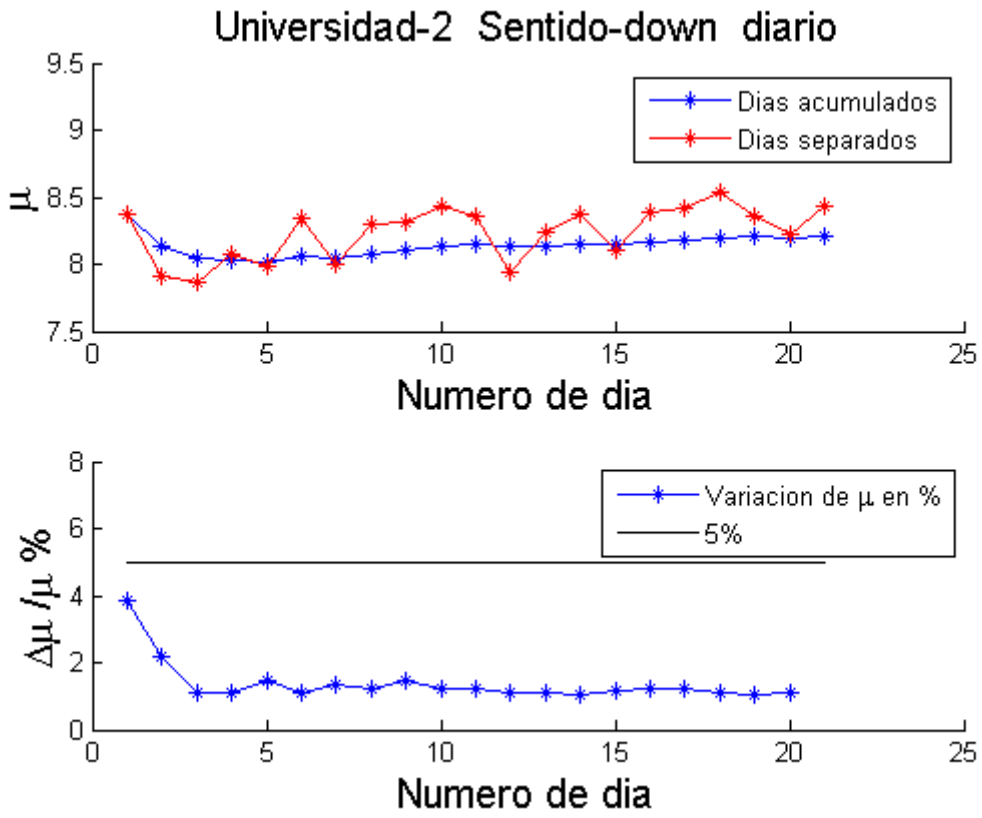

Figura 5-52: Estacionaridad parámetro mu en U2 – DOWN

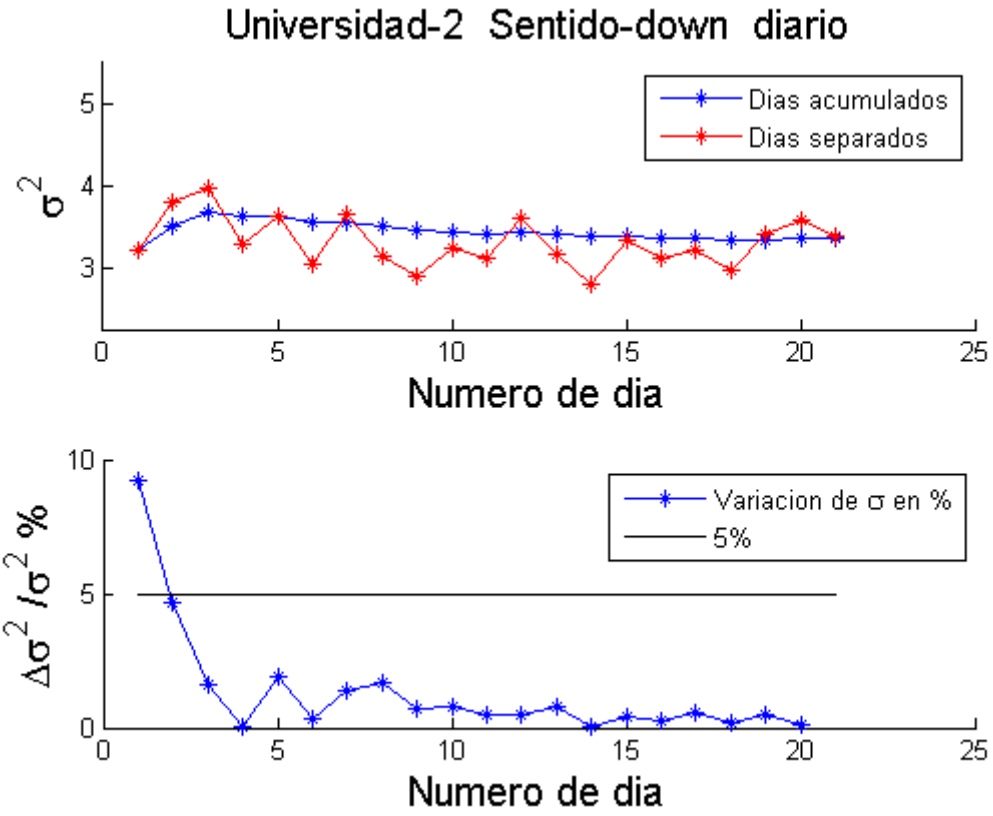

Figura 5-53: Estacionaridad parámetro sigma^2 en U2 – DOWN

En la primera figura, de nuevo, se denomina $X_N = X_{21}$ a la variable aleatoria que abarca los 21 días de tráfico. Todo lo comentado para la universidad $U1$ es aplicable en este caso. Resaltar de nuevo la aparición de wobbles de forma más acusada que para el caso ascendente e incluso superiores a los de la universidad $U1$.

Con estos resultados, se puede concluir que el tráfico en sentido descendente perteneciente a la universidad $U2$ se aproxima a una distribución LN estacionaria con parámetros $\mu \approx 8.2$ y $\sigma^2 \approx 3.4$.

Universidad U3

En este caso para el 71% de los bloques ($^{15}/_{21}$), el ajuste obtenido mediante la distribución LN es superior al obtenido mediante la GPD. A continuación se muestra, el ajuste visual log-log CCDF y la evolución paramétrica a lo largo del tiempo para este subconjunto de bloques:

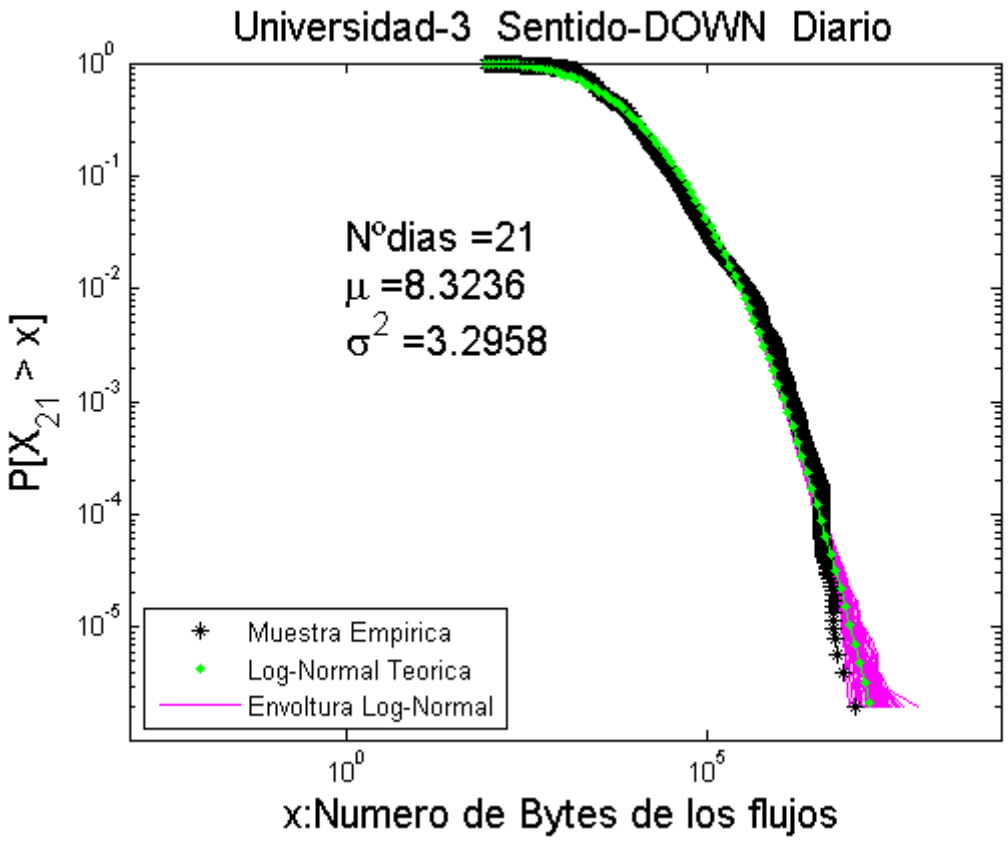

Figura 5-54: Ajuste visual log-log CCDF LN U3 con 21 días - DOWN

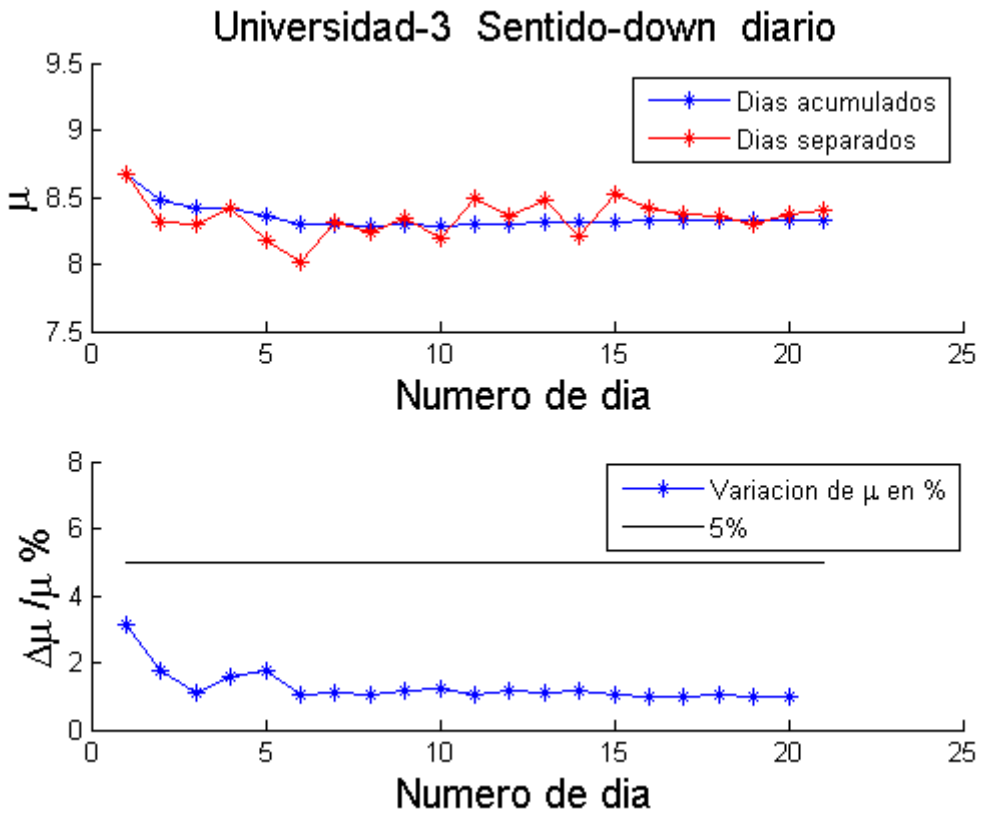

Figura 5-55: Estacionaridad parámetro mu en U3 – DOWN

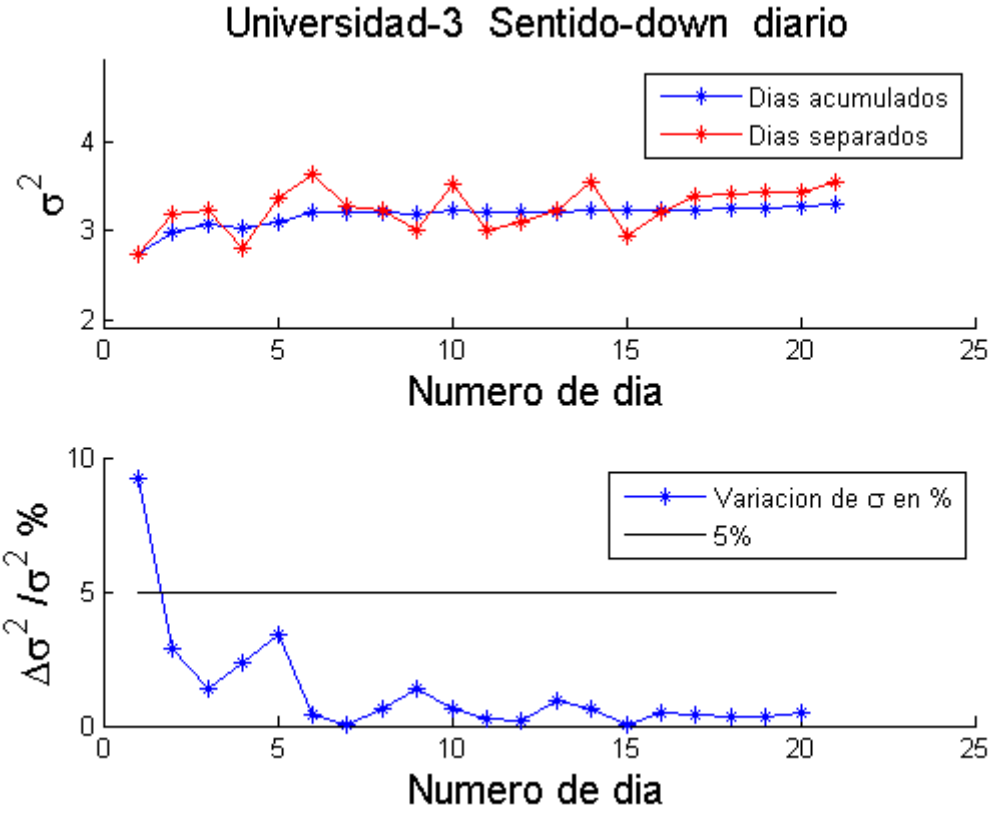

Figura 5-56: Estacionaridad parámetro sigma^2 en U3 – DOWN

En la primera figura, de nuevo, se denomina $X_N = X_{21}$ a la variable aleatoria que abarca los 21 días de tráfico. Todos los comentarios realizados en las universidades anteriores son aplicables en este caso. Hay que resaltar, con respecto a los anteriores, los wobbles que presenta la distribución empírica parecen ser más largos, es decir, el número de oscilaciones alrededor de la distribución teórica es menor ya que abarcan mucho mayor rango.

Con estos resultados, se puede concluir que el tráfico en sentido descendente perteneciente a la universidad $U3$ se aproxima a una distribución LN estacionaria con parámetros $\mu \approx 8.3$ y $\sigma^2 \approx 3.3$.

5.3.2.2 DIVERSIDAD ESPACIAL

Al igual que ocurre en el sentido ascendente, se ha observado que, los perfiles de tráfico observados, determinados en este caso por la dupla $[\mu, \sigma^2]$ que caracteriza la distribución log-Normal, son diferentes de un caso a otro, a pesar de que las redes han sido escogidas con propiedades intrínsecas muy similares. De nuevo, la mayor diferencia se produce entre la $U1$ [7.8 , 3.8] con las otras dos que poseen resultados similares, [8.2 , 3.4] y [8.3 , 3.3]. Por esto, se concluye de igual forma que en el caso ascendente, observando que existe un factor de diversidad espacial en las medidas.

6. CONCLUSIONES Y TRABAJO FUTURO

6.1 CONCLUSIONES

Una vez finalizado el análisis de los flujos de las trazas de tráfico reales provenientes de RedIRIS, se han alcanzado dos tipos de conclusiones, las que se refieren a la caracterización de los flujos propiamente dichos y las que se refieren a cómo son analizados.

Las primeras se pueden resumir en los siguientes puntos:

- Parece que, en sentido ascendente, existen evidencias suficientes de que la distribución que sigue el tamaño de los flujos del tráfico Web se puede modelar mediante la función Generalizada de Pareto, es decir, son de cola pesada. Esta distribución además abarca todo el rango de valores presentando poca curvatura.

- El tamaño de los flujos de tráfico Web en sentido ascendente presenta un carácter estacionario necesitando poco tiempo para presentar estabilidad. Desde el punto de vista del índice de cola α, presenta un factor de diversidad espacial, es decir, las medidas recopiladas en una red generalmente no son válidas para otra que tenga características intrínsecas similares.

- La distribución que sigue el tamaño de los flujos en sentido ascendente, apoya la hipótesis de que el tráfico de red presenta el fenómeno de autosimilaridad (SS) y dependencia a larga escala (LRD) debido a que, la duración de éstos es probable que sea de cola pesada.

- Parece que para índices de cola $\alpha < 1.5$, a medida que se dispone de un número de muestras mayor, las distribuciones empíricas se van aproximando versiones truncadas de las distribuciones teóricas de Pareto.

- Parece que, en sentido descendente, existen evidencias de que la distribución que sigue el tamaño de los flujos del tráfico Web se aproxima más al modelo de la distribución log-Normal que al de la Generalizada de Pareto. Las evidencias, en este caso, no son tan "fuertes" como en el caso del sentido ascendente. Este tráfico parece presentar una curvatura que con la distribución Generalizada no se puede modelar, pero tampoco es totalmente evidente que la log-Normal si lo haga.

- El tamaño de los flujos de tráfico Web en sentido descendente, desde el punto de vista del modelo de la log-Normal, presenta unas características similares al sentido ascendente en lo referente a las propiedades de diversidad temporal y espacial.

- La distribución que sigue el tamaño de los flujos en sentido descendente, no aporta ninguna explicación a los fenómenos de SS y LRD observados en el tráfico de red. Esto no quiere decir que lo contradiga, ya que, a pesar de que el tamaño de los flujos no sea de cola pesada, su duración temporal podría serlo.

- Por último, concluir que, a pesar de que las diferencias existentes entre ambos sentidos del tráfico son de tipo cualitativo más que cuantitativo, ambos poseen unas oscilaciones de tipo sistemático (denominados wobbles en el proyecto) que no parecen ser producidas por la variabilidad del muestreo.

Las segundas son las siguientes:

- La distribución Generalizada de Pareto, al ser más versátil que la distribución Pura, parece más apropiada para el análisis de características de cola pesada en el tráfico. No requiere una determinación heurística de un valor umbral inicial.

Gracias a la umbralización, puede caracterizar una muestra a partir de distintos rangos de valores comprobando cuales siguen el comportamiento esperado gracias al parámetro extra σ.

- No se debe descartar la utilización de distribuciones de cola "ligera" como la log-Normal porque se ha comprobado que pueden modelar mejor el tráfico que las de cola pesada.

- La curva de Lorenz y el coeficiente de Ginni parecen ser un método bastante apropiado para discernir entre la distribución GPD y la LN. Dentro de este proyecto, se ha comprobado que los resultados obtenidos mediante su utilización, son bastante más satisfactorios que otros métodos propuestos por otros autores, como pueden ser la m-agregación de muestras propuesta en [8] o el denominado test de curvatura en [12].

6.2 TRABAJO FUTURO

Una vez finalizado el análisis, se puede ver que el abanico de posibilidades que ofrece el estudio de flujos a partir de trazas de tráfico reales es muy amplio. Se van a plantear, por una parte, una serie de propuestas que buscan profundizar el análisis a partir del camino marcado por este proyecto, y, por otra, unas que abren nuevas líneas de investigación que guardan cierta relación con este trabajo.

Las primeras se pueden resumir en los siguientes puntos:

- Buscar, principalmente para el caso del sentido descendente del tráfico, un modelo más complejo que la distribución log-Normal, como puede ser un modelo híbrido Pareto-logNormal utilizado por otros autores, utilización de mixturas o tratar de definir una distribución doble-Pareto

Genralizada (la doble Pareto ya existe) que modele de forma más aproximada las medidas.

- Encontrar, en lo referido a los 8 métodos de cálculo de parámetros del algoritmo EPM aquí propuestos, alguna forma de establecer los pares de estadísticos de orden que sea eficiente en todos los casos y no sea necesario recurrir al posterior cálculo de distancias que conlleva cierto coste computacional.

- Modelar los wobbles, oscilaciones sistemáticas que parecen poseer las distribuciones subyacentes, mediante alguna de las distribuciones de cola pesada que aparecen en [10].

- Realizar un análisis similar sobre el puerto 443 (HTTPS[41]) que parece seguir la misma distribución y buscar, para otros importantes que no lo hacen como el 53 (DNS[42]) , nuevos modelos.

Las segundas poseen unas posibilidades prácticamente infinitas, a continuación se muestran algunas:

- Analizar la duración de los tiempos de los flujos en lugar de su tamaño en Bytes, para poder estudiar la relación que guardan ambas magnitudes.

- Buscar un modelo generativo para los flujos, atendiendo a la pila TCP/IP, que explique porque se distribuyen de esta manera.

[41] **HTTPS**: Del inglés, HyperText Transfer Protocol Secure. Este protocolo usa por defecto el protocolo 443 y utiliza un canal cifrado, más apropiado para el tráfico de información sensible que en el protocolo HTTP.

[42] **DNS**: Del inglés, Domain Name System.

- Realización de simulaciones de redes de colas, en las cuáles, se apliquen distribuciones GP truncadas para comprobar los efectos que produce respecto a las GP “ideales”.

REFERENCIAS

[1] Sally Floyd and Vern Paxson, **Dificulties in Simulating the Internet**, IEEE/ACM Transaction on Networking, vol. 9, no. 4, pp. 392-403, Aug. 2001.

[2]Vern Paxson and Sally Floyd, **Wide area traffic: the failure of Poisson modeling**, IEEE/ACM Transactions on Networking (TON) Volume 3, Pages: 226 – 244, 1995.

[3] Kihong Park, Gi Tae Kim, and Mark E. Crovella. **On the relationship between file sizes, transport protocols, and self-similar network traffic**. In Proceedings of the Fourth International Conference on Network Protocols (ICNP'96), pages 171-180, October 1998.

[4] Walter Willinger, Murad S. Taqqu, Robert Sherman, and Daniel V. Wilson. **Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level.** IEEE/ACM Transactions on Networking. 5(1):71-86. February 1997

[5] Kihong Park, Gi Tae Kim, Mark E. Crovella. **The effects of traffic self-similarity on TCP performance.** Technical report, Boston University Computer Science Departament.

[6] RedIRIS,"Qué es RedIRIS", **http://www.rediris.es/rediris/index.en.html.**

[7] José Luis García Dorado, Javier Aracil Rico**, Analysis and characterization of Internet traffic measurements: The case of RedIRIS,** June 2008.

[8] Mark E. Crovella, Murad S. Taqqu and Azer Bestavros, **Heavy-Tailed Probability Distributions in the World Wide Web**, In A Practical Guide To Heavy Tails, editors R. J. Adler, R.E. Feldman, M. S. Taqqu. Chapter 1, pages 3-26, Chapman and Hall, 1998.

[9] F. Hernández-Campos, J. S. Marron, G. Samorodnitsky and F. D. Smith, **Variable Heavy Tail Duration in Internet Traffic, Part I: Understanding Heavy Tails**, Proceedings of the 40th Allerton Conference in Communications, Control and Computing, October, 2002.

[10] F. Hernández-Campos, J. S. Marron, G. Samorodnitsky and F. D. Smith, **Variable Heavy Tail Duration in Internet Traffic, Part II: Theoretical Implications**, Proceedings of the 40th Allerton Conference in Communications, Control and Computing, October, 2002.

[11] Allen B. Downey, **The structural cause of file size distributions***,* MASCOTS '01*, 2001.* Available at http://rocky.wellesley.edu/downey/filesize/

[12] Allen B. Downey, **Evidence for long-tailed distributions in the internet**, ACM SIGCOMM Internet Measurement Workshop, November 2001.

[13] Allen B. Downey, **Lognormal and Pareto Distributions in the Internet***,* Computer Communications, May 2006.

[14] Mark E. Crovella and Azer Bestavros, **Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes**, IEEE/ACM Transactions on Networking, 5(6):835-846, 1997.

[15] Jan Hannig, J. S. Marron, Gennady Samorodnitsky, F. D. Smith, **Log-normal durations can give long range dependence**, Mathematical statistics and applications: Festschrift for Constance van Eeden (Beachwood, OH: Institute of Mathematical Statistics, 2003) pages 333-344

[16] M. Mitzenmacher, **A Brief History of Generative Models for Power Law and Lognormal Distributions**, Proceedings of the Thirty-Ninth Annual Allerton Conference on Communication, Control, and Computing, pages 182-191, 2001

[17] W. Gong, Y.Liu, V. Misra, and D. Towsley. **On the tails of Web file size distributions,** Proceedings of the Thirty-Ninth Annual Allerton Conference on Communication, Control, and Computing, pages 192-201, 2001.

[18] Enrique Castillo and Ali S. Hadi, **Fitting the Generalized Pareto Distribution to Data**, Journal of the American Statistical Association, Vol. 92, No. 440 (Dec., 1997), pp. 1609-1620.

[19] Mark E. Crovella and Lester Lipsky, **Long-Lasting Transient Conditions in Simulations with Heavy-Tailed Workloads**, Proceedings of the 1997 Winter Simulation Conference. pp.1005--1012.

GLOSARIO

Bufferring

Almacenamiento temporal de paquetes en las colas de los elementos de red.

Burst

Periodo de máxima actividad relativa dentro una transmisión de red, denominado ráfaga.

Caching

Almacenamiento en cache. Elemento de almacenamiento de datos transparente cuya funcionalidad es entregarlos de forma más rápida en futuras solicitudes.

CCDF

Función de distribución acumulada complementaria. Del inlgés, Complementary Cumulative Distribution Function.

CDF

Función de distribución acumulada. Del inlgés, Cumulative Distribution Function.

DNS

ECCDF

Función de distribución acumulada complementaria empírica, que se obtiene a partir de muestras. Del inglés, Empirical Complementary Cumulative Distribution Function.

ECDF

Función de distribución acumulada empírica, que se obtiene a partir de muestras. Del inglés, Empirical Cumulative Distribution Function.

EPM

Algoritmo o método que sirve para obtener los parámetros de la distribución generalizada de Pareto a partir de muestras, basándose en la utilización de percentiles. Del inglés, Elemental Percentile Method.

GPD

Distribución generalizada de Pareto. Del inglés, Generalized Pareto Distribution.

Ginni

Nombre del coeficiente que mide el nivel de desigualdad relativa que posee una variable aleatoria, a partir de la curva de Lorenz.

HPCN-UAM

Grupo de investigación de redes de computadoras de la UAM. Del inglés, High Preformance Computer and Networking.

HTTP

Protocolo utilizdo en cada transacción de la Worl Wide Web. Utiliza por defecto el puerto 80. Del inglés, Hypertext Transfer Protocol.

HTTPS

Protocolo basado en HTTP, destinado a la transferencia seguda de datos de hipertexto, es decir, es la versión segura de HTTP. Utiliza por defecto el puerto 443. Del inglés, Hypertext Transfer Protocol Secure.

Hurst

Nombre que recibe el parámetro que mide el nivel de autosimilaridad de fenómenos estocásticos.

IP

Protocolo no orientado a conexión para la comunicación de datos a través de una red de paquetes conmutados. Protocolo principal de la capa de red de la pila TCP/IP. Del inglés, Internet Protocol.

ISP

Proveedor de servicios de Internet. Del inglés, Internet Service Provider.

LAN

Red de computadoras de área local. Del inglés, Local Area Network.

LD test

Método que trata de determinar si una secuencia de variables aleatorias pertenece al dominio de atracción de distribuciones estables con varianza infinita. Del inglés, Limit Distribution test.

LN

Distribución log-Normal.

Log-likelihood

Función que se maximiza en el método de estimación por máxima verosimilitud (MLE).

Lorenz

Nombre de la curva que representa la distribución relativa de una variable aleatoria dentro de un dominio.

LRD

Fenómeno observado en el tráfico denominado dependencia a larga escala. Del inglés, Long Range Dependence.

MLE

Estimación por máxima verosimilitud, método estadístico de estimación paramétrica. Del inglés, Maximum Likelihood Estimation.

MOM

Método de los momento, método estadístico de estimación paramétrica. Del inglés, Method Of Moments.

Overfitting

Sobreajuste. Ocurre en estadística cuando un modelo es excesivamente complejo, posee demasiados grados de libertad en relación con la cantidad de datos disponible.

PPD

Distribución Pura de Pareto. Del inglés Pure Pareto Distribution.

PWM

Método estadístico de estimación paramétrica a partir de los momentos de la variable aleatoria. Del inglés, Probability Weighted Moemnts.

QoS

Calidad de servicio. Del inglés, Quality of Service.

QQ-plot

Representación gráfica de los cuantiles de una variable aleatoria teórica o empírica frente a los de otra. Del inglés, Quantil-Quantil plot.

RedIRIS

RedIRIS es la red española para Interconexión de los Recursos InformáticoS de las universidades y centros de investigación.

SS

Fenómeno estocástico de autosimilaridad. Del inglés, Self-Similarity.

TCP

Protocolo orientado a conexión fiable de extremo a extremo. Protocolo principal de la capa de transporte de la pila TCP/IP. Del inglés, Transmission Control Protocol.

UDP

Protocolo no orientado a conexión fiable de extremo a extremo basado en el intercambio de datagramas. Protocolo perteneciente a la capa de transporte de la pila TCP/IP. Del inglés, User Datagram Protocol.

WAN

Red de computadoras de área extensa. Del inglés, Wide Area Network.

Wobble

Oscilación sistemática que se produce en la distribución de las variables aleatorias que caracterizan distintas métricas del tráfico de red.

WWW

Sistema de documentos de hipertexto o hipermedios enlazados y accesibles a través de Internet.

PRESUPUESTO

- **EJECUCIÓN MATERIAL**

Compra de ordenador personal (Software incluido).....2.000 €

Alquiler de impresora láser durante 6 meses.....50 €

Material de oficina.....150 €

Total de ejecución material.....2.200 €

- **GASTOS GENERALES**

16 % sobre Ejecución Material.....352 €

- **BENEFICIO INDUSTRIAL**

6 % sobre Ejecución Material.....132 €

- **HONORARIOS PROYECTO**

640 horas a 15 € / hora.....9600 €

- **MATERIAL FUNGIBLE**

Gastos de impresión.....60 €

Encuadernación.....200 €

- **SUBTOTAL DEL PRESUPUESTO**

Subtotal Presupuesto.....12060 €

- **I.V.A. APLICABLE**

18% Subtotal Presupuesto……..2170.8 €

- **TOTAL PRESUPUESTO**

Total Presupuesto……..14231.8 €

Madrid, noviembre de 2010

El Ingeniero Jefe de Proyecto

Fdo.: Gonzalo Polo Vera

Ingeniero Superior de Telecomunicación

PLIEGO DE CONDICIONES

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, caracterización de flujos a partir de trazas de tráfico reales, aplicación a RedIRIS. En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

Supuesto que la utilización industrial de los métodos recogidos en el presente proyecto ha sido decidida por parte de la empresa cliente o de otras, la obra a realizar se regulará por las siguientes:

CONDICIONES GENERALES

1. La modalidad de contratación será el concurso. La adjudicación se hará, por tanto, a la proposición más favorable sin atender exclusivamente al valor económico, dependiendo de las mayores garantías ofrecidas. La empresa que somete el proyecto a concurso se reserva el derecho a declararlo desierto.

2. El montaje y mecanización completa de los equipos que intervengan será realizado totalmente por la empresa licitadora.

3. En la oferta, se hará constar el precio total por el que se compromete a realizar la obra y el tanto por ciento de baja que supone este precio en relación con un importe límite si este se hubiera fijado.

4. La obra se realizará bajo la dirección técnica de un Ingeniero Superior de Telecomunicación, auxiliado por el número de Ingenieros Técnicos y Programadores que se estime preciso para el desarrollo de la misma.

5. Aparte del Ingeniero Director, el contratista tendrá derecho a contratar al resto del personal, pudiendo ceder esta prerrogativa a favor del Ingeniero Director, quien no estará obligado a aceptarla.

6. El contratista tiene derecho a sacar copias a su costa de los planos, pliego de condiciones y presupuestos. El Ingeniero autor del proyecto autorizará con su firma las copias solicitadas por el contratista después de confrontarlas.

7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirvió de base para la contratación, a las modificaciones autorizadas por la superioridad o a las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con arreglo a los cuales, se harán las modificaciones y la valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados. Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna clase, salvo en los casos de rescisión.

8. Tanto en las certificaciones de obras como en la liquidación final, se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.

9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.

10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata, se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos precios convenidos por uno u otro procedimiento, se sujetarán siempre al establecido en el punto anterior.

11. Cuando el contratista, con autorización del Ingeniero Director de obras, emplee materiales de calidad más elevada o de mayores dimensiones de lo estipulado en el proyecto, o sustituya una clase de fabricación por otra que tenga asignado mayor precio o ejecute con mayores dimensiones cualquier otra parte de las obras, o en general, introduzca en ellas cualquier modificación que sea beneficiosa a juicio del Ingeniero Director de obras, no tendrá derecho sin embargo, sino a lo que le correspondería si hubiera realizado la obra con estricta sujeción a lo proyectado y contratado.

12. Las cantidades calculadas para obras accesorias, aunque figuren por partida alzada en el presupuesto final (general), no serán abonadas sino a los precios de la contrata, según las condiciones de la misma y los proyectos particulares que para ellas se formen, o en su defecto, por lo que resulte de su medición final.

13. El contratista queda obligado a abonar al Ingeniero autor del proyecto y director de obras así como a los Ingenieros Técnicos, el importe de sus respectivos honorarios facultativos por formación del proyecto, dirección técnica y administración en su caso, con arreglo a las tarifas y honorarios vigentes.

14. Concluida la ejecución de la obra, será reconocida por el Ingeniero Director que a tal efecto designe la empresa.

15. La garantía definitiva será del 4% del presupuesto y la provisional del 2%.

16. La forma de pago será por certificaciones mensuales de la obra ejecutada, de acuerdo con los precios del presupuesto, deducida la baja si la hubiera.

17. La fecha de comienzo de las obras será a partir de los 15 días naturales del replanteo oficial de las mismas y la definitiva, al año de haber ejecutado la provisional, procediéndose si no existe reclamación alguna, a la reclamación de la fianza.

18. Si el contratista al efectuar el replanteo, observase algún error en el proyecto, deberá comunicarlo en el plazo de quince días al Ingeniero Director de obras, pues transcurrido ese plazo será responsable de la exactitud del proyecto.

19. El contratista está obligado a designar una persona responsable que se entenderá con el Ingeniero Director de obras, o con el delegado que éste designe, para todo relacionado con ella. Al ser el Ingeniero Director de obras el que interpreta el proyecto, el contratista deberá consultarle cualquier duda que surja en su realización.

20. Durante la realización de la obra, se girarán visitas de inspección por personal facultativo de la empresa cliente, para hacer las comprobaciones que se crean oportunas. Es obligación del contratista, la conservación de la obra ya ejecutada hasta la recepción de la misma, por lo que el deterioro parcial o total de ella, aunque sea por agentes atmosféricos u otras causas, deberá ser reparado o reconstruido por su cuenta.

21. El contratista, deberá realizar la obra en el plazo mencionado a partir de la fecha del contrato, incurriendo en multa, por retraso de la ejecución siempre que éste no sea debido a causas de fuerza mayor. A la terminación de la obra, se hará una recepción provisional previo reconocimiento y examen por la dirección técnica, el depositario de efectos, el interventor y el jefe de servicio o un representante, estampando su conformidad el contratista.

22. Hecha la recepción provisional, se certificará al contratista el resto de la obra, reservándose la administración el importe de los gastos de conservación de la misma hasta su recepción definitiva y la fianza durante el tiempo señalado como plazo de garantía. La recepción definitiva se hará en las mismas condiciones que la provisional, extendiéndose el acta correspondiente. El Director Técnico propondrá a la Junta Económica la devolución de la fianza al contratista de acuerdo con las condiciones económicas legales establecidas.

23. Las tarifas para la determinación de honorarios, reguladas por orden de la Presidencia del Gobierno el 19 de Octubre de 1961, se aplicarán sobre el denominado en la actualidad “Presupuesto de Ejecución de Contrata” y anteriormente llamado ”Presupuesto de Ejecución Material” que hoy designa otro concepto.

CONDICIONES PARTICULARES

La empresa consultora, que ha desarrollado el presente proyecto, lo entregará a la empresa cliente bajo las condiciones generales ya formuladas, debiendo añadirse las siguientes condiciones particulares:

1. La propiedad intelectual de los procesos descritos y analizados en el presente trabajo, pertenece por entero a la empresa consultora representada por el Ingeniero Director del Proyecto.

2. La empresa consultora se reserva el derecho a la utilización total o parcial de los resultados de la investigación realizada para desarrollar el siguiente proyecto, bien para su publicación o bien para su uso en trabajos o proyectos posteriores, para la misma empresa cliente o para otra.

3. Cualquier tipo de reproducción aparte de las reseñadas en las condiciones generales, bien sea para uso particular de la empresa cliente, o para cualquier otra aplicación, contará con autorización expresa y por escrito del Ingeniero Director del Proyecto, que actuará en representación de la empresa consultora.

4. En la autorización se ha de hacer constar la aplicación a que se destinan sus reproducciones así como su cantidad.

5. En todas las reproducciones se indicará su procedencia, explicitando el nombre del proyecto, nombre del Ingeniero Director y de la empresa consultora.

6. Si el proyecto pasa la etapa de desarrollo, cualquier modificación que se realice sobre él, deberá ser notificada al Ingeniero Director del Proyecto y a criterio de éste, la empresa consultora decidirá aceptar o no la modificación propuesta.

7. Si la modificación se acepta, la empresa consultora se hará responsable al mismo nivel que el proyecto inicial del que resulta el añadirla.

8. Si la modificación no es aceptada, por el contrario, la empresa consultora declinará toda responsabilidad que se derive de la aplicación o influencia de la misma.

9. Si la empresa cliente decide desarrollar industrialmente uno o varios productos en los que resulte parcial o totalmente aplicable el estudio de este proyecto, deberá comunicarlo a la empresa consultora.

10. La empresa consultora no se responsabiliza de los efectos laterales que se puedan producir en el momento en que se utilice la herramienta objeto del presente proyecto para la realización de otras aplicaciones.

11. La empresa consultora tendrá prioridad respecto a otras en la elaboración de los proyectos auxiliares que fuese necesario desarrollar para dicha aplicación industrial, siempre que no haga explícita renuncia a este hecho. En este caso, deberá autorizar expresamente los proyectos presentados por otros.

12. El Ingeniero Director del presente proyecto, será el responsable de la dirección de la aplicación industrial siempre que la empresa consultora lo estime oportuno. En caso contrario, la persona designada deberá contar con la autorización del mismo, quien delegará en él las responsabilidades que ostente.