Análisis de la relación entre la intensidad del tráfico de datos y el número de alumnos en universidades españolas

Ignacio Gutiérrez[1], Jesús Martínez[1], Pedro María Santiago[1],
José Luis García-Dorado[1], Jorge E. López de Vergara[1], Javier Aracil[1]
Francisco Jesús Monserrat[2], Esther Robles[2] y Tomás P. de Miguel[2].

[1] Networking Research Group
Escuela Politécnica Superior, Universidad Autónoma de Madrid
Calle Francisco Tomás y Valiente, 11, 28049 Madrid
E-mail: proyecto.dior@uam.es

[2] RedIRIS
Edificio Bronce, Plaza de Manuel Gómez Moreno, s/n
28020 Madrid

Abstract *The capacity planning of existing telecommunication networks is a fundamental value in the design of the forthcoming Internet that satisfies the ever-increasing user demands. However, the capacity planning of Internet links remains challenging, and not many planning tools have been proposed. This can be partly due to the fact that Internet traffic exhibits extreme variability. Furthermore, as new technologies are being incorporated to the network, previous capacity planning analysis become outdated. This paper provides an approach to analyze the relationship between the data traffic and the number of students in the RedIRIS network. This analysis can be useful to estimate the needed bandwidth based on the number of users of the network. For this, the network traffic of several universities with similar number of students has been analyzed with different statistics tests.*

1. Introducción

El dimensionado de las redes de comunicaciones es una tarea esencial a la hora de su despliegue, de forma que se pueda planificar la capacidad necesaria de dicha red para atender las necesidades de sus usuarios. Sin embargo, dimensionar redes de datos que soporten tráfico de Internet no es una actividad que actualmente cuente con herramientas adecuadas, debido en parte a las características altamente variantes de dicho tráfico. Además, la aparición de nuevas aplicaciones y tecnologías de soporte hacen que dimensionados pasados no sean útiles a medida que evolucionan las redes.

En este contexto, el proyecto DIOR (Dimensionado de redes IP y redes Ópticas: aplicación a los accesos a la red académica española RedIRIS) pretende encontrar relaciones entre las características de una población y el uso de ancho de banda que realizan. Estas características pueden ser de muchos tipos: número de usuarios que forman la población, número de usuarios que tienen alguna característica en común, número de equipos que forman una red, datos de utilización de la red durante periodos largos de tiempo, etc. Se pretende identificar cuáles de estas posibles características son las que determinan de forma más directa el uso de la red. Dado que la red a analizar es la red académica española RedIRIS, se deben utilizar los patrones propios de este tipo de redes, es decir, número de alumnos, número de profesores, personal de administración y servicios, tipo de estudios ofertados por el centro (carreras técnicas, ciencias sociales,...), etc. Además se pretende que estas estimaciones valgan para caracterizar cuál es la evolución previsible del tráfico en dicha red. Con este análisis, se espera poder dimensionar dichos accesos de tal modo que su vida útil sea razonable, y permita un mejor ajuste del sobredimensionado de los mismos.

Este artículo presenta una primera aproximación dentro del citado proyecto. Se pretende determinar a partir de datos empíricos si el volumen de tráfico de datos es función únicamente del número de alumnos de las universidades. En concreto, en este estudio se analiza si el tráfico de datos medio por estudiante de las universidades es el mismo independientemente de la universidad. Para ello, se estudiará el volumen de datos de entrada real de cinco universidades españolas a las que llamaremos: Universidad 1, Universidad 2,..., Universidad 5, cada una de las cuales tiene una población en torno a los 10000 estudiantes (ver Tabla 1).

Tabla 1: Datos de población de las universidades estudiadas.

Universidad	alumnos
Universidad 1	8461
Universidad 2	12431
Universidad 3	12931
Universidad 4	11260
Universidad 5	10596

El resto del documento queda dividido de la siguiente manera: En la sección 2 se resume el estado del arte. En la siguiente sección se muestra la metodología seguida en el artículo. En la sección 4 se presentan los datos de tráfico disponibles, así como cuál ha sido su tratamiento. A continuación, en la sección 5 se muestran los resultados del análisis realizado. Por último, en la sección 6 se exponen algunas conclusiones y se presentan los trabajos actuales y futuros dentro del contexto de este proyecto.

2. Estado del arte

El dimensionado de redes es un tema al que se le ha prestado relativo poco interés en los últimos años a pesar de su importancia. En general la solución al mismo ha consistido en sobredimensionar la red de forma notable, como se demuestra en [1], o aceptar estimaciones muy imprecisas como válidas. El dimensionando de redes es un tema amplio que abarca muchos campos y que no puede ser tratado como un único bloque. De hecho, dentro de esta temática general podemos encontrar diversas aproximaciones con dispar desarrollo.

En primer lugar podríamos citar el desarrollo de técnicas para la correcta monitorización de redes como las propuestas en [2]. Por otro lado están los esfuerzos orientados en el desarrollo de técnicas que permitan capturar de forma precisa estadísticas del tráfico que atraviesa en red. Ejemplo de esto último son los trabajos presentados en [3] y [4]. Éstos presentan distintas técnicas capaces de capturar correctamente las características de los flujos que atraviesan una red, aprovechando que la información que incluye un flujo puede ser muy detallada, en comparación con otras herramientas como MRTG [5], que apenas aportan la cantidad de bytes que atraviesan un nodo (los flujos son capaces de capturar la direcciones IP, puertos, protocolos, etc.). A partir de estos conceptos en [6] son capaces de estimar la matriz de tráfico de una red de forma muy precisa. En [7] también se presentan herramientas para estimar la matriz de tráfico basada, no en flujos, sino en la cantidad de tráfico que atraviesa un nodo. Sin embargo, los resultados son, aparentemente, muy inferiores.

Dentro del dimensionado de redes también tienen cabida los trabajos que pretenden estimar la utilización futura de una red basándose en los datos históricos de utilización de la misma durante un largo periodo de tiempo. Este tipo de predicciones ya han sido utilizadas frecuentemente en otros campos, para estimar la demanda de electricidad o petróleo. En el caso de las telecomunicaciones se pueden citar a este respecto artículos como [8].

Aunque este artículo se basa en los trabajos anteriormente citados, el objetivo que se plantea aquí es distinto. Pretendemos ser capaces de estimar el ancho de banda que necesita una universidad a partir de sus características. En este sentido apenas se han encontrado trabajos. Se puede citar a [9], donde se utiliza una relación lineal entre el incremento de usuarios y el uso de ancho de banda. Sin embargo esta relación es insuficiente, como se verá más adelante en este articulo.

3. Metodología

Para llevar a cabo el presente estudio se han empleado los siguientes métodos estadísticos: ANOVA (*ANalysis Of VAriance*, Análisis de Varianza) y Kruskal-Wallis, ambos para decidir si las muestras aleatorias de k poblaciones tienen o no la misma media. El test ANOVA exige que las muestras sigan una distribución normal y que tengan la misma varianza. Para comprobarlo, se realizan respectivamente el test de Shapiro-Wilks y el test de Levene [10, 11].

3.1. Test ANOVA

El test ANOVA está basado, como su nombre indica, en el análisis de las varianzas de un conjunto de poblaciones. Sean μ_1, μ_2, $\ldots \mu_k$ las medias de las k poblaciones. Entonces se tiene el siguiente contraste de hipótesis:

$$\begin{aligned} H_0 &= \mu_1 = \mu_2 = ... = \mu_k \\ H_1 &= \text{al menos dos de ellas difieren} \end{aligned}$$

La característica común de las poblaciones que se estudian se llama factor, que tiene varios niveles representados por las poblaciones. En nuestro caso, el factor es la comunidad universitaria y los niveles son cada universidad en particular (Universidad 1, Universidad 2,...). La comparación entre poblaciones se hace escogiendo una cantidad numérica, llamada variable de respuesta, que se mide para cada elemento de la población, siendo en nuestro caso el tráfico de datos medio diario por estudiante. Para determinar si las medias de la población difieren, ANOVA compara la variación de la media de las poblaciones teniendo en cuenta la variabilidad interna en cada muestra. El test hace una relación entre los dos tipos de variaciones.

$$\text{test estadistico} = \frac{\text{variación entre muestras}}{\text{variación en cada muestra}}$$

Este estadístico sigue una distribución F de Schnedecor cuando la hipótesis nula es correcta. Los valores grandes de F rechazan hipótesis nulas. Para poder aplicar el test ANOVA es necesario que se cumplan las dos hipótesis siguientes:

1. Las varianzas de las k poblaciones son iguales.[1]
2. Las k poblaciones siguen una distribución normal.[2]

3.2. Test de Shapiro-Wilks

Como ya se ha anticipado, el test de Shapiro-Wilks decide si una muestra sigue una distribución normal o no. Se tiene el siguiente contraste de hipótesis:

H_0 = La muestra sigue una distribución normal.
H_1 = La muestra no la sigue.

Dada la muestra aleatoria simple de tamaño n $\{x_1, x_2, \ldots, x_n\}$ que se supondrá ordenada de mayor a menor, se calcula el siguiente estadístico de contraste:

$$W = \frac{1}{ns^2}\left(\sum_{i=1}^{h} a_{in} x_{n-i+1} - x_i\right)^2$$

donde s^2 es la varianza muestral,

$$h = \begin{cases} \frac{n}{2} \text{ si } n \text{ es par} \\ \\ \frac{n-1}{2} \text{ si } n \text{ es impar} \end{cases}$$

y las a_{in} se encuentran tabuladas en los manuales. Se rechaza la normalidad cuando el estadístico es menor que el valor de las tablas de la bibliografía.

3.3. Test de Levene

El test de Levene se utiliza para comprobar que existe igualdad de varianzas en las k muestras. Se tiene el siguiente contraste de hipótesis:

$H_0 = \sigma_1 = \sigma_2 = \ldots = \sigma_k$
$H_1 = \sigma_i \neq \sigma_j$para al menos un par (i, j)

Dada una muestra aleatoria Y con N elementos, dividida en k grupos, se calcula el siguiente estadístico:

$$L = \frac{(N-k)\sum_{i=1}^{k} n_i(\bar{Z}_i - \bar{Z})}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Z_{ij} - \bar{Z}_i)^2}$$

donde:

N_i es el tamaño del grupo i
$Z_{ij} = |X_{ij} - \bar{Y}_i|$
$\bar{X}_i$ es la media del grupo i
$\bar{Z}_i$ es la media de los grupos Z_{ij}
$\bar{Z}$ es la media de todos los Z_{ij}

El test rechaza la hipótesis nula si

$$L > F_{\alpha;k-1;n-k}$$

donde α es el nivel de significación y F la distribución de Schnedecor.

3.4. Test de Kruskal-Wallis

El test de Kruskal-Wallis es la alternativa no paramétrica al test ANOVA. Decide si las k muestras provienen de la misma población, es decir, siguen la misma distribución, y tienen la misma media y la misma varianza. Sin embargo, tan sólo requiere que las muestras sigan distribuciones contínuas, lo cual es fácilmente comprobable con el test de Shapiro-Wilks. El contraste de hipótesis es el siguiente:

H_0 = Las k muestras provienen de la misma población.
H_1 = Al menos dos provienen de poblaciones distintas.

Se supondrá que todas las observaciones (en total N) de las k muestras se encuentran ordenadas de menor a mayor y a cada una de las observaciones se le asigna un rango (1 para la menor, 2 para la siguiente, ... y N para la mayor). El estadístico es el siguiente:

$$H = \frac{12}{N(N+1))}\sum_{i=1}^{k}\frac{R_i^2}{N_i} - 3(N+1)$$

donde R_i es la suma de los rangos de las observaciones correspondientes a la muestra i, desde $i = 1$ hasta k y N_i es el tamaño del grupo i. Se rechaza H_0 cuando $H > \chi^2_{k-1,1-\alpha}$.

4. Análisis

Las trazas analizadas han sido extraídas de los registros de los routers de RedIRIS para los nodos de entrada de cada universidad. Dado que tan solo se tienen datos del enlace de entrada/salida de los nodos, se ha decidido elegir universidades de tamaño pequeño o medio para asegurarnos de que los estudiantes son los principales generadores de tráfico. En otros casos, por ejemplo, podrían existir centros de investigación que replicaran gran cantidad de datos a centros de cálculo externos a la universidad, alterando así la muestra. Las muestras han sido elegidas tomando el tráfico de datos

[1] Para comprobarlo puede utilizarse el test de Levene.
[2] Para comprobarlo puede utilizarse el test de Shapiro-Wilks o un contraste χ^2.

diario medio por estudiante de todos los martes y miércoles entre el 10 de febrero de 2004 y el 29 de junio de 2005, excluyendo vacaciones de verano y de navidad, resultando así alrededor de 84 muestras por universidad. No han sido elegidos el resto de días por ser un tráfico menos constante debido a fiestas de carácter autonómico o local (algunas universidades no tenían clase alguno de esos días), o coincidían en algunas horas con el fin de semana (el tráfico suele decaer a mitad del viernes y no se recupera hasta el lunes siguiente). En las figuras 1 y 2 se pueden ver las muestras de la Universidad 1 y la Universidad 3. El hueco central sin puntos corresponde al periodo estival de vacaciones del año 2004. Se observa que existe una serie de días en los que el tráfico es considerablemente inferior al resto. Esto se puede deber, por ejemplo, a caídas de algún enlace. Estos puntos que se encuentran muy alejados de la media son los llamados *outliers*. Se realizó un estudio de estos puntos y su eliminación de la muestra no cambiaba los resultados obtenidos. Por ello, no se detalla dicho tratamiento en este artículo.

5. Resultados

A continuación se presentan los resultados de la aplicación de los métodos estadísticos presentados en la sección 3 a los datos de tráfico de las distintas universidades.

5.1. Aplicación del test de Shapiro-Wilks

El primer contraste realizado ha consistido en comprobar, como exige el test ANOVA anteriormente referenciado, que las poblaciones siguen una distribución normal. Para ello, se ha aplicado el test estadístico de Shapiro-Wilks (ver Tabla 2).

En el test de Shapiro-Wilks, se acepta H_0 al nivel de significación del 5 % para las universidades 1, 2, 4 y 5 rechazándo H_0 para la universidad 3. Por tanto, todas las poblaciones a excepción de la de la Universidad 3 siguen una distribución normal y sobre ésas, por tanto, se ha llevado a cabo el resto del análisis.

Tabla 2: Resultado Test Shapiro-Wilks.

Universidad	**Estadístico**	**Acepta H_0**
Universidad 1	0,9752	ACEPTA
Universidad 2	0,9791	ACEPTA
Universidad 3	0,917	RECHAZA
Universidad 4	0,9709	ACEPTA
Universidad 5	0,9786	ACEPTA

5.2. Aplicación del test de Levene

A continuación se ha realizado el test de Levene para verificar la otra hipótesis necesaria para el test ANOVA, igualdad de varianzas. Para cualquier nivel de confianza ha resultado que las varianzas son distintas si tomamos todas las universidades juntas (estadístico $F_{0,05;3;332} = 23{,}1547$ con probabilidad asociada para ese estadístico inferior a 10^{-4}, menor que 0,05, por lo que se rechaza H_0). Observando las varianzas muestrales (ver Tabla 3), para algunas universidades eran de un factor de magnitud del doble que en otras, lo que es un indicativo claro, aparte del test de Levene, de que la varianza no es homogénea para todas las poblaciones.

Tabla 3: Varianzas de muestras sin normalizar para el Test de Levene.

Universidad	**Varianza muestral**
Universidad 1	274,6726
Universidad 2	1608,6333
Universidad 4	256,9303
Universidad 5	557,2590

Una explicación de que las varianzas sean tan heterogéneas es que unas universidades tienen más ancho de banda que otras. Para evitar estos efectos, se han normalizado las muestras dividiendo el tráfico entre el ancho de banda del canal. De esta manera, se elimina el caso en que un canal con un ancho de banda grande tenga unas variaciones mayores que otro con un ancho de banda pequeño. Se vuelve a realizar el test, mostrando el resultado en la Tabla 4. El estadístico resultante es $F_{0,05;3;332} = 90{,}9123$, con probabilidad asociada para ese estadístico inferior a 10^{-4}, menor que 0,05, por lo que se rechaza H_0.

Tabla 4: Varianzas de muestras normalizadas para el Test de Levene.

Universidad	**Varianza muestral**
Universidad 1	0,0114
Universidad 2	0,1609
Universidad 4	0,0006
Universidad 5	0,0003

5.3. Aplicación de los test ANOVA y de Kruskal-Wallis

Dado que no se cumplen las hipótesis del test ANOVA, se ha realizado el test de Kruskal-Wallis, que también comprueba que las poblaciones son similares sin exigir dichas hipótesis. Para un nivel de significación del 0,05, se rechaza la hipótesis nula. Esto es, que todas las muestras no provienen de la misma población. En la Fig. 3 puede verse la fuerte variación entre las poblaciones. Se representa con una línea horizontal (separando las formas poligonales) la mediana, y un polígono hexagonal delimita los cuartiles superior e inferior.

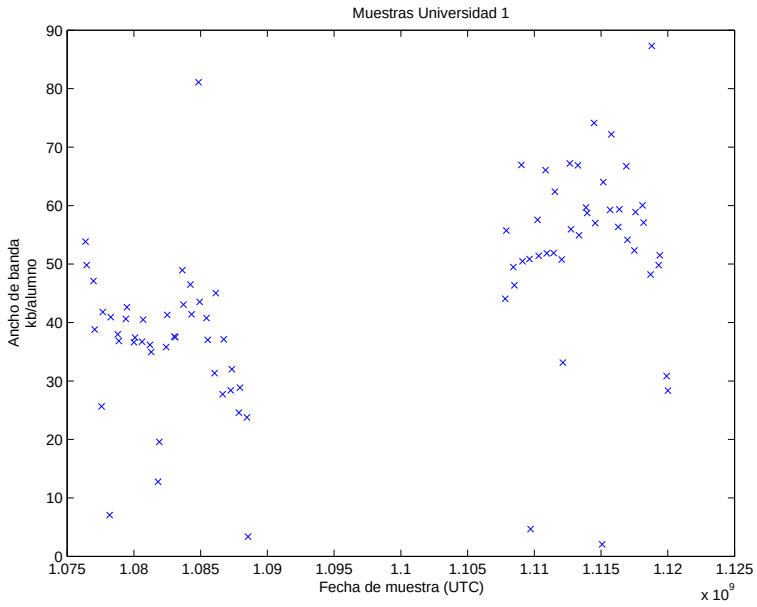

Figura 1: Tráfico por alumno en la Universidad 1 durante el periodo de tiempo estudiado.

Por último la línea punteada sigue hasta unir todos los valores. Se aprecia así que las únicas poblaciones que contienen una media similar son Universidad 4 y Universidad 5. De hecho, solamente estas dos muestras pasan de manera favorable el test ANOVA. Dado que son solo dos poblaciones, se ha considerado que este hecho no es suficiente para sacar conclusiones.

Los resultados del test se muestran en la Tabla 5: la primera columna representa la suma de los cuadrados (SS, *Sum of Squares*); la segunda los grados de libertad del test (df, *degrees of freedom*), la tercera la media cuadrática (MS, *Mean Square*) y la cuarta el valor χ^2. Se obtiene una probabilidad asociada al estadístico de 0 y, por tanto, se rechaza H_0. Luego, las poblaciones no siguen la misma distribución.

6. Conclusiones y trabajo futuro

A partir del análisis realizado se ha concluido que el tráfico medio por estudiante no es el mismo para todas las universidades estudiadas, o lo que es lo mismo, que otros factores deben influir determinantemente en el tráfico medio de las universidades. Podrían ser la proporción de profesores y PAS respecto al resto de miembros de la comunidad universitaria, el número de terminales con acceso a la Red, el tipo de titulaciones que se imparten en cada universidad, el número y tipo de centros de investigación y/o de cálculo, o incluso la capacidad de los enlaces y equipos de conmutación. En la Tabla 6 se muestra el número de profesores de cada universidad así como el porcentaje de carreras técnicas en las cinco universidades tomadas como modelo. Se comprueba que estos nuevos datos no son similares entre todas ellas cuando, recordemos, sí lo eran en cuanto al número de alumnos. Si aceptamos como razonable que aquellas con mayor número de carreras técnicas utilizan más el acceso a Internet, o, más evidente, que a mayor número de profesores existirá mayor demanda, se contrasta que la suposición de estimar el ancho de banda necesitado, únicamente, por el número de alumnos era insuficiente.

Esto sugiere una nueva investigación de mayor complicación teórica pues requiere el uso de funciones probabilísticas en varias variables. Este trabajo, ya en marcha, esta basado en la utilización de técnicas de análisis de datos multivariantes como lo son el PCA (*Principal Component Analisys*, Análisis de Componentes Principales), o el análisis multiresolución con *waveletes*. Ambas técnicas ya han sido utilizadas en estudios de extracción de características de redes, tal y como se indica en [8] o [13]. El Análisis de Componentes Principales o PCA permite detectar aquellos patrones que realmente determinan el comportamiento de la red reduciendo los datos al número de dimensiones óptimas, descartando aquellos que no superan un umbral mínimo de influencia. Por otro lado, técnicas como los *wavelets* eliminan redundancias temporales y permiten el tratamiento de gran cantidad de datos al submuestrear los mismo de forma óptima en el plano temporal y de frecuencias.

Una conclusión secundaria pero importante es que por lo general, para las poblaciones estudiadas, el tráfico medio por estudiante sigue una distribución normal, salvo en el caso de la Universi-

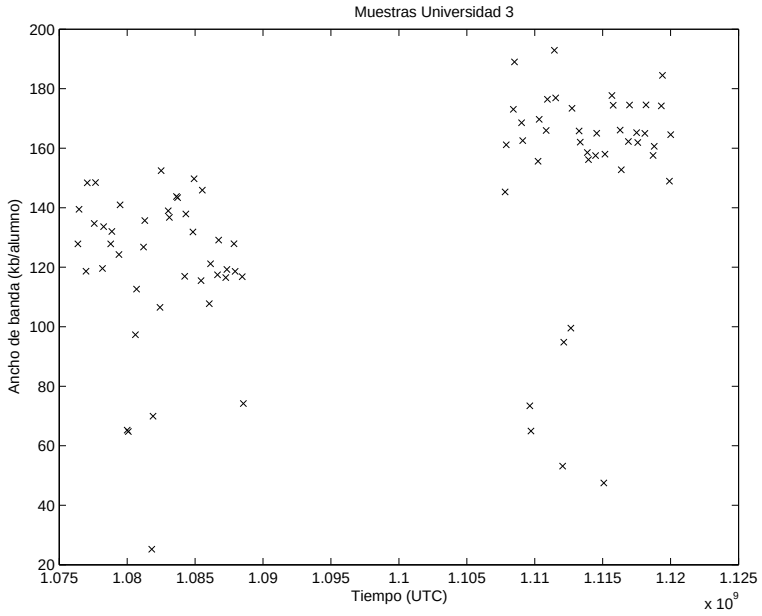

Figura 2: Tráfico por alumno en la Universidad 3 durante el periodo de tiempo estudiado.

Tabla 5: Resultados del test de Kruskal-Wallis.

	SS	**df**	**MS**	**Resultado χ^2**
Grupos	$3{,}50948 \cdot 10^6$	4	877371,2	238,17
Error	$2{,}66448 \cdot 10^6$	415	6420,4	
Total	$6{,}17397 \cdot 10^6$	419		

Tabla 6: Porcentaje de carreras técnicas y número de profesores por universidad [12]

Universidad	**Procentaje carreras técnicas**	**Número de profesores**
Universidad 1	49,7	651
Universidad 2	12,7	1118
Universidad 3	32,2	1030
Universidad 4	40,6	1108
Universidad 5	32,4	908

dad 3. Debido a que la mayor parte de tests estadísticos para variables continuas se basan en que la población siga una distribución normal el hallazgo puede ser importante para futuros estudios.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el Ministerio de Educación y Ciencia a través del proyecto DIOR (TEC2006-03246), y por la Comunidad de Madrid a través del programa de becas de excelencia.

Referencias

[1] Y. d'Halluin, P.A. Forsyth, and K.R. Vetzal, "Managing capacity for telecommunications networks under uncertainty," *IEEE/ACM Transactions on Networking (TON)*, vol. 10, pp. 579 – 587, 2002.

[2] D. López, J. López de Vergara, L. Bellido, and D. Fernandez, "Monitorización de una red académica mediante netflow," in *Actas de las XIV Jornadas Telecom I+D 2004, Madrid*, 2004.

[3] N. Duffield, C. Lund, and M. Thorup, "Estimating flow distributions from sampled flow statistics," *Transactions on Networking*, vol. 13, no. 5, pp. 933–946, Oct 2005.

[4] C. Estan and G. Varghese, "New directions in traffic measurement and accounting: Focusing on the elephants, ignoring the mice," *ACM Transactions on Computer Systems*, Ago 2003.

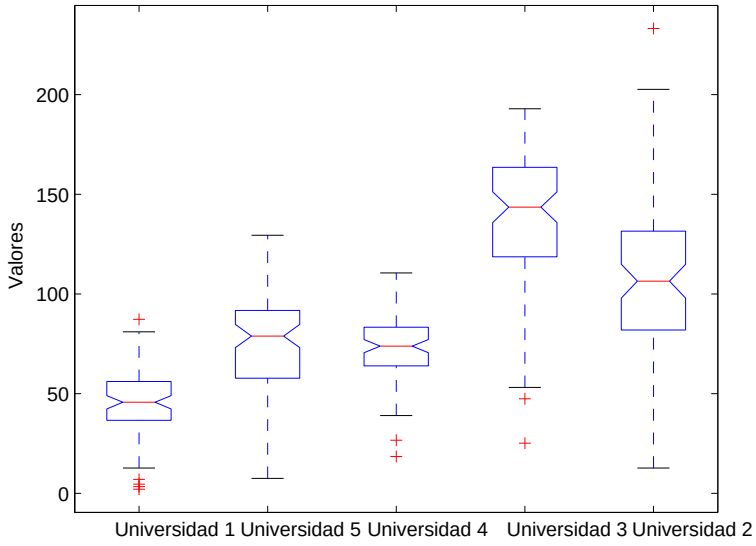

Figura 3: Gráfica de Kruskal-Wallis

[5] T Oetiker, "MRTG - the multi router traffic grapher," in *Proceedings of the Twelfth Systems Administration Conference (LISA .98)*, 1998.

[6] Anja Feldmann, Albert G. Greenberg, Carsten Lund, Nick Reingold, Jennifer Rexford, and Fred True, "Deriving traffic demands for operational ip networks: methodology and experience," *IEEE/ACM Transactions on Networking (TON)*, vol. 9, pp. 265 – 280, Jun 2001.

[7] N. Benameur and J. W. Roberts, "Traffic matrix inference in ip networks," in *10th International Telecommunication Network Strategy and Planning Symposium*, 2003.

[8] K. Papagiannaki, N. Taft, Zhi-Li Zhang, and C. Diot, "Long-term forecasting of internet backbone traffic," *IEEE Transactions on Neural Networks*, vol. 16, no. 5, pp. 1045–1124, Sep. 2005.

[9] R.E. Ahmed and S.H. Bakry, "New topology designs for the future expansion of the academic network of the gulf countries," *International Journal of Network Management*, vol. 7, pp. 18 – 32, Jan 1997.

[10] Nicholas R. Farnum Jay L. Devore, *Applied statistics for engineers and scientists*, Duxbury Pr., 2004.

[11] Daniel Peña Sánchez de Rivera, *Fundamentos de estadistica*, Ed. Alianza, 2001.

[12] Consejo de Coordinación Universitaria, "Estadística básica del personal al servicio de las universidades," 2006.

[13] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. Kolaczyk, and N. Taft, "Structural analysis of network traffic flows," in *Proceedings of ACM SIGMETRICS*, 2004.