UTILIZACIÓN COMBINADA DE MÉTODOS EXPLORATORIOS Y CONFIRMATORIOS PARA EL ANÁLISIS DE LA ACTIVIDAD ANTIBACTERIANA DE LA CEFALOSPORINA (PARTE I)

Yunier E. Tejeda Rodríguez*, Valia Guerra Ones**, Jesús E. Sánchez García** y Ramón Carrasco Velar
* Universidad de las Ciencias Informáticas, MIC
** Instituto de Cibernética, Matemática y Física, CITMA

ABSTRACT
The paper presents a strategy to reduce the dimension using the Robust Principal Component Analysis (PRCA) in combination with the matrix CUR decomposition. The results are applied in a practical problem for the fit of a regression model using partial least squares (PLS) so that in the modeling stage will have a deeper understanding of the sample. The first part of the work contains the problem statement and a brief description of the methods used. The second describes the results and gives a discussion of them.

KEYWORDS: Principal Component Analysis, Robust Principal Component Analysis, CUR Matrix Decomposition

MSC: 62P10

RESUMEN
En el trabajo se presenta una estrategia para reducir la dimensión mediante el uso del Análisis de Componentes Principales Robusto (APCR) en combinación con la descomposición matricial CUR. Los resultados se aplican en un problema práctico para el ajuste de un modelo de regresión mediante el uso de los mínimos cuadrados parciales (PLS) de manera tal que en la etapa de modelación se tenga un conocimiento más profundo de la muestra. La primera parte del trabajo contiene el planteamiento del problema, así como una breve descripción de los métodos empleados. En la segunda, se describen los resultados y se da una discusión de los mismos.

1. INTRODUCCIÓN

La bioinformática es una nueva disciplina que surge como respuesta al creciente aumento de los volúmenes de datos biológicos y estructurales. La misma se relaciona con la biología, la bioquímica, la química, la farmacología, la informática y las tecnologías de la información con vistas al análisis, organización y distribución de la información biológica. Esto ha propiciado grandes avances en las investigaciones biomédicas, como por ejemplo: en el diagnóstico, tratamiento y prevención de diversas enfermedades, lo cual incide en una mejoría de la calidad de vida. Es por eso que el desarrollo de la industria farmacéutica guarda una estrecha relación con la bioinformática, la cual permite la búsqueda, cada vez más eficiente, de nuevos fármacos más eficaces, específicos y con menores efectos secundarios, para el tratamiento de diversas patologías.
En los últimos años la industria farmacéutica ha reorientado sus investigaciones y prestado más atención a la utilización de métodos computacionales que relacionan la estructura química con la actividad biológica y que pueden dividirse en dos categorías: los Métodos SAR (**Escalona *et al.*,** 2008) y los Métodos QSAR (**Carrasco,** 2008).

Según Escalona ***et al.*** (2008), el método QSAR tradicional incluye el tratamiento estadístico de los datos por métodos multivariados, entre los cuales se encuentran el análisis de regresión, el análisis de conglomerados (Cluster analysis) y el análisis de componentes principales, entre otras técnicas estadísticas. En ellos se valora la actividad biológica como variable dependiente del conjunto de descriptores moleculares que constituyen las variables independientes.

Entre los problemas actuales de la modelación y el diseño de fármacos se encuentra el tratamiento de los compuestos antibacteriales. Se ha encontrado a nivel mundial que las bacterias desarrollan constantemente resistencia al empleo de los antibióticos. Por esa razón, es también necesaria la renovación del parque de medicamentos para combatir estas patologías infecciosas.

Las cefalosporinas son compuestos antibacteriales pertenecientes a la familia de los β-lactámicos. Estas se parecen a las penicilinas desde el punto de vista estructural. Según algunos autores como Frere ***et al.*** (1989), han planteado que el establecimiento de relaciones entre la estructura química y la actividad biológica de las cefalosporinas, (como compuestos del tipo de las β-lactamas), es algo imposible de realizar debido a la complejidad de dichas relaciones.

2. PLANTEAMIENTO DEL PROBLEMA

En esta investigación se plantea la necesidad de hacer uso de técnicas exploratorias con vista a lograr una evaluación preliminar de la calidad de los datos, así como de características de las variables, con respecto a su relación e importancia en el contexto de la descripción estructural de las cefalosporinas, de manera tal que en la etapa de modelación se tenga un conocimiento más profundo de la muestra.

2.1. Cefalosporinas

2.1.1. Generalidades

Las cefalosporinas son compuestos antibacteriales pertenecientes a la familia de los □-lactámicos. Todas las cefalosporinas se derivan de la cefalosporina C, un antibiótico natural producido por la cepa de *Cephalosporium acremonium* aislado por primera vez en 1945. Las cefalosporinas se parecen a las penicilinas (MOELLERING, 1995) en que ambas tienen un anillo □-lactámico, así como que en lugar del anillo de 5 miembros de tiazolidina, presentan un anillo de dihidrotiazina (Figura 1).

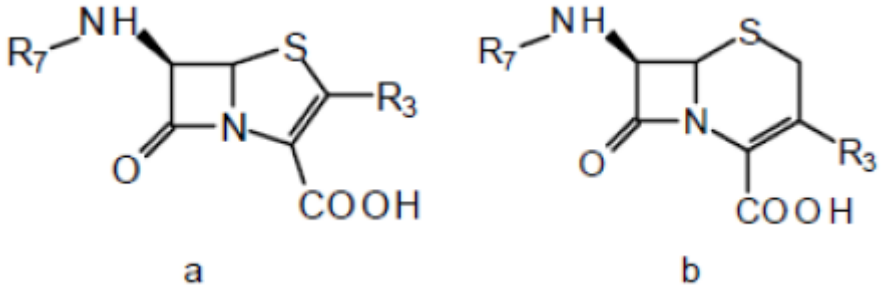

Figura 1. Esquema general de la estructura de a) penicilinas y, b) cefalosporinas

Las cefalosporinas pueden clasificarse según diferentes criterios, tales como su metabolismo y estabilidad a la acción de β-lactamasas (O'CALLAGHAN, 1979), la sustitución de la cadena lateral R2 (BRYSKIER, PROCYK & LABRO, 1990), sus propiedades farmacocinéticas (BRYSKIER *et al.*, 1990), (KARCHMER, 2000) o sus propiedades microbianas, principalmente relacionadas con su espectro antibacterial (BRYSKIER, PROCYK & LABRO, 1990). A partir de la combinación de los diferentes criterios, se han establecido grupos o generaciones de este antibiótico: El grupo I de las cefalosporinas incluye moléculas con la mayor actividad contra bacterias Gram positiva, el grupo II tiene la mayor actividad contra bacterias Gram negativa, el grupo III contra Pseudomonas aeruginosa y el grupo IV contra bacterias anaeróbicas (WILLIAMS, 1987).

2.1.2 Descripción de los datos

En esta investigación se tienen 104 compuestos clasificados en cuatro grupos (o generaciones) mencionados anteriormente. Como variables independientes se utilizaron 97 descriptores divididos en cinco grupos dados por: descriptores topológicos, topográficos, híbridos, químico-cuánticos y mixtos (TEJEDA, 2011). Por tanto se tiene una matriz de datos con 104 filas representando los compuestos y 97 columnas representando los descriptores.

2.1.3. Análisis de Componentes Principales Robusto

Una deficiencia del Análisis de Componentes Principales (ACP) clásico (MARDIA ***et al.,*** 1979) es que se debe tener mucho cuidado con los valores aberrantes (*outliers*) ya que estos son capaces de incrementar artificialmente la varianza haciendo que las componentes principales sean atraídas por ellos. Esto, sin duda, es

una característica no deseada en la reducción de la dimensión que aparece principalmente con la estimación clásica de las componentes principales. Por tal motivo, surge el método de ACP robusto (ROUSSEEUW, 1984) el cual determina las direcciones estimando la varianza por una manera robusta en lugar de la varianza clásica. **En VARMUZA & FILZMOSER** (2008) se resumen las características esenciales del ACP robusto que serán las pautas a seguir en este epígrafe.

Básicamente, se distinguen dos tipos de *outliers*: los puntos de palanca (*leverage*) y los *outliers* ortogonales**.** La Figura 2 muestra datos tridimensionales en un caso en el que las dos primeras componentes se utilizan para aproximarlos.

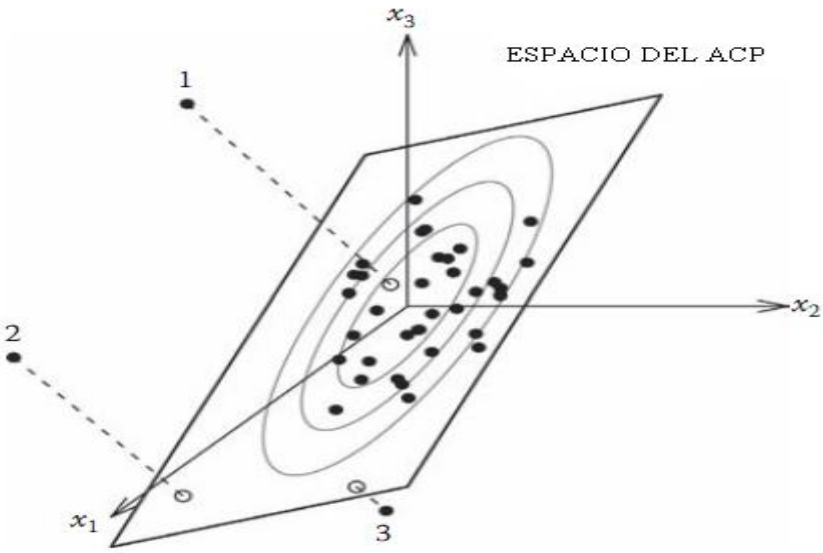

Figura 2. Visualización de los diferentes tipos de *outliers* que pueden influenciar en el ACP clásico

- El punto 1 muestra una distancia ortogonal grande al espacio ACP (*outliers* ortogonal).
- El punto 2 tiene una distancia ortogonal grande así como de puntaje. A *outliers* de este tipo se le llama puntos de *leverage* malo.
- El punto 3 tiene una distancia de puntaje alta y una distancia ortogonal pequeña. A *outliers* de este tipo se les llama puntos de *leverage* bueno.

Para la detección de los *outliers* se calculan las distancias de puntajes y las distancia ortogonales de cada observación, luego se grafican junto con las frontera críticas permitiendo identificar las observaciones regulares de los *outliers* (**TEJEDA**, 2011).

2.1.4 Descomposición CUR

El Análisis de Componentes Principales funciona impecablemente cuando el número de variables es relativamente pequeño (10 a 20). Pero si la cantidad de variables es muy grande usualmente se hace difícil la interpretación de los vectores propios con vista a la formación de grupos o, en general, a la clasificación de los individuos según su disposición en el espacio generado por las primeras componentes principales.
La descomposición matricial aleatoria CUR da una factorización de la matriz de los datos que resulta de utilidad para mejorar la interpretación de los resultados en un análisis exploratorio de datos (MAHONEY & DRINEAS**,** 2009). A continuación se describe brevemente en qué consiste esta técnica.

Las descomposiciones matriciales CUR, al igual que la descomposición en valores singulares (SVD) (GOLUB & VAN LOAN, 1996), permiten obtener aproximaciones matriciales de menor rango para una matriz de datos. Sin embargo, estas descomposiciones expresan los datos en términos de un número pequeño de columnas y/o filas lo que resulta de más fácil interpretación.

El término de descomposición CUR se emplea actualmente para nombrar aquellas descomposiciones en las que una matriz A_{mxn} se aproxima por el producto de tres matrices *C*, *U* y *R* donde *C* y *R* contienen algunas columnas y filas de *A*, respectivamente. Se conocen varias descomposiciones CUR que se diferencian en las cotas de error obtenidas y en el criterio para elegir las columnas y filas que forman las matrices *C* y *R*, (STEWART, 1999), (GOREINOV & TYRTYSHNIKOV, 2001), (FRIEZE, KANNAN & VEMPALA, 2004), (DRINEAS, KANNAN, MAHONEY, 2006).
La descomposición matricial CUR propuesta en (MAHONEY & DRINEAS**,** 2009) para mejorar el análisis exploratorio de datos consiste en construir C y R a partir de la determinación de un factor de importancia para cada columna de la matriz de datos.

Las columnas y filas de la matriz se seleccionan aleatoriamente según la distribución de probabilidad establecida por los factores de importancia, los que se interpretan como sensores de la influencia de cada columna en la mejor aproximación de menor rango de la matriz de datos.
De esta forma el factor de importancia normalizado de la j-ésima columna se define como

$$\pi_j = \frac{1}{k}\sum_{p=1}^{k}(v_j^p)^2 \qquad (1)$$

el vector π de componentes π_j es un vector de distribución de probabilidad.

A partir del vector de probabilidad definido por la Ec. 1, el algoritmo COLUMNSELECT (MAHONEY & DRINEAS, 2009) selecciona columnas de una matriz de datos *A* con un parámetro de rango k y un parámetro de error ε.

El resultado teórico más importante que avala el algoritmo establece que con probabilidad al menos del 99%, esta elección de columnas satisface que

$$\|A - P_C A\|_F \leq \left(1 + \frac{\varepsilon}{2}\right)\|A - A_K\|_F$$

donde P_C denota la matriz de proyección sobre el espacio columna generado por C y A_k es la matriz de rango k más próxima a *A* en norma de Frobenius. En (DRINEAS, MAHONEY & MUTHUKRISHNAN, 2008) se encuentra la demostración del resultado.

De esta forma el resultado garantiza que si *A* es una matriz cercana a una matriz de rango k entonces, con alta probabilidad, el subespacio generado por las columnas de *A* está próximo al subespacio generado por las columnas de *C*.

La información contenida en la matriz C de la descomposición CUR es suficiente para los fines propuestos, pues permiten seleccionar las variables más importantes según el algoritmo COLUMNSELECT.

Para el cálculo del vector de distribución que contiene los factores de importancia correspondientes a cada una de las columnas de la matriz de datos se utilizó el código de uso público Algoritmo CUR[1].

3. RESULTADOS Y DISCUSIÓN

3.1. ACP robusto

Problema 1: Detección de *outliers*.
En el trabajo original (CARRASCO, 2008), con la aplicación de la Regresión Lineal Múltiple se detectó un conjunto de cinco *outlliers*. Se decidió entonces la aplicación del ACP robusto para corroborar la presencia de los mismos.

Se realizó un ACP para la exploración de los datos obteniéndose 3 componentes que explicaban el 75% de la variabilidad de los datos. El hecho de que solo tres componentes explicaran el 75%, a partir de una matriz de

[1] AlgoritmoCUR en lenguaje MATLAB disponible en http://www.cs.rpi.edu/~boutsc/files/AlgorithmCUR.m

datos con 97 variables además de los cinco *outlliers* detectados en CARRASCO (2008) confirmaban la presencia de *outlier* en los datos. Por esta razón, se aplicó un ACP robusto.

Al realizar el ACP robusto utilizando el procedimiento de la proyección de búsqueda (VARMUZA & FILZMOSER, 2008) se obtienen siete componentes principales que explican el 74% de la variabilidad de los datos. Estos resultados se muestran en la Tabla 1.

Tabla 1. Resumen de las componentes principales

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Desviación estándar	14,014	10,688	6,618	6,516	5,976	5,338	5,319
Proporción de varianza	0,295	0,172	0,066	0,064	0,054	0,043	0,043
Proporción acumulada	0,295	0,467	0,533	0,597	0,650	0,693	0,736

Problema 2: Identificación de los *outliers.*

Para la identificación de los *outliers* se utilizaron siete componentes principales calculándose la distancia de puntaje así como la ortogonal para cada compuesto. En los gráficos de diagnóstico que se muestran en la Figura 3 se puede apreciar tres tipos de *outliers*: 10 *outliers* ortogonales, 10 puntos de *leverage* malo y 12 puntos de *leverage* bueno; las líneas discontinuas de cada gráfico representan las fronteras críticas. En **TEJEDA** (2011) se encuentran estos *outliers.*

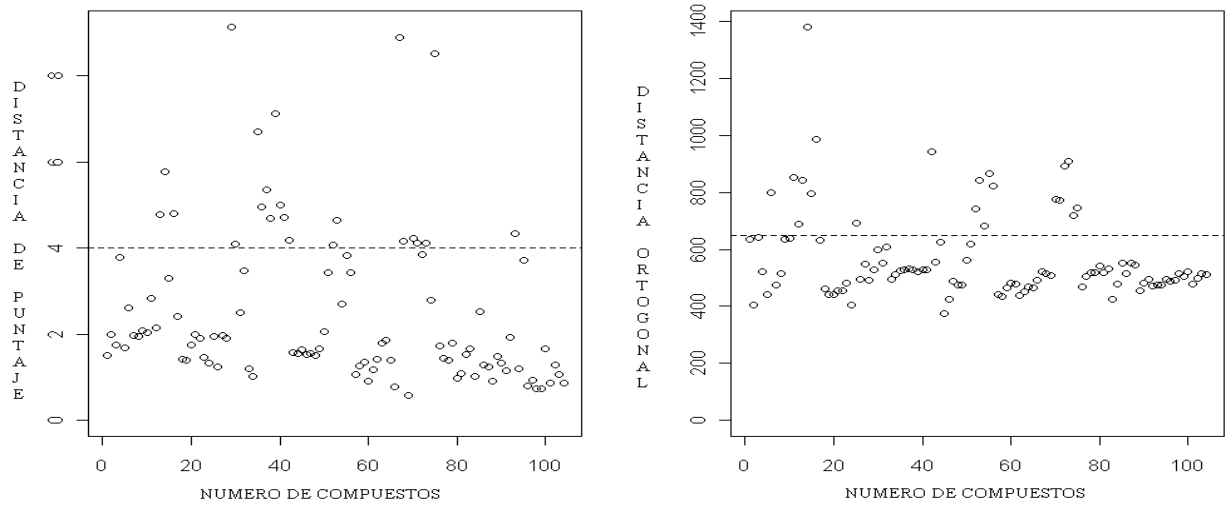

Figura 3. Gráficos de diagnóstico

3.2. Análisis de componentes principales (clásico)

Problema 3: ACP sobre una matriz con menos filas que columnas.
Una vez detectados y eliminados los *outliers*, mediante el ACP robusto se procedió a realizar nuevamente un ACP. La eliminación de los *outliers* de la matriz de datos implicó que dicha matriz redujera sus filas a 72 manteniendo las 97 columnas. Ante esta situación, se utilizó la descomposición del valor singular de la matriz de datos estandarizada para realizar el ACP ya que no se puede utilizar la función *princom* del software R (R Development Core Team, 2010) para calcular las componentes principales cuando la matriz de datos tiene menos filas que columnas. De la Tabla 2 se observa que tres componentes principales son suficientes pues explican el 75% de la variabilidad de los datos. En la Figura 4 se puede apreciar también que tres componentes principales son suficientes.

Tabla 2. Resumen de las componentes principales

	Comp .1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Desviación estándar	60,207	35,253	17,864	15,983	14,659	13,244	12,710
Proporción de varianza	0,526	0,180	0,046	0,037	0,031	0,025	0,023
Proporción acumulada	0,526	0,707	0,753	0,790	0,821	0,847	0,870

Para la interpretación de los resultados reportados en el Anexo 1 (**TEJEDA,** 2011), el análisis se condujo hasta el nivel de los grupos de variables y se observó que la primera componente está dominada por las variables topográficas y topológicas, la mayoría de ellas con el mismo signo, salvo un pequeño grupo. La segunda componente reproduce, en líneas generales, las relaciones de la primera. La tercera componente representa mayormente la reactividad química dada por los descriptores químico-cuánticos.

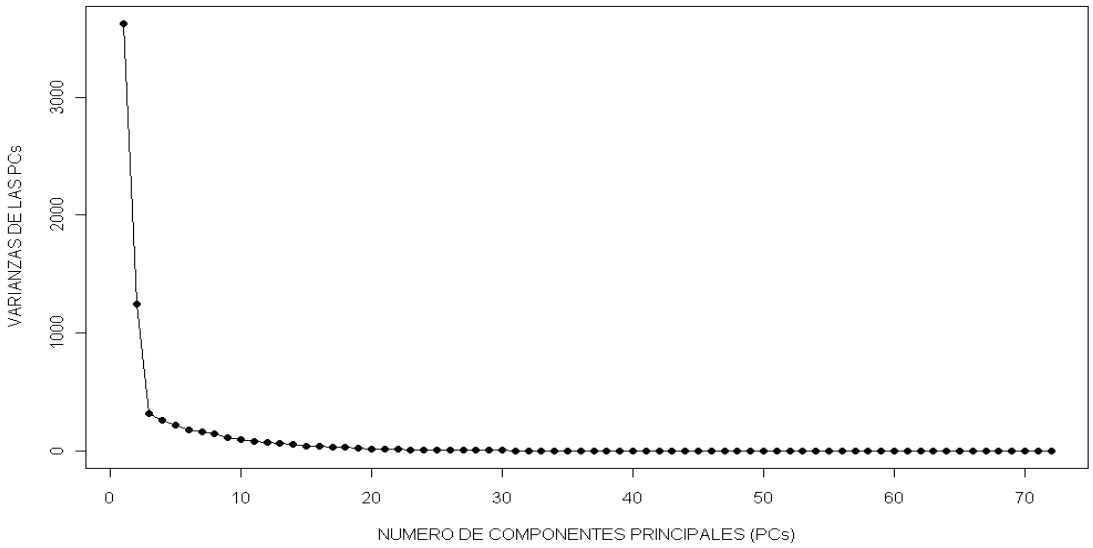

Figura 4. Diagrama de segmentación del ACP clásico

Está claro que la dificultad de trabajar con muchas variables radica fundamentalmente en la interpretación de las componentes. En este caso particular se aprovechó el hecho de que había agrupamientos dados por el usuario.

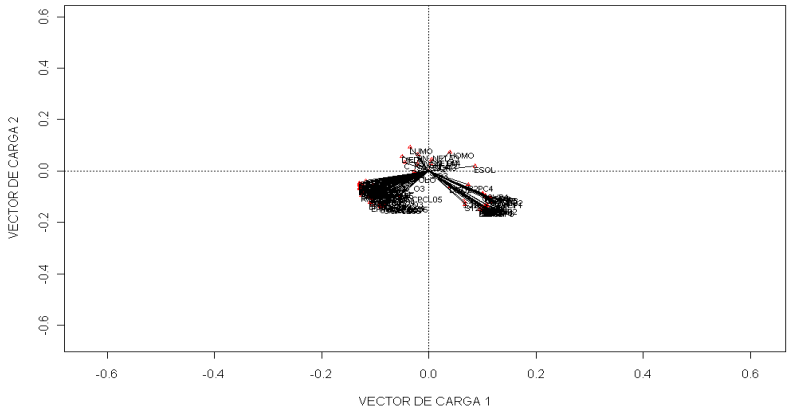

Figura 5. Gráfico de los vectores de cargas

Con respecto a la disposición de los compuestos en el sistema de referencia generado por las 2 primeras componentes, se aprecia una estructura por capas que no responde a las generaciones mencionadas en la descripción de las cefalosporinas. Esta configuración está en estudio por parte del usuario con vistas a profundizar en los resultados.

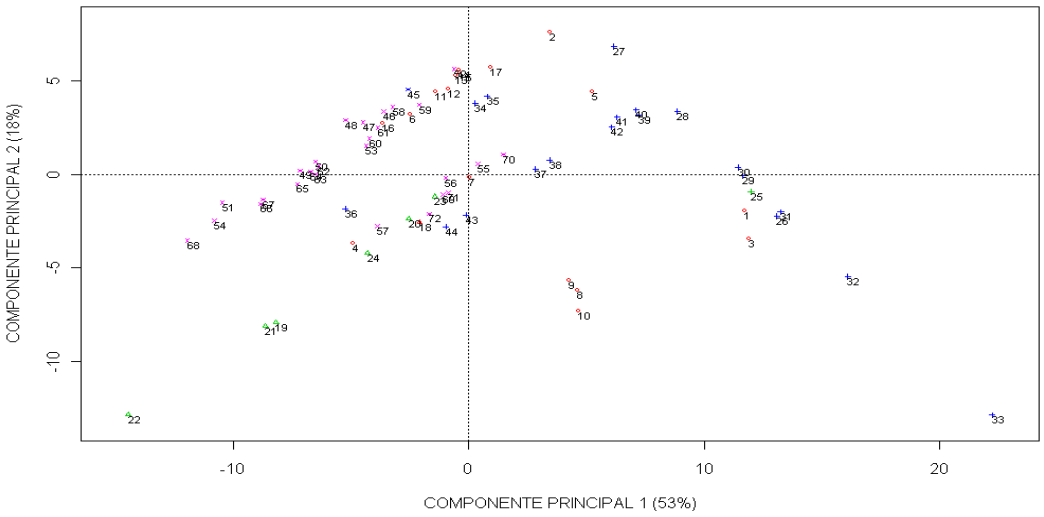

Figura 6. Gráfico de las componentes principales

3.3. Aplicación de la descomposición matricial CUR

Problema 4: Interpretación de las componentes principales.
Como se había mencionado anteriormente, la presente investigación consta de 97 descriptores, agrupados en 5 categorías, por lo que es lógico tratar de buscar otro método de análisis de datos que ayude a una mejor determinación de cuáles son las variables que mayor importancia tienen en la conformación de la nube de puntos transformada.

En esta investigación se utiliza la descomposición matricial CUR propuesta en (MAHONEY & DRINEAS, 2009) que consiste en construir C a partir de la determinación de un factor de importancia para cada columna de la matriz de datos.

Los resultados de la aplicación del CUR dieron como importantes las siguientes variables:
34, 21 y 45, en primer lugar; de igual forma, pueden considerarse como relevantes la variable 64, así como el grupo formado por la 8, 9 y 10.

Al compararse con los resultados del ACP se vio que la variable 34 es una de las importantes en la primera componente, mientras que la 21 resultó ser importante en la segunda y tercera componentes. Asimismo, la 45 resultó encontrarse entre las más importantes de la tercera componente.

De igual forma se analizaron la 64 y el grupo de 8, 9 y 10 y se comprobó que resultaban ser importante en, por lo menos 1 de las 3 componentes.

Estos resultados son una indicación de un posible sentido en el que se puede utilizar la Descomposición CUR con el ACP en el caso de muchas variables, en el que se hace muy difícil llegar a tener una idea para la interpretación. Es interesante ver que CUR da la importancia en un sentido global, mientras que el ACP lo hace dependiente de la o las componentes en las cuales aparecen las variables.

A modo de resumen, puede decirse que en este caso se obtuvo un buen grado de concordancia y que permitió brindar al usuario una interpretación más profunda, pues se pudo dar una interpretación al nivel de grupos de variables y esto se completó con el señalamiento de variables a tener en cuenta de manera especial.

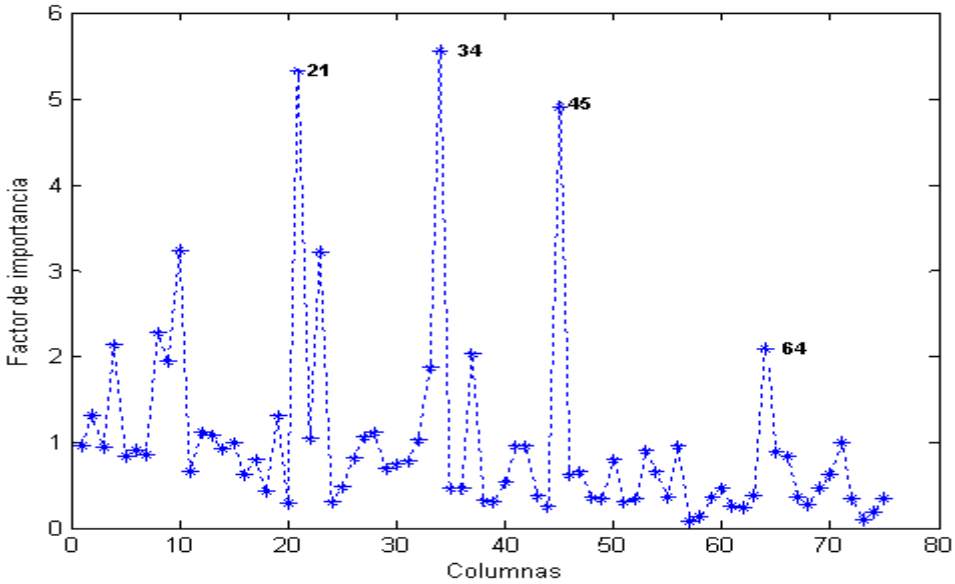

Figura 7. Factores de importancia por columnas para la matriz de datos.

4. CONCLUSIONES:

1. Se presentó una combinación de métodos para el análisis exploratorio
2. La descomposición matricial CUR es una posibilidad para interpretar los resultados combinado con el ACP
3. Las cefalosporinas no están clasificadas según su generación

RECEIVED JUNE 2011
REVISED SEPTEMBER 2011

REFERENCIAS:

[1] BRYSKIER, A., PROCYK, T., TREMBLAY, D. ***et al.*** (1990): The pharmacokinetics of cefodizime following intravenous and intramuscular administration of a single dose of 1.0 g**. J Antimicrob Chemother**, 26(Suppl C):59-63

[2] BRYSKIER, A., PROCYK, T. & LABRO MT. (1990): Cefodizime, a new 2-aminothiazolyl cephalosporin: physicochemical properties, toxicology and structureactivity relationships**. J Antimicrob Chemother**, 26(Suppl C):1-8

[3] CARRASCO, R. (2008): **Nuevos descriptores atómicos y moleculares para estudios de estructura-actividad: Aplicaciones**. Tesis en opción al grado de Doctor en Ciencias Químicas, La Habana, 141 .

[4] DRINEAS, P., KANNAN, R. & MAHONEY, M. W. (2006): Fast Monte-Carlo algorithms for matrices III:Computing a compressed approximate matrix decomposition. **SIAM J Comput**, 36, . 184–206

[5] DRINEAS, P., MAHONEY, M. W. & MUTHUKRISHNAN, S. (2008): Relative-error CUR matrix decompositions, **SIAM J Matrix Anal Appl**, 30, . 844-881

[6] ESCALONA, J. C., CARRASCO, R., PADRÓN, J. A. (2008): **Diseño racional de fármacos**. Editorial Universitaria, La Habana, 45

[7] FRERE, J. M., JORIS B., VARETTO, L., CRINE (1989): **Molecular Pharmacology**,Wiley, N. York.

[8] FRIEZE, A., KANNAN, R. & VEMPALA, S. (2004): Fast Monte-Carlo algorithms for finding low-rank approximations. ***J. ACM,*** 51**,** 1025–1041

[8] GOLUB, G. H. & VAN LOAN C. F. (1996): **Matrix Computations**. Johns Hopkins University Press, Baltimore

[9] GOREINOV, S. A. & TYRTYSHNIKOV, E. E. (2001): The maximum-volume concept in approxima-tion by low-rank matrices. **Contemporary Mathematics**, 280,. 47–51

[10] KARCHMER AW. (2000): Cephalosporins. In: Mandell GL, Bennett JE, Dolin R, editors. **Principles and practice of infectious diseases**. Philadelphia: Churchill Livingstone;. 274-299

[11] MAHONEY, M. W. & DRINEAS, P. (2009): CUR matrix decompositions for improved data analysis, **PNAS**, 106,. 697-702

[12] MARDIA, K.V., KENT, J. T. & BIBBY, J.M. (1979**): Multivariate Analysis**, Academic Press, Londres

[13] MOELLERING, R.C**.** (1995): **Oral Cephalosporins**, Karger, Suiza

[14] O'CALLAGHAN CH. (1979): Description and classification of the newer cephalosporins and their relationship with the established compounds. **J Antimicrob Chemother**; 5, 635-671

[15] R DEVELOPMENT CORE TEAM (2010): R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

[16] ROUSSEEUW, P. J. (1984): Least median of squares regression, **Journal of the American Statistical Association**, 79, pp.871-880

[17] STEWART, G. W. (1999): Four algorithms for the efficient computation of truncated QR approximations to a sparse matrix, **Numer. Math**, 83, pp.313-323

[18] **TEJEDA, Y.** (2011): **Utilización combinada de métodos exploratorios y confirmatorios para el análisis de la actividad antibacteriana de la cefalosporina**. Tesis en opción al grado de Máster en Ciencias Matemáticas, Universidad de La Habana, 64 pp.

[19] VARMUZA, K. & FILZMOSER, P. (2008): **Introduction to multivariate statistical analysis in chemometrics,** CRC Press, Boca Raton

[20] WALSH, C. (2003): **Antibiotics: Actions, origins, resistance**, DC: ASM Press, Washington, 345 pp.

[21**] WILLIAMS, J. D.** (1987): Classification of cephalosporins. **Drugs**; 34(Suppl 2),15-22