

# Caracterización de Productos Software con Métricas no redundantes

M<sup>a</sup> Esperanza Manso Martínez (\*)

[manso@infor.uva.es](mailto:manso@infor.uva.es)

Yania Crespo González-Carvajal (\*)

[yania@infor.uva.es](mailto:yania@infor.uva.es)

José Javier Dolado Cosín (✦)

[dolado@si.ehu.es](mailto:dolado@si.ehu.es)

(\*) Dpto. de Informática de la U. de Valladolid

(✦) Dpto. de Lenguajes y Sistemas de la U. del País Vasco

## Resumen

El objetivo de este estudio exploratorio es obtener un subconjunto de métricas de diagramas de UML, que resuman la información que proporcionan un conjunto inicial de 21 métricas, las cuales miden diferentes aspectos de atributos de los diagramas como tamaño, herencia, etc. La técnica fundamental utilizada en el estudio fue el Análisis de Componentes Principales (ACP). El estudio se ha realizado sobre 78 diagramas de clases reales, procedentes del producto desarrollado en una empresa, y las mediciones de las 21 métricas se obtuvieron mediante el módulo “Quality Assurance” de la herramienta CASE Together. La principal conclusión del estudio es que las 20 dimensiones o métricas, consideradas, se pueden reducir a cuatro factores o dimensiones, de los cuales el factor1 contiene información, fundamentalmente, de la herencia y complejidad de la herencia, el factor2 recoge la información relativa al acoplamiento, el factor3 recoge la cohesión, y el factor4 recoge la visibilidad de las clases. Además la métrica PPKGM no está correlacionada con las otras métricas, por lo que se excluyó del estudio de ACP.

Palabras clave: *diagramas de UML, métricas, análisis exploratorio de datos, análisis de componentes principales.*

## 1. Introducción

La medición de atributos internos del diseño de un sistema de software, tiene interés desde diferentes puntos de vista. Es bien conocida su utilización, junto con los modelos de calidad, para predecir medidas de los atributos externos de la calidad, así como el papel de variables independientes en los diferentes modelos del esfuerzo de desarrollo [2, 9, 19]. Además sirven como guías en las fases iniciales de desarrollo del producto para detectar problemas, que de otra forma no se conocerían hasta las etapas finales [3, 9], lo que permitirá modificar los elementos software, con el correspondiente beneficio en tiempo y costes. En otros estudios se trata de relacionar el uso adecuado de la herencia con la mantenibilidad [3, 5, 15], o de explicar la modificabilidad a partir de métricas de atributos internos [10], etc.

Las herramientas que permiten el almacenamiento de medidas del sistema y/o del proceso, son un instrumento imprescindible en estos estudios, y en general, van a estar integradas en el proceso de desarrollo. Sin embargo, estas herramientas proporcionan gran cantidad de información procedente de mediciones, que pueden formar parte de las variables independientes o dependientes de los estudios experimentales. Una parte importante de esta información puede ser redundante, que en términos estadísticos es equivalente a decir que muchas de las métricas pueden estar muy correlacionadas. Esto justifica el interés de analizar la información inicial para eliminar la redundancia. En los estudios experimentales en Ingeniería del Software [2, 4], como en los de otras disciplinas, este problema se resuelve utilizando ciertas técnicas estadísticas, como el Análisis de Componentes Principales (ACP).

El presente documento está estructurado de la siguiente forma, el apartado 1.1 contiene los objetivos y la justificación del trabajo, en la sección 2 se expone un resumen de las técnicas estadísticas utilizadas, en las secciones 3 y 4 se presentan la realización del estudio y el análisis de los resultados observados. Finalmente se resumen las conclusiones y el trabajo futuro en la sección 5.

## 2.1 Objetivos y justificación del estudio

Los estudios experimentales que trabajan con una gran cantidad de variables independientes o explicativas (medidas internas del diseño orientado a objetos en este caso), siguen una metodología en el análisis de los datos que comprende desde la descripción y resumen de los mismos, hasta análisis exploratorios para identificar posibles relaciones entre las variables, o análisis para clasificar los objetos del estudio [12].

En este trabajo el enfoque se hace sobre el análisis exploratorio para seleccionar las variables que contengan más información, con el fin de caracterizar los diagramas de UML procedentes de un producto desarrollado en una empresa.

El objetivo de este trabajo, siguiendo el modelo GQM [1] es:

Realizar un estudio *exploratorio de diagramas de clases* para *caracterizarlos* con respecto a un *conjunto de métricas internas*, desde el punto de vista de *los desarrolladores de software* en el contexto de *las organizaciones que desarrollan software*.

Una técnica estadística que permite conseguir este objetivo es el Análisis de Componentes Principales (ACP), del que se da una somera descripción a continuación.

## 2. Técnicas estadísticas utilizadas

### 2.1 Análisis de Componentes Principales (ACP)

Este tipo de análisis permite, en general, reducir la dimensión de un espacio inicial definido por  $k$  variables asociadas a un determinado fenómeno, de forma que nos proporciona  $p$  variables (factores) que son combinación lineal de las  $k$  de partida, con la peculiaridad de que:

- Los  $p$  factores son ortogonales, o lo que es igual, su covarianza (Cov) es cero:

$$\text{Cov}(f_i, f_j) = 0, \forall i, j(1..p)$$

- Los  $p$  factores,  $f_i$ , recogen la máxima variación de las variables iniciales  $v_j$ :

$$\max\left(\sum_{i=1}^p \text{var}(f_i) / \sum_{j=1}^k \text{var}(v_j)\right)$$

El ACP [12,13] se puede utilizar desde dos puntos de vista, para explorar o para confirmar, en este caso el punto de vista es exploratorio. La utilización de ACP supone una ventaja si:

1. Se ha reducido notablemente el número de variables inicial y/o
2. Los  $p$  factores nuevos tienen una interpretación inteligible, en términos de las variables iniciales u otras características ocultas.

### 2.2 Examen de la matriz de correlaciones y prueba de Barlett

Si partimos de variables no correlacionadas el ACP no tendría sentido, ya que precisamente, la reducción de la dimensión del nuevo espacio se basa en eliminar información redundante o, lo que es igual, variables muy correlacionadas. Por ello se han utilizado las matrices de correlaciones de Pearson y Spearman [16, 17], con los  $p$ -valores asociados para contrastar la no correlación entre pares de variables, ( $H_0: \rho_{x,y} = 0$ ) y el test de esfericidad de Barlett [12] para contrastar la hipótesis  $H_0$ : no hay correlación entre  $k$  variables.

### 3. Realización del estudio

Se ha utilizado el módulo “Quality Assurance”, distribuido con la versión 4.2 de la herramienta CASE Together, para obtener las mediciones en los objetos del estudio, y el paquete de software SPSS versión 11.0 [18] para analizar los resultados observados.

#### 3.1 Conjunto de variables a considerar

El módulo “Quality Assurance” de la herramienta CASE Together proporcionó las mediciones de los diagramas de clases de UML. En la Tabla 1 se resumen las métricas seleccionadas, así como alguno de los atributos de calidad con los que pueden estar relacionados, o su utilidad como guías de diseño [9, 11, 19]. Entre ellas se encuentran las métricas de Chidamber & Kemerer (C&K) [7] y otras de amplia difusión.

Tabla 1. Métricas de diagramas de clases, de secuencia y de colaboración de UML.

Acrónimo	Nombre de la métrica	Atributos de calidad /Guías de diseño
CBO	Acoplamiento entre objetos (C&K)	Reusabilidad
DAC	Acoplamiento por abstracción de datos	Guía de diseño
DOIH	Profundidad de herencia (C&K)	Modificabilidad
LOCOM1	Falta de cohesión de los métodos (C&K)	Reusabilidad
LOCOM2	Falta de cohesión de los métodos	Complejidad
LOCOM3	Falta de cohesión de los métodos (Henderson-Sellers)	Reusabilidad
MNOP	Nº máx. de parám. entre las operaciones de una clase	Reusabilidad
NOA	Nº de atributos heredados	Guía de diseño
NOAM	Nº de métodos añadidos	Guía de diseño
NOC	Nº de clases	Tamaño, Complejidad
NOCC	Nº de hijos de una clase	Facilidad de Pruebas
NOM	Nº de miembros	Guía de diseño
NOO	Nº de operaciones	Guía de diseño, Tamaño,
NOOM	Nº de métodos heredados	Guía de diseño, Tamaño,
NORM	Nº de métodos remotos	Guía de diseño, Tamaño
PPKGM	% de miembros del paquete	Tamaño
PPRIVM	% de miembros privados	Guía de diseño, Tamaño
PPROTM	% de miembros protegidos	Guía de diseño, Tamaño
PPUBM	% de miembros públicos	Guía de diseño Facilidad de Pruebas
RFC	Respuesta de una clase (C&K)	Complejidad
WMPC2	Nº de métodos y parámetros de los mismos	Complejidad

#### 3.2 Objetos del estudio

La métricas de la Tabla 1 se observaron en 78 diagramas procedentes de una aplicación de Gestión de gran tamaño.

### 3.3 Estudio previo de las correlaciones entre variables

Cuando se estudiaron los coeficientes de correlación de Pearson y de Spearman se observaron resultados muy parecidos, con ligeras variaciones en CBO, PPKGM y PPRIVM. La métrica PPKGM es la que tuvo unos resultados claros de no correlación con el resto de las métricas, lo que se puede explicar teniendo en cuenta los valores observados (Tabla 3). De acuerdo con ellos se ha realizado el ACP excluyéndola.

En cuanto a la prueba de Bartlett, examinando el resultado de la Tabla 2 se puede concluir que da significativa al nivel 0,05, por lo que se puede rechazar la hipótesis nula  $H_0$ : no hay correlación en el conjunto de 20 métricas de diagramas de UML, y seguir con el ACP. Además la medida de adecuación de Kaiser-Meyer-Olkin (KMO) es mayor que 0,5, nivel recomendado como límite para aceptar que los factores del ACP explican bastante variación inicial.

Tabla 2: KMO y prueba de Bartlett

Medida de adecuación muestral de Kaiser-Meyer-Olkin.	0,773
Prueba de esfericidad de Bartlett (Chi-cuadrado aproximado)	1266,469
<i>Grados de libertad</i>	<i>120</i>
<i>Significación</i>	<i>,000</i>

De este modo se hace el ACP para 20 métricas evaluadas en 78 diagramas, eligiendo como método de selección de las componentes que el autovalor sea mayor que 1. Se elige además una rotación ortogonal que facilite la interpretación de los resultados.

## 4. Análisis de los resultados

### 4.1 Descripción de las métricas

En la Tabla 3 se presenta un resumen de los estadísticos: percentiles ( $P_k$ ), media y desviación estándar (D.E.), para las métricas consideradas. Como se puede apreciar las métricas de falta de cohesión LCOM1, LCOM2 y LCOM3 son las que tienen una media y percentiles mayores. En cuanto a la dispersión de los datos (D.E.), es mayor que la media en todas las métricas excepto CBO, DOIH, LCOM2 y LCOM3, MNOP, PPROTM y PPRIVM.

Es de resaltar que la métrica PPKGM obtuvo cero en la mayoría de las observaciones, lo que es coherente con el significado de la métrica, ya que el n° de miembros de un paquete parece que no tiene por qué estar relacionado con las otras variables. Esto dará lugar a la exclusión de esta variable en el estudio de ACP.

### 4.2 Selección del número de factores

En la Tabla 4 se puede ver que con 4 factores, o dimensiones, se explica el 91,204% de la varianza total, en los dos supuestos, con y sin rotación. Puesto que la interpretación de los componentes obtenidos inicialmente no era fácil, por ejemplo el componente primero estaba muy correlacionado con casi todas las métricas, se realizó una rotación con el fin de obtener nuevas dimensiones (factores) que dependan fuertemente de subconjuntos disjuntos de variables. Como era de esperar, el primer factor pierde algo de peso a favor de los otros (ver la Tabla 4) y las nuevas dimensiones (Tabla 5) se pueden interpretar de la forma siguiente:

- La primera dimensión recoge información sobre herencia y complejidad de la herencia, teniendo en cuenta que ésta engloba a la de las clases que componen la jerarquía de herencia.
- La segunda dimensión recoge información fundamentalmente de la interacción de la clase con el exterior por medio del acoplamiento.

- La tercera dimensión está claramente relacionada con la falta de cohesión, Si bien es verdad que falta LCOM1, tampoco es muy representativa de este atributo, y presenta más problemas de interpretación como medida de la cohesión que las LCOM2 y LCOM3 [8, 11].
- La cuarta dimensión esta relacionada con la visibilidad de los métodos, pues claramente depende positivamente de PPUBM y negativamente de PPRIVM (y si tuviera relación con PPROTM sería negativa).

Tabla 4: Varianza total explicada

Componente	Sumas de las saturaciones al cuadrado de la extracción Total	% de la varianza	% acumulado	Suma de las saturaciones al cuadrado de la rotación Total	% de la varianza	% acumulado
1	13,069	65,346	65,346	8,495	42,476	42,476
2	2,574	12,871	78,217	6,254	31,268	73,744
3	1,445	7,226	85,443	1,996	9,981	83,725
4	1,152	5,761	91,204	1,496	7,479	91,204

Tabla 3. Descripción de las métricas

Métricas	P <sub>25</sub>	P <sub>50</sub>	P <sub>75</sub>	Media	D. E.
CBO	4,00	7,00	14,25	10,60	7,35
DAC	,75	1,00	1,25	2,40	4,42
DOIH	2,00	2,00	4,00	3,56	2,06
<i>LOCOM1</i>	<i>18,00</i>	<i>28,00</i>	<i>43,00</i>	<i>70,28</i>	<i>126,97</i>
<i>LOCOM2</i>	<i>78,75</i>	<i>88,00</i>	<i>100,00</i>	<i>84,16</i>	<i>18,98</i>
<i>LOCOM3</i>	<i>97,75</i>	<i>101,00</i>	<i>134,00</i>	<i>110,53</i>	<i>38,98</i>
MNOP	1,00	2,00	3,00	3,07	1,96
NOA	1,00	1,00	7,00	6,39	10,52
NOAM	2,75	3,50	12,25	15,66	22,00
NOC	2,00	4,00	22,25	25,28	77,95
NOCC	,00	,00	26,00	10,82	29,94
NOM	6,75	7,50	16,50	21,75	29,15
NOO	5,75	6,00	12,50	16,82	21,79
NOOM	2,00	5,00	5,00	3,47	3,82
NORM	1,75	3,50	20,50	14,06	16,82
<i>PPKGM</i>	<i>,00</i>	<i>,00</i>	<i>,00</i>	<i>0,83</i>	<i>3,78</i>
PPRIVM	5,00	14,50	29,50	19,99	18,97
PPROTM	5,50	24,50	37,50	24,65	18,85
PPUBM	37	56,50	67,00	54,71	20,35
RFC	27,00	29,50	51,00	34,37	18,71
WMPC2	9,00	9,50	27,75	38,92	63,11

Tabla 5. Matriz de componentes rotados

Métricas	Componentes			
	1	2	3	4
<i>CBO</i>	,429	<i>,819</i>	,165	4,445E-02
<i>DAC</i>	<i>,907</i>	,374	,132	1,265E-02
<i>DOIH</i>	<i>,706</i>	,439	4,766E-02	,175
<i>LOCOM1</i>	,551	<i>,725</i>	,236	-7,746E-02
<i>LOCOM2</i>	,127	4,834E-02	<i>,964</i>	-2,769E-02
<i>LOCOM3</i>	,532	9,139E-02	<i>,732</i>	,239
<i>MNOP</i>	,485	<i>,793</i>	,171	,106
<i>NOA</i>	<i>,801</i>	,560	,145	1,768E-02
<i>NOAM</i>	<i>,909</i>	,383	7,190E-02	3,724E-02
<i>NOC</i>	<i>,894</i>	,237	,220	,161
<i>NOM</i>	<i>,860</i>	,492	,106	-1,074E-02
<i>NOO</i>	<i>,867</i>	,472	,117	6,034E-03
<i>NOOM</i>	<i>,942</i>	-3,092E-03	,285	-7,785E-03
<i>NORM</i>	,476	<i>,832</i>	,137	2,680E-02
<i>RFC</i>	,464	<i>,820</i>	,169	-7,782E-02
<i>WMPC2</i>	<i>,918</i>	,362	9,876E-02	3,647E-02
<i>PPROTM</i>	-4,138E-02	<i>-,817</i>	,319	-,314
PPUBM	8,580E-02	,397	-2,273E-02	<i>,897</i>

## 5. Conclusiones y Trabajo futuro

En este trabajo se ha utilizado ACP como técnica exploratoria para reducir el espacio de 21 dimensiones (métricas) observadas en 78 diagramas, eliminando información redundante. Esto ha permitido seleccionar 4 dimensiones que recogen el 91,204% de la variación inicial. La exploración de la matriz de correlaciones ha permitido descartar a priori PPKGM, ya que no estaba correlacionada con ninguna de las métricas del estudio, y por lo tanto podría ser una dimensión per se.

El ACP permitió identificar claramente 4 dimensiones o factores con el significado siguiente:

- El factor 1 corresponde a la información sobre herencia y su complejidad, entendiendo que ésta engloba a la de las clases individuales.
- El factor 2 corresponde a la dimensión del acoplamiento.
- El factor 3 recoge la información relativa a la cohesión.
- Y el factor 4 es la dimensión de la visibilidad de los métodos.

Estos cuatro factores reflejan lo que podemos considerar cuatro pilares básicos del diseño orientado a objetos: buen diseño de la herencia, máxima cohesión, mínimo acoplamiento, y respeto de los principios de ocultación de información: principio abierto-cerrado y principio de caja negra.

Este análisis exploratorio es el primer paso de un estudio, en el que se sondeará la relación entre estas componentes y variables dependientes de interés, relacionadas con el proceso de desarrollo del software. Para las siguientes fases del estudio, se dispondrá de los mismos diagramas de UML utilizados en esta fase exploratoria, y otros productos reales, procedentes de diferentes dominios, que se podrán incorporar en el futuro. Los sujetos de los experimentos serán estudiantes de Ingeniería Informática, procedentes de diferentes universidades españolas, y profesionales de empresas, con el fin de extender la validez de las conclusiones tanto como sea posible.

## 6. Agradecimientos

Nuestro agradecimiento a la empresa Tecsidel (nos proporcionó el material del estudio). Esta investigación forma parte del proyecto DOLMEN, CYCIT-TIC 2000-1673-C06-05 y del proyecto CICYT-TIC1143-C03-01

## 7. Bibliografía

1. Basili V. & Rombach H. 1988. The TAME project: Towards improvement-oriented software environments. *IEEE Transactions on Software Engineering* 14(4).
2. Briand L.C. & Wüst J. 2002. Empirical studies of quality models. *To be published in Advances in Computers Academic Press Zelkowitz (ed)*.
3. Briand L., Morasca S., Basili V. 1999. Defining and validating measures for object-based high level design. *IEEE Transactions on Software Engineering*. Vol 25, nº 5 September.
4. Briand L. & Melo W. & Wüst J. 2001. Assessing the applicability of fault-proneness models across object-oriented software projects. *IEEE Transactions on Software Engineering*.
5. Brito e Abreu F & Melo W. 1996. Evaluating the impact of object-oriented design on software quality. *IEEE Transactions on Software Engineering March*.
6. Cartwright M & Shepperd M. 2002. An Empirical view of inheritance. *ESERG TR98-02*.
7. Chidamber S.R. & Kemerer 1994. A metrics suite for object oriented design. *IEEE Transactions on Software Engineering*. Vol 20, nº 6 June.
8. Churcher N.I. & Shepperd M.J. , 1995. Comments on " A metrics suite for object oriented design" *IEEE Transactions on Software Engineering Vol 21, no 3 March*.
9. Fenton Norman E. Pfleeger Shari L. 1997. Software Metrics. A Rigorous & Practical Approach. second De. International Thomson Publishing Inc.
10. Genero Bocco M. 2002. Defining and validating metrics for conceptual models. Tesis doctoral. Escuela Superior de Informática. Dpto. de Informática. U. De Castilla-La Mancha
11. Hogan J. 1997. An analysis of OO software metrics. <http://www.dcs.warwick.ac.uk/jer/>
12. Kleinbaum D.G. Et al 1987. Applied regression analysis and other multivariate methods, second ed. Duxbury Press
13. Lebart L. et al. 1995. Tratamiento estadístico de los datos. Marcombo
14. Nesi P. & Querci T. 1998. Effort estimation and prediction of object-oriented systems . *Journal of systems and software* 42.
15. Poels G., Dedene G. 2001. Evaluating the effect of inheritance on the modificability of object-oriented business domain models *IEEE Transactions on Software Engineering*. 2001
16. Siegel S. 1985. Estadística no paramétrica Ed Trillas.
17. Snedecor G. & Cochran W.G 1989. Statistical Methods 8ª ed Iowa State University Press.
18. SPSS 11.0. 2001. Syntax Reference Guide. Chicago. SPSS Inc.
19. Zuse H. 1998. A framework of software measurement. Berlin, New-York W. de Gruyter.