Estudio del Modelo Paramétrico DMCoMo de Estimación de Proyectos de Explotación de Información

Pytel, P., Tomasello, M., Rodriguez, D., Pollo-Cattaneo, F., Britos, P., García-Martínez, R.

Grupo Investigación en Sistemas de Información. Departamento Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús.
Grupo de Estudio en Metodologías de Ingeniería de Software. Facultad Regional Buenos Aires. Universidad Tecnológica Nacional.
Grupo de Investigación en Explotación de Información. Sede Andina (El Bolsón). Universidad Nacional de Río Negro.
ppytel@gmail.com, fpollo@posgrado.frba.utn.edu.ar, paobritos@gmail.com, rgarcia@unla.edu.ar

Resumen. En este trabajo se presenta un estudio inicial del comportamiento del método paramétrico de estimación de proyectos de explotación de información DMCoMo, para lo cual se realiza una introducción sobre las características del método de estimación, se delimita el problema presentando el diseño experimental y los resultados obtenidos para finalizar con la puntualización de algunas conclusiones parciales.

Palabras Claves. DMCoMo. Estimación de proyectos. Explotación de Información.

1. Introducción

Al comenzar la gestión de todo proyecto de software es necesario realizar las actividades que se denominan planificación del proyecto. Esto incluye calcular el costo del proyecto para lo cual es necesario realizar una estimación: del trabajo a ejecutar, de los recursos necesarios y del tiempo que transcurrirá desde el comienzo hasta el final de su realización [1]. Dada las diferencias que existen entre un proyecto convencional de construcción de software y un proyecto de explotación de información, los métodos usuales de estimación no son aplicables ya que los parámetros a ser utilizados son de naturalezas diferentes [5]. Por ejemplo, las herramientas de estimación de software convencional como COCOMO II [2] o PRICE-S [3] utilizan como parámetros la cantidad de líneas de código, la experiencia del equipo de trabajo, características de la plataforma de desarrollo, entre otras. Sin embargo, en proyectos de explotación de información existen otras características que deben ser consideradas para la estimación, como por ejemplo, cantidad de fuentes de información, nivel de integración de los datos, el tipo de problema a ser resueltos, entre las más representativas de este tipo de proyectos.

En [4] se propone un método analítico de estimación para proyectos de explotación de información el cual se denomina Matemático Paramétrico de Estimación para Proyectos de Data Mining (en inglés Data Mining Cost Model, o DMCoMo). El método DMCoMo es un modelo de estimación de la familia de COCOMO que permite estimar los meses/hombre que serán necesarios para desarrollar un proyecto de explotación de información desde su concepción hasta su puesta en marcha. Los modelos de estimación analíticos se definen a través de la aplicación de métodos de regresión en datos históricos para obtener relaciones matemáticas entre las variables (también llamadas factores de costo) representadas en ecuaciones matemáticas.
Este trabajo es una continuación de [7] y tiene como objetivo informar sobre los primeros resultados obtenidos del estudio del método de estimación DMCoMo. Para ello primero se realiza una introducción sobre las características del método de estimación DMCoMo (sección 2), luego se delimita el problema (sección 3) presentando el diseño experimental (sección 4) y los resultados obtenidos (sección 5) finalizando con la puntualización de algunas conclusiones parciales (sección 6).

2. Método de Estimación DMCoMo

DMCoMo define una serie de factores de costo para realizar la estimación los cuales están vinculados a las características más importantes de los proyectos de explotación de información. Estos se clasifican en seis categorías que se describen brevemente con sus factores de costo relacionados en la tabla 1. Una vez que los valores de los factores de costo son definidos, se ingresan en las ecuaciones matemáticas suministradas por el método. DMCoMo dispone de dos fórmulas, una que utiliza 23 factores de costo como variables (ver tabla 2) que puede ser utilizada cuando el proyecto está bien definido y otra de 8 factores de costo como variables (ver tabla 3) que puede utilizarse cuando no todos los datos del proyecto se encuentran definidos. Como resultado de ingresar los valores a la ecuación correspondiente, se obtiene la cantidad de meses/hombre correspondiente al proyecto.

3. Delimitación del Problema

El problema identificado es motivado por la propia limitación señalada en [4] en la que se aclara que DMCoMo se considera confiable para estimar el esfuerzo de proyectos de explotación de información que se encuentren en el rango de esfuerzo de 90 a 185 meses/hombre. Si el esfuerzo del proyecto se encuentra fuera de este rango, el comportamiento del método es desconocido.
En este contexto el objetivo de este trabajo es obtener información sobre el comportamiento del modelo para distintos tamaños de proyectos de explotación de información [5] con una particular focalización en proyectos pequeños que son los que usualmente requieren las PyMEs [6].

Tabla 1. Categorías y Factores de Costo utilizados por DMCoMo

CATEGORÍA	DESCRIPCIÓN	FACTOR DE COSTO
RELACIONADOS A LOS DATOS	Agrupa los factores de costo que tienen que ver con la cantidad, la calidad de los datos a tratar en el proyecto de explotación de información.	– Cantidad de Tablas (NTAB) – Cantidad de Tuplas de las Tablas (NTUP) – Cantidad de Atributos de las Tablas (NATR) – Grado de Dispersión de los Datos (DISP) – Porcentaje de valores NULL (PNUL) – Grado de Documentación de las Fuentes de Información (DMOD) – Grado de Integración de Datos Externos (DEXT)
RELACIONADOS A LOS MODELOS	Incluye todos aquellos factores de costo que tienen que ver con los modelos que hay que generar y que tienen en cuenta el volumen de datos que se va a utilizar para generar los modelos, la disponibilidad de técnicas para generar los modelos y la dificultad del mismo.	– Cantidad de Modelos a ser Creados (NMOD) – Tipo de Modelos a ser Creados (TMOD) – Cantidad de Tuplas de los Modelos (MTUP) – Cantidad y Tipo de Atributos por cada Modelo (MATR) – Cantidad de Técnicas Disponibles para cada Modelo (MTEC)
RELACIONADOS AL DESARROLLO DE LA PLATAFORMA	Agrupa los factores de costo que tienen que ver con las características de los almacenes de datos y su localización.	– Cantidad y Tipo de Fuentes de Información Disponibles (NFUN) – Distancia y Medio de Comunicación entre Servidores de Datos (SCOM)
RELACIONADOS A LAS TÉCNICAS Y HERRAMIENTAS	Agrupa las características de las técnicas y herramientas de explotación de información que se van a utilizar en el proyecto. Estas características se centran principalmente en el nivel de formación que requieren, la amigabilidad de las mismas y el número de técnicas que soportan.	– Herramientas Disponibles para ser Usadas (TOOL) – Grado de Compatibilidad de las Herramientas con Otros Software (COMP) – Nivel de Formación de los Usuarios en las Herramientas (NFOR)
RELACIONADOS AL PROYECTO	Agrupa aquellas características relativas a los departamentos y a las localizaciones para las que se desarrolla el proyecto de explotación de información. También incluye características acerca de la documentación que es necesario generar durante la realización del proyecto.	– Cantidad de Departamentos Involucrados en el Proyecto (NDEP) – Grado de Documentación que es necesario generar (DOCU) – Cantidad de Sitios donde se realizará el Desarrollo y su Grado de Comunicación (SITE)
RELACIONADOS AL EQUIPO DE TRABAJO	Incluye aquellos factores relacionados con el equipo de trabajo que participa en el proyecto (dirección, implementadores, expertos, etc.). Estos factores evalúan el conocimiento y la capacidad requerida para llevar a cabo cada una de las tareas del proyecto.	– Grado de Familiaridad con el Tipo de Problema (MFAM) – Grado de Conocimiento de los Datos (KDAT) – Actitud de los Directivos (ADIR)

Tabla 2. Fórmula de 23 factores

MM23 = 78,752 + 2,802 x NTAB + 1,953 x NTUP + 2,115 x NATR
+ 6,426 x DISP + 0,345 x PNUL + (- 2,656) x DMOD
+ 2,586 x DEXT + (-0,456) x NMOD + 6,032 x TMOD
+ 4,312 x MTUP + 4,966 x MATR + (-2,591) x MTEC
+ 3,943 x NFUN + 0,896 x SCOM + (-4,615) x TOOL
+ (-1,831) x COMP + (-4,689) x NFOR
+ 2,931 x NDEP + (-0,892) x DOCU + 2,135 x SITE
+ (-0,214) x KDAT + (-3,756) x ADIR
+ (-4,543) x MFAM

Tabla 3. Fórmula de 8 factores

MM8 = 70,897 + 2,368 x NTAB
+ 2,885 x NATR
+ 4,792 x DISP
+ 2,713 x DEXT
+ 7,257 x TMOD
+ 4,615 x MATR
+ (-3,842) x NFOR
+ (-3,275) x MFAM

4. Diseño Experimental

El estudio experimental se realizó siguiendo el diseño planteado en [7] aplicando el método de simulación Monte Carlo [8] en el que se propone el siguiente protocolo:

Paso 1: Desarrollo de un banco de pruebas donde se generan los datos de proyectos de explotación de información con características (es decir los valores de los factores de costo) determinadas aleatoriamente dentro de un marco definido por el experimentador (aplicando una clasificación a partir del tamaño de los proyectos definidos como pequeños, medianos y grandes)

Paso 2: Para cada lote de datos de proyecto de explotación de información generado, aplicar las fórmulas de estimación del método DMCoMo descriptas en la tabla 2 y en la tabla 3.

Paso 3: Integrar estadísticamente la información obtenida, analizar los resultados obtenidos para cada tamaño de proyecto y formular conclusiones.

5. Resultados Obtenidos

En esta sección se presentan los resultados obtenidos del estudio experimental el cual consistió en generar los datos de 10.000 proyectos por cada tamaño definido. En la tabla 4 se indica la estadística obtenidas por cada una de las fórmulas y tamaño de proyecto. De esta tabla se puede ver que la media de la fórmula que utiliza los 23 factores de costo (representada por la variable MM23) en proyectos medianos es un 52% más grande que la de proyectos pequeños y la media de los proyectos grandes es un 42% más grande que los proyectos medianos (o sea un 117% con respecto a los proyectos pequeños). Por otro lado, la fórmula que utiliza sólo 8 factores de costo (variable MM8) posee un crecimiento menor por escalones de aproximadamente el 22% con respecto al tamaño anterior. Entonces debido a que la fórmula de 23 factores de costo crece aproximadamente el doble con respecto a la otra fórmula al variar el tamaño del proyecto, se podría decir que es más conservadora.

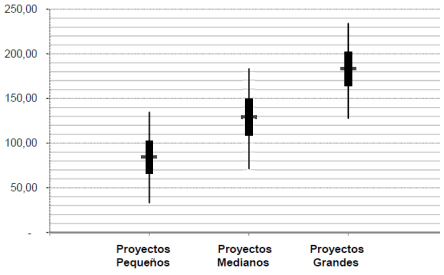
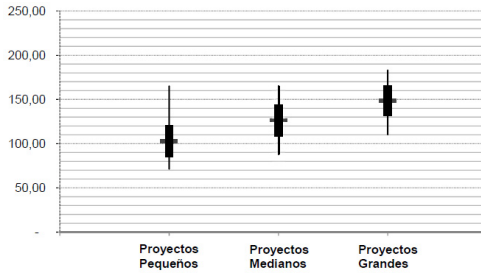
Tabla 4. Media por fórmula y tamaño de proyecto

MEDIA (en meses/hombre)	MM23	MM8
Proyectos Pequeños	84,41	102,59
Proyectos Medianos	128,89	126,30
Proyectos Grandes	183,45	148,51

Para realizar un análisis más detallado del comportamiento de las fórmulas por tamaño de proyecto se utilizan gráficos Boxplot (ver tabla 5). Estos gráficos permiten ver en un único gráfico los datos correspondientes a los límites superior e inferior (valores máximo y mínimo), el desvío máximo (media más la desviación estándar) y mínimo (media menos la desviación estándar) y la media de los resultados obtenidos en el

experimento. Al realizar la primera observación de los gráficos de la tabla 5, se nota que los costos de ambas fórmulas poseen un solapamiento entre sí, siendo mayor para la fórmula de 8 factores de costo (variable MM8) que la de 23 factores de costo (MM23). Además se puede observar que los valores de MM23 poseen costos estimados dispersos entre 33,16 meses/hombre (valor mínimo para proyectos pequeños) y 234,14 meses/hombre (valor máximo para proyectos grandes) y los valores de MM8 se encuentran comprendidos entre 71,45 y 183,09 meses/hombre. Como se dijo anteriormente, esto se debe a que la función de MM23 es más conservadora que MM8. Al observar en los gráficos el desvío estándar donde están contenidos la mayor cantidad de los proyectos (más del 70% de los proyectos para cada muestra de 10.000 proyectos) se confirma este comportamiento.

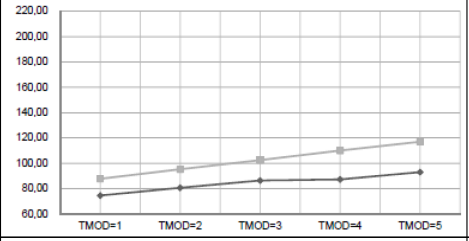
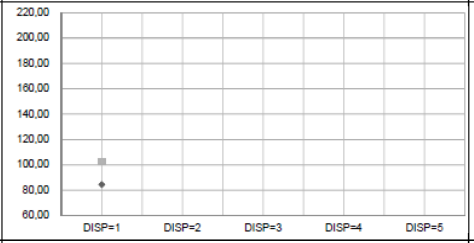
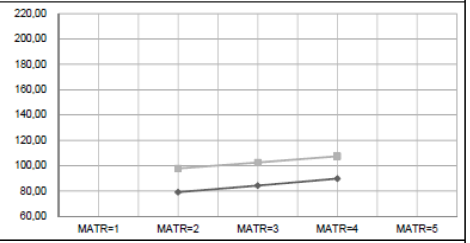
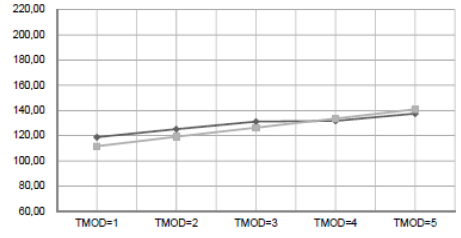
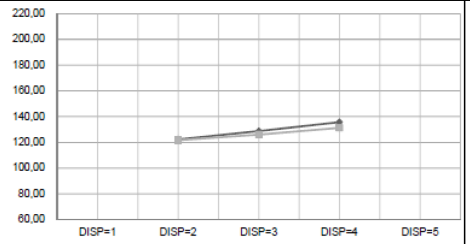
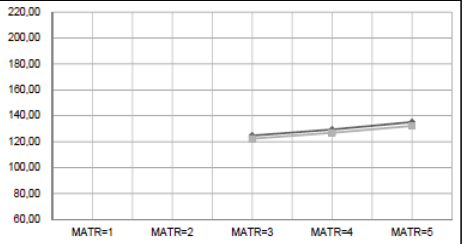
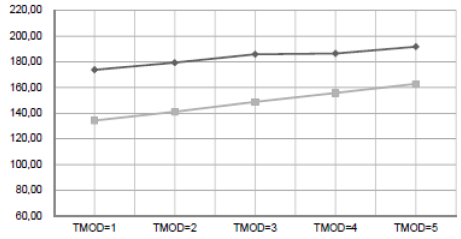
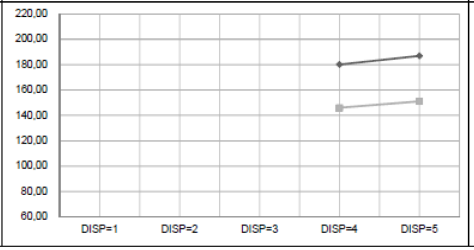
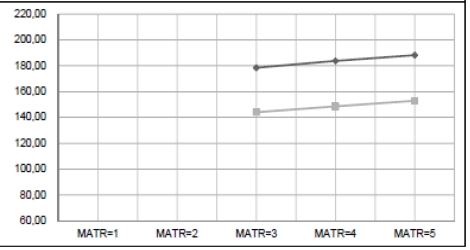
Tabla 5. Gráficos Boxplot por fórmula

MM23	MM8
	

Luego de este análisis preliminar se seleccionaron los tres factores de costo de DMCoMo que producen el mayor impacto en el incremento del esfuerzo estimado (por poseer los mayores coeficientes positivos, es decir los factores de costo TMOD, DISP y MATR) para generar los gráficos de la tabla 6 donde se muestra como la variación de sólo un factor de costo (columnas de la tabla) afecta a las fórmulas (línea negra para MM23 y línea gris para MM8) dependiendo del tamaño del proyecto (filas de la tabla). Al analizar estos gráficos por columna (o sea por factor de costo) se puede observar que:

- Para el factor de costo 'Tipo de Modelo a Ser Creado' (TMOD):
  - En los Proyectos Pequeños el crecimiento de ambas fórmulas es el mismo para Modelos Descriptivos (TMOD=1, TMOD=2 y TMOD=3) por ser las rectas paralelas, pero en los Modelos Predictivos (TMOD=4 y TMOD=5) los valores estimado por MM23 no crecen igual que MM8. Además se puede ver que los valores estimados por MM8 son siempre superiores a los de MM23.
  - En los Proyectos Medianos los valores estimados por ambas fórmulas son muy cercanos, siendo casi idénticos para los Modelos Predictivos.
  - En los Proyectos Grandes, sucede algo similar que en los proyectos pequeños: para los Modelos Descriptivos el crecimiento de ambas fórmulas es similar pero para proyectos Predictivos los valores estimado por MM23 no crecen en la misma medida que MM8 (los cuales son siempre inferiores a los de MM23).

Tabla 6 Gráficos de variación de las variables que aumentan el costo estimado donde 

	Variando sólo TMOD	Variando sólo DISP	Variando sólo MATR
Proyectos Pequeños			
Proyectos Medianos			
Proyectos Grandes			

En general al analizar el factor de costo TMOD se podría decir que construir un sistema de explotación de información con Modelos Predictivos posee un costo estimado mayor que utilizando Modelos Descriptivos.

- Para el factor de costo 'Grado de Dispersión de Datos' (DISP):
  - En los Proyectos Pequeños solamente pueden tomar un valor (dispersión menor al 20%) donde el valor estimado por MM8 es superior al de MM23.
  - En los Proyectos Medianos solamente toma dos valores (correspondientes a una dispersión entre 20 y 80%). Los valores estimados por ambas fórmulas son muy cercanos siendo más similares a medida que la dispersión es menor (o sea a medida que el valor de DISP es menor).
  - En los Proyectos Grandes solamente toma dos valores (correspondientes a una dispersión mayor al 60%). Se ve que las dos rectas son casi paralelas por lo que ambas fórmulas crecen de la misma manera pero, al contrario que con proyectos pequeños, los valores estimados por MM23 son superiores a los de MM8.

  Por lo tanto se podría al analizar este factor de costo decir que utilizar datos con mayor dispersión de valores trae aparejado un mayor costo estimado.
- Para 'Cantidad y Tipos de los Atributos por cada Modelo' (MATR):
  - En los Proyectos Pequeños el crecimiento de los valores estimados por ambas fórmulas es similar. Además se puede ver que para estos proyectos los valores estimados por MM8 son siempre superiores a los estimados por MM23.
  - En los proyectos medianos los valores estimados por ambas fórmulas son muy cercanos.
  - En los Proyectos Grandes se mantiene el crecimiento similar de ambas fórmulas pero, al contrario que para los proyectos pequeños, los valores estimados por MM8 son siempre inferiores a los estimados por MM23.

  En general se podría decir que utilizar fuentes de datos con mayor cantidad y tipos de atributos trae aparejado un mayor costo estimado.

Por otro lado, al analizar los gráficos de la tabla 6 por fila (o sea por tamaño de proyecto) se puede ver que:

- Para Proyectos Pequeños, los valores estimados de ambas fórmulas poseen valores más extremos con la variación del factor de costo TMOD que con respecto a DISP y MATR donde se ve una diferencia casi constante.
- Para los Proyectos Medianos, la diferencia entre MM8 y MM23 es casi inexistente para los factores de costo DISP y MATR mientras que para TMOD posee una distancia mayor, sobre todo cuando toma valores correspondientes a Modelos Descriptivos.
- En los Proyectos Grandes se mantiene el crecimiento similar de ambas fórmulas pero, al contrario que para los proyectos pequeños, los valores estimados por MM8 son siempre inferiores a los estimados por MM23.

Por último, para finalizar el análisis de los resultados obtenidos, se seleccionaron otros tres factores de costo, en este caso los que producen el mayor impacto en la disminución del esfuerzo estimado en el DMCoMo (por poseer los mayores coeficientes negativos, factores de costo NFOR, TOOL y MFAM) para generar los gráficos de la tabla 7 con la misma estructura que la tabla anterior.

Al analizar estos gráficos por factor de costo se puede observar que:

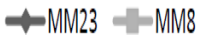
- Para ‘Nivel de Formación de los Usuarios en las Herramientas’ (NFOR):
  - En los Proyectos Pequeños el crecimiento de los valores estimados por ambas fórmulas es similar. Además se puede ver que para estos proyectos los valores estimados por MM8 son siempre superiores a los estimados por MM23.
  - En los Proyectos Medianos los valores estimados por ambas fórmulas son muy cercanos, tomando valores casi idénticos cuando NFOR=4 (‘Se requiere conocimiento en técnicas de explotación de información y conocimiento experto en las herramientas’) y NFOR=5 (‘Se requiere conocimiento experto en técnicas de explotación de información y en las herramientas’).
  - En los Proyectos Grandes se mantiene el crecimiento similar de ambas fórmulas, pero, al contrario que para los proyectos pequeños, los valores estimados por MM8 son siempre inferiores a los estimados por MM23.

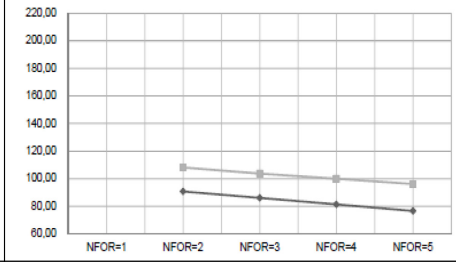
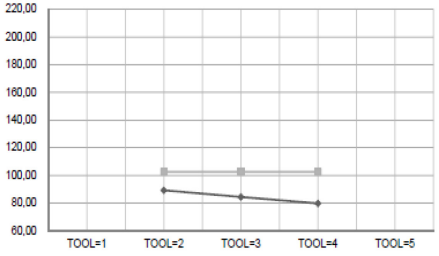
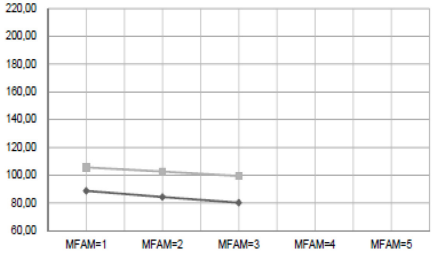
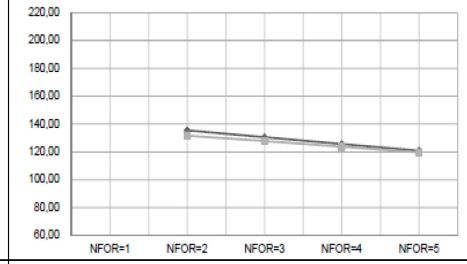
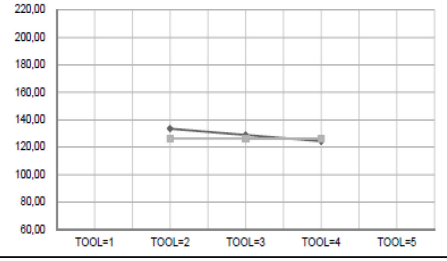
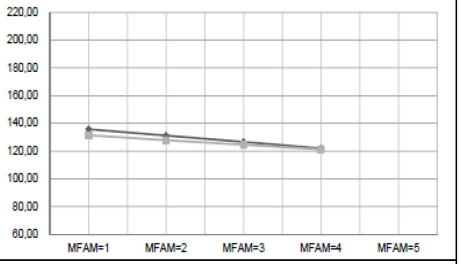
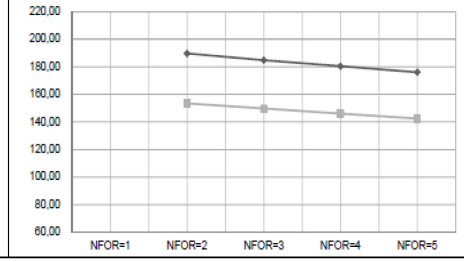
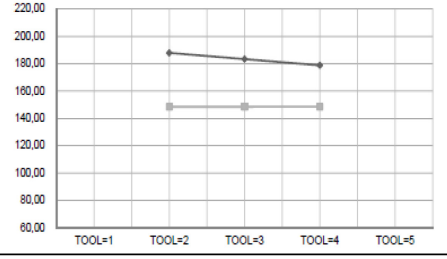
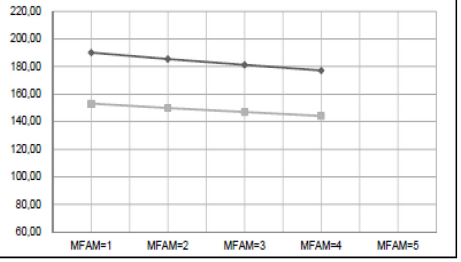
  Así, parecería que el costo estimado disminuye a media que se requiere mayor conocimiento experto en técnicas de en técnicas de explotación de información y en las herramientas (o sea, mayor valor del factor de costo). Esto indicaría que al requerirse mayor experiencia para manejar las herramientas, el equipo de trabajo podrá finalizar el proyecto en menos tiempo que un equipo sin experiencia con herramientas que poseen asistentes para su uso.
- Para ‘Herramientas Disponibles para Ser Usadas’ (TOOL) el crecimiento de MM23 está afectado por la variación de los valores del factor de costo, pero no afecta a MM8 que posee un valor constante para cualquier valor de TOOL.

  Además parecería que el costo estimado por MM23 disminuye a media que se usa menos cantidad de herramientas (o sea, a medida que el valor del factor de costo crece). Esto no tiene mucho sentido porque indicaría que es preferible no utilizar ninguna herramienta para realizar el proyecto en menos tiempo y menor costo.
- Para ‘Grado de Familiaridad con el Tipo de Problema’ (MFAM):
  - En los Proyectos Pequeños el crecimiento de los valores estimados por ambas fórmulas es similar. Además se puede ver que para estos proyectos los valores estimados por MM8 son siempre superiores a los estimados por MM23.
  - En los Proyectos Medianos los valores estimados por ambas fórmulas son muy cercanos, tomando valores casi idénticos cuando MFAM=3 (‘El equipo ha trabajado en tipos de proyectos igual al del nuevo proyecto pero con datos diferentes’) y MFAM=4 (‘El equipo ha trabajado en tipos de proyectos igual al del nuevo proyecto pero nunca en el mismo ambiente’).
  - En los Proyectos Grandes se mantiene el crecimiento similar de ambas fórmulas, pero, al contrario que para los proyectos pequeños, los valores estimados por MM8 son siempre inferiores a los estimados por MM23.

  En general parecería que el costo estimado disminuye a media que el equipo de trabajo posee menos experiencia trabajando en proyectos de explotación de información (o sea, a medida que el valor del factor de costo crece). Esto no tiene sentido porque indicaría que es preferible tener un equipo de trabajo con menos conocimiento y experiencia para realizar el proyecto en menos tiempo y menor costo.

Finalmente, al analizar los gráficos de la tabla 7 por tamaño de proyecto se puede ver que sin importar el tamaño del proyecto las fórmulas MM23 y MM8 tienen un comportamiento similar para los tres factores de costo analizados, presentado en todos los casos tanto valores extremos (máximos y mínimos) como pendientes similares.

Tabla 7. Gráficos de variación de las variables que diminuyen el costo estimado donde 

	Variando sólo NFOR	Variando sólo TOOL	Variando sólo MFAM
Proyectos Pequeños			
Proyectos Medianos			
Proyectos Grandes			

6. Conclusiones

A partir del análisis del experimento realizado se concluye que la fórmula de 23 factores de costo de DMCoMo es más conservadora por crecer casi el doble que la de 8 factores de costo al aumentar el tamaño de proyecto a estimar. Del estudio detallado de los factores de costo, se puede indicar que para los proyectos medianos el costo estimado por ambas fórmulas es muy similar (casi idéntico en algunos casos), pero esto no sucede en los otros tamaños de proyectos. En los proyectos pequeños los valores estimados por la fórmula de 8 factores de costo son siempre superiores, mientras que en los proyectos grandes sucede lo contrario. Ambas fórmulas poseen un comportamiento diferente con respecto al tipo de modelo (descriptivos o predictivos) a desarrollar en el proyecto, obteniendo un costo estimado mayor para los modelos predictivos. En el caso de herramientas disponibles y el grado de familiaridad con el tipo de problema, se detecta una incongruencia entre el comportamiento de las fórmulas y la realidad. Según DMCoMo el costo estimado parecería reducirse al poseer menor cantidad de herramientas disponible y menor experiencia trabajando en proyectos similares, lo cual se contradice con el sentido común. Queda como futura línea de investigación, estudiar la variación de los restantes factores de costo para lograr así un análisis completo del comportamiento de DMCoMo.

7. Financiamiento

Las investigaciones que se reportan en este artículo han sido financiadas parcialmente por el Proyecto de Investigación 33A105 del Departamento de Desarrollo Productivo y Tecnológico de la Universidad Nacional de Lanús, por el Proyecto de Investigación 40B065 de la Universidad Nacional de Rio Negro - Sede Andina (El Bolsón), y por el Proyecto 25C126 de la Facultad Regional Buenos Aires de la Universidad Tecnológica Nacional.

8. Referencias

1. Pressman, R. 2004. *Software Engineering: A Practitioner's Approach*. Mc Graw Hill.
2. Boehm, B., Abts, C., Brown, A., Chulani, S., Clark, B., Horowitz, E., Madachy, R., Reifer, D., Steece, B. 2000. *Software Cost Estimation with COCOMO II*, Prentice- Hall.
3. LLC PRICE Systems. 1998. *PRICE S Reference Manual Version 3.0*, Lockheed-Martin.
4. Marbán, O. 2003. *Modelo Matemático Paramétrico de Estimación para Proyectos de Data Mining (DMCoMo)*. Tesis de Doctorado en Informática. Universidad Politécnica de Madrid.
5. Rodríguez, D., Pollo-Cattaneo, F., Britos, P., García-Martínez, R. (2010). *Estimación Empírica de Carga de Trabajo en Proyectos de Explotación de Información*. Anales del XVI CACIC. Pág. 664-673. ISBN 978-950-9474-49-9.
6. García-Martínez, R., Lelli, R., Merlino, H., Cornachia, L., Rodriguez, D., Pytel, P., Arboleya, H. (2011). *Ingeniería de Proyectos de Explotación de Información para PYMES*. Proceedings XIII WICC. Pág. 253-257. ISBN 978-950-673-892-1.
7. Pytel, P., Tomasello, M., Rodríguez, D.; Arboleya, H., Pollo-Cattaneo. M., Britos, P., García-Martínez, R. (2011). *Estimación de Proyectos de Explotación de Información. Estudio Comparado de Modelos Analíticos y Empíricos*. Proceedings XIII WICC. Pág. 295-299. ISBN 978-950-673-892-1.
8. Kalos, M.H.; Whitlock P.A. (1986) *Monte Carlo Methods. Vol I. Basics*. Wiley & Sons.