

Comparison of Estimation Metrics for Information Mining Projects

Comparación de Métricas de Estimación para Proyectos de Explotación de Información

Pablo Pytel¹, Paola Britos², Ramón García-Martínez³

¹ Programa de Doctorado en Ciencias Informáticas. Facultad de Informática. Universidad Nacional de La Plata. Argentina. ppytel@gmail.com

² Grupo de Investigación en Explotación de Información. Laboratorio de Informática Aplicada. Universidad Nacional de Río Negro. Argentina. paobritos@gmail.com

³ Grupo Investigación en Sistemas de Información. Departamento Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús. Argentina. rgarcia@unla.edu.ar

INFORMACIÓN DEL ARTÍCULO

Tipo de artículo: Investigación

Historia del artículo

Recibido: 23/04/2012

Correcciones: 29/05/2012

Aceptado: 31/05/2012

Palabras clave

Aseguramiento de calidad, métricas métodos de estimación de esfuerzo, explotación de información, ingeniería en software.

Categories and Subject Descriptors

D.2.8 [Software Engineering]: Metrics – Process metrics.

General Terms

Software Engineering, Validation

Keywords

Quality assurance, metrics, effort estimation method, information mining, software engineering.

ABSTRACT

Software Quality Assurance Software is a "protection activity" that applies through the whole process of software engineering. Among other mechanisms, it includes the metric measurement of the product and project. On the other hand, to improve processes of Information Mining projects it is necessary to register quality metrics as the models for assessing project productivity that can be used to establish targets for improvement. These models need the estimation of the activities effort at the beginning of the project. In this context, this paper has the objective of analysing two existing estimation methods for Information Mining Projects, a method oriented to relatively large-sized projects and another oriented to small-sized projects which are normally required by the Small and Medium Enterprises (SMEs).

RESUMEN

El Aseguramiento de la Calidad del Software es una "actividad de protección" que se aplica a lo largo de todo el proceso de ingeniería software incluyendo entre otros los mecanismos de medición sobre métricas del producto y del proyecto. Por otro lado, para mejorar los procesos correspondientes al desarrollo de Proyectos de Explotación de Información se identifica la necesidad de registrar métricas de calidad como los modelos para evaluar la productividad del proyecto que permiten establecer objetivos de mejora. Para ello es necesario estimar los esfuerzos al comienzo del proyecto. De esta manera, en este contexto, este trabajo tiene el objetivo de analizar dos métodos de estimación existentes para Proyectos de Explotación de Información, uno orientado a proyectos con tamaño relativamente grande y otro orientado a proyectos pequeños que son los normalmente requeridos por las Pequeñas y Medianas Empresas (PyMEs).

1. INTRODUCCIÓN

El Aseguramiento de la Calidad del Software (SQA) es una "actividad de protección" que se aplica a lo largo de todo el proceso de ingeniería software [1] incluyendo entre otros los mecanismos de medición sobre métricas del producto y del proyecto. Estas métricas comprenden un amplio rango de actividades que incluyen el aseguramiento y control de calidad, modelos de fiabilidad, modelos para evaluación de ejecución y modelos para evaluar la productividad del proyecto. Al conocer el estado actual de desarrollo, permite establecer objetivos de mejora [2].

Por otro lado, la Explotación de Información es una sub-disciplina de la informática relativamente nueva y vinculada a la Inteligencia de Negocio [3]. Esta disciplina engloba un conjunto de técnicas encaminadas a la extracción de conocimiento procesable, implícito en el almacén de datos de la organización. Dicho conocimiento es previamente desconocido y puede resultar útil para algún proceso [4, 5]. Al intentar llevar adelante diferentes Proyectos de Explotación de Información con un alto grado de previsibilidad y

calidad se utilizan distintos modelos de producción y metodologías [6]. Esto incluye el registro de métricas para suministrar información relevante a tiempo y para intentar mejorar tanto los procesos como los productos. Dentro de las métricas a utilizar se destaca la medición del esfuerzo del proyecto para evaluar la productividad del proyecto. Para ello se debe realizar estimaciones al comienzo del proyecto, las cuales son comparadas con los valores reales del proyecto a su finalización. De esta manera se destaca la necesidad de contar con métodos de estimación de esfuerzo confiables para Proyectos de Explotación de Información.

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. Por ejemplo, las herramientas de estimación de software convencional como COCOMO II [7] o PRICE-S [8] 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. Por otro lado, en [9] 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). Pero según sus autores este método se puede considerar confiable en un rango de esfuerzo muy superior al de los proyectos pequeños que son los que usualmente requieren las Pequeñas y Medianas Empresas [10].

Por lo tanto, en este contexto este trabajo tiene el objetivo de analizar dos métodos de estimación existentes para Proyectos de Explotación de Información. Para ello primero se realiza un análisis del método de estimación DMCoMo (sección 2), luego se describe un nuevo método de estimación orientado a las Pequeñas y Medianas Empresas (PyMEs) propuesto por los autores (sección 3) realizando una comparación de ambos métodos utilizando datos de proyectos reales (sección 4). Finalmente se indican algunas conclusiones y las futuras líneas de trabajo (sección 5).

2. ANÁLISIS DEL MÉTODO DMCoMo

En esta sección se realiza el análisis del método de estimación Matemático Paramétrico de Estimación para Proyectos de Data Mining (en inglés Data Mining Cost Model, o DMCoMo). Primero se realiza una breve descripción del método (sección 2.1) se delimita el problema detectado (sección 2.2) y se presentan los resultados del análisis realizado (2.3) con sus conclusiones (2.4).

2.1 Descripción del Método DMCoMo

El método DMCoMo [9] es un modelo analítico de estimación de la familia de COCOMO [7] 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. DMCoMo define 23 factores de costo los cuales están vinculados a las características más importantes de los proyectos de explotación de información. Estos factores de costo se clasifican en seis categorías que se indican 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 (Tabla 2), una que utiliza 23 factores de costo como variables (fórmula denominada como MM23) que puede ser utilizada cuando el proyecto está bien definido y otra de 8 factores de costo como variables

(fórmula MM8) 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.

Tabla 1. Categorías y Factores de Costo de DMCoMo

CATEGORÍA	FACTOR DE COSTO
Relacionados a los Datos	<ul style="list-style-type: none"> ▪ 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	<ul style="list-style-type: none"> ▪ 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	<ul style="list-style-type: none"> ▪ 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	<ul style="list-style-type: none"> ▪ 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	<ul style="list-style-type: none"> ▪ 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	<ul style="list-style-type: none"> ▪ Grado de Familiaridad con el Tipo de Problema (MFAM) ▪ Grado de Conocimiento de los Datos (KDAT) ▪ Actitud de los Directivos (ADIR)

Tabla 2. Fórmulas de DMCoMo

$MM23 = 78,752 + 2,802 \times NTAB + 1,953 \times NTUP + 2,115 \times NATR + 0,345 \times PNUL + (-2,656) \times DMOD + 2,586 \times DEXT + (-0,456) \times NMOD + 6,032 \times TMOD + (-4,543) \times MFAM + 4,312 \times MTUP + 4,966 \times MATR + (-2,591) \times MTEC + 3,943 \times NFUN + 0,896 \times SCOM + (-4,615) \times TOOL + (-1,831) \times COMP + (-4,689) \times NFOR + (-3,756) \times ADIR + 2,931 \times NDEP + (-0,892) \times DOCU + 2,135 \times SITE + (-0,214) \times KDAT$
$MM8 = 70,897 + 2,368 \times NTAB + (-3,275) \times MFAM + 2,885 \times NATR + 4,792 \times DISP + (-3,842) \times NFOR + 2,713 \times DEXT + 7,257 \times TMOD + 4,615 \times MATR$

2.2 Problema Detectado del Método DMCoMo

El problema identificado es motivado por la propia limitación señalada en [9] debido a las características de los 15 proyectos utilizados para la validación del método. Los autores declaran 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 (es decir 7,5 a 15,42 años/hombre). Si el esfuerzo del proyecto se encuentra fuera de este rango, el comportamiento del método es desconocido.

Sin embargo, por experiencia se conoce que los proyectos de explotación de información utilizados en las PyMEs poseen un esfuerzo mucho menor a los 90 meses/hombre. Para realizar una revisión inicial del comportamiento de DMCoMo se ha utilizado un proyecto de tamaño pequeño.

A partir de las características del proyecto, se definen los valores de los factores de costo y se calculan los esfuerzos correspondientes a las fórmulas (ver Tabla 3). El objetivo del proyecto era la detección de evidencias de causalidad entre Satisfacción General e Internet, y fue desarrollado por 3 personas en 4 meses (es decir que el esfuerzo total es 12 meses/hombre). Como se puede ver, el método DMCoMo sobreestima el esfuerzo requerido para el proyecto en aproximadamente un 630%, con un error para la fórmula MM23 de 64,56 meses/hombre, y de 63,86 meses/hombre para MM8.

Tabla 3. Revisión inicial de DMCoMo

Factor de Costo	Valor	Factor de Costo	Valor
ADIR	1	NFOR	3
COMP	3	NFUN	3
DEXT	2	NMOD	3
DISP	1	NTAB	0
DMOD	5	NTUP	1
DOCU	5	PNUL	2
KDAT	2	SCOM	4
MATRn	3	SITE	3
MATRt	1	TMOD	1
MFAM	4	TOOL	1
MTEC	2	NATR	1
MTUP	2	NDEP	2
Esfuerzo Real del Proyecto = 12 meses/hombre			
Esfuerzo Estimado por MM23 = 76,56 meses/hombre			
Esfuerzo Estimado por MM8 = 75,86 meses/hombre			

Por lo tanto, se considera necesario realizar un estudio detallado del comportamiento de DMCoMo con una particular focalización en proyectos pequeños. Para ello se ha utilizado el método de simulación Monte Carlo [11] generando en forma pseudo-aleatoria un banco de pruebas con los datos de 30.000 proyectos de explotación de información (distribuidos en forma equitativa en tres rangos de tamaño: Proyectos Pequeños, Medianos y Grandes). Luego se han aplicado las fórmulas MM23 y MM8 definidas por DMCoMo para calcular los esfuerzos estimados (Datos disponibles en <http://tinyurl.com/DMCoMoData>).

2.3 Resultados del Análisis Realizado para DMCoMo

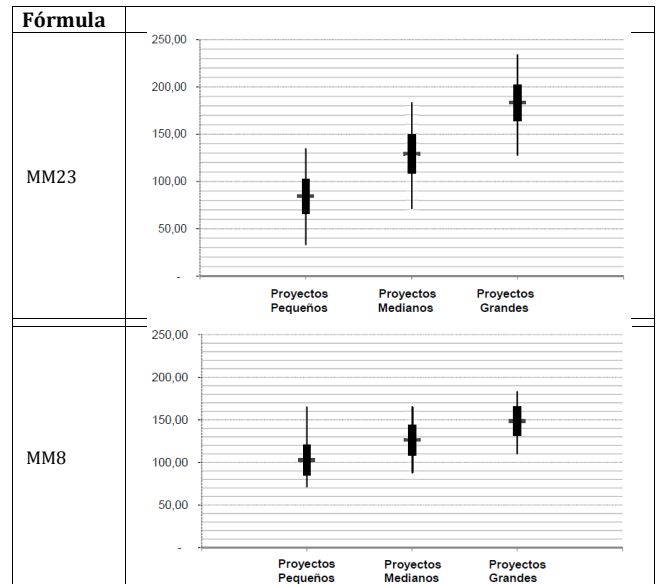
A partir del banco de prueba con los proyectos generados en forma pseudo-aleatoria se procede a realizar un análisis estadístico de los mismos. En la Tabla 4 se indica la media estadística obtenida por cada una de las fórmulas de DMCoMo y tamaño de proyecto. De esta tabla se puede ver que la media de la fórmula 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 MM8 posee un crecimiento menor por escalones de aproximadamente el 22% con respecto al tamaño anterior.

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.

Tabla 5. Gráficos Boxplot por fórmula



Al realizar la primera observación de estos gráficos, 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. Entonces, debido a que al variar el tamaño del proyecto la fórmula MM23 crece aproximadamente el doble con respecto a MM8, se puede decir que es más conservadora. 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.

2.4 Conclusiones para DMCoMo

A partir de los análisis realizados 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 MM8 siempre son superiores a los estimados por MM23, mientras que en los proyectos grandes sucede lo contrario. En todos los casos los mínimos valores estimados son de 33,16 meses/hombre para MM23 y 71,45 meses/hombre para MM8. Por lo que se puede concluir que DMCoMo siempre tiende a

sobreestimar los esfuerzos de los proyectos. Esto significa que a pesar de que se podría aplicar para proyectos medianos y grandes, el método no es recomendable para proyectos pequeños.

3. MÉTODO DE ESTIMACIÓN PARA PyMEs

Debido a la necesidad detectada de contar con un método de estimación fiable para proyectos de tamaño pequeño, los autores han propuesto un nuevo método de estimación analítico con énfasis en las características de las PyMEs. Para ello primero se identifican y describen las principales características de los proyectos en PyMEs que son utilizadas para identificar los factores de costo del método propuesto y luego se procede a especificar la fórmula mediante regresión. Tanto para la definición del método propuesto y su posterior validación se ha utilizado información de 44 proyectos reales de explotación de información (<http://tinyurl.com/proy-PYMES>) que fueron recolectados por investigadores del Grupo de Investigación en Sistemas de Información del Departamento de Desarrollo Productivo y Tecnológico de la Universidad Nacional de Lanús (GISI-DDPyT-UNLa), investigadores del Grupo de Estudio en Metodologías de Ingeniería de Software de la Facultad Regional Buenos Aires de la Universidad Tecnológica Nacional (GEMIS-FRBA-UTN), e investigadores del Grupo de Investigación en Explotación de Información de la Sede Andina (El Bolsón) de la Universidad Nacional de Río Negro (SAEB-UNRN).

Debe notarse que todos estos proyectos fueron realizados utilizando la metodología CRISP-DM [12], por lo que el método propuesto se considera confiable para proyectos de explotación de información a ser desarrollados con dicha metodología.

3.1 Principales Características de Proyectos de Explotación de Información en PyMEs

Según el informe de las PyMEs y el reporte de espíritu empresarial [13] de la Organización para la Cooperación y el Desarrollo Económico (OCDE) "las PyMEs constituyen la forma dominante de organización empresarial en todos los países de todo el mundo, representando más del 95% y hasta el 99% de la población de empresas según el país". Sin embargo, a pesar de que es conocida la importancia de las PyMEs, no existe un criterio universal para identificarlas. Dependiendo de cada país y región se utilizan diferentes criterios cuantitativos y cualitativos para reconocer a una organización como PyME.

De esta forma, en Latinoamérica cada país tiene una definición diferente [14]: Argentina identifica como PyME a las empresas autónomas que poseen una facturación menor a US\$ 20.000 por año (monto máximo que depende de la actividad realizada); Brasil incluye a todas las compañías con 500 empleados o menos mientras que Colombia considera como PyME a las empresas que poseen hasta 200 empleados y activos menores a los US\$ 6.500. En este sentido la Organización Internacional para la Estandarización

(más conocida como International Organization for Standardization o ISO) ha reconocido la necesidad de especificar definir los perfiles de ciclos de vida para proyectos de ingeniería en software en pequeñas entidades (denominadas en inglés 'Very Small Entities' o VSE) y se encuentra trabajando en el estándar ISO/IEC 29110 [15]. El término VSE fue definido por el grupo de trabajo 24 de SO/IEC JTC1/SC7 como cualquier "empresa, organización, departamento o proyecto que cuenta con a lo más 25 personas" [16].

A partir de estas definiciones y nuestra propia experiencia, en este trabajo contempla que un proyecto de explotación de información para PyMEs se encuentra demarcado como un proyecto realizado en una empresa de 250 empleados o menos donde los gerentes de alto nivel (por lo general los propietarios de la empresa) necesitan obtener conocimiento no trivial extraído de las bases de datos disponibles para resolver un problema de negocio específico, sin riesgos especiales en juego. Como normalmente los miembros de la empresa no tienen los conocimientos necesarios, el proyecto es realizado por consultores especializados contratados para llevar adelante el proyecto. Desde nuestra experiencia podemos restringir al equipo del proyecto en un máximo de 25 personas (incluyendo tanto los consultores subcontratados y al personal de la empresa involucrada) para realizar el proyecto en menos de un año.

Las primeras tareas de un proyecto de explotación de información son similares a las de un proyecto de desarrollo de software convencional, dado que se deben educir las necesidades y deseos de los interesados (*stakeholders*) de la organización. Pero, en estos proyectos, además se necesita conocer las fuentes de información disponibles en la organización por lo que es preciso relevar los repositorios existentes junto con su estructura. Como estos repositorios suelen no estar correctamente documentados, es necesario entrevistar también a los expertos en datos de la organización. Ya que estos expertos son escasos y con poca disponibilidad, será necesario entonces requerir a su buena disposición (y la de sus jefes) para que participen en las sesiones de educación para identificar las características de la organización y de los repositorios a ser utilizados.

Por otro lado, la infraestructura de la Tecnologías de la Información y la Comunicación (TIC) de las PyMEs es analizada. En [17] se indica que en Latinoamérica más del 70% de las PyMEs cuentan con una infraestructura informática, pero sólo el 37% posee servicios automatizados y/o software propio para realizar sus actividades. En general, hacen uso de aplicaciones comerciales (sobre todo manejadores de planillas de cálculo y de documentos) para registrar la información comercial y operativa. Esto significa que los repositorios a ser utilizados en el proyecto estarán implementados en diferentes formatos y tecnologías. Aunque estos repositorios no suelen ser grandes (normalmente no superan el millón de registros), las tareas de

preparación de datos (es decir, limpieza, formateo e integración de los datos) tendrán un esfuerzo considerable. Este esfuerzo puede reducirse si se dispone de una herramienta software que las posee ya implementadas. En ese caso no se necesitará desarrollar un software específico para realizarlas.

3.2 Identificación de los Factores de Costo del Método de Estimación Propuesto para PyMEs

Teniendo en cuenta las características de los proyectos de Explotación de Información para PyMEs que se indican en la sección 3.1, se definen ocho factores de costos. Se han definido pocos factores de costo, ya que como se demuestra en [18], al momento de crear un nuevo método de estimación es preferible ignorar muchos de los datos no significativos para evitar que el modelo sea demasiado complejo y por lo tanto poco práctico. De esta manera se busca eliminar las variables tanto irrelevantes como dependientes, y además reducir la varianza y el ruido. Los factores de costo han sido seleccionados teniendo en cuenta las tareas más críticas de la metodología CRISP-DM: en [19] se indica que actualmente la construcción de los modelos de minería de datos y buscar los patrones es bastante simple, pero el 90% de los esfuerzos del proyecto están incluidos en el pre-procesamiento de los datos (es decir la fase de 'Preparación de los Datos' de CRISP-DM).

A partir de nuestra experiencia, las otras tareas críticas se relacionan con la fase de 'Comprensión del Negocio' (entre las que se destacan el entendimiento del negocio y la identificación de los goles del proyecto). Los factores de costos propuestos se encuentran agrupados en tres grupos dependiendo de su naturaleza:

Factores de costo relacionados al proyecto:

- *Tipo de objetivo de explotación de información (OBTY).* Este factor de costo analiza el objetivo del proyecto de Explotación de Información considerando el tipo de proceso a ser aplicado para obtenerlo de acuerdo a la definición realizada en [20]. Los posibles valores de este factor de costo se indican en la Tabla 6.

Tabla 6. Valores del factor de costo OBTY

Valor	Descripción
1	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida.
2	Se desea dividir los datos disponibles en grupos sin poseer una clasificación conocida previamente.
3	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente.
4	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre un comportamiento o la identificación de una clase conocida.
5	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre la identificación de una clase desconocida previamente.

- *Grado de apoyo de los miembros de la organización (LECO).* El grado de apoyo y participación de los miembros de la organización se analiza viendo si la alta gerencia (normalmente los dueños de la PyME), la gerencia media (supervisores y/o jefes de área) y/o el resto del personal están dispuestos a ayudar al equipo de trabajo para comprender el negocio y los

datos. Se sobreentiende que si un proyecto de explotación de información fue contratado, por lo menos la alta gerencia va a apoyar el mismo. Los posibles valores de este factor de costo se indican en la Tabla 7.

Tabla 7. Valores del factor de costo LECO

Valor	Descripción
1	Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto.
2	Sólo los directivos poseen buena disposición para colaborar en el proyecto mientras que el personal es indiferente al proyecto.
3	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto mientras que la gerencia media y el personal es indiferente.
4	Sólo la alta gerencia posee buena disposición para colaborar en el proyecto pero la gerencia media no desea colaborar.

Factores de costo relacionados a los datos:

- *Cantidad y tipo de los repositorios de datos disponibles (AREP).* Se analizan los repositorios de datos disponibles (es decir sistemas gestores de bases de datos, planillas de cálculos, documentos entre otros). En este caso interesa saber tanto la cantidad de repositorios disponibles (públicos o privados de la organización) como la tecnología en que se encuentran implementadas. No interesa conocer la cantidad de tablas que posee cada repositorio dado que se entiende que la integración de los datos dentro de un repositorio es relativamente sencilla (sobre todo al utilizar sistemas gestores de bases de datos por poder ser realizada con un comando *query*). Sin embargo, dependiendo de la tecnología, la complejidad de las tareas de integración de los datos puede ser mayor o menor. Criterios recomendados:
 - Si todos los repositorios están implementados con la misma tecnología, entonces se consideran como compatibles para la integración.
 - Si todos los repositorios permiten exportar los datos en un formato común, entonces pueden ser considerados como compatibles para la integración al realizar la integración con estos datos exportados.
 - Por otro lado, si existen repositorios que no están en forma digital (es decir impreso en papel) se considera que la tecnología será no compatible pero el método de estimación no puede predecir el tiempo requerido para realizar la digitalización de esta información ya que esto puede variar de acuerdo a muchos factores externos (como son la longitud, diversidad, formato entre otros).

Los posibles valores de este factor se indican en la Tabla 8.

Tabla 8. Valores del factor de costo AREP

Valor	Descripción
1	Sólo 1 repositorio disponible
2	Entre 2 y 4 repositorios con tecnología compatible para la integración
3	Entre 2 y 4 repositorios con tecnología no compatible para la integración
4	Más de 5 repositorios con tecnología compatible para la integración
5	Más de 5 repositorios con tecnología no compatible para la integración

- **Cantidad de tuplas disponibles en la tabla principal (QTUM).** Este factor de costo considera la cantidad total de tuplas (registros) disponibles en la tabla principal utilizada para aplicar los algoritmos de minería de datos. Los posibles valores de este factor de costo se indican en la Tabla 9.

Tabla 9. Valores del factor de costo QTUM

Valor	Descripción
1	Hasta 100 tuplas en la tabla principal
2	Entre 101 y 1.000 tuplas en la tabla principal
3	Entre 1.001 y 20.000 tuplas en la tabla principal
4	Entre 20.001 y 80.000 tuplas en la tabla principal
5	Entre 80.001 y 5.000.000 tuplas en la tabla principal
6	Más de 5.000.000 tuplas en la tabla principal

- **Cantidad de tuplas disponibles en tablas auxiliares (QTUA).** Esta variable considera la cantidad aproximada de tuplas (registros) disponibles en las tablas auxiliares (si existieran) que son utilizadas para agregar información a la tabla principal (como es la tabla que define las características del producto a partir de su identificador en la tabla de ventas). Estas tablas auxiliares normalmente suelen tener menos registros que la tabla principal. Los posibles valores de este factor se indican en la Tabla 10.

Tabla 10. Valores del factor de costo QTUA

Valor	Descripción
1	No se utilizan tablas auxiliares
2	Hasta 1.000 tuplas en las tablas auxiliares
3	Entre 1.001 y 50.000 tuplas en las tablas auxiliares
4	Más de 50.000 tuplas en las tablas auxiliares

- **Nivel de conocimiento sobre los datos (KLDS).** Considera el nivel de documentación existente sobre los repositorios de datos. Es decir, se analiza si existe un documento donde se indique la tecnología en que están implementados, los campos que componen sus tablas y la forma en que los datos son creados, modificados, y/o eliminados. En caso de que esta información no se encuentre, será necesario realizar reuniones con los expertos (encargados de la administración y mantenimiento de los repositorios). Los valores de este factor se indican en la Tabla 11.

Tabla 11. Valores del factor de costo KLDS

Valor	Descripción
1	Todas las tablas y repositorios están correctamente documentados
2	Más del 50% de las tablas y repositorios están correctamente documentados y existen expertos en los datos disponibles para explicarlos
3	Menos del 50% de las tablas y repositorios están correctamente documentados pero existen expertos en los datos disponibles para explicarlos
4	Las tablas y repositorios no están documentadas pero existen expertos en los datos disponibles para explicarlos
5	Las tablas y repositorios no están documentados y existen expertos en los datos pero no están disponibles para explicarlos
6	Las tablas y repositorios no están documentados y no existen expertos en los datos para explicarlos

Factores de costo relacionados a los recursos:

- **Nivel de conocimiento y experiencia del equipo de trabajo (KEXT).** Analiza la capacidad del equipo de trabajo que se ocupará del proyecto. El equipo de trabajo contratado para realizar el proyecto debe

tener un mínimo conocimiento y experiencia en el desarrollo de proyectos de explotación de información. No obstante pueden poseer o no experiencia en proyectos similares en el mismo tipo de negocio y los datos a ser utilizados. Por lo tanto se debe evaluar el conocimiento y experiencia previa en proyectos anteriores similares al que se está llevando a cabo con respecto al tipo de negocio, los datos a ser utilizados y los objetivos que se esperan lograr. Los valores de este factor se indican en la Tabla 12.

Tabla 12. Valores del factor de costo KEXT

Valor	Descripción
1	El equipo ha trabajado en tipos de organizaciones y con datos similares para obtener los mismos objetivos
2	El equipo ha trabajado en tipos de organizaciones similares pero con datos diferentes para obtener los mismos objetivos
3	El equipo ha trabajado en otros tipos de organizaciones y con datos similares para obtener los mismos objetivos
4	El equipo ha trabajado en otros tipos de organizaciones y con datos diferentes para obtener los mismos objetivos
5	El equipo ha trabajado en tipos de organizaciones diferentes, con datos diferentes y otros objetivos

- **Funcionalidad de las herramientas disponibles (TOOL).** Finalmente, este factor de costo evalúa las características de las herramientas disponibles para realizar el proyecto. Para ello se analiza tanto las funcionalidades de preparación de los datos como los algoritmos de minería de datos que posee implementadas. Sus posibles valores de este factor de costo se indican en la Tabla 13.

Tabla 13. Valores del factor de costo TOOL

Valor	Descripción
1	La herramienta posee funciones tanto para el formateo e integración de los datos (permitiendo importar más de una tabla de datos) como para aplicar las técnicas de minería de datos.
2	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, y permite importar más de una tabla de datos en forma independiente.
3	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de minería de datos, pero sólo permite importar una tabla de datos.
4	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos, y permite importar más de una tabla de datos.
5	La herramienta posee funciones sólo para aplicar las técnicas de minería de datos y sólo permite importar una tabla de datos.

3.3 Especificación de la Ecuación del Método de Estimación Propuesto para PyMEs

Una vez que los factores de costo fueron definidos, se han utilizado para caracterizar 34 proyectos reales de explotación de información recolectados junto con su esfuerzo real (provistos por colegas como se ha indicado anteriormente). Un método de regresión lineal multi-variante [21] fue aplicado sobre estos datos para obtener una ecuación lineal de la forma utilizada por los métodos de la familia COCOMO. Como resultado del proceso de regresión se obtiene la fórmula indicada en la Tabla 14.

Tabla 14. Fórmula del método propuesto para PyMEs

$$PEM = 0.80 \text{ OBTY} + 1.10 \text{ LECO} - 1.20 \text{ AREP} - 0.30 \text{ QTUM} - 0.70 \text{ QTUA} + 1.80 \text{ KLDS} - 0.90 \text{ KEXT} + 1.86 \text{ TOOL} - 3.30$$

donde:

- PEM es el esfuerzo estimado por el método de estimación para PyMEs (en meses/hombre)
- OBTY, LECO, AREP, QTUM, QTUA, KLDS, KEXT y TOOL son los valores correspondientes de los factores de costo definidos en las tablas 8 a 16.

4. COMPARACIÓN DE LOS MÉTODOS

Para comparar el comportamiento del método DMCoMo (descrito en la sección 2) y el método de estimación orientado para PyMEs (descrito en la sección 3) se utiliza la información de otros 10 proyectos reales recolectados (provistos por colegas como se ha indicado anteriormente). Con esta información primero se ha calculado los esfuerzos por el método DMCoMo mediante la definición de cada uno de los factores de costo y su aplicación en las fórmulas MM23 y MM8. De la misma manera, se realiza el mismo procedimiento para calcular el esfuerzo para el método propuesto para PyMEs con fórmula PEM definida en la sección 3. Con los esfuerzos ya calculados se completa la Tabla 15 donde se muestra por cada proyecto su esfuerzo real (E_{fR}), los esfuerzos calculados por el método DMCoMo (MM8 y MM23) y por el método propuesto para PyMEs (PEM) indicando también los errores absolutos (es decir la diferencia entre el esfuerzo real y el calculado por cada método) y los errores relativos (ErRel que es calculado como el error absoluto dividido por el esfuerzo real del proyecto).

Para mayor claridad, esta comparación se refleja en un gráfico boxplot (Figura 1) donde el comportamiento del esfuerzo real y los calculados se muestran indicando los valores mínimos y máximos, el rango del desvío estándar y el valor medio para cada uno.

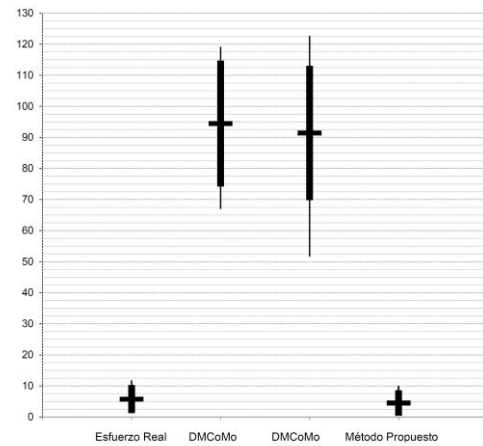


Fig. 1: Comportamiento del Esfuerzo Real, el método DMCoMo y el Método de Estimación

Tabla 15. Comparación de los esfuerzos calculados (en meses/hombre)

#	E _{fR}	DMCoMo						Método para PyMEs		
		MM8	E _{fR} - MM8	ErRel MM8	MM23	E _{fR} - MM23	ErRel MM23	PEM	E _{fR} - PEM	ErRel PEM
P1	2,41	84,23	-81,82	-3.400%	94,88	-92,47	-3.843%	2,58	-0,17	-7,2%
P2	7,00	67,16	-60,16	-859%	51,84	-44,84	-641%	6,00	1,00	14,3%
P3	1,64	67,16	-65,52	-3.991%	68,07	-66,43	-4.047%	1,48	0,16	9,8%
P4	3,65	118,99	-115,34	-3.160%	111,47	-107,82	-2.954%	1,68	1,97	54,0%
P5	9,35	110,92	-101,57	-1.087%	122,52	-113,17	-1.211%	9,80	-0,45	-4,8%
P6	11,63	80,27	-68,65	-590%	81,36	-69,73	-600%	5,10	6,53	56,1%
P7	6,73	96,02	-89,29	-1.328%	92,49	-85,76	-1.275%	3,78	2,95	43,8%
P8	5,40	116,87	-111,47	-2.064%	89,68	-84,28	-1.561%	4,88	0,52	9,6%
P9	8,38	97,63	-89,26	-1.066%	98,74	-90,36	-1.079%	8,70	-0,33	-3,9%
P10	1,56	105,32	-103,75	-6.640%	103,13	-101,56	-6.500%	1,08	0,48	30,9%
Error Medio		88,68						1,46		
Varianza del Error		380,28						428,99		

Al analizar los resultados de la Tabla 15, se puede observar que el error promedio es muy grande (86 meses/hombre, o 7 años/hombre, para ambas fórmulas) con un desvío estándar de aproximadamente ± 20 meses/hombre respectivamente, DMCoMo siempre tiende a sobreestimar el esfuerzo del proyecto (por lo que los valores de error son siempre negativos) con una proporción mayor al 590% (menor diferencia para el proyecto #6 y fórmula MM8).

Por otro lado, al analizar los resultados del método de estimación para PyMEs (PEM) en la tabla 15, se puede observar que produce un error promedio de aproximadamente 1,46 meses/hombre con un desvío estándar para el error de aproximadamente ± 2 meses/hombre. Para poder analizar en forma más detallada este método se muestra en la Figura 2 un gráfico boxplot mostrando el comportamiento del esfuerzo real y el calculado por PEM.

De este segundo gráfico se nota que el método propuesto tiende a generar estimaciones inferiores a las reales con una diferencia general de un mes/hombre (el

promedio del esfuerzo real es de 5,77 meses/hombre y la del método propuesto es de 4,51 meses/hombre).

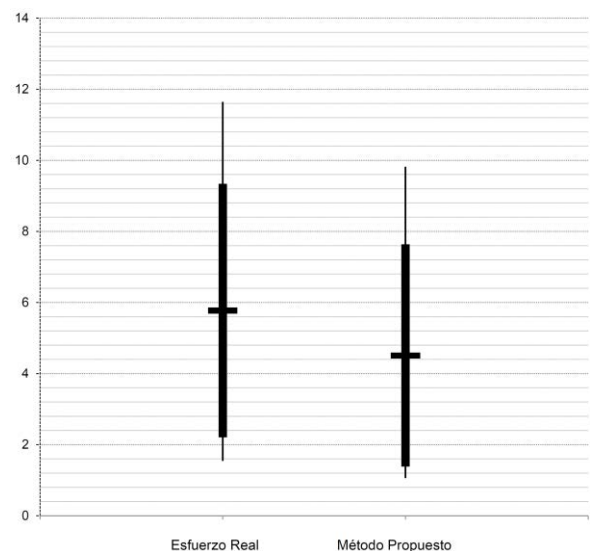


Fig. 2: comportamiento del Esfuerzo Real y el Método de Estimación

Finalmente, si el esfuerzo real y el estimado para cada proyecto se comparan con diagrama de barras (Figura 3), se puede observar que el método propuesto no es completamente exacto en sus estimaciones:

- Los proyectos #1, #3, #5, #8 y #9 tienen un esfuerzo estimado con un error absoluto menor a un mes/hombre y un error relativo menor al 10%.
- Los proyectos #4, #6 y #7 tienen un error relativo mayor al 35% (y menor al 60%) con un error absoluto máximo de 7 meses/hombre (en el caso del proyecto #6).
- Por último, los proyectos #2 y #10 tienen un error relativo entre 10% y 35% con un error máximo a un mes/hombre (proyecto #2).

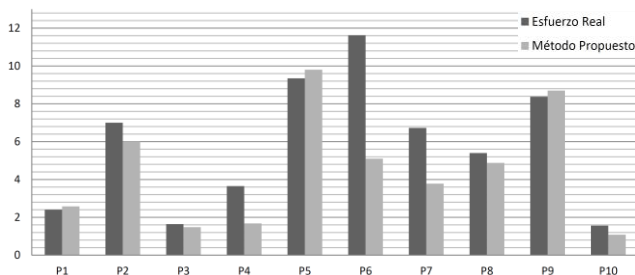


Fig. 3: Comparación del eEfr y el estimado

5. CONCLUSIONES

Dentro de los mecanismos de Aseguramiento de Calidad se incluyen las métricas. Entre ellas se incluyen los modelos para evaluar la productividad del proyecto que permiten establecer objetivos de mejora. Los Proyectos de Explotación de Información no escapan de dicha necesidad. Para ello se debe realizar estimaciones al comienzo del proyecto, las cuales son comparadas con los valores reales del proyecto a su finalización. De esta manera se destaca la necesidad de contar con métodos de estimación de esfuerzo confiables para Proyectos de Explotación de Información de tamaño pequeño.

Este trabajo describe y analiza dos métodos de estimación existentes para Proyectos de Explotación de Información: el método de estimación DMCoMo y el otro un método de estimación orientado a las PyMEs. Del estudio del método DMCoMo se puede ver que este método genera estimaciones superiores a los 5 años/hombres, umbral de esfuerzo muy superior al real vinculado a proyectos de PyMEs.

Al comparar este método con los esfuerzos de proyectos pequeños se puede ver que tiende a sobrestimar el esfuerzo en casi seis veces más el real. Por lo tanto se puede indicar que este método sólo puede ser utilizado para proyectos con tamaño grande. Por otro lado, el método de estimación orientado a PyMEs produce resultados más precisos que DMCoMo para proyectos pequeños. Aunque el comportamiento general de este método tiende a ser similar al de los proyectos reales, tiene una leve inclinación a calcular un esfuerzo inferior al real. De todas formas se destaca que el error medio es de aproximadamente 1,5 meses/hombre y que para los datos de validación utilizados el 50% de las

estimaciones poseen un error relativo menor al 10%. Estos errores se podrían deber a la existencia de factores que están afectando el cálculo del esfuerzo que no han sido considerados por el método de estimación.

6. AGRADECIMIENTOS

El trabajo de investigación presentado en este artículo ha sido parcialmente financiado por los proyectos 33A105 and 33B102 de la Universidad Nacional de Lanús, por los proyectos 40B133 y 40B065 de la Universidad Nacional de Río Negro, y el proyecto EIUTIBA11211 de la Universidad Tecnológica Nacional Facultad Regional Buenos Aires. Además, los autores desean agradecer a los investigadores que han provisto la información de proyectos reales en PyMEs utilizados en este trabajo.

7. REFERENCIAS

- [1] Pressman, R. 2005. Ingeniería de Software: Un enfoque práctico. McGraw-Hill.
- [2] Kan, S. H.; Parrish, J. & Manlove, D. 2001. In-process metrics for software testing. IBM Systems Journal, 40(1): 220-241.
- [3] Burstein, F. et al. 2008. Business Intelligence. Handbook on Decision Support Systems 2, Intl Handbooks on Information Systems, 175-193. Springer.
- [4] Ferreira, J. E.; Takai, O. K. & Pu, C. 2005. Integration of business processes with autonomous information systems: a case study in government services. Proceedings Seventh IEEE International Conference on E-Commerce Technology, 2005. System Sciences, 471-474.
- [5] Kanungo, S. 2005. Using Process Theory to Analyze Direct and Indirect Value-Drivers of Information Systems. Proceedings 38th Annual Hawaii International Conference on System Sciences, 2005, 231-240
- [6] García-Martínez, R. et al 2011. In Zapata, J. C. M. et al. (Eds.), Towards an Information Mining Engineering. Software Engineering, Methods, Modeling and Teaching. Editorial Universidad de Medellín, 83-99.
- [7] Boehm, B. W. et al. 2000. Software Cost Estimation with Cocomo II. Prentice Hall.
- [8] PRICE System. 1997. PRICE-S Reference Manual Version 3.0. Lockheed-Martin.
- [9] Marbán, O.; Menasalvas, E. & Fernández-Baizán, C. 2008. A cost model to estimate the effort of data mining projects (DMCoMo). Information Systems, 33(1), 133-150.
- [10] García-Martínez, R.; Lelli, R. & Merlino, H. 2011. Ingeniería de Proyectos de Explotación de Información para PYMES. Proceedings XIII Workshop de Investigadores en Ciencias de la Computación, 253-257.
- [11] Kalos, M. H. & Whitlock, P. A. 1986. Monte Carlo Methods. Vol I. Basics. Wiley & Sons.
- [12] Chapman P. et al. 2000. CRISP-DM 1.0 Step-by-step data mining guide. The CRISP-DM consortium.
- [13] OECD. 2005. OECD SME and Entrepreneurship Outlook 2005. OECD Publishing.
- [14] Álvarez, M. & Durán, J. 2009. Manual de la Micro, Pequeña y Mediana Empresa. Una contribución a la mejora de los sistemas de información y el desarrollo de las políticas públicas. CEPAL - Naciones Unidas.
- [15] ISO. 2011. ISO/IEC DTR 29110-1 Software Engineering - Lifecycle Profiles for Very Small Entities (VSEs) - Part 1: Overview. Inter. Organization for Standardization.



- [16] Laporte, C.; Alexandre, S. & Renault, A. 2008. Developing International Standards for VSEs. IEEE Computer, 41(3): 98-110
- [17] Ríos, M. D. 2006. El Pequeño Empresario en ALC, las TIC y el Comercio Electrónico. ICA.
- [18] Chen, Z. et al. 2005. Finding the right data for software cost modeling. Software, IEEE, 22(6), 38-46.
- [19] Domingos, P. et al. 2006. 10 challenging problems in data mining research. International Journal of Information Technology & Decision Making, 5(4), 597-604.
- [20] García-Martínez, R. et al. 2011. Information Mining Processes Based on Intelligent Systems. Proceedings II International Congress on Computer Science and Informatics, 87-94.
- [21] Weisberg, S. 1985. Applied Linear Regression. Wiley.