Revista de Procesos y Métricas de las Tecnologías de la Información (RPM) ISSN: 1698-2029
VOL. 1, Nº 2, Agosto 2004, 3-12
Asociación Española de Sistemas de Informáticos (AEMES)

ESTIMACIÓN DE VARIABLES EN PROYECTOS DE DESARROLLO DE SOFTWARE (PDS)

Javier Aroba Páez
Dpto. Ingeniería Electrónica, Sistemas Informáticos y Automática. Universidad de Huelva.
Ctra. Huelva-La Rabida s/n. Huelva , Spain.
E-Mail : aroba@diesia.uhu.es

Isabel Ramos Román, Jose C. Riquelme Santos
Dpto. Lenguajes y Sistemas Informáticos. Universdad de Sevilla.
Avda. Reina Mercedes s/n. Sevilla, Spain.
E-Mail: isabel.ramos@lsi.us.es ; riquelme@lsi.us.es

Abstract: The application of some Data Mining techniques to Software Development Projects (SDP) numeric databases (obtained through the simulation), facilitates the obtaining of qualitative information about the project evolution. Many of these techniques are descriptive, like the clustering, for which reason we don't have the capacity, a priori, to forecast results (project variables) from a new data set (project attributes) of a SDP. In this paper we try some statistic methods to forecast these variables from a new set of values of the attributes, so that the obtained results can be compared. The goal is to check if the project variables can be forecasted without having to simulate the whole project, and with low margin of error.

Resumen: La aplicación de determinadas técnicas de Data Mining sobre bases de datos numéricas de proyectos de desarrollo de software (PDS), nos permite, entre otras cosas, obtener información cualitativa sobre la evolución del proyecto. Muchas de estas técnicas son descriptivas, como por ejemplo el clustering, por lo que no se tiene a priori capacidad de predicción de resultados (variables del proyecto) a partir de nuevos datos (atributos del proyecto) de PDS. Para estimar estas variables a partir de nuevos valores en los atributos, se proponen en esta investigación diversas técnicas, de forma que se puedan comparar los resultados obtenidos. El objetivo es comprobar si se pueden estimar las variables de un proyecto, a partir de una nueva serie de valores de atributos, sin tener que simular todo el proyecto, y con unos márgenes de error bajos.

Palabras Clave:Regresión lineal, k-vecinos más cercanos, estimación de proyectos, predicción, minería de datos.

1. DESCRIPCIÓN DEL PROBLEMA

El proceso de estimación del software es un área en el que se están proponiendo métodos y técnicas desde hace 15 años. La mayoría de los procesos existentes describen cómo aplicar un método simple de predicción, que normalmente está basado en uno o más modelos algorítmicos.

Boehm [Boehm, 1981] proporciona una amplia visión del proceso de estimación de costes de software e incluye el establecimiento de los objetivos de la estimación, la planificación de la estimación así como la recopilación de estimaciones realizadas mediante varios métodos.

En líneas generales, *estimar* consiste en determinar el valor de una variable desconocida a partir de otras conocidas, o de una pequeña cantidad de valores conocidos de esa misma variable. La estimación forma parte de la inferencia estadística.

Si enfocamos todos los conceptos estadísticos referentes a estimación (estadístico de una muestra, parámetro, estimador, población, muestra) hacia el caso particular de la estimación del software, tendremos que los parámetros a estimar son: el tamaño del proyecto, el esfuerzo para realizar el mismo, el coste y el tiempo que se tardará en desarrollarlo. La población serían los proyectos similares realizados. La muestra podrían ser los

proyectos similares realizados en nuestra compañía, y los estimadores los valores de los estadísticos correspondientes a los parámetros de la muestra. Un posible estimador para el tamaño del proyecto podría ser la media del tamaño de los proyectos de la muestra.

Es importe pensar en una predicción como en rango mas que como un simple número. Una estimación o predicción no es un objetivo, sino una valoración probabilística. El valor que se obtiene de una estimación es el centro del rango. DeMarco [DeMarco, 1982] fue uno de los primeros en explicar que una estimación es una predicción que es igualmente probable que esté por encima o por debajo del resultado real.

Para comprender lo que esto significa, consideremos una situación en la cual queremos estimar el tiempo que nos llevará finalizar un proyecto. Imaginemos que pudiéramos registrar los valores reales de finalización de un gran número de proyectos, que hayan implementado el mismo conjunto de requerimientos. Entonces podríamos ser capaces de dibujar la función de densidad del tiempo *t* necesario para finalizar el proyecto con los requerimientos establecidos. Esta función de densidad suele ser una distribución normal, aunque no necesariamente ha de ser siempre así.

Los ingenieros de software frecuentemente interpretan mal el concepto de estimación, considerándole, como el número mas pequeño de meses en el cual el proyecto puede ser completado.

1.1. Precisión en la Estimación

Una vez que el proyecto de software está terminado, se tiene la oportunidad de comparar los valores reales con las estimaciones que se hicieran con anterioridad. Supongamos que V_f es el valor estimado y V_a el valor real. El error relativo (RE) en la estimación se calcula como la diferencia entre el valor real y el estimado, dividido por el valor real.

Frecuentemente, se necesita obtener el error relativo de un conjunto de estimadores. Por ejemplo, usualmente se desea saber si las predicciones de esfuerzo son precisas para un grupo de proyectos desarrollado. Para ello utilizaremos el error relativo medio para n proyectos $\overline{RE}$. También es posible calcular la magnitud de estos mismos valores, considerando el valor absoluto, es decir MRE = |RE|, para los mismos n proyectos.

Si la media de la magnitud del error relativo es pequeña, entonces las estimaciones son buenas. Esta noción se usa para definir la métrica *calidad de predicción*. Para un conjunto de *n* proyectos, *i* es el numero de ellos cuya magnitud media del error relativo es menor o igual que *q*.

$$pred(q) = \frac{i}{n}$$

La métrica `pred` proporciona una indicación del grado de ajuste para un conjunto de datos, basado en el valor de RE obtenido para cada dato. A modo de ilustración, si pred (0,25)=0,4, entonces nosotros se puede decir que el 40% de los valores de ajuste caen dentro del 25% de sus correspondientes valores actuales. En términos de evaluación del rendimiento de un modelo dado, se podría considerar un buen modelo aquel que consiguiera que MRE<=0,25 y *pred* (0,25)>=0,75.

DeMarco sugiere el uso de un método denominado *factor de calidad de la estimación* (estimating quality factor EQF) para evaluar la precisión del proceso [DeMarco, 1982]. Con esta técnica, las estimaciones se realizan repetidamente a lo largo de todo el proceso, y se van aproximando al valor real según avanza el proyecto. EQF varía desde cero hasta infinito. Para una serie de estimaciones, DeMarco sugiere un valor de 4 (el cual corresponde a una estimación dentro del 25% del valor real del esfuerzo de desarrollo), que debe ser relativamente fácil de retener, y los estimadores de coste deben guardar valores de 8 o superior.

1.2. Principios de Estimación

La estimación del software es un tema abordado de forma general por todos los autores dedicados a la ingeniería del software [Pressman, 1998], [Humphrey, 1995], [Jones, 1996]. En este sentido, suelen coincidir que a la hora de realizar una estimación, hay que tener en consideración los siguientes principios básicos:

A. Aplicar la cantidad de recursos correcta para crear y refinar las estimaciones. Para determinar el nivel de detalle hay que considerar:

- La magnitud del proyecto
- Riesgo de las estimaciones inexactas
- Incertidumbres del proyecto

B. La estimación de recursos requerida para un escenario dado no puede cambiarse arbitrariamente. Las características que afectan a la precisión requerida de la estimación pueden ser:

- Riesgo inherente al proyecto
- Fiabilidad de la información usada
- Efectividad del proceso de estimación

C. Reestimar con frecuencia. A medida que evoluciona el proyecto se dispone de más información que confirmará o refutará las estimaciones originales. Esa información servirá de base para realizar estimaciones más exactas del resto del proyecto.

El propósito para el que se requiere una estimación es el principal motivo por el que se realizan predicciones sobre las variables de un proyecto. Las principales causas que hacen necesarias las estimaciones son:

- Investigación de la factibilidad del desarrollo o compra de un sistema nuevo.
- Investigación del impacto del cambio en la funcionalidad de un sistema existente.
- Determinación del personal de soporte necesario, para del desarrollo del proyecto.
- Cuantificar el coste económico del sistema nuevo.

2. TÉCNICAS DE ESTIMACIÓN PROPUESTAS

La obtención de información cualitativa a partir de bases de datos numéricas de proyectos de desarrollo de software (PDS), puede lograrse con la utilización de técnicas de Data Mining. Muchas de estas técnicas son descriptivas, como por ejemplo el clustering, por lo que no se tiene a priori, capacidad de predicción de resultados (variables del proyecto) a partir de nuevos datos (atributos del proyecto) de PDS. Para estimar estas variables a partir de nuevos valores en los atributos, se van a utilizar en esta investigación diversas técnicas de forma que se puedan analizar y comparar los resultados obtenidos.

El objetivo principal en esta investigación es comprobar si se pueden estimar las variables de un proyecto, a partir de una nueva serie de valores de atributos, sin tener que simular todo el proyecto, y con unos márgenes de error bajos. Para ello se van a utilizar las técnicas siguientes:

2.1 Algoritmo wk-NN : k-vecinos más cercanos ponderados

En los últimos años, muchos investigadores han buscado soluciones eficientes para el problema de los vecinos más cercanos. Los trabajos más recientes en este campo son los de Lin y Yang [Lin, 2001], y Maneewongvatana [Maneewongvatana, 2001].

El algoritmo propuesto, wk-NN es una modificación del k-NN (k-Nearest Neighbor), donde se pondera el peso (*w*) de cada atributo, asociando dicho peso con el coeficiente de correlación de cada atributo con cada variable del PDS.

El algoritmo wk-NN de forma general puede enunciarse de la siguiente forma: *"dada una colección de datos etiquetados y un nuevo dato en un espacio métrico m-dimensional, encontrar el dato que dista menos del nuevo dato"*.

Los datos con los que va trabajar el algoritmo (wk-NN) en esta investigación son de la forma $X\ (x_1, x_2, ..., x_n)$, donde la dimensión de X es igual al número de atributos del proyecto en cuestión. El criterio utilizado en el algoritmo propuesto, para calcular la distancia ponderada entre dos puntos X y Z , de una base de datos de PDS es:

$$d(X,Z) = \sqrt{\sum_{i=1}^{n} w_i (x_i - z_i)^2} \qquad (1)$$

Como ya se ha comentado anteriormente, el peso w_i se calculará a partir de los coeficientes de correlación (ς) , de cada atributo x_i respecto a cada variable del proyecto y_i. En concreto a cada atributo a_j le asociaremos un peso w_j igual a la media de los coeficientes de correlación de dicho atributo respecto a cada una de las n variables del proyecto, es decir:

$$w_j = \sum_{i=1}^{n} \varsigma(a_j, y_i) \Big/ n \qquad (2)$$

Donde w_j es el peso asociado al atributo a_j , y calculado como la media de los coeficientes de correlación de ese atributo respecto a cada variable y_i del proyecto.

El objetivo es predecir, sin tener que simular el proyecto de nuevo, el valor estimado para las variables, dada un nuevo conjunto de valores en los atributos $A\ (a_1, a_2, ..., a_n)$.

La técnica de los *k*-vecinos más cercanos establece que cada variable estimada, para una nueva serie de datos , se calcula de la siguiente forma:

$$\hat{y}_i = \frac{y_{i1} + y_{i2} + ... + y_{ik}}{k} \qquad (3)$$

Donde la estimación de una variable para una nueva serie de datos *A*, se calcula como la media de la suma de los valores en dicha variable de los *k* datos vecinos mas cercanos (y_{ik} : valor de la variable y_i en el vecino más cercano *k-esimo*) al nuevo dato, en orden de cercanía, según la formula de distancia definida anteriormente.

En esta investigación después de realizar numerosas pruebas se ha optado por un k=3. En este

sentido hay varios estudios que analizan esta cuestión, destacando el realizado por Wettschereck [Wettschereck, 1995] quien propone tres métodos para elegir un buen valor de *k*.

2.2 Modelos de Regrsión lineal múltiple

El modelo de regresión lineal simple es un método sencillo para analizar la relación lineal entre dos variables cuantitativas. Sin embargo, en la mayoría de los casos lo que se pretende es predecir una respuesta en función de un conjunto más amplio de atributos (por ejemplo bases de datos de PDS), siendo necesario considerar el modelo de regresión lineal múltiple como una extensión de la recta de regresión. En esta investigación, la regresión lineal múltiple (MLR) se va a utilizar como técnica de estimación, por tanto se trata de predecir el valor de cada variable *y* como función lineal de una familia de *m* atributos (x_1, x_2, ..., x_m), a partir de una muestra de tamaño *n* cuyas ocurrencias se ordenan matricialmente de la forma mostrada en la figure1 .

$$
\begin{array}{l}
((y_{1,}, x_{11}, x_{12}, ..., x_{1m}), \\
\quad (y_{2,}, x_{21}, x_{22}, ..., x_{2m}), \\
\quad ..... \\
\quad (y_{n,}, x_{n1}, x_{n2}, ..., x_{nm}))
\end{array}
$$

Figure 1: Matriz de Atributos y variable

siendo y_i la *i*-ésima variable y $x_{i,j}$ el *j*-ésimo atributo asociado a la variable *i*.

Así las cosas, se trata de ajustar los datos a un *modelo lineal* de la forma:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_m x_{im} + \alpha_i$$

Bajo las siguientes hipótesis:

- Los residuos α_i son normales de media 0 y varianza común desconocida σ^2; además, estos residuos son independientes.
- El número de variables explicativas (*m*) es menor que el de observaciones (*n*); esta hipótesis se conoce con el nombre de *rango completo*.
- No existen relaciones lineales exactas entre las variables explicativas.

Algunos de los principales trabajos que hacen referencia a esta técnica como método de predicción en ingeniería del software son los de Dolado [Dolado, 1998] que describe las posibilidades de la regresión lineal, la programación genética y las redes neuronales como técnicas de estimación en proyectos de desarrollo de Software; y Krishnamoorthy [Krishnamoorthy, 2002] que presenta uno de los últimos trabajos relacionados con la predicción de información utilizando regresión lineal.

2.3 Modelos de Regresión no lineal múltiple

Al igual que en el caso de regresión lineal, se trata de predecir el valor de cada variable *y* como *función no lineal* de una familia de *m* atributos (x_1, x_2, ..., x_m), a partir de una muestra de tamaño *n* cuyas ocurrencias se ordenan matricialmente (figure 1). Así las cosas, se trata de ajustar los datos a un modelo *no lineal*, que puede tener por ejemplo las siguientes formas:

$$y_i = \beta_1 x_{i1}^{\delta 1} + \beta_2 x_{i2}^{\delta 1} + ... + \beta_m x_{im}^{\delta 1} + \beta_0$$

$$y_i = Exp\,(\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_m x_{im} + \beta_{0)}$$

La idea básica sería igual que para el caso de Regresión Lineal, siendo la diferencia fundamental la forma de la función que queremos obtener. Obviamente, al tratarse de formas de función no lineales, los cálculos se hacen mucho más complejos que en el caso lineal, pudiéndose darse el caso de no poder aproximar la variable *y* mediante determinadas formas de función no lineales.

3. DESCRIPCIÓN DE LAS BASES DE DATOS DE PDS

La aparición de Simuladores de Proyectos de Software (SPS) tales como Stella, Vensim, PowerSim, etc, ha significado un avance significativo en la gestión y estimación de proyectos. Uno de nuestros primeros trabajos estuvo centrado en el desarrollo de un SPS [Ramos, 1998] que permitiera simular el comportamiento de un proyecto utilizando la mitad de atributos que el modelo que aparece en [Abdel, 1991].

Los modelos dinámicos para PDS incluyen una serie de parámetros y tablas que nos permiten definir las políticas de gestión que pueden ser aplicadas en dichos proyectos, tanto las relacionadas con el entorno del proyecto (número de tareas, tiempo, coste, número de técnicos, complejidad del producto, etc.) como las relacionadas con la organización de desarrollo, y el nivel de madurez de la organización.

Una vez definidos los parámetros del modelo, el gestor de proyecto debe decidir cuáles son las variables que se van a analizar. Las opciones mas habituales son las variables que definen el desarrollo del proyecto: tiempo de entrega, coste, productividad media de desarrollo, etc. Cada vez que se asigna un valor a cada parámetro del modelo, éste queda completamente definido y mediante su simulación se obtendrán unos valores para estas variables. Para generar un conjunto de casos de entrenamiento, el responsable del proyecto debe escoger un rango para cada uno de los parámetros del modelo. A continuación, el SPS genera aleatoriamente un valor en cada uno de esos intervalos.

A cada tupla de parámetros así definidos le corresponderá entonces una tupla de valores para las variables resultado de la simulación. De esta forma se genera un registro en una base de datos, con los valores de los parámetros y los valores obtenidos para las variables del proyecto que se deseen analizar. Repitiendo este proceso un número determinado de veces se puede obtener un fichero de entrenamiento, que se podrá utilizar para extraer información acerca del PDS, o para realizar estimaciones sobre las variables del proyecto.

En esta investigación se han utilizado dos bases de datos cuantitativas, obtenidas tras la simulación de un PDS usando las políticas de gestión de personal y de tiempo de entrega siguientes:

1- Contratación rápida Con restricción en el tiempo de entrega (CrCrTi): Los atributos relacionados con la gestión de personal toman valores dentro del intervalo considerado como rápido, y el atributo relacionado con la gestión del tiempo de entrega toma valores dentro de los intervalos de plazo fijo y plazo moderado, es decir el aplazamiento en el tiempo de entrega no puede superar al tiempo estimado en más de un 20%

2- Contratación rápida Sin restricción en el tiempo de entrega (CrSrTi): Los atributos relacionados con la gestión de personal toman valores dentro del intervalo considerado como rápido, y el atributo relacionado con la gestión del tiempo de entrega se puede mover dentro de todo el rango posible de valores. La descripción de los atributos que se van a analizar relacionados con políticas de gestión de personal es:

ATRIBUTO	***DESCRIPCIÓN(Unidad)*** ***BD [Rango]*** ***[Concepto que representa]***
ASIMDY	Retraso medio de adecuación de los técnicos nuevos que se ha incorporado al proyecto (días) CrCrTi, CrSrTi [5,15] [Integración en el proyecto]
TRNSDY	Retraso medio en la salida de los técnicos del proyecto (días) CrCrTi, CrSrTi [5, 10] [Despido de técnicos]
HIREDY	Retraso medio de incorporación de los técnicos nuevos en el proyecto (días) CrCrTi, CrSrTi [5 , 10] [Contratación de técnicos]

Tabla 1: Atributos relacionados con Gestión de personal

La descripción del atributo que vamos a analizar relacionado con política de gestión del tiempo de entrega es:

MXSCDX	Aplazamiento máximo permitido en el tiempo de entrega respecto del estimado (%) CrCrTi [1.0, 1.2] CrSrTi [1.0, 50] [Retraso de Tiempo de entrega]

Tabla 2: Atributos relacionados con Tiempo de entrega

La descripción de las variables que se desean estimar en las bases de datos de proyectos, es la siguiente:

VARIABLE	***DESCRIPCIÓN*** ***(Unidad)*** ***[Concepto que representa]***
COSTE	Esfuerzo necesario para realizar el proyecto (técnicos-días) [Coste]
TIEMPO	Tiempo de desarrollo (días) [Tiempo de entrega]
CALIDAD	Calidad del producto final (errores/tareas) [Calidad]

Tabla 3: Variables de un PDS

4. RESULTADOS OBTENIDOS

Los rangos de valores en los que se mueven las variables que deseamos estimar (coste, tiempo y calidad) en las bases de datos que vamos a usar para realizar las estimaciones son:

VARIABLE	***Base de Datos [Rango de Valores] (Amplitud)***	
COSTE	CrCrTi [1709 , 3395] (1686)	CrSrTi [1693 , 2415] (722)
TIEMPO	CrCrTi [349.5 , 354.3] (5)	CrSrTi [349.5 , 361.5] (12)
CALIDAD	CrCrTi [0.235 , 0.661] (0.43)	CrSrTi [0.236 , 0.567] (0.33)

Tabla 3.1: Rangos de valores de las Variables

Las medidas de error que se han utilizado para medir la bondad de cada una de las estimaciones realizadas son:

- Error A: Error Absoluto medio
- Error B: Error Relativo medio
- Error C: Error Relativo disperso medio. La expresión de este medida de error es:

$$ErrorC = \frac{\sum_{i=1}^{n} \frac{|Y_i - \hat{Y}_i|}{|Y_i - \bar{Y}_i|}}{n}$$

Para el cálculo de los pesos w_i utilizados en el algoritmo wk-NN, se ha realizado un análisis de correlación sobre las bases de datos con las que vamos a trabajar, obteniendo los siguientes resultados:

CrCrTi Correlation	**COSTE**	**TIEMPO**	**CALIDAD**
Asimdy	0.868	0.416	-0.447
Hiredy	-0.008	0.293	-0.256
Mxscdx	-0.36	0.344	-0.331
Trnsdy	-0.016	-0.074	0.128

Tabla C1: Análisis de Correlación CrCrTi

Se observa como el atributo más influyente sobre las variables es *Asimdy*, lo que significa que la adecuación de los técnicos al proyecto es un atributo relevante para obtener los objetivos del proyecto. A continuación le sigue en orden de relevancia *Mxscdx*, lo que quiere decir que en un proyecto con restricción en el tiempo de entrega como este, el porcentaje de aplazamiento es determinante. El retraso en la contratación (*Hiredy*) influye de alguna forma sobre el tiempo y la calidad, pero muy poco sobre el coste del proyecto.

CrSrTi Correlación	**COSTE**	**TIEMPO**	**CALIDAD**
Asimdy	0.986	0.154	-0.821
Hiredy	0.042	0.251	-0.224
Mxscdx	-0.078	-0.034	0.074
Trnsdy	0.049	0.032	-0.04

Tabla C2: Análisis de Correlación CrSrTi

En esta tabla podemos ver como el atributo más influyente sobre las variables vuelve a ser *Asimdy*, sobre todo sobre el coste y la calidad, aunque en este caso no influye tanto sobre el tiempo de entrega. En esta base de datos (CrSrTi) *Mxscdx* no es un atributo muy influyente sobre las variables, como era de esperar, ya que se trata de un proyecto donde no hay restricción en el tiempo de entrega. Por su parte *Hiredy* sigue teniendo prácticamente a misma influencia y sobre las mismas variables que en el caso anterior.

4.1. MLR: Modelos de regresión lineal múltiple

Los errores cometidos al estimar el valor de las variables de cada una de las tuplas de las distintas bases de datos, usando la técnica de MLR (Regresión Lineal Multivariante) y utilizando la forma de función lineal: *Variable* = **a** Asimdy + **b** Hiredy + **c** Mxscdx + **d** Trnsdy + **e**, son los siguientes:

A) BD: Contratación rápida, Con restricción en el tiempo de entrega (CrCrTi): Las ecuaciones lineales obtenidas para cada una de las variables del proyecto, son:

Coste = 123.7 Asimdy – 23.4 Hiredy –
2915.3 Mxscdx + 9.19 Trnsdy + 4321.9

Tiempo = 0.0717 Asimdy + 0.102 Hiredy +
3.222 Mxscdx – 0.0148 Trnsdy + 344.77

Calidad = -0.010 Asimdy – 0.011 Hiredy – 0.39 Mxscdx + 0.0045 Trnsdy + 1.10

Los coeficientes obtenidos para cada atributo en las distintas ecuaciones corroboran los resultados obtenidos en la Tabla C1. En la ecuación del *Coste* se puede ver que los atributos con mayor coeficiente son precisamente los más relevantes respecto de esta variable: Asimdy y Mxscdx.

En las ecuaciones del *Tiempo* y la *Calidad* se observa que, excepto en Trnsdy, los coeficientes obtenidos son del mismo orden aproximadamente, dato que concuerda con los coeficientes de correlación obtenidos en estas variables. Trnsdy es un atributo poco relevante en las tres variables, lo que se comprueba con los coeficientes tan bajos obtenidos en las tres ecuaciones.

El error cometido al estimar el valor de cada variable para cada una de las tuplas de la base de datos CrCrTi es el mostrado en la tabla 4:

CrCrTi MLR	**Estimación COSTE**	**Estimación TIEMPO**	**Estimación CALIDAD**
Error A	88.58	0.30	0.04
Error B	4.00	0.08	9.63
Error C	1.08	1.30	1.49

Tabla 4: Resultados MLR CrCrTi

B) BD: Contratación rápida, Sin restricción en el tiempo de entrega (CrSrTi): Las ecuaciones lineales obtenidas para cada una de las variables del proyecto son:

Coste = 63.54 Asimdy – 5.74 Hiredy + 0.0066 Mxscdx + 0.0623 Trnsdy + 1389

Tiempo = 0.698 Asimdy + 0.377 Hiredy + 0.00252 Mxscdx – 0.011 Trnsdy + 341

Calidad = - 0.0221 Asimdy – 0.0088 Hiredy + 0.0001 Mxscdx + 0.0002 Trnsdy + 0.739

Al igual que en en el caso anterior, los coeficientes obtenidos corroboran los resultados obtenidos en la Tabla C2. En la ecuación del *Coste* se puede ver que el atributo con mayor coeficiente es precisamente el más relevante respecto de esta variable: Asimdy

En las ecuaciones del *Tiempo* y la *Calidad* se observan diferencias respecto al caso anterior. En este caso los atributos relevantes son Asimdy e Hiredy, hecho que se comprueba al ver el grado de los coeficientes en las ecuaciones. Cabe destacar en el caso de la *Calidad*, que Asimdy, como ocurre con el Coste, es con mucha diferencia el atributo más relevante. Trnsdy sigue siendo un atributo poco relevante sobre las tres variables, lo que se comprueba con los coeficientes tan bajos obtenidos en las tres ecuaciones.

El error cometido al estimar el valor de cada variable para cada una de las tuplas de la base de datos CrSrTi es el mostrado en la tabla 5:

CrSrTi MLR	**Estimación COSTE**	**Estimación TIEMPO**	**Estimación CALIDAD**
Error A	24.76	1.49	0.03
Error B	1.24	0.42	8.31
Error C	0.36	1.89	1.29

Tabla 5: Resultados MLR CrSrTi

Se observa en ambas tablas (4 y 5), que si analizamos el error absoluto (error A) la calidad es el atributo mejor estimado ya que es la variable con el rango de valores de menor amplitud. Sin embargo, analizando el error relativo (error B) es el Tiempo el mejor estimado, mientras que teniendo en cuenta el promedio de valores de cada variable (error C) es el Coste, la variable mejor estimada.

4.2. MNLR: Modelos de regresión no lineal múltiple

A) Forma de Función Polinómica

La forma de función polinómica que deseamos calcular es: *Variable* = **a** $\text{Asimdy}^{\mathbf{a1}}$ + **b** $\text{Hiredy}^{\mathbf{b1}}$ + **c** $\text{Mxscdx}^{\mathbf{c1}}$ + **d** $\text{Trnsdy}^{\mathbf{d1}}$ + **e**.

A.1) BD: Contratación rápida, Con restricción en el tiempo de entrega (CrCrTi): Las ecuaciones no lineales polinómicas obtenidas para cada una de las variables del proyecto, son:

Coste = $34.27\text{Asimdy}^{1.41} – 0.04\ \text{Hiredy}^{3.49} – 0.1311\ \text{Mxscdx}^{2.48} + 515\ \text{Trnsdy}^{0.09} + 2055.1$

Tiempo = $17.61\ \text{Asimdy}^{0.03} + 105.18\ \text{Hiredy}^{0.006} + 112.8\ \text{Mxscdx}^{0.03} + 111.29\ \text{Trnsdy}^{-0.001}$

Calidad = $6.05\ \text{Asimdy}^{-0.01} + 2.54\ \text{Hiredy}^{-0.03} + 3.86\ \text{Mxscdx}^{-0.11} + 0.42\ \text{Trnsdy}^{0.072} – 12.02$

El error cometido al estimar el valor de cada variable para cada una de las tuplas de la base de datos CrCrTi es el mostrado en la tabla 6:

CrCrTi MNLR P	**Estimación COSTE**	**Estimación TIEMPO**	**Estimación CALIDAD**
Error A	87.23	0.30	0.05
Error B	3.94	0.09	9.93
Error C	0.93	1.31	1.50

Tabla 6: Resultados MNLR CrCrTi Polinómica

Los coeficientes obtenidos para cada atributo en las distintas ecuaciones corroboran, al igual que en el caso lineal, los resultados obtenidos en la Tabla C1. En la ecuación del *Coste* se puede ver que los atributos con mayor coeficiente y exponente a la vez, son precisamente los más relevantes respecto de esta variable: Asimdy y Mxscdx.

En las ecuaciones del *Tiempo* y la *Calidad* se observa que, excepto en Trnsdy, los coeficientes y exponentes obtenidos son del mismo orden aproximadamente, dato que concuerda con los coeficientes de correlación obtenidos en estas variables. Trnsdy es un atributo poco relevante en las tres variables, lo que se comprueba con los coeficientes tan bajos obtenidos en las tres ecuaciones.

A.2) BD: Contratación rápida, Sin restricción en el tiempo de entrega (CrSrTi): Las ecuaciones no lineales polinómicas obtenidas para cada una de las variables del proyecto, son:

$$Coste = 2.32\ Asimdy^{2.11} - 2.30\ Hiredy^{1.28} - 59.5\ Mxscdx^{0.020} + 66.28\ Trnsdy^{0.10} + 1661.22$$

$$Tiempo = 0.17\ Asimdy^{1.44} + 49.09\ Hiredy^{0.049} + 229.7\ Mxscdx^{0.00013} + 61.9\ Trnsdy^{-0.0007}$$

$$Calidad = 3.76 Asimdy^{-0.06} + 0.35 Hiredy^{-0.14} + 0.024\ Mxscdx^{0.084} + 0.038\ Trnsdy^{0.125} - 3.320$$

Al igual que en en el caso anterior, los coeficientes obtenidos corroboran los resultados obtenidos en la Tabla C2. En la ecuación del *Coste* se puede ver que el atributo con mayor coeficiente y exponente es precisamente el más relevante respecto de esta variable: Asimdy

En las ecuaciones del *Tiempo* y la *Calidad* se observan diferencias respecto al caso anterior. En este caso los atributos relevantes son Asimdy e Hiredy, hecho que se comprueba al ver el grado de los coeficientes en las ecuaciones. Cabe destacar en el caso de la *Calidad*, que Asimdy, como ocurre con el Coste, es con mucha diferencia el atributo más relevante. Trnsdy sigue siendo un atributo poco relevante sobre las tres variables, lo que se comprueba con los coeficientes tan bajos obtenidos en las tres ecuaciones

El error cometido al estimar el valor de cada variable para cada una de las tuplas de la base de datos CrSrTi es el mostrado en la tabla 7:

CrSrTi MNLR P	**Estimación COSTE**	**Estimación TIEMPO**	**Estimación CALIDAD**
Error A	13.02	1.41	0.04
Error B	0.66	0.40	9.08
Error C	0.12	1.98	2.19

Tabla 7: Resultados MNLR CrSrTi Polinómica

B) Forma de Función Exponencial

La forma de función exponencial que deseamos calcular es: ***Variable*** = **EXP (a** Asimdy + **b** Hiredy + **c** Mxscdx + **d** Trnsdy **+ e)**

B.1) BD: Contratación rápida, Con restricción en el tiempo de entrega (CrCrTi): Las ecuaciones no lineales Exponenciales obtenidas para cada una de las variables del proyecto son:

$$Coste = EXP(0.055\ Asimdy - 0.01\ Hiredy - 1.37\ Mxscdx + 0.004\ Trnsdy + 8.69)$$

$$Tiempo = EXP(0.0002\ Asimdy + 0.0003\ Hiredy + 0.009\ Mxscdx - 0.00004 Trnsdy + 5.84)$$

$$Calidad = EXP(-0.01\ Asimdy - 0.02\ Hiredy - 0.072\ Mxscdx + 0.008\ Trnsdy + 0.40)$$

El error cometido al estimar el valor de cada variable para cada una de las tuplas de atributos de la base de datos CrCrTi es el mostrado en la tabla 8:

CrCrTi MNLR E	**Estimación COSTE**	**Estimación TIEMPO**	**Estimación CALIDAD**
Error A	69.60	0.30	0.10
Error B	3.18	0.09	9.72
Error C	0.70	1.38	1.69

Tabla 8: Resultados MNLR CrCrTi Exponencial

B.2) BD: Contratación rápida, Sin restricción en el tiempo de entrega (CrSrTi): Las ecuaciones no lineales Exponenciales obtenidas son:

$$Coste = EXP(0.03\ Asimdy - 0.003\ Hiredy - 0.000002\ Mxscdx + 0.000048\ Trnsdy + 7.28)$$

$$Tiempo = EXP(0.002\ Asimdy + 0.001 Hiredy + 0.000007 Mxscdx - 0.00003 Trnsdy + 5.83)$$

Calidad = EXP (-0.047 Asimdy – 0.017 Hiredy + 0.00012Mxscdx + 0.001Trnsdy – 0.197)

El error cometido al estimar el valor de cada variable para cada una de las tuplas de la base de datos CrSrTi es el mostrado en la tabla 9:

CrSrTi MNLR E	**Estimación COSTE**	**Estimación TIEMPO**	**Estimación CALIDAD**
Error A	19.36	1.51	0.03
Error B	0.97	0.43	8.58
Error C	0.26	2.20	2.25

Tabla 9: Resultados MNLR CrSrTi Exponencial

Al igual que ocurría con la regresión lineal, se puede observar en las tablas 6 y 7: MNLR Polinómica y en las tablas 8 y 9: MNLR Exponencial, que si analizamos el error absoluto (error A) la calidad es el atributo mejor estimado ya que es la variable con el rango de valores de menor amplitud. Sin embargo, analizando el error relativo (error B) es el Tiempo el mejor estimado, mientras que teniendo en cuenta el promedio de valores de cada variable (error C) es el Coste, la variable mejor estimada.

4.3. wk-NN: k-vecinos más cercanos ponderados

Los errores cometidos al estimar cada variable para cada una de los tuplas de las distintas bases de datos, usando la técnica de los wk-vecinos más cercanos, considerando los pesos w_i (coeficientes de correlación) en el cálculo de las distancias son los siguientes:

A) BD: Contratación rápida, Con restricción en el tiempo de entrega (CrCrTi):

CrCrTi wk-*NN*	**Estimación COSTE**	**Estimación TIEMPO**	**Estimación CALIDAD**
Error A	31.81	0.039	0.029
Error B	1.363	0.011	6.515
Error C	0.372	0.115	1.446

Tabla 10: Resultados wk-*NN* CrCrTi

En esta tabla (10) se observa que si analizamos el error absoluto (error A) la calidad es el atributo mejor estimado ya que es la variable con el rango de valores de menor amplitud. Sin embargo, analizando el error relativo (error B) es el Tiempo el mejor estimado, mientras que teniendo en cuenta el promedio de valores de cada variable (error C) es el Coste, la variable mejor estimada.

B) BD: Contratación rápida, Sin restricción en el tiempo de entrega (CrSrTi):

CrSrTi *wk-NN*	**Estimación COSTE**	**Estimación TIEMPO**	**Estimación CALIDAD**
Error A	11.30	0.169	0.023
Error B	0.558	0.048	6.038
Error C	0.143	0.291	1.147

Tabla 11: Resultados wk-*NN* CrSrTi

En esta otra tabla (11) se observa que si analizamos el error absoluto (error A) la calidad es el atributo mejor estimado ya que es la variable con el rango de valores de menor amplitud. Sin embargo, analizando el error relativo (error B) es el Tiempo el mejor estimado, mientras que teniendo en cuenta el promedio de valores de cada variable (error C) es el Coste, la variable mejor estimada.

5. CONCLUSIONES

Se observa como el tiempo es una variable que se logra estimar con una precisión muy grande con todas las técnicas aplicadas, sea cual sea la medida de error utilizada, mientras que la calidad es la peor estimada. Llama la atención los valores de error obtenidos en la estimación del *Coste* al medirlo con el error A. Esto se debe a que el error absoluto, es una medida muy sensible al rango de los valores que manipula.

Se puede comprobar como la mejor técnica de estimación de las presentadas, es la de los k-vecinos, ponderados con pesos (wk-NN), con la que se obtienen mejores resultados en la estimación de variables en las dos bases de datos utilizadas.

Si al realizar una simulación de un PDS, y obtener por tanto la base de datos asociada, llegara un nuevo conjunto de datos (atributos), y se desea conocer cual sería el valor de las variables (coste, tiempo y calidad) que le correspondería, sin necesidad de volver a simular todo el proyecto, el gestor podría aplicar cualquiera de las tres técnicas propuestas, en especial la de los k-vecinos, obteniendo así una información bastante aproximada.

A continuación se muestra una comparativa de los errores de estimación obtenidos para cada variable y con cada una de las técnicas propuestas:

CrCrTi		Error A	Error B	Error C
MLR	Coste	88.58	4.00	1.08
	Tiempo	0.30	0.08	1.30
	Calidad	0.04	9.63	1.49
MNLR Polinómico	Coste	87.23	3.94	0.93
	Tiempo	0.30	0.09	1.31
	Calidad	0.05	9.93	1.50
MNLR Exponencial	Coste	69.60	3.18	0.70
	Tiempo	0.30	0.09	1.38
	Calidad	0.10	9.72	1.69
wk-NN	Coste	31.81	1.36	0.37
	Tiempo	0.039	0.011	0.115
	Calidad	0.029	6.515	1.446

Tabla 12: Resultados CrCrTi

Analizando los resultados mostrados en esta tabla (12), podemos concluir que considerando el orden de técnicas mostrado, el error cometido en las distintas estimaciones, va decreciendo, de forma que con wk-NN se obtiene una reducción del error de más del 50% respecto a MLR, sea cual sea la variable estimada.

CrSrTi		Error A	Error B	Error C
MLR	Coste	24.76	1.24	0.36
	Tiempo	1.49	0.42	1.89
	Calidad	0.03	8.31	1.29
MNLR Polinómico	Coste	13.02	0.66	0.12
	Tiempo	1.41	0.40	1.98
	Calidad	0.04	9.08	2.19
MNLR Exponencial	Coste	19.36	0.97	0.26
	Tiempo	1.51	0.43	2.20
	Calidad	0.03	8.58	2.25
wk-NN	Coste	11.30	0.558	0.143
	Tiempo	0.169	0.048	0.291
	Calidad	0.023	6.038	1.147

Tabla 13: Resultados CrSrTi

Por otro lado, con los resultados mostrados en esta tabla (13), podemos concluir que considerando el orden de técnicas mostrado, el error cometido en las distintas estimaciones, no es del todo decreciente como en el caso CrCrTi. En este caso podemos observar que el error obtenido con la técnica MNLR Exponencial es mayor que con MNLR Polinómico. Además con esta última técnica se obtiene una mejora del error respecto a MLR mucho mayor que en el caso anterior, alcanzando casi el 50%. Al igual que en el caso anterior, la técnica que proporciona los mejores resultados vuelve a ser wk-NN, con una mejora respecto a MLR de mas del 50% en todas las variables estimadas.

AGRADECIMIENTOS

Este trabajo ha sido respaldado por la agencia española de investigación CICYT, bajo concesión TIC2001-1143-C03-02.

REFERENCIAS

Abdel-Hamid T., Madnick S., 1991. "Software Project Dynamics: an integrated approach". Prentice-Hall.

Boehm B.W., 1981. *Software Engineering Economics.* Prentice-Hall, Inc. Englewood Cliffs, New Jersey.

DeMarco T.,1982. *Controlling Software Projects.* Yourdon Press, Prentice-Hall. Englewood Cliffs, New Jersey.

Jeffery D.R., Lawrence M.J., 1985. *Managing programming productivity.* Journal of Systems and Software, v.5 n.1, p.49-58.

Ramos I., Ruiz M., 1998. "A Reduced Dynamic Model to Make Estimations in the Initial Stages of a Software Development Project". In: Hawkings C et al (ed), INSPIRE III. Process Improvement through Training and Education, Pp.:172-185. London.

Pressman R. 1998. Ingeniería de Software, Un enfoque práctico. Cuarta edición. Mc Graw Hill.

Dolado J.J, Fernandez L. 1998. Genetic programming, neural networks and linear regression in software project estimation. In INSPIRE III, Process Improvement through training and education. C. Hawkins, M.Ross, G. Staples, J.B. Thompson (Eds.), The British Computer Society, pp. 157-171.

Krishnamoorthy K, Moore B. 2002. Combining information for prediction in linear regression, Metrika, 56, 73-81.

Wettschereck C. 1995. A study of distance-based machine learning algorithms. PhD thesis, Oregon State University.

Humphrey W S. 1995 A Discipline for Software Engineering. Addison-Wesley.Reading, Massachusetts.

Jones C. 1996 Applied Software Measurement 2º Edition. McGraw-Hill. NewYork.

Lin K, Yang C. 2001. The ANN-tree: An index for efficient approximate nearest neighbor search. Proceedings of the Seventh International Conference on Database Systems for Advanced Applications.

Maneewongvatana K.S., Mount D.M, 2001. The Analysis of a Probabilistic Approach to Nearest Neighbor Searching.' in Proc. 7th Workshop on Algorithms and Data Structures (WADS 2001), pp. 276-286.