УДК 004.054

БОЛДАК А.А.
НЕВДАЩЕНКО М.В.
ДЕМЬЯНЕНКО М.В.

СИСТЕМА ФАКТОРОВ ДЛЯ КОМПАРАТИВНОГО АНАЛИЗА ПРОГРАММНЫХ ПРОЕКТОВ

В статье разрабатывается система факторов, вычисляемых на основе параметрической модели COCOMO, которая может применятся в рамках компаративного анализа программных проектов. Факторы определяются на основе результатов корреляционного анализа параметрических оценок COCOMO для выборки из 156 программных проектов. Редукция параметрического пространства с помощью метода главных компонент подтверждает корректность определения разработанной системы факторов.

In this paper, we develop a system of factors, calculated on the basis of a parametric model COCOMO, which can be applied in the comparative analysis of software projects. Factors are determined on the basis of the results of correlation analysis of COCOMO parametric estimates for a sample of 156 software projects. Reduction of the parameter space using the method of principal components confirms the correctness of determination of the developed system factors.

ВВЕДЕНИЕ

Программный проект (ПП) как процесс предполагает разработку программного продукта надлежащего качества при условии ограничений на сроки и ресурсы.

Сложностью планирования этого процесса и необходимостью объективной оценки его состояния объясняется интерес к использованию различных моделей параметрической оценки ПП[1-4]. Одной из таких моделей, которая широко используется при планировании ПП, является модель COCOMO (Constructive Cost Model), предложенная Барри Боэмом в 1981 году[2].

В соответствии с этой моделью для вычисления затрат, необходимых для выполнения ПП (в человеко-месяцах) используется формула:

$$effort = a \cdot KLOC^{b} \cdot \prod_{i=1}^{15} X^{i}, \qquad (1)$$

где *KLOC* – количество строк программного кода (тыс. стр. кода); *a*, *b* – константы, значения которых зависят от категории ПП, X^{i} – критерии затрат (метрики).

В этой модели используется разбиение программных проектов на 3 категории:

- organic – небольшие ПП, разработанные группой людей до 5 человек;
- embedded – ПП в рамках сложных взаимосвязанных операционных (встроенных) аппаратных средств и (или) программного обеспечения;
- semi-detached – ПП, которые занимают промежуточное положение между organic и embedded.

Такое разделение является важным, поскольку позволяет уточнить значения констант *a* и *b* (табл. 1).

Табл. 1. Значения констант a и b в зависимости от категории ПП

Категория ПП	a	b
organic	3.2	1.05
embedded	2.8	1.2
semi-detached	3.0	1.12

Метрики X^{i} для ПП сведены в табл. 2 и могут принимать одно из 6 значений: vl (very low) – неудовлетворительное, l (low) – удовлетворительно, n (normal) – нормальное, h (high) – хорошее, vh (very high) – очень хорошее, xh (extra high) – отличное. Для каждого из этих значений определены соответствующие числовые эквиваленты[2].

Ниже определяется система факторов, вычисляемых на основе метрик COCOMO, которая может эффективно использоваться при компаративном анализе ПП.

Табл. 2. Критерии затрат ПП

Категория критериев	Обозн.	Описание критерия
Критерии ПП	RELY	Требуемая надежность ПП
	DATA	Размер базы данных
	CPLX	Сложность ПП
Критерии техники	TIME	Ограничения по быстродействию
	STOR	Ограничения по оперативной памяти
	VIRT	Изменяемость виртуальной машины – степень, в которой изменяется аппаратура и ПО (ОС, СУБД и т.п.)
	TURN	Цикл обращения к аппаратуре
Критерии исполнителей	ACAP	Квалификация аналитика
	AEXP	Опыт работы в данной области
	PCAP	Квалификация программиста
	VEXP	Опит работы с виртуальной машиной (ОС)
	LEXP	Опыт работы с языком программирования
Критерии проекта	MODP	Использование современных методик программирования
	TOOL	Использование инструментальных средств
	SCED	Ограничение сроков разработки

ПОСТАНОВКА ЗАДАЧИ

В работе используются данные о 156 ПП, представленные в [5], которые содержат значения 15 метрик, а также количество строк кода (KLOC) и реальные затраты (REFF). Эти данные были дополнены уровнем планируемых затрат (PEFF), расчитанным по формуле (1). Эти данные можно представить в виде:

$$X_i = \left\langle x_i^1, x_i^2, ..., x_i^m \right\rangle, \qquad (2)$$

где X_i – оценка i-го ПП; $x_i^1, x_i^2, ..., x_i^m$ – значения параметров оценки ПП; m – количество параметров оценки ПП, которое в нашем случае равно 18.

Существует рад причин обусловливающих применение методов многомерного статистического анализа (МСА) для описанных выше данных. Во-первых, исходные параметры X^i в силу тех или иных причин являются взаимосвязанными. Незнание характера и степени этих связей может существенно искажать реальную оценку ПП. Указанные характеристики взаимосвязей определяются с помощью методов корреляционного анализа. Во-вторых, для удобства и наглядности компаративного анализа ПП необходимо свести исходные метрики к нескольким скрытым (латентным) факторам. В данном случае оценка ПП может быть упрощенно представлена не 18 параметрами, а несколькими, наиболее существенными факторами. Для этих целей используют метод главных компонент.

ИСПОЛЬЗОВАНИЕ МСА В РАМКАХ КОМПАРАТИВНОГО АНАЛИЗА ПП

Поскольку, все параметры $X^i, i = \overline{1,18}$ выражены в интервальных шкалах, то степень линейной связи между отдельными параметрами можно оценить с помощью коэффициента корреляции Пирсона [6]:

$$r_{k,l} = \frac{\sum_{i=1}^{N}(x_i^k - \overline{X^k}) \cdot (x_i^l - \overline{X^l})}{\sqrt{\sum_{i=1}^{N}(x_i^k - \overline{X^k})^2} \cdot \sqrt{\sum_{i=1}^{N}(x_i^l - \overline{X^l})^2}},$$

где $r_{k,l}$ – коэффициент корреляции параметров X^l и X^k; $\overline{X^l} = \frac{\sum_{i=1}^{N} x_i^l}{N}; \overline{X^k} = \frac{\sum_{i=1}^{N} x_i^k}{N}$ – средние значения параметров X^k и X^l; N – количество ПП.

Из [6, 7] следует, что коэффициенты корреляции, которые по модулю больше 0,70, говорят о сильной связи параметров (при этом коэффициенты детерминации > 50%, т.е. один признак определяет другой более, чем наполовину). Коэффициенты корреляции, которые по модулю меньше 0,7, но больше 0,5, говорят о связи средней силы (при этом коэффициенты детерминации меньше 50%, но больше 25%). Наконец, коэффициенты корреляции, которые по модулю меньше 0,5, говорят о слабой связи параметров (при этом коэффициенты детерминации меньше 25%).

Коэффициенты корреляции для параметров с сильной и средней степенью связи можно представить в виде графа, вершинами которого являются параметры, а ребра соот-

ветствуют корреляционной связи высокого и среднего уровней.

Такой граф для исследуемых данных представлен на рис. 1. Как видно, множество параметров разбивается на 7 групп:

1. RELY, CPLX, TIME и STOR.
2. KLOC, REFF и PEFF.
3. LEXP, VEXP и VIRT.
4. DATA и TURN.
5. ACAP и PCAP.
6. MODP и TOOL.
7. SCED.

Каждой группе в соответствие может быть поставлена одна интегральная оценка (фактор).

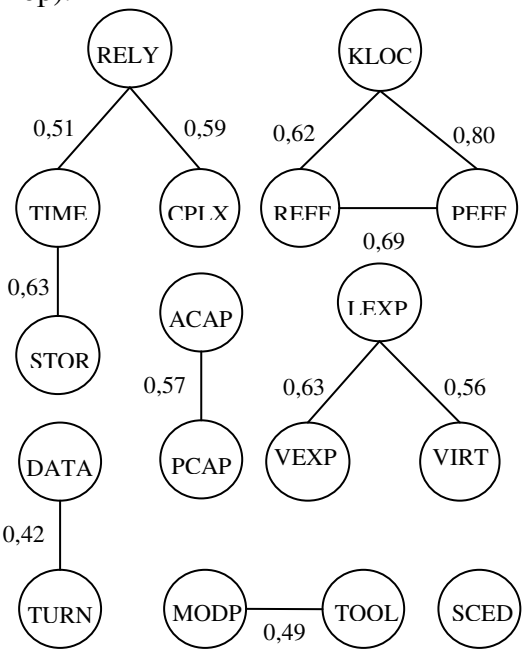

Рис. 1. Граф корреляционных связей

ПРИМЕНЕНИЕ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ ДЛЯ АНАЛИЗА МЕТРИК ОЦЕНКИ ПП

Использование метода главных компонент позволяет выявить минимальное число скрытых (латентных) факторов, отображающих основные свойства исходных 18 параметров, и одновременно уменьшить степень зависимости этих факторов от своих остаточных случайных компонент.

Осуществим поиск нового ортонормированного базиса векторов-компонент [8,9], определив разложение матрицы X в виде:

$$X = T \cdot P^T + E, \qquad (2)$$

где T – матрица счетов размерностью $n \times m'; (m' \le m)$; P – матрица нагрузок размерностью $m' \times m$; (m' - размерность пространства параметров; m - количество главных компонент, выбранных для проецирования); E – матрица остатков.

Определение главных компонент (факторов) связано с вычислением собственных векторов ковариационной матрицы [9,10], определяемой как:

$$C = (c_{ij}, c_{ij} = \mathrm{cov}(X^i, X^j)), i = \overline{1,m},\ j = \overline{1,m}, \qquad (3)$$

где $\mathrm{cov}(X^i, X^j) = \dfrac{\sum_{k=1}^{n} (x_k^i - \overline{X^i}) \cdot (x_k^j - \overline{X^j})}{n-1}$ – ковариация признаков X^i и X^j.

Из работы [10] следует, что сумма дисперсий параметров равна сумме дисперсий всех факторов, и они упорядочены в соответствии с долями их дисперсий. Поэтому анализируя изменение относительной доли дисперсии, вносимой первыми m' факторами можно определить число факторов, которое целесообразно оставить для последующего рассмотрения. При этом для выбора достаточного числа $m' \le m$ факторов часто используют кумулятивную дисперсию [10]:

$$D_i = \frac{\sum_{j=1}^{i} \lambda_j}{m}, i = \overline{1,m},$$

где $\lambda_j, j = \overline{1,m}$ – собственные числа ковариационной матрицы C.

В соответствии с критериями Кайзера [9] (см. рис. 2) и критериями Кэттеля [11] (табл. 3) для последующего исследования достаточно оставить семь факторов. Это позволит представить около 77% данных о ПП (табл. 3).

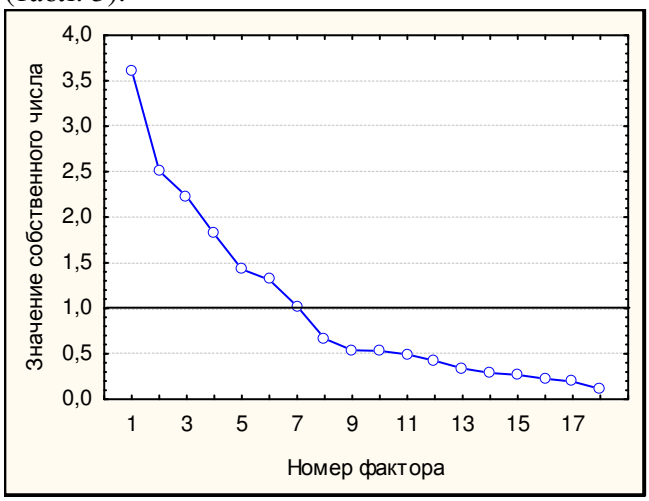

Рис. 2. График собственных чисел факторов

Матрица нагрузок P из формулы (2) задает отображение базиса исходного пространства X^i в пространство факторов T^j, в котором можно оценить значимость каждой из метрик.

Табл. 3. Таблица собственных значений и дисперсий факторов

Номер фактора	Собственное значение	Общая дисперсия, %	Суммарное собственное значение	Суммарная дисперсия, %
1	3,61	20,08	3,61	20,08
2	2,51	13,94	6,12	34,02
3	2,23	12,37	8,35	46,39
4	1,82	10,08	10,17	56,47
5	1,43	7,94	11,59	64,41
6	1,31	7,29	12,91	71,70
7	1,02	5,67	13,93	77,36

Проанализировав матрицу нагрузок в пространстве 18 метрик и семи факторов для исходных данных, можно сделать следующие выводы:

- метрики RELY, CPLX, TIME и STOR, с коэффициентами нагрузок 0.82, 0.77, 0.75 и 0.74 соответственно, составляют фактор 1 – требования к надежности, сложности, быстродействию и объемам оперативной памяти;
- метрики KLOC, REFF и PEFF, с коэффициентами нагрузок 0.91, 0.80 и 0.91, составляют фактор 2 – трудоемкость ПП и затраты на его производство;
- метрики LEXP и VEXP, с коэффициентами нагрузок равными 0.87, определяю фактор 3 – использование традиционных технологий;
- метрики DATA и TURN, с коэффициентами 0.74 и 0.84, определяют фактор 4 – требования к аппаратному обеспечению;
- метрики ACAP и PCAP, с коэффициентами 0.81 и 0.84, определяют фактор 5 – квалификация персонала;
- метрики MODP и TOOL, с коэффициентами 0.75 и 0.86, определяют фактор 6 – использование современных средств разработки программного обеспечения;
- метрика SCED, с коэффициентом нагрузки 0.92, определяет фактор 7 - ограничения на время разработки.

Согласованность результатов корреляционного и факторного анализов дает возможность предложить модель параметрической оценки ПП, суть которой заключается в использовании интегральных показателей. Для получения этих показателей необходимо провести корреляционный анализ и построить граф корреляционных связей. При этом число компонент связности соответствует числу интегральных показателей. Поскольку коэффициенты нагрузок параметров в каждом факторе примерно одинаковы, то каждый из показателей определяется как среднее арифметическое значений параметров, соответствующих компоненте связности.

ЗАКЛЮЧЕНИЕ

Установлено, что в рамках компаративного анализа ПП может быть использована система факторов, в которую входят требования к надежности, сложности, быстродействию и объемам оперативной памяти, трудоемкость ПП и затраты на его производство, использование традиционных технологий, требования к аппаратному обеспечению, квалификация персонала, использование современных средств разработки программного обеспечения, ограничения на время разработки.

Числовые значения этих факторов можно определять как средние арифметические параметров оценки COCOMO. Для этих целей необходимо провести корреляционный анализ и построить граф корреляционных связей для выборки ПП.

СПИСОК ЛИТЕРАТУРЫ

1. Andrew Stellman, Jennifer Greene. Applied Software Project Management // Sebastopol, MA: O'Reilly Media. – 2005. – 308 p.
2. Barry Boehm. Software engineering economics // Englewood Cliffs. – 1981. – 308 p.

3. Lawrence H. Putnam, Ware Myers. Five core metrics : the intelligence behind successful software management // Dorset House Publishing. – 2003. – 328 p.
4. Albrecht A. J. Measuring Application Development Productivity // Proceedings of the Joint SHARE, GUIDE, and IBM Application Development Symposium. – 1979. – P.83-92.
5. Данные для модели COCOMO // Режим доступа: http://unbox.org/wisp/trunk/cocomo/data, свободный. – Загл. с экрана. – Яз. англ.
6. Айвазян С. А. и др. Прикладная статистика: Основы моделирования и первичная обработка данных. Справочное изд. // С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. — М.: Финансы и статистика, 1983. — 471с.
7. Айвазян С.А. и др. Прикладная статистика: Исследование зависимостей: Справ, изд. // С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин; Под ред. С. А. Айвазяна. – М.: Финансы и статистика, 1985. – 487 с, ил.
8. Lindsay I Smith. A tutorial on Principal Components Analysis. – 2002. – 208 p.
9. Kaiser H. F. The application of electronic computers to factor analysis // Educational and Psychological Measurement. – 1960. – P.141-151.
10. Прикладная статистика: Классификации и снижение размерности: Справ, изд. // С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин; Под ред. С. А. Айвазяна. – М.: Финансы и статистика, 1989. – 607 с: ил.
11. Cattell R. B. The scree test for the number of factors // Multivariate Behavioral Research. – 1966. – P.245-276.