Computer Science and Application 计算机科学与应用, 2017, 7(5), 428-437
Published Online May 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.75052

Hans汉斯

The Effort Distribution of Software Development Phases

Yong Wang, Jing Zhang

College of Information Science and Engineering, Ocean University of China, Qingdao Shandong
Email: markwy@126.com

Received: May 7th, 2017; accepted: May 21st, 2017; published: May 24th, 2017

Abstract

In order to effectively control the software development process, understanding the distribution rules of different life cycle phases is needed. This paper analyzes the effort distribution of development phases on the basis of a large-scale real software project data set-EDS. It is found that the phase effort is consistent with the normal distribution, and the effort distribution of New Development type and Enhancement type is consistent. There are significant differences in the distribution patterns between the Other Projects & Services type and the other three development types. As the duration of the project grows, the effort of Produce phase is on the rise and the Implementation phase effort is declining. The results of the study are quite different from the traditional results based on the individual project or small-scale data sets, which have a good effect on software project effort management and schedule control.

Keywords

Software Project Management, Development Phases, Normal Distribution, Schedule Control

软件开发阶段成本分布研究

王　勇，张　敬

中国海洋大学信息科学与工程学院，山东　青岛
Email: markwy@126.com

收稿日期：2017年5月7日；录用日期：2017年5月21日；发布日期：2017年5月24日

摘　要

为了对软件开发过程进行有效控制，需要了解软件生命周期不同开发阶段的成本分布规律。本文通过分

文章引用: 王勇, 张敬. 软件开发阶段成本分布研究[J]. 计算机科学与应用, 2017, 7(5): 428-437.
https://doi.org/10.12677/csa.2017.75052

析一个大规模真实软件工程项目数据集-EDS，对不同开发阶段成本的分布规律进行了全面地揭示，发现各开发阶段成本符合正态分布，New Development类型和Enhancement类型的项目成本分布一致，Other Projects & Services类型和其他三种开发类型项目成本分布存在显著差异。随着项目持续时间的增加，Produce阶段成本呈上升趋势，Implement阶段成本呈下降趋势。本文的研究结果与传统基于单个项目或小规模数据集上得出的结果有较大不同，对软件项目成本的管理和进度控制有较好的促进作用。

关键词

软件项目管理，开发阶段，正态分布，进度控制

1. 引言

软件开发过程中合理的成本分配是促进软件项目规划的基础[1] [2] [3] [4]，做好软件项目在开发过程中的成本分配工作，有利于软件项目的成本管理和进度控制，但是与成本估算的其他研究相比，成本分配极少引起研究者的关注。软件开发阶段成本分布研究，可以提高人们对成本分配的理解[5]，从而对成本进行有效控制，改善软件项目开发过程。

现有的研究中，Norden [6]认为瑞利曲线符合人员分配随着时间变化的趋势，而且发现在尾端部分的过长曲线没有实际意义。MacDonell [7]认为对于特定的开发活动，不存在一种“标准成本比例”，并且肯定了简单的定量分析方法和局部数据的价值。杨叶[8] [9]等人研究了 CSBSG 数据库中的项目成本数据，在开发类型、软件规模和开发团队规模等方面分析了软件成本分布的情况，还单独对维护类型的工业项目成本进行了分布模式的研究，认为在多种因素的影响下，成本分布变化较大。Heijstek [10]为提高对项目动态分配资源方面的理解，研究了 RUP 模式下成本分布情况，在项目开发早期提供成本分配的参考意见。

在这些已有的研究工作中，有按照传统瀑布模型和 COCOMO 模型进行成本分布的研究，也有按照 RUP 模式进行的分布研究，重点对某一工业类型项目的成本分布研究[9] [11]。但是这些研究的项目数大都在一百以内，缺少对大规模数据集及多种开发类型、不同持续时间项目的阶段成本分布研究。本文以课题合作伙伴提供的大规模软件项目数据集 EDS 为依据，从数据集总体以及多种开发类型、不同项目持续时间等方面研究了软件项目开发阶段成本的分布情况和相应差异，能够有效指导项目经理对项目资源进行分配。

2. 数据预处理

2.1. 数据集概述

本文在大规模的历史项目数据集-EDS 进行研究，数据集共包含 14054 个软件项目，各类信息数据上万条，是目前世界上规模最大的软件工程数据集之一。其中，项目集属性众多，如表 1 所示。原始数据集中包含多种开发方法学，每种方法学都具备相应细化的开发活动阶段。同时，数据集中的项目来源于多种行业，例如电信业、制造加工业、政府机构等；开发类型也多种多样，既有新开发的软件项目，也有对软件产品的维护项目；项目所属国家包含美国、新西兰、法国等；项目所采用的开发团队规模也大小不一；项目开发时间从上世纪七十年代至今。

数据集中各开发阶段所包含的详细开发活动，如表 2 所示。其中 Define & Analyze 阶段对应于问题

Table 1. Partial property set of the dataset
表 1. 数据集部分属性集列表

Attributes	Description
项目持续时间	每个项目所占用的时间跨度，单位：月
总成本	项目工作量总成本，单位：人小时
阶段成本	各开发阶段所花费的成本，单位：人小时
开发类型	项目所属开发类型，例如新开发类型、增强类型等
工业类型	项目所属行业，例如电信通信业、制造加工业、财政业等
开发时间	项目各类开发阶段的开发时间，单位：年月
开发语言	C、C++、SQL、HTML、SAS、JAVA、ASP、PASCAL、Oracle 等语言
所属地区	亚太区、美国、加拿大等地区
所属国家	美国、加拿大、津巴布韦等非洲、亚太地区国家
开发团队规模	单人开发，单个开发团队，多个开发团等
项目复杂度	低级、中级、高级
功能点数目	项目开发过程中的功能点，包含未调整的功能点、调整的功能点等方面
开发方法学	项目所使用开发方法学，共 46 种，不同方法学包含的项目子阶段不同
开发工具类型	项目开发时使用的各类工具，例如网页开发工具、测试工具、文本分析等

Table 2. Development activities included in each development phase of the data set
表 2. 数据集各开发阶段所包含的开发活动

Phases	Activities included
Define & Analyze	Define, Document current environment, Refine & Analyze Requirements, Planning.
Design	Design System, Design Application, Technical Design, Business Design.
Produce	Development, Integration, Produce application.
Optimize	User Acceptance Test, Software Test-System Test, System Test, Testing.
Implement	Release control, Release Application, Implement Application, Production Preparation, Deploy.

定义和需求分析阶段；Design 阶段对应于系统设计和详细设计阶段；Produce 阶段对应于编码阶段；Optimize 阶段对应于测试阶段；Implement 阶段对应于实施交付阶段。

2.2. 数据筛选

为了获得实验数据集，对原始数据集中的项目进行筛选，步骤如下：

1) 在原始数据集 14,054 个项目中，筛选出通用方法学 ID 为 20 的项目，得到 8540 个项目；
2) 将包含开发阶段成本数据的项目筛选出来，共 6879 个项目；
3) 按照生命周期中的开发阶段进行成本汇总，得到 6879 个项目的成本数据；
4) 筛选出包含完整生命周期的项目，共计 2588 个项目；
5) 去除这些项目中各开发阶段成本中存在零值的项目数据，剩余 2570 个项目。

2.3. 归一化处理

如表 3，各开发阶段的均值和中位数差别较大。由于项目规模、开发类型等因素的影响，软件项目各

生命周期阶段成本数值大小各异，不利于研究软件项目的阶段成本分布，因此，对数据进行归一化处理。

归纳统一样本的统计分布性，可以更清晰地分析软件项目的阶段成本分布情况。本文将阶段成本数统一转化为各阶段比例关系，即各开发阶段成本所占项目开发总成本的比例，如表 4。

由表 4 可知，阶段成本进行归一化后，均值和中位数之间的关系，与未进行归一化之前的数据关系大有不同：经过归一化之后的成本中位数和均值的差异不明显，在 Produce 阶段仅仅相差 0.26%，而 Design 阶段相差 1.56%，Define & Analyze 阶段相差 2.63%，Implement 阶段相差最大是 3.71%，进行归一化之后的数据更加直观合理。对阶段成本进行归一化操作，转化为各开发阶段成本所占该项目总成本的比例，得到实验数据集。

3. 实验结果与分析

3.1. 数据集总体阶段成本分布

根据项目阶段的成本信息绘制出各阶段成本分布的频数直方图和概率密度估计曲线，如图 1。其中，左坐标表示成本分布的频数，右坐标表示成本分布的概率密度估计大小。图像表明，Define & Analyze 阶段、Design 阶段、Produce 阶段成本的概率密度曲线光滑，类似正态分布的趋势，尤其是 Produce 阶段的成本分布曲线与正态分布曲线一致性较高，呈现正态分布的完整曲线，Optimize 阶段，Implement 阶段的概率密度曲线光滑，大致呈现正态分布曲线的半峰形态，向右滑动。同时，Produce 阶段的分布范围比较广泛，大致在 10%~70%的范围内，呈现比较大的跨度，而 Define & Analyze 阶段、Design 阶段、Optimize 阶段、Implement 阶段的分布大致局限在 5%~30%范围内。这对于未来项目的成本分配在一定程度上提供了参考意见。

3.2. 各开发类型项目阶段成本分布

开发类型主要包括四类：New Development、Enhancement、Maintenance、Other Projects & Services，各开发类型对应 Type_ID 如表 5。表 6 是四种开发类型项目信息，New Development 类型花费的成本是最

Table 3. The statistical description of software project effort
表 3. 软件项目成本数据统计描述

	max	median	mean	min
Total Effort	308339.2	1757.5	4588.79	5
Define & Analyze	31283.5	207.5	631.27	0.25
Design	35677	286	871.88	0.75
Produce	160351	754	2082.23	0.5
Optimize	58432.7	143.7	573.62	0.25
Implement	49386	99	429.78	0.25

Table 4. The statistical description of the development phase effort ratio
表 4. 各开发阶段成本比例统计描述

	Define & Analyze	Design	Produce	Optimize	Implement
max	94.73%	77.15%	94.96%	90.74%	99.33%
median	12.47%	17.32%	44.47%	8.79%	5.66%
mean	15.1%	18.78%	44.61%	12.13%	9.37%
min	0.036%	0.048%	0.058%	0.011%	0.001%

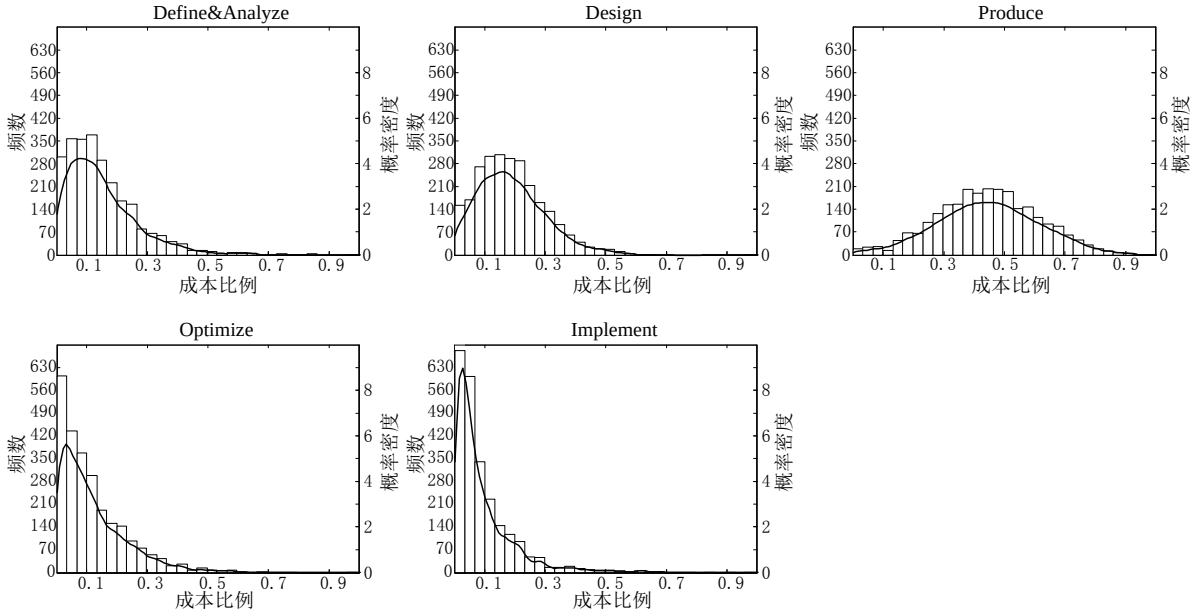


Figure 1. The frequency histogram and probability density curve of each development phaseeffort distribution
图 1. 各阶段成本分布频数直方图和概率密度估计曲线

Table 5. The corresponding ID of each development type
表 5. 开发类型对应 ID

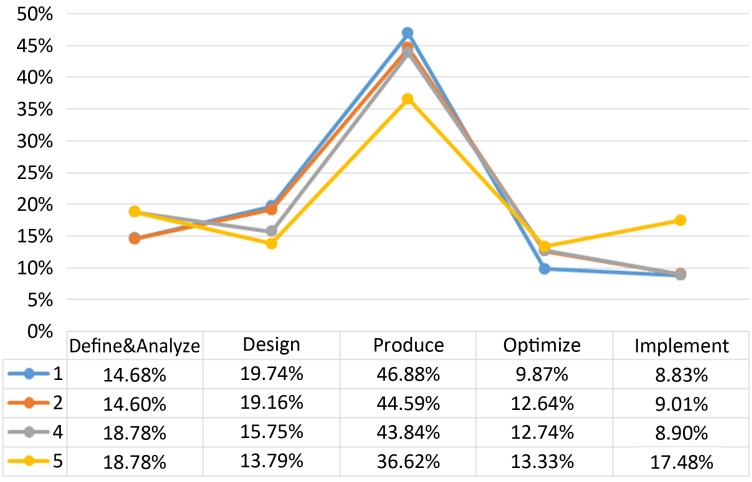
Type_ID	Type_ description
1	New Development
2	Enhancement
4	Maintenance
5	Other Projects&Service

Table 6. The project information of each development type
表 6. 各开发类型项目信息

	1	2	4	5
Total Effort	8043.2	3781.39	3112.42	3957.19
项目数	510	1761	174	124
项目占比	19.84%	68.52%	6.77%	4.82%

多，显著大于其他开发类型的项目成本，其他的三种类型的总成本在 3000 到 4000 人小时之间，其中 Maintenance 类型的项目总成本花费最少。

根据不同开发类型项目的阶段成本信息绘制成本分布折线图，如图 2。各开发类型项目 Design 阶段的成本约是 Produce 阶段的一半。在开发阶段的成本之间，可能存在重要的分布关系。各类型项目花费在 Implement 阶段的成本是最少的，而 Other Projects & Services 类型在 Implement 阶段的成本比例在所有开发类型中最高。New Development 类型和 Enhancement 类型的项目成本分布比较一致，Other Projects & Services 类型与其他类型项目在成本分布上有显著差别。New Development 类型项目在 Produce 阶段的成本较 Enhancement 类型项目高(约 2.29%)；Enhancement 类型项目在 Optimize 阶段成本比 New Development 类型项目较高(约 2.77%)，这和杨叶关于开发类型这一属性的研究结果[8]相同。出现这种差别的原因是



	Define&Analyze	Design	Produce	Optimize	Implement
1	14.68%	19.74%	46.88%	9.87%	8.83%
2	14.60%	19.16%	44.59%	12.64%	9.01%
4	18.78%	15.75%	43.84%	12.74%	8.90%
5	18.78%	13.79%	36.62%	13.33%	17.48%

Figure 2. The line chart of each development type project phase effort distribution
图 2. 各开发类型项目阶段成本分布折线图

Enhancement 类型的项目编码并不是对于整个项目编码，只是针对某一部分功能编码，而在测试时，进行的是功能模块单元测试以及整个系统的集成测试。因此，相对来说，Enhancement 类型在 Optimize 阶段的成本比例较高。另外，Enhancement 类型和 Maintenance 类型的分布，在后期开发阶段的成本分布折线几乎是重合的，主要区别在 Define & Analyze 阶段和 Design 阶段：Maintenance 类型在 Define & Analyze 阶段高出大约 4.18%，而 Enhancement 类型在 Design 阶段高出 3.31%，这是因为 Enhancement 类型在添加新功能时进行设计，与 Maintenance 类型侧重阶段不同。在成本规划过程中应注意开发类型对成本分配的影响。

3.3. 不同持续时间项目阶段成本分布

项目持续时间，指从软件项目开始一直到测试、交付等软件活动结束所占的时间跨度。本文中以月为基本单位进行度量。实验数据包含持续时间为 4 至 20 个月的项目，按照不同持续时间对项目进行分类。

如表 7，根据成本均值和中位数可知，项目总成本和项目持续时间呈正比例关系，以此数据画出项目持续时间与项目总成本散点图。同时，绘制相应的指数趋势线和线性趋势线，得出项目总成本随项目持续时间而变化的趋势情况，如图 3、图 4。

判定系数 R^2 可以衡量回归曲线拟合样本的优劣程度，其值越接近 1 说明拟合的效果越好。相比项目持续时间和项目总成本之间的线性增长关系(判定系数 R^2 为 0.8808、0.8975)，它们之间的指数增长趋势线(判定系数 R^2 为 0.9373、0.967)与数据之间的拟合效果更好，这一点不论是从趋势线的拟合程度还是判定系数 R^2 上，都可以得到体现：指数趋势线的相关程度大于线性趋势线的相关程度。项目持续时间和项目总成本之间的这种回归关系在项目管理和成本控制方面提供了重要的参考。

持续时间为 4 至 20 个月的项目阶段成本比例数据，如表 8 所示。

根据表 8 绘制阶段成本比例与项目持续时间折线图，更直观看到结果，如图 5。

项目持续时间在 4~10 个月时，各开发阶段的成本分布趋势比较平稳。随着项目持续时间的增加，在 10~20 个月之间，Define & Analyze 阶段、Design 阶段的变化趋势大致同步：Define & Analyze 阶段的相对花费成本上升或者下降时，Design 阶段随着 Define & Analyze 阶段的变化而变化。Produce 阶段成本随项目持续时间的增加呈上升趋势，Implement 阶段相对成本呈下降趋势，也就是说随着项目持续时间的增加，在 Produce 阶段的成本增多，在 Implement 阶段的成本减少。

Table 7. The software project information for duration from 4 to 20 months
表 7. 持续时间为 4~20 个月的软件项目信息

持续时间	项目数	项目占比	项目总成本(mean)	项目总成本(median)
4 个月	196	9.32%	1329.34	848.25
5 个月	278	13.22%	1722.26	1115.25
6 个月	295	14.03%	2526.11	1403
7 个月	260	12.36%	2773.65	1689.5
8 个月	228	10.84%	2752.01	1728.13
9 个月	188	8.94%	4028.32	2342.5
10 个月	147	6.99%	6639.77	3208.75
11 个月	114	5.42%	4335.25	2669.25
12 个月	87	4.14%	6511.53	3005.75
13 个月	78	3.71%	6945.49	3750.88
14 个月	47	2.23%	9983.63	6209.25
15 个月	41	1.95%	7901.69	6099
16 个月	44	2.09%	13227.12	7648.25
17 个月	35	1.66%	13951.51	7639.5
18 个月	27	1.28%	13823.34	8886.75
19 个月	18	0.86%	11397.53	7611
20 个月	20	0.95%	19400.28	11953.25

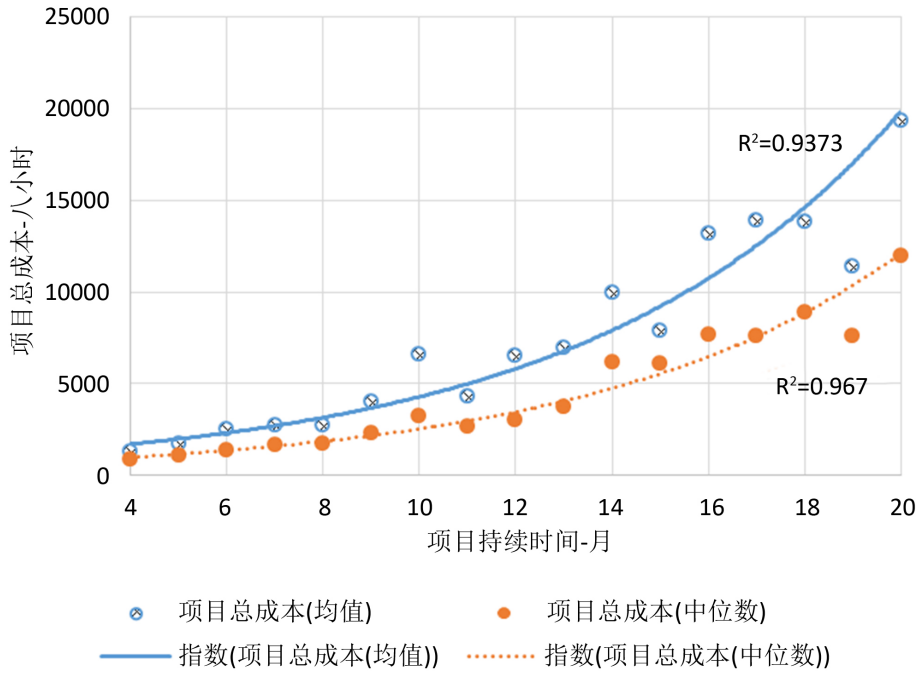


Figure 3. The exponential trend between the duration and project total effort
图 3. 持续时间与项目总成本指数趋势图

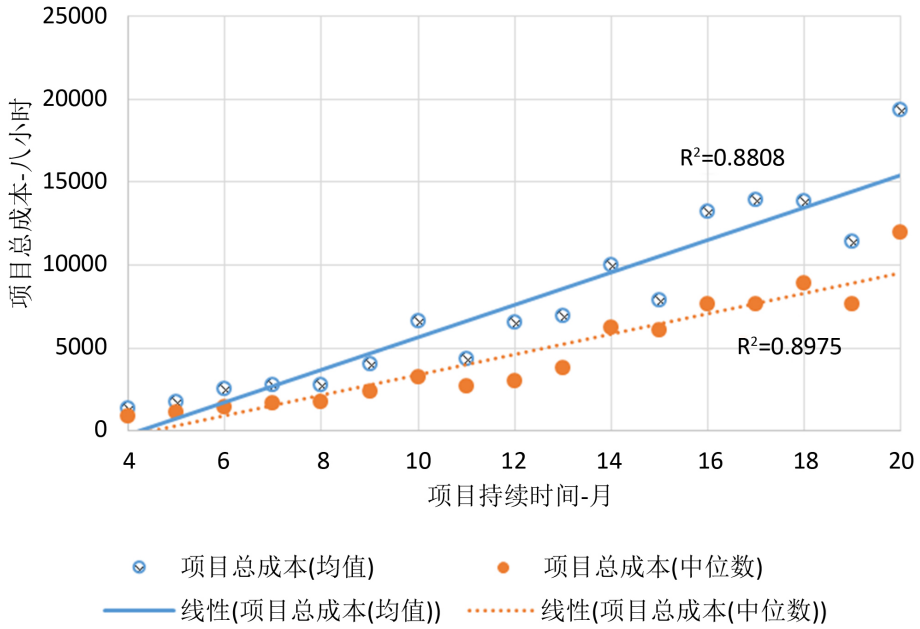


Figure 4. The Linear trend between the duration and project total effort
图 4. 持续时间与项目总成本线性趋势图

Table 8. The average phase effort ratio for project duration from 4 to 20 months
表 8. 持续时间为 4-20 个月项目的阶段成本比例(mean)

持续时间	Define & Analyze	Design	Produce	Optimize	Implement
4 个月	12.74%	16.96%	40.7%	8.99%	7.52%
5 个月	11.64%	18.69%	44.91%	8.47%	6.38%
6 个月	11.58%	17%	45.7%	9.58%	5.61%
7 个月	12.46%	17.63%	43.46%	10.33%	6.34%
8 个月	12.22%	17.95%	42.69%	10.07%	5.65%
9 个月	12.8%	19.29%	43.08%	9.68%	5.08%
10 个月	11.5%	19.64%	42.98%	10.24%	5.49%
11 个月	12.61%	16.03%	48.01%	7.71%	5.23%
12 个月	13.69%	17.62%	45.32%	8.5%	5.84%
13 个月	10.05%	14.77%	45.77%	8.24%	5.49%
14 个月	13.15%	19.06%	48.84%	6.43%	3.59%
15 个月	12.22%	20.07%	47.44%	8.37%	3.53%
16 个月	14.82%	17.42%	45.84%	7.43%	5.14%
17 个月	16.22%	19.63%	46.98%	7.29%	5.52%
18 个月	11.38%	15.2%	46.8%	7.26%	4.57%
19 个月	15.45%	20.2%	48.06%	7.37%	4.25%
20 个月	11.33%	11.78%	50.55%	11.13%	3.77%

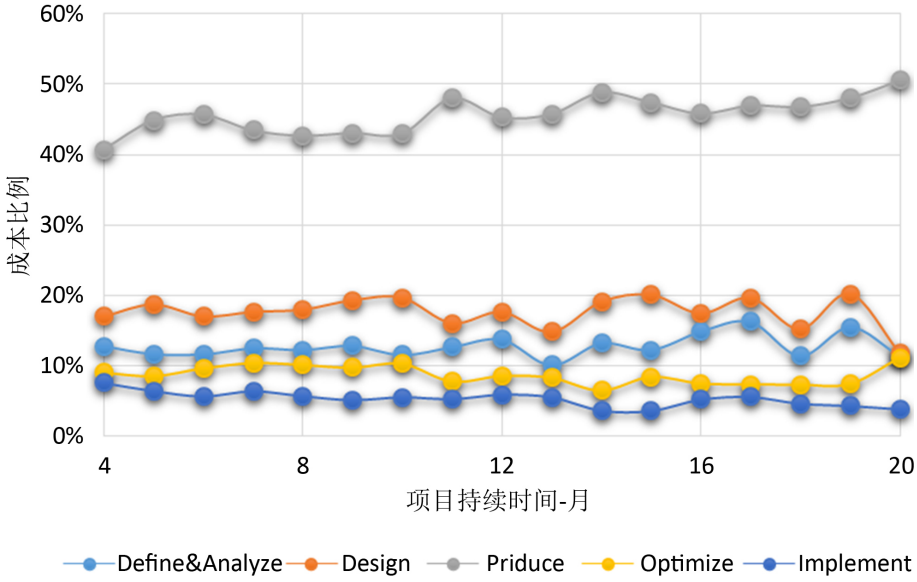


Figure 5. The trend of project phase effort ratio under different durations
图 5. 不同持续时间下阶段成本比例变化趋势

4. 结论

本文基于大规模软件项目数据集 EDS，研究了软件开发成本的阶段分布情况。本研究从数据集总体、多种开发类型以及不同项目持续时间等方面分析了项目成本分布情况。研究结果显示软件生命周期各开发阶段成本符合正态分布的规律；New Development 开发类型和 Enhancement 开发类型的项目成本分布一致性较高，而 Other Projects & Services 开发类型项目和其他开发类型项目成本分布存在显著差异；Produce 阶段成本随着项目持续时间增长而上升，相反，Implement 阶段成本随着项目持续时间增长而降低，同时在项目持续时间为 10~20 个月时，Define & Analyze 阶段、Design 阶段成本的变化趋势同步。本文的研究对阶段成本的建模预测提供了一定的理论基础作用，对各开发阶段的成本分配有重要的参考价值，在软件项目成本管理和进度控制方面提供了有效地指导。后续研究将继续探索其他因素对软件阶段成本分布的影响，如团队规模、软件大小等因素，为改进软件项目管理提供更丰富的指导意见。

基金项目

本论文得到国家自然科学基金面上项目(61170312)及软件工程国家重点实验室开发基金项目(SKLSE 2012-09-14)的支持。

参考文献 (References)

[1] Boehm, B.W., Horowitz, C., *et al*. (2000) Software Cost Estimation with COCOMOII. Prentice Hall, Upper Saddle River, 1-4.

[2] Reifer, D.J. and Consultants, R. (2004) Industry Software Cost, Quality and Productivity Benchmarks. *The DoD Software Tech News*, **7**, 3-4.

[3] Heijstek, W. and Chaudron, M.R.V. (2007) Effort Distribution in Model-Based Development.

[4] Boehm, B.W. (1981) Software Engineering Economics. Prentice-Hall, Upper Saddle River, 641-686.

[5] Boehm, B.W. and Papaccio, P.N. (1988) Understanding and Controlling Software Costs. *IEEE Transactions on Software Engineering*, **14**, 1462-1477.

[6] Norden, P.V. (1958) Curve Fitting for a Model of Applied Research and Development Scheduling. *IBM Journal of Research & Development*, **2**, 232-248.

[7] Macdonell, S.G. and Shepperd, M.J. (2003) Using Prior-Phase Effort Records for Re-Estimation during Software Projects. *Proceedings of the Ninth International Software Metrics Symposium*, Sydney, 3-5 September 2003, 73-86.

[8] Yang, Y., He, M., Li, M., *et al*. (2008) Phase Distribution of Software Development Effort. *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, Fraunhofer Cente, 8-10 October 2008, 61-69.

[9] Yang, Y., Li, Q., Li, M., *et al*. (2008) An Empirical Analysis on Distribution Patterns of Software Maintenance Effort. *Proceedings of the IEEE International Conference on Software Maintenance*, Beijing, 28 September-4 October 2008, 456-459.

[10] Heijstek, W. and Chaudron, M.R.V. (2008) Exploring Effort Distribution in RUP Projects. *Proceedings of the ACM/IEEE International Symposium on Empirical Software Engineering & Measurement*, Fraunhofer Cente, 8-10 October 2008, 359.

[11] Lucia, A.D., Pompella, E. and Stefanucci, S. (2003) Assessing the Maintenance Processes of a Software Organization: An Empirical Analysis of a Large Industrial Project. *Journal of Systems & Software*, **65**, 87-103.