다중회귀 분석을 이용한 소프트웨어 개발노력추정

정 혜 정† · 양 해 술†† · 신 석 규††† · 이 상 운††††

요 약

소프트웨어분야에서 성공적인 프로젝트를 완수하기 위해서는 프로젝트를 완수하는데 필요한 개발노력이 정확히 추정되어야 한다. 그러나 이러한 개발노력은 소프트웨어의 크기나 여러 가지 운영환경의 영향으로 인해 프로젝트에 따라서 총 개발 노력의 규모는 차이가 있다. 일반적으로 기존의 연구는 개발노력을 추정하기 위하여 소프트웨어 규모인 기능점수(FP ; Function Point)를 이용하였다. 본 연구를 위해서 1990년대에 개발된 789개의 소프트웨어 개발 프로젝트들에 관련된 데이터를 이용하였다. 실험을 통해서 개발노력에 영향을 미치는 변수를 조사하였다. 또한 변수사이에 선형적인 관계를 조사하기 위하여 다중회귀분석을 실시하였다. 이 경우 전체의 데이터를 이용하는 것이 아니라 프로젝트 인도비율(PDR ; Project Delivery Rate : Hours/FP)을 다단계로 나누어서 각 단계별로 개발노력에 영향을 미치는 변인을 찾아내고 가장 이상적인 회귀식을 도출하였다.

The Estimation of Software Development Effort Using Multiple Regression Method

Hye-Jung Jung† · Hae-Sool Yang†† · Seok-Kyoo Shin††† · Sang-Un Lee††††

ABSTRACT

To accomplish a project successfully, we have to estimate development effort accurately. But, development effort is different to software size and operation environment. Usually, we made use of function point for estimating development effort. In this paper, we make use of 789 project data. It is related to development projects in 1990's. We investigate the variable affecting development effort. Also, we exedcute the multiple linear regression analysis for looking linear relation about variables. We find the regression equation for multistage by dividing PDR that influenced development effort step by step.

키워드 : 기능점수(Function Point), 프로젝트 인도율(Project Delivery Rate), 개발노력(Development Effort), 다중회귀(Multiple Regression), 단계적선택법(Stepwise Method)

1. 서 론

성공적으로 프로젝트를 완수하기 위해서는 프로젝트를 완수하는데 요구되는 기간과 노력을 정확하게 추정하여야 한다. 기존의 연구에서는 소프트웨어의 노력 추정 모델들은 대체적으로 생명주기 전반에 걸쳐 투입되는 총 개발노력과 단위시간당 소요되어지는 노력 함수만을 제시하고 있다. 또한 Putman은 세부 단계별로 일정한 개발 노력 투입 비율을 제시하였으나 소프트웨어의 개발 노력이라는 것은 소프트웨어의 규모, 복잡도와 운영환경의 영향으로 인해 프로젝트 별로 투입되는 총 개발 노력의 규모에는 차이가 있으므로 개발 세부 단계별로 투입되는 노력의 규모도 프로젝트 마다 차이가 발생되어진다. 대형 프로젝트의 경우는 1% 만이 계획된 기간과 예상비용 한도 내에서 고객을 만족시키며 제품 개발을 완료하였다는 보고도 있으며 대부분의 대형 프로젝트들은 1년 이상 일정이 지연되고 초기 예상 비용의 많은 정도가 초과되어지고 있다고 한다. 그러므로 프로젝트를 관리하는 측면에서 소프트웨어의 개발 및 유지보수 비용을 줄이고자 체계적인 연구가 계속되어지고 있으며 계획단계에서 정확한 일정과 비용산정을 위해서 프로젝트를 관리할 때 발생되는 다양한 의사결정, 소요예산 및 개발 인원 분배와 계약 체결 여부에 신뢰할 만한 정보를 제공하고 있다. 개발 초기에 소프트웨어의 개발 노력이나 비용을 산정하기 위해서 주어진 실험데이터를 잘 설명할 수 있는 정확한 모델을 제시하는 것은 쉽지 않다. 그 이유는 개발노력과 비용에 영향을 미치는 환경적인 요인들이 불명확하기 때문이다. 소프트웨어의 개발노력이나 비용을 추정하려면 소프트웨어의 규모나 그 외 환경적인 요인들의 속성간의 작용에 의한 변화가 많을 것이다. 기존의 소프트웨어 규모를 추정하

※ 본 연구는 대학 IT 연구센터육성·지원사업의 연구결과로 수행되었음.
† 종신회원 : 평택대학교 정보통계학과 교수
†† 종신회원 : 호서대학교 벤처전문대학원 교수
††† 정 회 원 : 한국정보통신기술협회 IT시험연구소 소프트웨어 시험센터 센터장
†††† 정 회 원 : 국립 원주대학 여성교양과 교수
논문접수 : 2003년 11월 12일, 심사완료 : 2004년 10월 15일

기 위한 척도로는 LOC(Line Of Codes)와 FP(Function Point)를 많이 사용하였다.

소프트웨어의 규모 추정을 위해서 초기에는 주로 소프트웨어의 길이(Length)로써 규모를 추정한 LOC를 이용하였으나[8, 10], LOC는 소프트웨어의 개발 언어에 따라서 소프트웨어의 규모에 영향을 주며 코딩이 완료된 이후에야 크기를 추정할 수 있다는 어려운 문제점을 갖고 있다. 그러나 FP는 사용자에게 양도될 시스템의 기능에 의존하는 것으로 소프트웨어의 개발언어나 도구에 독립적이며 소프트웨어 개발 초기단계인 요구분석 단계에서 측정 가능한 잇점이 있다. 이러한 이유로 인하여 소프트웨어 규모인 기능점수에 따른 소프트웨어 개발 노력을 추정하기 위한 회귀분석 모델들은 많은 연구가 진행되어 있다. 그러나 이러한 연구결과는 특정업체에서 개발된 소프트웨어를 대상으로 하고 있으며 실험에 이용된 샘플수가 적어서 모든 개발 소프트웨어에 적용할 수 있는 모델은 아니다. 기능점수 FP를 이용하여 개발노력을 추정한 Albrecht et al.[1]은 IBM Data Processing Service에서 개발된 24개의 응용 프로그램에 대하여 단순회귀분석을 실시하였다. Kemerer는 ABC 회사에서 개발된 15개의 소프트웨어 개발 프로젝트에 대하여 단순회귀직선모델과 곡선회귀모델을 제시하였다.

개발 노력을 추정하기 위하여 FP를 이용한 기존의 연구는 소프트웨어 규모인 기능점수 FP를 독립변수로 하여 종속변수인 개발노력을 추정하는 회귀분석방법을 이용한 것이 대부분이다. 그러나 본 연구에서는 먼저 1990년대에 20여개국에서 다양한 개발환경에서 개발된 789개의 프로젝트 데이터를 갖고 있는 ISBSG(International Softwae Benchmarking Standards Group) Benchmark Release 6[13]을 이용하여 소프트웨어 개발노력에 영향을 미치는 변수를 찾고자 하였다. 표본의 크기가 크므로 프로젝트의 성질이나 크기에 따라서 개발노력이나 팀의 규모 등에 많은 차이가 있으므로 789개의 프로젝트를 세분화하였다. 세분화한 기준은 FP 1개를 개발하는데 소요되는 노력의 규모로 하였다. 이러한 속성은 개발업체의 개발능력 수준에 영향을 받는다고 할 수 있다. 즉 연구에 이용된 데이터는 소프트웨어 기능을 사용자에게 양도하는 비율인 프로젝트 인도율(PDR ; Project Delivery Rate ; Hours/FP)을 개발노력(시간)의 계수로 추정하였다. 즉 프로젝트 인도율이란 것은 프로젝트 개발에 소요된 전체 노력(시간)을 사용자에게 인도된 소프트웨어의 기능 규모(FP)로 나눈 값으로 정의하였다. 이는 기능점수 FP 1개를 개발하는데 소요되는 개발노력의 규모라고 할 수 있으며 본 연구에서는 프로젝트들을 프로젝트 인도율에 따라서 세분화하여 개발노력 추정을 시도하여 보았다. 본 연구에서는 ISBSG의 789개의 프로젝트 개발에 관련된 자료를 이용하여 개발노력을 추정하기 위해서 프로젝트를 유사그룹 별로 나누려고 한다. 유사그룹으로 나누기 위해서 그룹변수선정을 위해 실험하여 본 결과 프로젝트 인도율(PDR)의 크기에 따라서 그룹별 분석을 실시하는 것이 가장 설명력이 높은 회귀식을 찾을 수 있다는 것을 증명하였다. 따라서, 본 논문에서는 789개의 프로젝트 데이터를 PDR의 크기에 따라서 그룹화 시켰으며 PDR의 크기에 따른 분류는 표본의 크기를 고려하였다.

2장에서 개발노력 추정을 위해서 진행된 연구 내용을 간단히 소개하였다. 3장에서는 PDR을 범위에 따라서 세분화 시켜 개발노력에 영향을 미치는 요인들을 찾고 다중회귀 분석을 실시하여 개발노력 추정을 시도하였다. 다중회귀분석 결과 결정계수 값을 이용하여 가장 이상적인 모델을 제안하였다. 4장에서는 기존의 연구와 비교하고 앞으로의 연구 방향을 제시하였다.

2. 관련연구 및 문제점

소프트웨어의 측정 분야에 대한 연구는 30년 이상 계속적으로 수행되어 왔다. 그러나 소프트웨어의 개발 노력과 비용에 영향을 미치는 다양한 요인과 요인간의 속성들이 불명확하여 현재까지 정확한 예측모델을 제시하고 있지 못한 실정이다. 그러나 현재 전자정부구현 등을 비롯한 여러 분야에서 정보화에 대한 욕구가 증가되면서 소프트웨어 개발 사업은 점차 대규모화 되어가고 있으며 이러한 시점에서 소프트웨어를 성공적으로 개발하기 위해서는 개발 초기에 적정한 전문 인력을 투입하는 것이 중요한 과제이다. ISBSG의 보고에 의하면 프로젝트의 개발평균 비용은 초기 예산의 187%에 이르고 기한내에 완료된 프로젝트는 32%라고 발표하였다. 이 보고에 의하여 프로젝트를 시작하는 시점에서 세웠던 예산이나 프로젝트를 완료하는 기한을 제대로 예측하고 있지 못하다는 것을 쉽게 알 수 있다.

소프트웨어의 개발노력을 추정하기 위해서 초기에는 소프트웨어의 규모를 이용한 프로그램의 스텝수(LOC ; Line Of Code)를 이용하였다. LOC 방식으로 개발된 소프트웨어 개발 프로젝트는 주로 3세대 언어로 이루어졌으며, 개발환경으로는 메인프레임용 소프트웨어에 상당히 유효성이 있었으나 최근의 프로젝트들은 4세대 언어로 이루어져 기존의 LOC를 이용한 소프트웨어의 개발노력 추정에는 많은 어려움이 있다.

LOC를 이용한 소프트웨어 개발노력(E : Effort : Man-month)의 추정 모델은 기본적으로 식 (1)의 형태를 취한다.

$$E = a + b \times KLOC^{c}$$
$$E = a \times KLOC^{b} \quad (1)$$

식 (1) 추정식에서 개발노력 E는 Man-Months 또는 Man-Hours로 측정되어지고 KLOC(Thousands of Line Of Codes)는 최종코딩 된 소프트웨어의 라인수이며 a, b, c는 경험적으로 유도된 상수이다. 식 (1)에 제시된 것은 소프트웨어의 규모를 나타내는 LOC를 이용하여 개발노력을 추정하는 대표적인 모델로써 단순회귀분석을 실시한 것이다. 예를 들어 LOC를 이용한 기존의 연구 몇 가지는 아래와 같다.

$$E=5.2\times(KLOC)^{0.91} \quad (Walston-Felix\ model)$$
$$E=5.5+0.73\times(KLOC)^{1.16} \quad (Bailey-Basili\ model)$$
$$E=3.2\times(KLOC)^{1.05} \quad (Boehm\ simple\ model)$$
$$E=3.0\times(KLOC)^{1.12} \quad (Boehm\ average\ model)$$
$$E=2.8\times(KLOC)^{1.20} \quad (Boehm\ complex\ model)$$
$$E=5.288\times(KLOC)^{1.047} \quad \text{for } KLOC>9 \ (Doty\ model)$$

그러나 식 (1)과 같이 LOC를 이용한 소프트웨어 개발노력 추정 방법은 몇 가지의 문제점을 가지고 있으므로 현재는 기능점수 FP를 이용하여 개발노력 E를 추정하는 연구가 진행되고 있다. LOC를 사용하여 소프트웨어 규모를 추정할 경우에 문제점으로는 첫째, LOC에 대한 정확한 정의가 부족하다는 것과 둘째, LOC는 언어에 종속되어 있다는 것과 셋째, 요구분석이나 설계 단계에서 정확한 LOC의 추정이 어렵고 코딩이 완료된 후 추정이 가능하다는 것과 넷째, 소프트웨어의 기능성이나 복잡성을 고려하지 않고 단지 길이만을 고려하여 추정한다는 문제점을 가지고 있다. 기능점수 FP를 이용하여 개발노력 E를 추정하게 되면 첫째, 소프트웨어의 규모측정에 있어 신뢰성이 높다. 즉, 정해진 절차에 의해서 계산되어지므로 산출성이 높고 예측이 아닌 계산에 의해서 이루어지므로 신뢰성이 높아진다는 장점이 있다. 둘째 LOC가 개발초기 단계에서 적용하기 적절하지 않은데 반해서 기능점수 FP는 개발초기 단계에 적용이 가능하고 소프트웨어의 개발기술이나 환경, 언어에 의존하지 않는다는 장점을 가지고 있다. 기능점수 FP를 이용하여 개발노력 E를 추정하는 기본적인 모델은 식 (2)와 같다.

$$E=a+b\times FP \tag{2}$$

식 (2)는 독립변수로 기능점수 FP의 규모를 이용하여 종속변수인 개발노력 E를 추정하는 단순선형회귀식으로써 이때 a, b는 최소자승추정법(LSE : Least Square Estimates)에 의해서 추정되어져야 할 상수이다.

이것은 개발노력 E가 기능점수 FP와 단순선형회귀관계를 이루고 있다는 가정으로 세워진 모델이다. Albrecht et al.[1]은 IBM Data Processing Services에서 개발한 24개의 응용프로그램에 대하여 개발노력 E를 추정하기 위하여 식 (3)의 모델을 제시하였다.

$$E=-13.39+0.0545\times FP \tag{3}$$

그리고 Matson et al.[8]은 개발노력 E를 추정하기 위하여 기능점수 FP와의 관계를 비선형 회귀관계로 제곱근의 관계인 식 (4)를 제안하였다.

$$\sqrt{E}=1+0.00468\times FP \tag{4}$$

식 (3)과 식 (4)에 제시되어 있는 기능점수와 개발노력과의 관계는 식 (3)에서는 선형관계로 설명되어져 있고 식 (4)에서는 비선형의 관계로 설명되어 있다. 개발노력과 기능점수 간에 회귀관계를 밝히기 위해서 많은 연구가 진행되어졌으나 일반적으로 Albrecht et al.[1]이 제안한 식 (3)의 단순회귀모델에서 선형관계를 만족하지 않는 경우 선형관계로 회귀식을 변형하기 위해서 Box-Cox가 제안한 사다리꼴 재표현에 의한 방법을 선택한다. 식 (3)의 선형회귀관계에서 선형을 만족하지 못하는 경우에 식 (4)와 같이 제곱근 변환을 시도해서 선형관계로 자료를 변형하게 된다. 그러나 이것은 측정되어지는 실험데이터의 경향에 따라서 다르게 나타날 수 있으므로 주어지는 실험데이터에 맞게 회귀식의 유의성 검정을 통한 연구가 진행되어져야 할 것이다.

개발노력의 추정을 위한 개발노력과 기능점수에 대한 회귀식에서 적합성에 대한 검정을 위해서 먼저 각 모델의 제안식에 대한 잔차의 검정을 실시하여야 한다. 회귀식의 적합성에 대한 검정 방법으로는 첫째 분산분석(ANOVA : Analyis of Variance) 테이블을 이용하여 검정통계량을 이용하여 회귀식의 유의성을 검정할 수 있으며 둘째, 결정계수(Coefficient of determination, R^2)를 이용하는 방법이 있으나 결정계수의 경우는 정확한 평가의 기준을 설정할 수 없으므로 회귀식의 유의성 검정에 있어서 다소 주관적인 평가를 하게 된다. 그러나 본 논문에서는 회귀식의 유의성 검정을 분산분석 테이블로 확인하고 유의성이 있는 회귀식에 대하여 결정계수의 결과치를 이용하여 비교 연구하였다. 또한 회귀식의 유의성 검정에는 데이터에 대한 회귀식에서 잔차의 독립성, 등분산성, 정규성과 독립변수간의 독립성이라는 기본성질이 만족되어야 하며 이러한 가정에 대한 모든 조건이 만족되어야 주어진 데이터에 대한 회귀식이 유의하다는 결론을 내릴 수 있다

Kemerer[4]은 개발노력 E를 추정하기 위해서 기능점수 FP를 이용하여 15개의 소프트웨어 프로젝트에 대한 실험데이터를 이용하여 단순회귀모델로써 선형관계와 곡선관계식 (5)의 회귀식을 제안하였다.

$$\begin{aligned} E&=-121.57+0.3411\times FP \\ E&=-69.13+0.723\times FP-8.054\times10^{-4}\times FP^2 \\ &\quad +3.073\times10^{-7}\times FP^3 \\ E&=60.62+7.728\times10^{-8}\times FP^3 \end{aligned} \tag{5}$$

그러나 식 (5)에 제시된 모델들은 대체적으로 표본의 수가 작고 특정업체에 대한 데이터를 이용해서 얻어진 결과이므로 다양한 프로젝트에 대해서 획일적으로 적용하기에는 문제가 있다. Matson et al.[2] 은 대형업체로부터 획득된 104개의 프로젝트를 이용해서 개발노력 E와 기능점수 FP사이에 식 (6)의 회귀모델을 제안하였다.

$$\begin{aligned} E&=585.7+15.12\times FP \\ E&=2.51+1.00\times\ln(FP) \end{aligned} \tag{6}$$

제안된 식 (6)에서도 개발노력 E를 추정하기 위하여 기능점수 FP를 이용한 단순회귀관계로 설명하였으나 두 번째 수식에서는 기능점수 FP를 로그변환하여 단순회귀관계를 추정하였다. 본 연구도 데이터의 경향성을 먼저 파악하여 개발노력 E와 기능점수 FP사이에 선형회귀관계를 만족하지 못하고 두 개의 관계가 볼록관계로 형성되어졌을 때 종속변수와 독립변수사이에 단순선형회귀관계의 모델을 설정하기 위해서 설명변수에 로그변환을 시도하게 되는 것이다. 그러나 이러한 변환도 Box-Cox의 사다리꼴 재표현의 공식에 의하여 변환을 시도한다.

실험데이터가 얻어지면 그 자료를 중심으로 하여 플롯을 그려보고 두 변수의 관계를 파악하여 직선관계의 회귀식을 만들기 위해서는 변형을 시도하면 된다. 즉 개발노력 E의 추정이란 것은 식 (3)~식 (6)에서 제시한 것처럼 개발된 소프트웨어의 주어진 데이터에 따라서 획일적인 모델로 적용할 수는 없다.

그래서 본 논문에서는 이러한 개발노력과 기능점수사이에 획일적인 모델을 제안하기 보다는 개발노력에 영향을 미치는 요인을 찾아내고 이 요인에 대한 분석을 실시하는 것이 개발노력의 추정에 좀더 정확한 접근을 할 수 있으므로 3장에서는 PDR을 구간별로 나누어서 각 구간에 따라 개발노력에 영향을 미치는 변수를 찾고 변수들을 이용한 다중회귀분석을 실시하고 각 경우에 대한 결정계수값을 비교하여 보았다. 그리고 기본적인 회귀식의 관계에 대한 유의성 검정을 위해서 분산분석 테이블을 보고 잔차에 대한 검정을 실시하였다.

3. 소프트웨어 개발노력에 영향을 주는 요인

본 논문에서는 1990년대에 20여개국에서 다양한 개발환경에서 개발된 789개의 프로젝트 데이터를 갖고 있는 ISBSG Benchmark Release 6[13]을 이용하여 소프트웨어 개발노력에 영향을 미치는 변수를 찾고자 하였다. 789개의 프로젝트 데이터를 이용해서 각 변수별 상관계수를 구하여 보았다. 상관분석 결과 소프트웨어의 개발노력과 관계가 있는 변수들로 FP, 개발기간(D ; Duration ; Months), 팀규모(TS ; Team size), 팀레벨(L ; Level), 프로젝트 인도율(PDR)이 선택되었으며 선택되어진 변수를 이용하여 회귀분석을 실시하여 보았다.

조사 방식은 먼저 전체의 데이터를 이용하여 개발노력 E에 영향을 미치는 변수를 조사하였다. 전체의 데이터를 가지고 회귀분석을 실시하지 않고 789개의 프로젝트는 규모에 있어서 상당한 차이가 있으므로 이것을 프로젝트 인도율에 따라서 유사한 크기로 그룹화 하여 개발노력 E에 대한 선형회귀식을 구해서 가장 영향력을 미치는 변수를 찾아보았다.

FP를 중심으로 789개의 데이터를 몇 개의 그룹으로 나누어 회귀분석을 실시하여 보았으나 회귀식의 설명력이 높지 못하였다. 이러한 시도를 다른 변인들에 대하여서도 실시하여 보았으나 회귀식의 설명력이 그리 높지 않게 추정되었다. 본 논문에서는 여러번의 실험을 거쳐서 개발노력에 대한 회귀식을 세우는데 있어서 PDR에 따라서 그룹을 나누는 것이 독립변수와 종속변수 사이에 설명력을 가장 높게 나타낼 수 있는 이상적인 결과가 얻어짐을 밝혔다.

그리하여 개발노력은 PDR에 의해서 상대적으로 많은 변화가 있다는 것을 감안하여 PDR 크기에 따라서 구간을 나누어 개발노력 E와 5개 변인 사이에 회귀분석을 실시하여 보았으며 회귀식의 유의성은 결정계수 (R^2)로 판단하였다.

각 구간별로 변수선택은 회귀분석의 변수 선택법 중에서 단계적선택법(Stepwise Method)을 이용하였으며 선택의 기준은 유의수준(Significance Level) $\alpha = 0.05$를 기준으로 하였다.

PDR의 구간에 따라서 선택되어지는 변인은 거의 동일하였으며 각각의 구간별 결정계수를 기준으로 하여 기존의 연구와 비교 검정을 실시하였다.

전체 789개의 데이터를 이용하여 회귀식을 세워 보면 회귀분석의 설명력을 나타내는 결정계수의 값이 상당히 낮음을 알 수 있으므로 전체의 데이터를 설명하기에는 데이터를 특성별로 나누어 개발노력을 추정하는 것이 이상적이라는 것을 <표 1>을 보고 검정할 수 있다.

〈표 1〉 PDR에 따른 다중회귀분석결과

PDR	선택되어지는 변수	결정계수(R^2)
0<PDR≤1	E=-246.889+0.761FP-13.278D+425.010PDR	R^2=0.916
1<PDR≤2	E=115.061+1.373FP-13.12D	R^2=0.968
2<PDR≤3	E=24.488+2.362FP	R^2=0.985
3<PDR≤4	E=329+3.714FP-145.024L-54.771D	R^2=0.999
4<PDR≤5	E=36.03+4.314FP	R^2=0.994
5<PDR≤6	E=-76.832+5.698FP	R^2=0.998
6<PDR≤7	E=-129.829+6.86FP	R^2=0.999
7<PDR≤8	E=-161.892+7.174FP+29.508D	R^2=0.999
8<PDR≤9	E=57.102+8.184FP	R^2=0.999
9<PDR≤10	E=-62.277+9.656FP	R^2=0.999
10<PDR≤12	E=3.858+10.207FP+36.361TS	R^2=0.999
12<PDR≤14	E=423.797+13.709FP-80.679D	R^2=0.999
14<PDR≤16	E=231.425+14.632FP	R^2=0.999
16<PDR≤18	E=131.434+16.302FP	R^2=0.999
18<PDR≤20	E=371.584+19FP-59.442TS	R^2=0.999
20<PDR≤30	E=-1259.211+21.159FP+1268.518L	R^2=0.996
30<PDR≤40	E=432.322+32.462FP	R^2=0.994
40<PDR	E=-1917.659+1814.280D	R^2=0.963
전체	E=-6350.802+6.13FP+370PDR+232D+54TS+950L	R^2=0.708

*E : Effort=Man-Months, D : Duration=Months

<표 1>은 PDR에 따라서 각 그룹별로 실험데이터를 나누어 회귀분석을 실시한 결과이며 PDR의 값에 따라 그룹을 나눈 기준은 따로 두지 않고 표본의 수를 고려하여 표본의 수가 고르게 분배되도록 기준을 삼은 것이다.

<표 1>에서 모든 데이터를 사용하여 회귀분석을 실시한 결과 제안한 독립변수 5개를 이용하여 회귀분석을 실시한 결정계수의 값이 0.708의 설명력을 가지고 있는 것으로 실험되었다. 본 연구에서 실험데이터를 조건 없이 전체를 고려하여 개

발노력을 추정하기 위한 회귀식 보다는 PDR에 따라 나누어진 자료를 이용한 분석이 개발노력 추정에 훨씬 유의하다는 것을 쉽게 파악할 수 있다. <표 1>의 결과에서 PDR>40인 구간을 제외하고는 거의 전 구간에서 개발노력에 대한 설명력이 있는 변수로 기능점수 FP가 선택되어졌다.

즉 기존의 연구를 활용하여 FP를 이용하여 개발노력을 추정할 경우에도 프로젝트의 크기를 고려하지 않고 전체를 추정하기 보다는 프로젝트 인도율 등을 나누어서 평가의 기준을 설정 하면 훨씬 정확한 예측치를 구해낼 수 있을 것이다. Albrecht et al.[1], Matson et al.[8], Kermerer[4] 등은 각각의 업체에서 소수의 데이터를 이용하여 기능점수만을 이용해 개발노력을 추정하는 단순회귀식을 모델로 제시하였다. 그러나 이러한 관계에서 이상운 et al.[14]는 개발노력과 기능점수 FP 사이에는 단순회귀관계에서 다양한 지수회귀, 곡선회귀, 로그회귀, 누승회귀 등에 대한 관계를 세워서 모델을 설정하면 회귀분석의 설명력인 결정계수의 값이 높아짐을 증명하고 평가비교를 통하여 가장 우수한 모델을 제시하였다.

이상운 et al.[14]의 연구에서도 전체의 데이터를 이용하여 개발노력과 기능점수와의 관계를 구한 것이 아니라 PDR을 구간별로 나누어서 단순회귀, 곡선회귀, 누승회귀, 지수회귀, 로그회귀의 관계를 구하고 가장 결정계수값이 높은 회귀식을 선택하였다.

<표 2>는 이상운 et al.[14]가 제안한 기능점수를 이용한 개발노력의 단순회귀직선에서 설명력이 가장 우수한 모델의 제시값과 본 연구에서 제안한 다중회귀분석에서의 설명력인 결정계수를 비교한 테이블이다. 본 연구의 실험결과 단순회귀식에 의한 것보다 다중회귀분석을 실시하면 회귀식의 설명력인 결정계수의 값이 많이 상승되어짐을 확인할 수 있다.

<표 2>에서 단순회귀분석은 종속변수에는 개발노력을, 설명변수에는 기능점수 FP만을 이용한 결과이고 다중회귀분석은 앞에서 제시한 5개의 변수를 독립변수로 하여 단계적선택법(Stepwise Method)에 의해서 변수를 선택하는 다중회귀분석을 실시한 결과이다. <표 1>에서 PDR의 값에 따른 그룹별 종속변수인 개발노력을 설명하기 위한 설명변수의 선택사항은 <표 1>에 제시한 바 있다. <표 2>에서 단순회귀는 기능점수와 개발노력 사이에 설명력이 가장 우수한 관계에 있는 모델을 찾아내기 위하여 직선회귀, 곡선회귀, 누승회귀, 지수회귀, 로그회귀를 실시하여 보고 실시된 결과 중 가장 설명력이 우수한 모델을 제시하였다. 즉 기존에 연구된 결과에서 각 구간별로 가장 설명력이 높은 회귀관계의 모델에 대한 결정계수를 제시하였으며 <표 2>에서 ( )은 그때에 선택된 모델이다. 테이블에 나타난 결과를 살펴보면 전 구간에서 단순회귀를 통한 여러 가지 변형모형을 선택하는 회귀식 보다는 다중회귀를 통한 회귀관계가 개발노력을 설명하는 설명력이 높다는 것을 쉽게 알 수 있다.

단순회귀의 경우는 구간별로 개발노력과 기능점수 사이에 곡선이나 누승관계의 회귀식이 가장 높은 설명력을 가지고 있음을 알 수 있다.

〈표 2〉 회귀분석에서 결정계수의 비교

PDR	단순회귀분석	다중회귀분석
0<PDR≤1	0.721(누승)	0.856
1<PDR≤2	0.954(누승)	0.968
2<PDR≤3	0.979(누승)	0.985
3<PDR≤4	0.997(곡선)	0.999
4<PDR<5	0.997(누승)	0.999
5<PDR≤6	0.998(누승)	0.998
6<PDR≤7	0.999(곡선)	0.999
7<PDR≤8	0.999(곡선)	0.999
8<PDR<9	0.999(곡선)	0.999
9<PDR≤10	0.999(누승)	0.999
10<PDR<12	0.997(누승)	0.999
12<PDR≤14	0.999(곡선)	0.999
14<PDR≤16	0.999(곡선)	0.999
16<PDR≤18	0.998(누승)	0.998
18<PDR≤20	0.999(곡선)	0.999
20<PDR≤30	0.994(누승)	0.996
30<PDR≤40	0.994(누승)	0.994
40<PDR	0.918(누승)	0.963

위의 실험결과 개발노력에 여러 가지 변수를 예측하기에 가장 적합한 것은 PDR을 구간별로 나누어서 다중회귀분석을 실시하는 것이라는 것을 확인하였다.

또한 각 구간별 다중회귀를 실시하여 회귀식에 대한 유의성 검정을 한 결과 전구간에서 회귀식이 유의함을 확인할 수 있었으며 회귀식에 대한 기본 가정을 만족함을 확인하였다. 본 연구에서 기존의 연구가 기능점수 FP만을 이용하여 개발노력을 추정하기 위한 회귀식을 제안하였으나 <표 1>에서 제시한 것처럼 PDR의 구간값에 따라서 기능점수가 개발노력을 추정하기 위한 설명변수로 선택하지 않는 경우도 존재한다는 것이다. PDR의 구간에 따른 변화를 고려한다면 개발노력을 추정하기 위한 변수로 FP 만을 선택하고 전 데이터에 적용하는 것은 개발노력 추정에 있어서 오류를 범할 수 있다는 것이다. 본 연구에서 PDR의 구간별 회귀식의 유의성 검정을 모두 실시하여 보았으며 잔차에 대한 분석도 실시하였다. 또한 본 연구에 이용된 데이터는 실험자료의 수가 많기는 하나 각각의 문항에 결측치가 많아서 이 결측치에 대한 영향을 회귀식의 유의성 검정 방법과 함께 확인하기 위하여 결측치를 조사된 자료의 평균값으로 대체하여 회귀분석을 실시하였다. 이와 같이 평균값으로 대체한 경우 결정계수 값은 다소 상승되어 나타났으며 변수선택에 있어서 선택되어지는 변수에도 다소 차이가 있음을 확인할 수 있었다. 실험에서 PDR의 구간을 나누어서 개발 노력을 추정한 것은 개발 노력에 영향을 미치는 5개의 변수 모두를 구간별로 나누어서 실험하여 본 결과 PDR의 값에 따른 개발 노력 추정의 변화가 가장 크게 나타나서 PDR에 따른 구간별 변화추이를 연구하게 된 것이다.

단순회귀 관계와 비교하여 다중회귀 관계를 이용하여 개발노력에 대한 회귀분석을 실시한 결과 결정계수의 값에 가장 큰 상승을 나타내는 그룹은 PDR의 값이 $0<PDR\leq 1$ 사이였다. 회귀식의 유의성 검정에 대한 방법론을 제시하기 위하여 PDR이 $0<PDR\leq 1$인 구간에서 데이터를 이용하여 회귀모델을 세웠다. 먼저 본 실험결과 결정계수 $R^2=0.923$이고 본 실험에 사용된 데이터의 결측치는 평균값으로 대체하여 사용하였다.

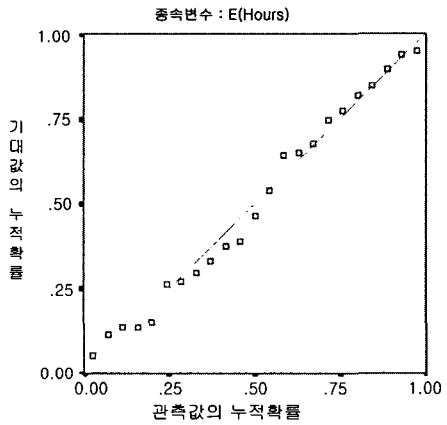
$$E=-153.113+0.53FP+669PDR-143.813L \quad (10)$$

식 (10)에서 제시된 다중회귀분석모델은 기존의 연구와 비교시 선택되어지는 변수에 있어서도 차이가 있음을 알 수 있다. 이 경우에 분산분석의 결과는 <표 3>과 같다.

〈표 3〉 PDR이 0<PDR≤1에서 데이터에 대한 분산분석테이블

변수	제곱합	자유도	F	P-value
회귀	4919074.2	3	36.634	0.000
잔차	850427.12	19		
합계	5769501.3	22		

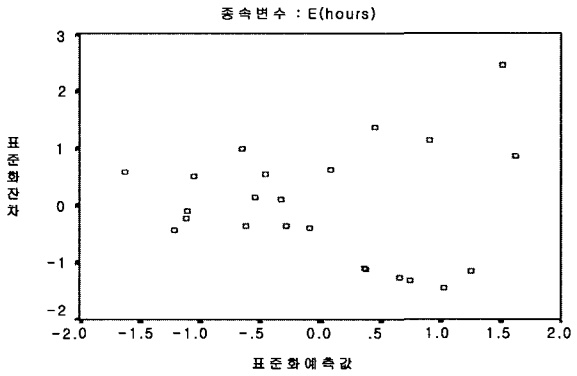
다음은 식 (10)에서 제시한 개발노력 추정을 위한 다중회귀 모델의 유의성을 확인하기 위하여 잔차 분석을 실시하여 보았다. 먼저 잔차의 정규성을 확인하기 위하여 잔차에 대한 정규확률 플롯(Normal Probability Plot)을 그려본 결과는 (그림 1)과 같다. (그림 1)의 정규 확률 플롯을 살펴보면 관측 값과 예측 값 사이에 거의 직선관계를 나타내고 있어 정규성의 성질을 만족함을 확인할 수 있다.



(그림 1) 정규확률플롯

다음은 잔차의 독립성과 등분산성을 확인하기 위하여 표준화된 잔차와 표준화된 예측 값 사이의 산점도가 (그림 2)에 제시되어 있다.

(그림 2)의 결과를 살펴보면 본 회귀식은 잔차의 등분산성과 독립성이 만족함을 확인할 수 있다. 개발노력 추정을 위한 다중회귀모델인 식 (10)의 회귀식의 기본 가정은 모두 만족됨을 확인하였다. 위의 실험결과를 통해서 개발노력의 추정은 기능점수 FP만을 정해놓고 추정하기 보다는 PDR의 값을 그룹으로 나누어 다중회귀 분석을 실시하는 것이 설명력이 높은 회귀식을 산출할 수 있다는 것을 증명하였다. 또한 PDR의 그룹별 값에 따라서 개발노력을 추정하는 것이 소프트웨어의 개발 규모라든지 기능점수에 따른 규모를 고려하여 추정하는 것보다 훨씬 정확한 회귀모델을 찾을 수 있었다. 그래서 본 논문에서는 789개의 개발 프로젝트 실험데이터에서 제시한 소프트웨어 프로젝트 팀 규모의 평균값과 중앙값을 조사해서 소프트웨어 개발 프로젝트에 참여하는 평균적인 팀규모를 확인하였다.



(그림 2) 잔차분석을 위한 플롯

자료를 분석하는데 있어서 중심적 경향에 대한 분석의 대표값으로 평균값과 중앙값을 이용한다. 그러나 탐색적 자료분석의 경우는 평균값의 경우 이상치에 영향을 많이 받게 되므로 이상치에 영향을 받지 않는 중앙값을 중심경향을 설명하는 자료의 대표값으로 이용한다. 본 연구에서 개발팀의 규모에 따른 변화를 확인하기 위하여 PDR의 구간값에 따라서 실험에 사용된 데이터의 평균과 중앙값을 모두 제시하였다. 또한 각 PDR의 구간별 이상치를 조사하여 본 결과 이상치가 존재하는 구간에서는 평균값이 기본 예상치보다 상당히 높게 제시되어 팀규모에 대한 분석은 탐색적자료분석의 측면에서 제시한 중앙값을 중심으로 분석하는 것이 정확한 예측을 위하여 필요하다는 것을 알 수 있었다. 평균값과 중앙값의 분석결과는 <표 4>에 제시되어 있다.

팀의 규모에 관한 연구는 업무를 수행하는데 있어서 작은팀으로 구성된 것이 큰팀으로 구성된 프로젝트보다 효율성이 높다는 것이 일반적으로 증명된 사실이다. 또한 최소한의 실질적인 팀 규모를 구성하는데 목표를 두고 있는데 이상적인 팀 규모를 결정한다는 것은 다른 외부변수와 여러 가지 환경적인 변수를 고려하여야 할 것이다.

<표 4>에서 PDR 구간별 팀 규모에 대한 값을 추정하여 보면 일반적으로 2~4명 정도가 적합하다는 것을 알 수 있으며 PDR이 커짐에 따라서 팀의 규모도 커지고는 있으나 PDR의 변화에 크게 변화를 보이고 있지는 않다. 위에서 제시한 PDR의 구간별 실험데이터의 수는 크지 않으며 IBM Data Processing Service에서 개발된 응용프로그램을 이용하여 연구한 Albrecht

et al.[1]의 연구도 24개의 데이터를 가지고 단순회귀관계를 연구하였고 Kemerer[4]은 ABC 회사에서 개발된 15개의 소프트웨어 프로젝트에 대해서 단순회귀직선과 곡선회귀 모델을 제시하였다. 기존의 연구에서 제시된 표본의 크기는 모두 그리 크지 않다.

<표 4>에서 제시한 것처럼 팀 규모에 대한 평균값은 상당한 차이를 보이고 있으며 특히 PDR이 14와 16사의에서 팀 규모에 대한 평균값이 35.26으로 조사되어 이 구간에서 발생된 이상치로 인한 영향이 크게 작용하므로 표본의 수가 적은 실험에서의 대표값으로는 평균보다는 중앙값이 합당하다고 보여진다. 팀규모에 대한 연구도 프로젝트의 성격이나 다른 환경적인 변수에 의해서 많은 변화가 있을 것이다. 이러한 변수들을 고려하여 팀 규모를 추정하기 위한 가장 이상적인 모델을 제시하기 위해서는 계속적인 연구가 진행되어져야 할 것이다.

〈표 4〉 PDR에 따른 팀의 규모에 대한 대표값

PDR	Mean	Median	Sample Size
0<PDR<1	2.92	2	13
1<PDR≤2	2.73	3	26
2<PDR<3	3.29	2.5	24
3<PDR<4	4.50	4	32
4<PDR<5	6.92	4	28
5<PDR≤6	4.88	4	25
6<PDR<7	6.20	5	20
7<PDR<8	5.58	6	17
8<PDR<9	7.86	6	15
9<PDR≤10	7.55	6	18
10<PDR<12	7.58	5	17
12<PDR≤14	8.91	7	24
14<PDR<16	35.26	7	19
16<PDR<18	7.09	5	11
18<PDR<20	7.92	8	13
20<PDR<30	10.19	6	21
30<PDR<40	8.22	8	9
40<PDR	17.83	18.5	6

4. 결론 및 향후 연구과제

기존에 개발노력을 추정하기 위하여 기능점수 FP만을 이용하여 회귀식을 제시하였다. 일차적으로 기능점수 FP를 이용하여 단순회귀관계를 구하여 보고 곡선회귀에서 누승회귀, 지수회귀, 로그회귀 등의 관계를 구하여 가장 설명력이 우수한 관계의 모델을 찾는 방법을 취하였다. 본 논문에서는 개발노력을 추정하기 위하여 먼저 개발노력을 가장 잘 설명해 주는 독립변수를 찾아서 상관관계를 구하여 보고 상관관계가 높은 5개의 변수를 독립변수로 선택하였다.

선택된 5개의 변수를 중심으로 PDR의 크기에 따라 데이터를 그룹화하고 각 구간별 다중회귀분석을 실시하였다. 이때의 변수선택은 단계선택법(Stepwise Method)을 이용하였으며 선택되어지는 변수를 통해서 비교를 실시하였다.

각 구간별 다중회귀분석을 통해서 추정된 회귀식의 결정계수를 확인하였다. 본 논문의 실험결과 전구간에서 기존에 제시된 단순회귀관계보다 다중회귀분석을 실시하면 회귀식의 설명력인 결정계수가 좋아진다는 것을 확인하였다. 기존의 연구에서 제시한 단순회귀관계의 여러 변형 모델보다도 다중회귀를 이용하여 개발노력을 추정하면 훨씬 설명력이 우수하다는 것을 증명하였다. 단순회귀보다는 다중회귀방법을 이용하게 되면 설명력은 높아지나 본 연구에서 PDR의 그룹별로 선택되어지는 변수에도 다소의 차이가 있다는 것이다. 특히 PDR의 구간에 따라서 개발노력을 추정하기 위하여 기존에 독립변인으로 제시된 기능점수가 선택되지 않는 구간도 있다는 것이다. 개발노력을 추정하기 위한 환경적인 변수에 대한 분석은 앞으로 더 연구되어져야 할 사항이다. 또한 본 연구에서는 <표 4>에 가장 이상적인 팀 규모를 제시하기 위하여 789개 프로젝트를 PDR의 값에 따라서 평균과 중앙값을 조사하였다. 평균의 경우는 각 구간별 팀 규모에 상당한 차이가 있다는 것을 쉽게 알 수 있으나 중앙값의 경우는 거의 큰 변동이 없음을 확인할 수 있었다. 통계적인 측면에서 자료에 대한 대표값으로 가장 많이 활용되고 있는 것이 평균값이기는 하나 평균은 이상치에 의한 영향을 쉽게 받을 수 있으므로 팀 규모에 대한 대표값으로는 중앙값이 이상적임을 확인하였다. 특히 본 연구에서 사용된 데이터는 PDR을 구간별로 나누어 분석한 결과 표본의 수가 그리 많지 않으므로 팀의 규모에 대한 예측은 평균보다 중앙값이 이상적인 결과를 줄 것이다.

본 연구를 통해서 앞으로의 연구과제는 개발노력에 영향을 미치는 변수를 찾아서 개발노력 추정에 다중회귀식을 이용하면 훨씬 설명력이 높아지므로 개발노력에 가장 큰 영향을 미치는 변수를 찾아내야 할 것이다. 또한 실험자료를 이용하여 팀의 규모를 추정하는데 있어서 중앙값과 평균값과의 관계를 더욱 상세히 밝혀내어 대표값으로서 적당한 통계치를 찾아내는 연구도 선행되어야 할 것이다. 모델을 연구하기 위해서 사용되는 데이터의 수가 모두 그리 많지 않으므로 소수의 표본에 대하여 가장 오류를 적게 하는 방법에 대한 연구도 선행되어야 할 것이다. 탐색적인 방법에서 자료를 접근하는 것으로 탐색적 자료분석이라는 통계적인 척도를 이용하여 개발노력을 추정하는 것도 앞으로 더 연구되어야 할 과제이다. 또한 개발노력을 추정하는데 있어서 제안된 회귀식을 세우는데 결측치에 대한 문제도 고려되어야 할 사항이다. 앞에서 제시한 것처럼 회귀분석 시에 결측치를 제거하지 않고 평균값으로 대체하여 사용하는 경우에 회귀식에 대한 설명력이 높아지면서 선택되는 변수에도 변화가 생겼음을 확인하였다. 그러나 소수의 자료를 분석하는데는 대표값으로 평균보다는 중앙값이 좋다는 것을 확인하였으므로 결측치에 대한 대체값으로 회귀분석을 실시할 경우 평균과 중앙값을 나누어서 결과에 대한 차이를 연구하여야 할 것이다.

참 고 문 헌

[1] A. J. Albercht, "Measuring Aplications Development Productivity," Proceeding of IBM Application Dev., Joint SHARE/ GUIDE Symposium, Monterey, CA., pp.83-92, 1979.

[2] A. J. Albrecht and J. E. Gaffney, "Software Function, Source Line Of Code and Development Effort Prediction : A Software Science Validation," IEEE Trans. on Software Eng., Vol.SE-9, No.6, pp.639-648, 1983.

[3] B. W. Boehm and P. N. Papaccio, "Understanding and controlling software cost," IEEE Trans. Software Eng., Vol.14, pp.1462-1477, 1988.

[4] C. F. Keremer, "An Empirical Validation of Software Cost Estimation Models," Communications of ACM, Vol.30, No.5, pp.416-429, 1987.

[5] D. Meyerhoff, B. Laibarra, R. V. D. Pouw Kraan, and A. Wallet, Software Quality and Software Testing in Internet Times, Springer, 2002.

[6] G. C. Low and D. R. Jeffery, "Function Point in the Estimation of the Software Process," IEEE Trans on Software Eng., Vol.16, pp.64-71, 1990.

[7] I. Jacobson, M. Christerson et al., "Object-oriented Software Engineering. A Use Case Driven Approach," Addison-Wesley, 1992.

[8] J. E. Matson, B. E. Barrett and J. M. Mellichamp, "Software Development Cost Estimation Using Function Points," IEEE Trans. on Software Eng., Vol.20, No.4, pp.275-287, 1994.

[9] K. Johnson, "Software Cost Estimation : Metrics and Models," Department of Computer Science University of Calgary, Albreta, Canada, http://sern.ucalgary.ca/courses/seng/621/W98/Johnsonk/cost.htm, 1998.

[10] L. A. Laranjeira, "Software size estimation of object-oriented systems." IEEE Trans. Software Eng., Vol.67, pp. 10-18, 1990.

[11] P. F. Velleman, "Definition and Comparison of Robust Nonlinear Data Smoothing Algorithms," American Statistical Association, Vol.75, pp.609-715, 1980.

[12] P. F. Velleman and D. C. Hoaglin, 'ABC of EDA,' Duxbury Press, 1981.

[13] ISBSG, "Worldwide Software Development -The Benchmark Release 6," Victoria, Australia International Software Benchmarking Standards Group, 2000.

[14] 이상운, 노명옥, 이부권, "프로젝트 인도율 그룹 분할 방법을 이용한 소프트웨어 개발노력 추정", 정보처리학회논문집, 제9권 제2호, pp.259-266, 2002.

[15] 이상운, "신경망을 이용한 소프트웨어 개발노력 추정", 정보처리학회논문집, 제8권 제3호, pp.241-246, 2001.

[16] 허명회, 문승호, '탐색적자료분석(EDA)', 자유아카데미, 2000.

정 혜 정

e-mail : jhjung@ptuniv.ac.kr

1988년 경북대학교 통계학과(이학사)

1991년 경북대학교 통계학과 대학원 (이학석사)

1994년 경북대학교 통계학과 대학원 (이학박사)

1995년~현재 평택대학교 정보통계학과 부교수

관심분야 : 소프트웨어공학, 소프트웨어 모형결정, 소프트웨어 품질특성, 소프트웨어 신뢰도 측정, 영상처리

양 해 술

e-mail : hsyang@office.hoseo.ac.kr

1975년 홍익대학교 전기공학과(학사)

1978년 성균관대학교 정보처리학과(석사)

1991년 日本 오사카대학교 정보공학과 S/W공학 전공(공학박사)

1999년~현재 호서대학교 벤처전문대학원 교수

2001년~현재 한국정보처리학회 부회장

2003년~현재 미국 ACIS 학회 Vice President

관심분야 : 소프트웨어공학(특히, S/W 품질보증과 품질평가, 품질감리와 컨설팅, OOA/OOD/OOP, CASE, SI), 컴퍼넌트관리, CBD기반기술, IT품질경영

신 석 규

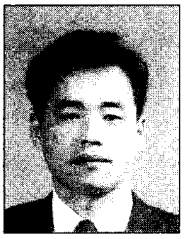
e-mail : skshin@tta.or.kr

1991년 서울산업대학교 재료공학과(학사)

1999년 충남대학교 대학원 전산학과 (석사과정수료)

2001년~현재 한국정보통신기술협회 IT 시험연구소 소프트웨어 시험센터 센터장

관심분야 : 소프트웨어공학(특히 S/W 품질시험과 품질평가, 품질감리와 BMT)



이 상 운

e-mail : sulee@sky.wonju.ac.kr

1983년 한국항공대학교 항공전자공학과(학사)

1995년 경상대학교 컴퓨터과학과(석사)

1998년 경상대학교 컴퓨터과학과(박사)

1992년~국방품질관리소 항공전자장비 및 소프트웨어 품질보증 담당

2003년 도립 강원전문대학 컴퓨터응용과 전임강사

2004년~현재 국립 원주대학 여성교양과 전임강사

관심분야 : 프로젝트 관리, 소프트웨어 개발 방법론, 소프트웨어 Metrics, 소프트웨어 시험, 소프트웨어 신뢰성, 아키텍쳐, 신경망, 뉴로-퍼지