UNIVERSITY OF SALERNO

GENETIC PROGRAMMING FOR SOFTWARE DEVELOPMENT EFFORT ESTIMATION

Federica Sarro
fsarro@unisa.it

25 July 2011

14th CREST Open Workshop : *Genetic Programming for Software Engineering*

Outline

- Background and Motivations
  - Software Development Effort Estimation
  - Effort Estimation with Search-Based Approaches
  - How to Assess Estimation Model Accuracy?
- Empirical Study: Influence of Fitness Function
  - Research Goals
  - GP Setting, Dataset Selection, Validation Method, and Evaluation Criteria
  - Results
- Preliminary Empirical Study: Multi-Objective Genetic Programming
  - Research Goals
  - MOGP Setting, Dataset Selection, Validation Method, and Evaluation Criteria
  - Results
- Conclusions

Outline

- Background and Motivations
  - Software Development Effort Estimation
  - Effort Estimation with Search-Based Approaches
  - How to Assess Estimation Model Accuracy?
- Empirical Study: Influence of Fitness Function
  - Research Goals
  - GP Setting, Dataset Selection, Validation Method, and Evaluation Criteria
  - Results
- Preliminary Empirical Study: Multi-Objective Genetic Programming
  - Research Goals
  - MOGP Setting, Dataset Selection, Validation Method, and Evaluation Criteria
  - Results
- Conclusions

Software Development Effort Estimation

- Software development effort estimation is meant to predict the human effort needed to realize a software project
  - effort usually quantified as person-hours or person-months
- Obtaining accurate estimates is a critical activity
  - for planning and monitoring software project development
  - for delivering the product on time and within budget
- Significant over or under-estimates expose a software project to several risks
  - addition of manpower to a late software project makes the project later (Brooks's Law)
  - cancellation of activities, such as documentation and testing, impacts on software quality and maintainability

Software Development Effort Estimation

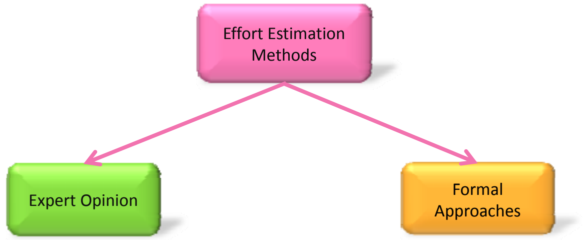
- Obtaining accurate estimates is a challenging activity
  - the estimation is needed early in the software lifecycle, when little information about the project is available

I NEED A BUDGET ESTIMATE FOR MY PROJECT, BUT I DON'T HAVE A SCOPE OR A DESIGN FOR IT YET.
Dilbert.com DilbertCartoonist@gmail.com
OKAY, MY ESTIMATE IS $3,583,729.
12-7-09 ©2009 Scott Adams, Inc./Dist. by UFS, Inc.
YOU DON'T KNOW ANYTHING ABOUT MY PROJECT.
THAT MAKES TWO OF US.

from http://www.dilbert.com/

- Several approaches have been proposed to support project managers in estimating software development effort

Approaches for estimating software development effort



Produce an estimate relying on the expertise of the company's practitioners

- low repeatability
- difficult to quantify and to determine those attributes that have been used to derive the estimate

Produce an estimate (or an estimation model) applying algorithms to factors that influence the development effort

- provide the estimation in an objective way
- classified in generic model based (e.g., COCOMO, COCOMOII) and specific (e.g. LR, MSWR, CART, CBR, NN) model based methods

Data-Driven Approaches

- A Data-Driven approach exploits data from past projects to estimate the effort for a new project
  - data consist of information about some relevant project features (i.e., cost drivers) and the effort actually spent to develop the projects

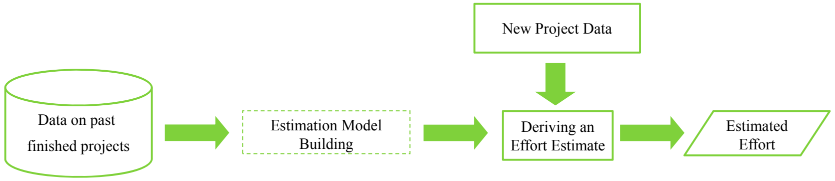

New Project Data
Data on past finished projects
Estimation Model Building
Deriving an Effort Estimate
Estimated Effort

Figure1. Sequence used when estimating effort using a data-driven approach *

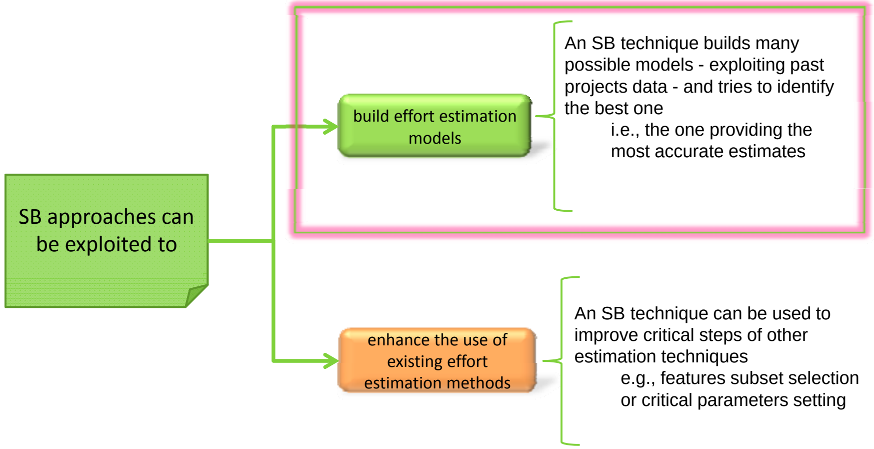
- Widely used techniques for model building are Linear Regression (LR) and Stepwise Regression (SWR)

* *Adapted from E. Mendes, "Web Cost Estimation and Productivity Benchmarking", ISSSE 2008, LNCS 5413 Springer 2009, pp. 194-222.*

Effort Estimation with Search-Based Approaches

- The effort estimation problem can be formulated as an optimization problem
  - we have to find among the possible estimation models *the most accurate*
- The use of Search-Based (SB) approaches has been suggested for effort estimation
  - the fitness function guides the search
    - it should be able to determine whether an estimation model leads to more accurate predictions than another

Effort Estimation with Search-Based Approaches



SB approaches can be exploited to

build effort estimation models

An SB technique builds many possible models - exploiting past projects data - and tries to identify the best one

i.e., the one providing the most accurate estimates

enhance the use of existing effort estimation methods

An SB technique can be used to improve critical steps of other estimation techniques

e.g., features subset selection or critical parameters setting

F. Ferrucci, C. Gravino, R. Oliveto, F. Sarro, "Using Evolutionary Based Approaches to Estimate Software Development Effort", in Evolutionary Computation and Optimization Algorithms in Software Engineering: Applications and Techniques, M. Chis, IGI Global

"…no matter what search technique is employed, it is the fitness function that captures the crucial information; it differentiates a good solution from a poor one, thereby guiding the search."[1]

"…each measure used to evaluate properties of interest can be used as fitness function."[2]

…in the effort estimation context several criteria have been proposed to evaluate models' accuracy…

Influence of Fitness Function…

(1) Harman, M., The current state and future of search-based software engineering. In Procs of IEEE FOSE 2007
(2) Harman,, M., Clark, J.A., Metrics Are Fitness Functions Too. In Procs of IEEE METRICS 2004

How to assess estimation model accuracy?

- Several evaluation criteria are employed for assessing the accuracy of effort estimation models
- The most commonly used are based on

$$\text{Absolute Residuals} = |\text{ActualEffort} - \text{EstimatedEffort}|$$

- Summary Measures
  - **MMRE** (Mean MRE)
  - **MdMRE** (Median MRE)
  - **Pred(25)** (Prediction at level 25): percentage of the estimates whose MRE < 25
  - **MEMRE** (Mean EMRE)
  - **MdEMRE** (Median EMRE)
    - **MRE** (Magnitude of Relative Error) $$MRE = \frac{|ActualEffort - EstimatedEffort|}{ActualEffort}$$
    - **EMRE** (Estimated MRE) $$EMRE = \frac{|ActualEffort - EstimatedEffort|}{EstimatedEffort}$$

How to assess estimation model accuracy?

- Different accuracy measures take into account different aspects of model performance[1,2]
  - MMRE measures poor performance
  - MEMRE is more sensitive to under-estimates
  - Pred(25) measures how well an estimation model performs
  - …
- There is no convergence of opinion on what is the best accuracy measure[3]
  - to compare different models and consistently derive the best one

(1) T. Menzies, Zhihao Chen, J. Hihn, K. Lum, Selecting Best Practices for Effort Estimation. IEEE TSE, 32(11)(2006)
(2) B. A. Kitchenham, L. M. Pickard, S.G MacDonell, M.J. Shepperd, What accuracy statistics really measure. IEE Procs. Software 148(3)(2002)
(3) T.Foss, E.Stensrud, B. Kitchenham, I. Myrtveit, A Simulation Study of the Model Evaluation Criterion MMRE. IEEE TSE 29(11)(2003)

How to assess estimation model accuracy?

- Different accura ccount different aspec 2
  - MMRE mea
  - MEMRE is m
  - Pred(25) me model performs
  - …
- There is no conve ence o pinion on what is the best accuracy m asure[3]
  - to compare di ferent models and consistently derive the best one

What accuracy measure can be used as fitness function?

(1) T. Menzies, Zhihao Chen, J. Hihn, K. Lum, Selecting Best Practices for Effort Estimation. IEEE TSE, 32(11)(2006)
(2) B. A. Kitchenham, L. M. Pickard, S.G MacDonell, M.J. Shepperd, What accuracy statistics really measure. IEE Procs. Software 148(3)(2002)
(3) T.Foss, E.Stensrud, B. Kitchenham, I. Myrtveit, A Simulation Study of the Model Evaluation Criterion MMRE. IEEE TSE 29(11)(2003)

How to assess estimation model accuracy?

- Some previous works exploited MMRE as fitness function[1,2]
  - one of the most widely used criterion
  - one of the most questioned
    - e.g., it does not consistently select the best from two competing models [3]
- Each measure used to evaluate properties of interest can be used as fitness function[4]
  - the choice of the evaluation criterion can be a managerial issue
- Using Genetic Programming (GP) project managers can select their preferred evaluation criterion as fitness function
  - the search for the estimation model is driven by such a criterion

(1) C.J. Burgess, M. Lefley, Can genetic programming improve software effort estimation? A comparative evaluation. IST 43(14) (2001)
(2) M.Lefley, M.J. Shepperd, Using Genetic Programming to Improve Software Effort Estimation Based on General Data Sets. GECCO 2003
(3) T.Foss, E.Stensrud, B. Kitchenham, I. Myrtveit, A Simulation Study of the Model Evaluation Criterion MMRE. IEEE TSE 29(11)(2003)
(4) M.Harman, J.A.Clark, Metrics Are Fitness Functions Too. IEEE METRICS 2004

How to assess estimation model accuracy?

- Some previous wor ss function[1,2]
  - one of the most w
  - one of the mo
    - e.g., it does n eting models [3]
- Each measure u nterest can be used as fitness fu
  - the choice of the evalu n be a managerial issue
- Using Genetic Progr ming (GP) project managers can select their preferred ev uation criterion as fitness function
  - the search for the estimation model is driven by such a criterion

Does the choice of the fitness function impact on the accuracy of the effort estimation models built with GP?

(1) C.J. Burgess, M. Lefley, Can genetic programming improve software effort estimation? A comparative evaluation. IST 43(14) (2001)
(2) M.Lefley, M.J. Shepperd, Using Genetic Programming to Improve Software Effort Estimation Based on General Data Sets. GECCO 2003
(3) T.Foss, E.Stensrud, B. Kitchenham, I. Myrtveit, A Simulation Study of the Model Evaluation Criterion MMRE. IEEE TSE 29(11)(2003)
(4) M.Harman, J.A.Clark, Metrics Are Fitness Functions Too. IEEE METRICS 2004

Outline

- Background and Motivations
  - Software Development Effort Estimation
  - Effort Estimation with Search-Based Approaches
  - How to Assess Estimation Model Accuracy?
- Empirical Study: Influence of Fitness Function
  - Research Goals
  - GP Setting, Dataset Selection, Validation Method, and Evaluation Criteria
  - Results
- Preliminary Empirical Study: Multi-Objective Genetic Programming
  - Research Goals
  - MOGP Setting, Dataset Selection, Validation Method, and Evaluation Criteria
  - Results
- Conclusions

Empirical Study: Research Goals

- RG$_1$: How the choice of the fitness function impact on the accuracy of the estimation models built with GP?
  - Does GP effectively optimize the criterion employed as fitness function?
  - Are there any differences in using different fitness functions?
- RG$_2$: Is GP more effective than widely used effort estimation methods?
  - Manual Stepwise Regression (MSWR), Case-Based Reasoning (CBR), Mean and Median of Effort

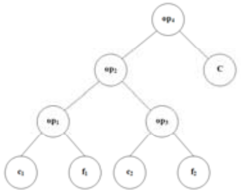
Empirical Study: GP Setting (1)

- A solution consists of an estimation model described by an equation

$$\boldsymbol{Effort = c_1\, op_1 f_1\, op_2 \ldots\, op_{2n-2}\, c_n\, op_{2n-1} f_n\, op_{2n}}\, C$$

where c_i represents the coefficient of the ith project feature
f_i represents the value of the ith project feature
$op_i \in \{+, -, \cdot, /, \mathrm{f_i} \wedge \mathrm{c_i}, \ln(\mathrm{f_i})\}$
C represents a constant
$Effort > 0$

- encoded as a binary tree of fixed depth
  - leaves: features and coefficients
  - internal nodes: mathematical operators



Empirical Study: GP Setting (2)

- Initial Population
  - 10V random trees, where V is the number of project features contained in the dataset
- Genetic Operators
  - crossover randomly selects the same point of cut in parent trees and swaps the corresponding subtrees
  - mutation randomly selects a node in a tree and replaces its value with a new one
- Selection
  - Roulette Wheel Selection for parent selection
  - Tournament Selection for survival selection
- Termination Criteria
  - GP is stopped after 1000V generations or if the fitness value of the best solution does not change after 100V generations
- Execution Number
  - we performed 10 runs considering as final prediction model the one that had the fitness value closest to the average value achieved in the 10 runs on training sets

Empirical Study: GP Setting (3)

- The experimented fitness functions

Accuracy Measure	Employed Fitness Function
MMRE	1/MMRE
Pred(25)	Pred(25)
MdMRE	1/MdMRE
MEMRE	1/MEMRE
MdEMRE	1/MdEMRE
MMRE e Pred(25)	Pred(25)/MMRE
MdMRE e Pred(25)	Pred(25)/MdMRE
MEMRE e Pred(25)	Pred(25)/MEMRE
MdEMRE e Pred(25)	Pred(25)/MdEMRE

We experimented with the above accuracy measures as fitness function to analyze the impact on the estimation accuracy of the constructed models

The observation that different accuracy measures take into account different aspects of predictions accuracy suggested us to investigate also the effectiveness of some combinations of those accuracy measures

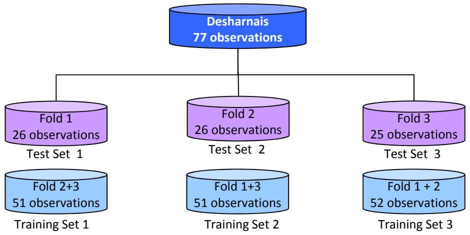
Empirical Study: Dataset Selection

Table 1. A summary of the employed datasets selected from the PROMISE repository

	Dataset	Description	Observations	Employed Features
Single-Company	Desharnais	Software projects derived from a Canadian software house	77	7
	Maxwell	Software projects coming from one of the biggest commercial bank in Finland	62	17
	Telecom	Data about enhancement projects for a U.K. telecommunication product	18	2
Cross-Company	China	Projects developed by Chinese software companies	499	5
	Finnish	Data collected by the TIEKE organizations on projects from different Finnish software companies	38	4
	Kemerer	Data on large business applications collected by a national computer consulting and services firm, specialized in the design and development of data-processing software	15	1
	Miyazaki	Data on projects developed in 20 companies by Fujitsu Large Systems Users Group	48	3

Empirical Study: Validation Method

- We applied a 3-fold[1] cross validation randomly partitioning the original datasets into
  - 3 training sets for model building
  - 3 test sets for model evaluation


Desharnais
77 observations

Fold 1
26 observations
Test Set 1

Fold 2
26 observations
Test Set 2

Fold 3
25 observations
Test Set 3

Fold 2+3
51 observations
Training Set 1

Fold 1+3
51 observations
Training Set 2

Fold 1 + 2
52 observations
Training Set 3

Figure 2. 3-fold cross validation for the Desharnais dataset

(1) The folds were made publicly available to allow for replications

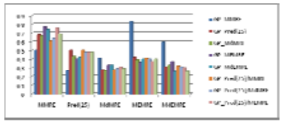
Empirical Study: Evaluation Criteria

- To assess whether the selected criterion was optimized we employed
  - the summary measure used as fitness function
- To assess the overall estimation accuracy[1] we employed
  - MMRE, Pred(25), MdMRE, MEMRE, MdEMRE
  - Boxplots of absolute residuals
  - Wilcoxon Test (α=0.05) to analyze whether there is significant difference between the absolute residuals
    - since the absolute residuals were not normally distributed and the data was naturally paired

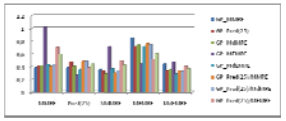
(1) Kitchenham, B., Pickard, L. M., MacDonell, S. G., Shepperd, M. J., What accuracy statistics really measure, IEE Procs Software (2001)

Empirical Study: Results
Influence of the fitness function (1)

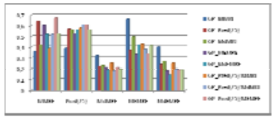
Desharnais



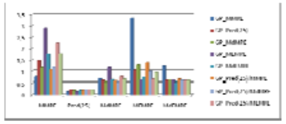
Maxwell



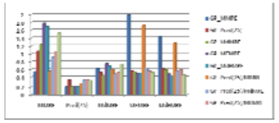
Telecom



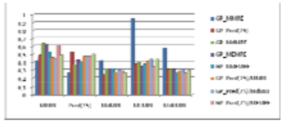
China



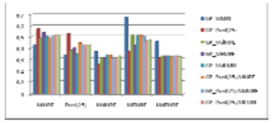
Finnish



Kemerer



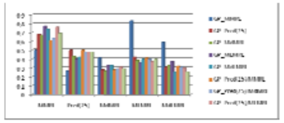
Miyazaki



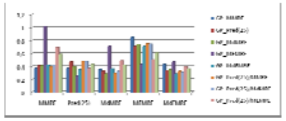
Results on Training Sets…
… to assess the models' ability to fit data

Empirical Study: Results
Influence of the fitness function (1)

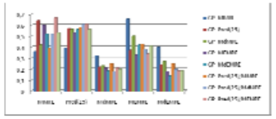
Desharnais



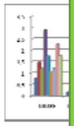
Maxwell



Telecom

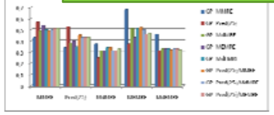


China



Our Running Example is the Desharnais dataset
Note that the observations we will make hold also for the other datasets

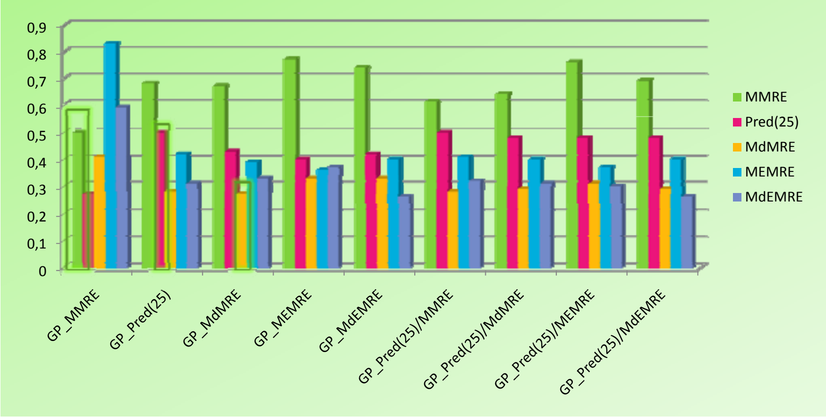
Miyaza



Results on Training Sets…
… to assess the models' ability to fit data

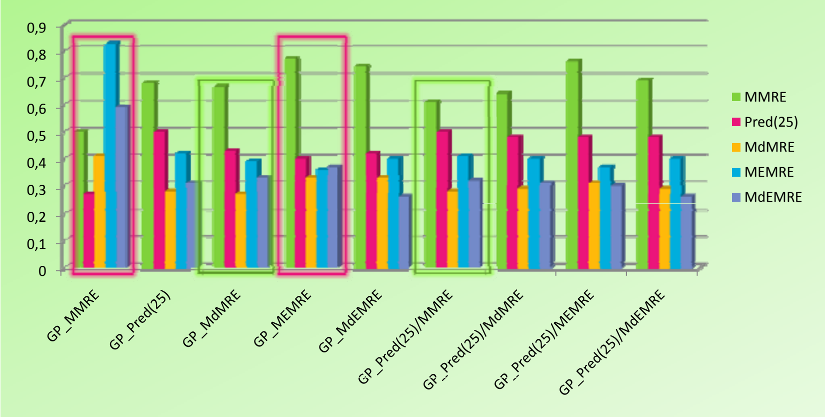
Empirical Study: Results
Influence of the fitness function (3)

Figure 3. Performance of using GP with different fitness functions in terms of MMRE, Pred(25), MdMRE, MEMRE, and MdEMRE on Desharnais dataset (TRAINING SETS)



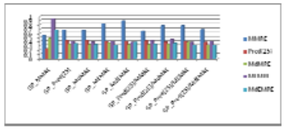
Empirical Study: Results
Influence of the fitness function (3)

Figure 3. Performance of using GP with different fitness functions in terms of MMRE, Pred(25), MdMRE, MEMRE, and MdEMRE on Desharnais dataset (TRAINING SETS)

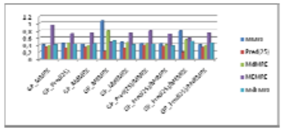


Empirical Study: Results
Influence of the fitness function (4)

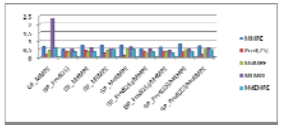
Desharnais



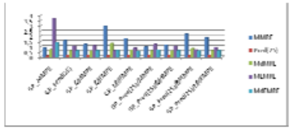
Maxwell Test Sets



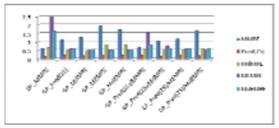
Telecom



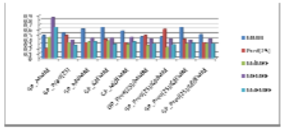
China



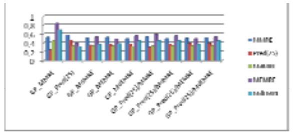
Finnish



Kemerer



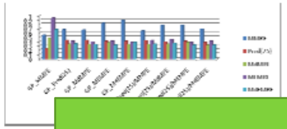
Miyazaki



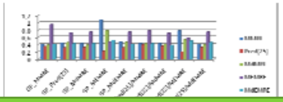
Results on Test Sets…
… to assess the models' predictive capability

Empirical Study: Results
Influence of the fitness function (4)

Desharnais



Maxwell Test Sets

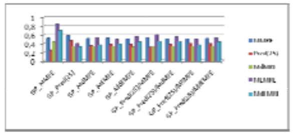


Telecom

China

Our Running Example is again the Desharnais dataset
Note that the observations we will make hold also for the other datasets

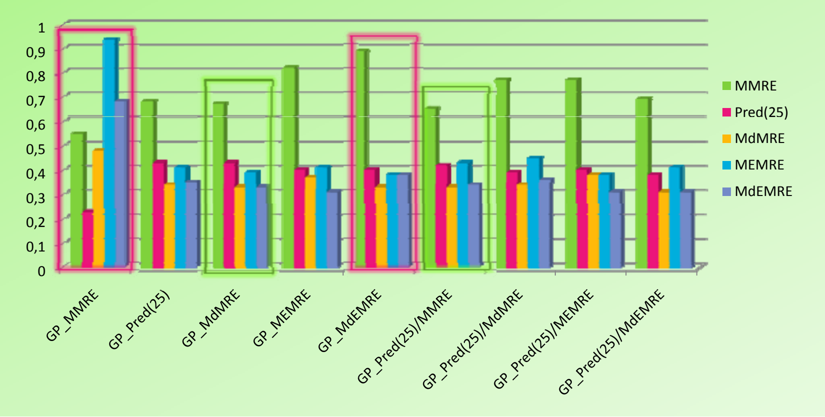
Miyazaki



Results on Test Sets…
… to assess the models' predictive capability

Empirical Study: Results
Influence of the fitness function (4)

Figure 4. Performance of using GP with different fitness functions in terms of MMRE, Pred(25), MdMRE, MEMRE, and MdEMRE on Desharnais dataset (TEST SETS)



Empirical Study: Results
Influence of the fitness function (5)

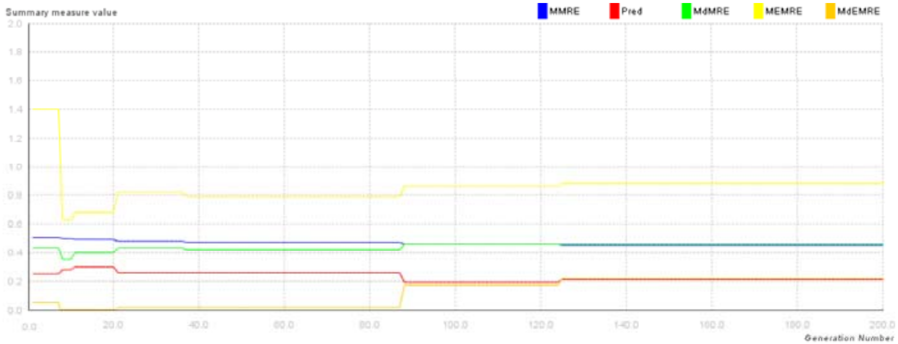


Figure 5. An excerpt of the trend of summary measures during the evolution process when MMRE is used as fitness function (Desharnais dataset)

Empirical Study: Results
Influence of the fitness function (6)

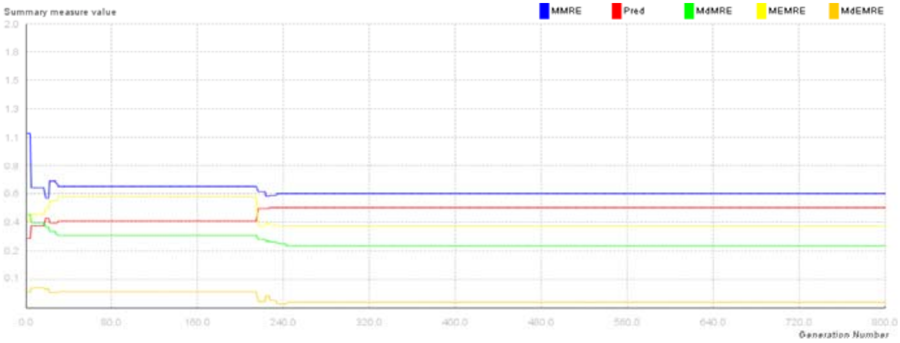


Figure 6. An excerpt of the trend of summary measures during the evolution process when MdMRE is used as fitness function (Desharnais dataset)

Empirical Study: Results
Influence of the fitness function (7)

Figure 7. Results of the Wilcoxon test comparing different fitness functions (Desharnais dataset)

(training set)

	MMRE	Pred(25)	MdMRE	MEMRE	MdEMRE	Pred(25)/MMRE	Pred(25)/MdMRE	Pred(25)/MEMRE	Pred(25)/MdEMRE
Pred(25)/MdEMRE	X				X				
Pred(25)/MEMRE	X								
Pred(25)/MdMRE	X				X				
Pred(25)/MMRE	X				X				
MdEMRE									
MEMRE									
MdMRE	X				X				
Pred(25)	X								
MMRE									

(test set)

	MMRE	Pred(25)	MdMRE	MEMRE	MdEMRE	Pred(25)/MMRE	Pred(25)/MdMRE	Pred(25)/MEMRE	Pred(25)/MdEMRE
Pred(25)/MdEMRE	X			X					
Pred(25)/MEMRE									
Pred(25)/MdMRE									
Pred(25)/MMRE	X			X			X		
MdEMRE									
MEMRE									
MdMRE	X			X	X		X	X	X
Pred(25)	X			X	X		X		
MMRE									

Each "x" means that "the fitness function indicated on the corresponding row provides significantly less absolute residuals than the fitness function indicated on the corresponding column"

Empirical Study: Results
Influence of the fitness function (8)

Table 2. Influence of the fitness function: a summary

(training set)

Dataset	Best Fitness Functions	Worst Fitness Functions	(1)	(2)
Desharnais	Pred(25)/MMRE MdMRE Pred(25)	MMRE MdEMRE	YES	NO
Finnish	Pred(25) Pred(25)/MEMRE	MMRE MEMRE, MdEMRE	YES	NO
Kemerer	Pred(25)/MdMRE Pred(25)/MMRE Pred(25)	MMRE MEMRE	YES	NO
Miyazaki	Pred(25) Pred(25)/MEMRE MEMRE	MMRE	YES	NO
Telecom	Pred(25)/MdMRE Pred(25)/MMRE Pred(25), MdMRE	MMRE	YES	NO
China	Pred(25)/MdMRE MdMRE Pred(25)	MEMRE MMRE	YES	NO
Maxwell	Pred(25)/MdMRE MdMRE Pred(25)(25)/MMRE	MMRE MEMRE	YES	NO

(test set)

Dataset	Best Fitness Functions	Worst Fitness Functions	(1)	(2)
Desharnais	MdMRE Pred(25)/MMRE Pred(25)	MMRE MEMRE	YES	NO
Finnish	Pred(25) Pred(25)/MEMRE	MMRE MEMRE, MdEMRE	YES	NO
Kemerer	Pred(25)/MdMRE Pred(25)/MMRE Pred(25)	MMRE MdMRE MEMRE	YES	NO
Miyazaki	Pred(25) Pred(25)/MEMRE MEMRE	MMRE	YES	NO
Telecom	Pred(25)/MdMRE Pred(25)/MMRE Pred(25), MdMRE	MMRE	YES	NO
China	Pred(25)/MdMRE MdMRE Pred(25)	MEMRE MMRE	YES	NO
Maxwell	Pred(25)/MdMRE MdMRE Pred(25)/MMRE ,Pred(25)	MMRE MEMRE	YES	NO

(1) Does using MMRE as fitness function negatively impact on MEMRE value and viceversa?
(2) Does using MdMRE as fitness function negatively impact son MdEMRE value and viceversa?

Empirical Study: Results
GP vs. other estimation methods (1)

Figure 8. Comparison based on Summary Measures (Desharnais dataset)

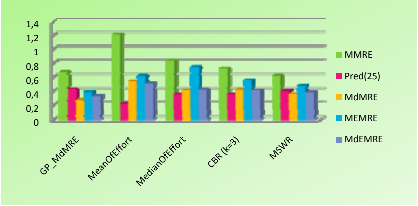


Table 3. Wilcoxon Test (p-value) (Desharnais dataset)

<	MeanOfEffort	MedianOfEffort	CBR	MSWR
GP_MdMRE	**0**	**0.002**	**0.009**	0.093

- GP achieved the best results in terms of summary measures
- Absolute residuals achieved by GP were significantly less than those achieved by MeanOfEffort, MedianOfEffort, and CBR
- There was not statistical significant difference between GP and MSWR

Empirical Study: Results
GP vs. other estimation methods (2)

Table 4. Wilcoxon Test (p-value)

Dataset	<	MeanOfEffort	MedianOfEffort	CBR	MSWR
Desharnais	GP_MdMRE	**0**	**0.002**	**0.009**	0.093
Finnish	GP_Pred(25)	**0**	**0.001**	**0.046**	0.337
Miyazaki	GP_Pred(25)	**0**	**0**	**0.006**	**0.034**
Maxwell	GP_Pred(25)/MdMRE	**0**	**0.001**	0.057	0.691
Telecom	GP_Pred(25)/MdMRE	**0.037**	**0.01**	**0.041**	0.82
China	GP_Pred(25)/MdMRE	**0**	**0**	**0**	0.817
Kemerer	GP_Pred(25)/MdMRE	**0.017**	**0.025**	0.295	0.147

How the choice of the fitness function impact on the accuracy of the estimation models built with GP?

GP optimizes the criterion selected as fitness function

Using MMRE or MEMRE is the worst choice for the overall accuracy

- Using MMRE negatively impacts on MEMRE value and viceversa
- Significantly worse results with respect to the ones achieved using the other fitness functions

Pred(25), 1/MdMRE, Pred(25)/MMRE, Pred(25)/MdMRE can be more promising as fitness function

- Estimates significantly better than those obtained with CBR
- The fitness functions based on the combination of two criteria often provided better estimates than fitness functions based on a single criterion
  - Pred(25)/MMRE (Pred(25)/MdMRE) takes into account good and poor model performance aspect → complex multi-objective approaches might be a viable way to improve the overall accuracy

Outline

- Background and Motivations
  - Software Development Effort Estimation
  - Effort Estimation with Search-Based Approaches
  - How to Assess Estimation Model Accuracy?
- Empirical Study: Influence of Fitness Function
  - Research Goals
  - GP Setting, Dataset Selection, Validation Method, and Evaluation Criteria
  - Results
- Preliminary Empirical Study: Multi-Objective Genetic Programming
  - Research Goals
  - MOGP Setting, Dataset Selection, Validation Method, and Evaluation Criteria
  - Results
- Conclusions

Preliminary Empirical Study: Research Goals

- RG_1: Is Multi-Objective Genetic Programming effective to address the effort estimation problem?
- RG_2: Do the objectives employed in the definition of the fitness function impact on estimation accuracy?
- RG_3: Is the increasing of complexity determined by the use of MOGP paid back by an improvement of performance?

Preliminary Empirical Study: MOGP Setting (1)

- We designed and experimented a Multi-Objective Genetic Programming (i.e., MOGP)
  - an adaptation to GP of the Non dominated Sort Genetic Algorithm-II (NSGA-II)
  - same GP setting except for
    - an objective vector is considered instead of a single function and the fitness assignment is based on the dominance deep according to NSGA-II
    - selection operators perform according to the non-dominance and crowding distance criteria
    - the final solution is selected from the pareto front by using an "a priori" decision maker which provides a complete order between the Pareto optimal solutions according to the following expression
      - Pred(25)/(O_1,+ … + O_n), where O_i is the value of a measure belonging to the objective vector and to the set {MMRE, MEMRE, MdMRE, MdEMRE}

Preliminary Empirical Study: MOGP Setting (2)

- □ Different objective vectors were employed as multi-objective functions

Name	Employed Objective Vector
MOGP1	[1/MMRE, Pred(25), 1/MdMRE, 1/MEMRE, 1/MdEMRE]
MOGP2	[1/MMRE, Pred(25), 1/MdMRE]
MOGP3	[Pred(25), 1/MMRE]
MOGP4	[Pred(25), 1/MdMRE]

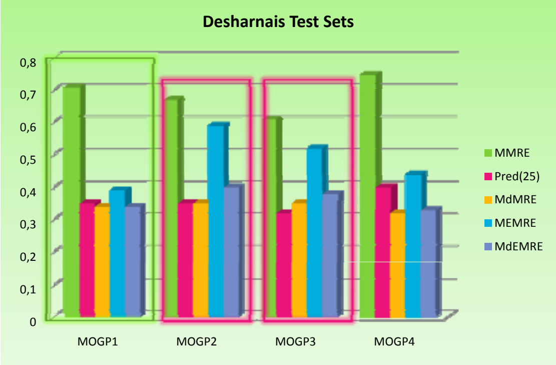
Preliminary Empirical Study: Dataset Selection, Validation Method, and Evaluation Criteria

- □ Employed Dataset

Dataset	Description	Observations	Employed Features
Desharnais	Software projects derived from a Canadian software house	77	7
Miyazaki	Data on projects developed in 20 companies by Fujitsu Large Systems Users Group	48	3

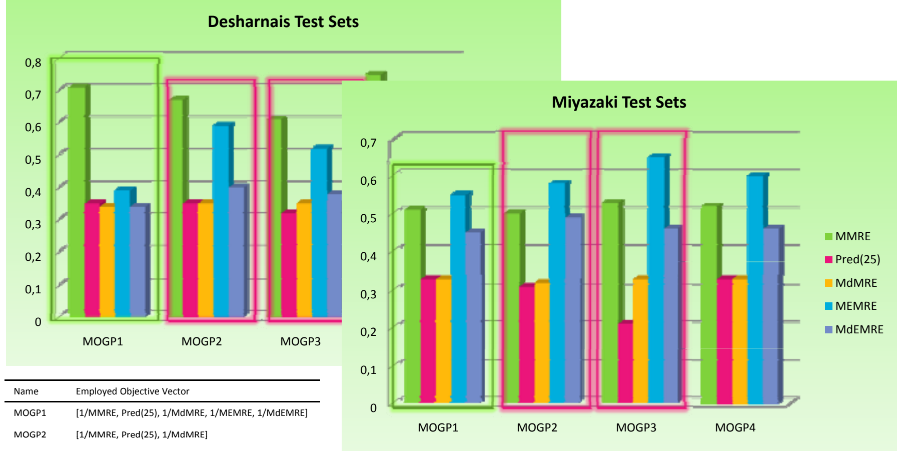
- □ Validation Method
  - 3-fold cross-validation
- □ Evaluation criteria
  - summary measures and statistical significance test

Preliminary Empirical Study: Results Influence of the objective vector (1)



Name	Employed Objective Vector
MOGP1	[1/MMRE, Pred(25), 1/MdMRE, 1/MEMRE, 1/MdEMRE]
MOGP2	[1/MMRE, Pred(25), 1/MdMRE]
MOGP3	[Pred(25), 1/MMRE]
MOGP4	[Pred(25), 1/MdMRE]

Preliminary Empirical Study: Results Influence of the objective vector (1)



Name	Employed Objective Vector
MOGP1	[1/MMRE, Pred(25), 1/MdMRE, 1/MEMRE, 1/MdEMRE]
MOGP2	[1/MMRE, Pred(25), 1/MdMRE]
MOGP3	[Pred(25), 1/MMRE]
MOGP4	[Pred(25), 1/MdMRE]

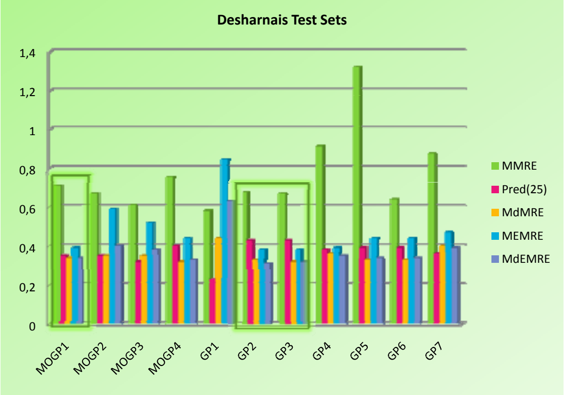
Preliminary Empirical Study: Results
Influence of the objective vector (2)

Table 5. p-values of the Wilcoxon test comparing the considered MOGPs (on the test sets)

Dataset	<	MOGP1	MOGP2	MOGP3	MOGP4
Desharnais	MOGP1	-	**0.019**	0.124	0.102
	MOGP2	0.981	-	0.927	0.847
	MOGP3	0.876	0.073	-	0.635
	MOGP4	0.898	0.153	0.365	-
Miyazaki	MOGP1	-	0.324	**0.033**	**0.011**
	MOGP2	0.676	-	0.179	0.142
	MOGP3	0.967	0.821	-	0.971
	MOGP4	0.989	0.858	**0.029**	-

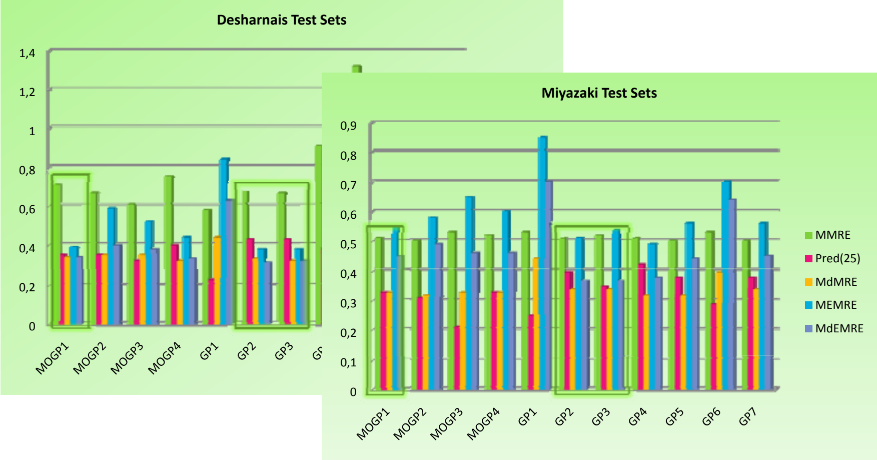
Null hypothesis: "the use of m_i does not provide better absolute residuals than using m_j", where m_i and m_j are two experimented multi-objective functions

Preliminary Empirical Study: Results
Comparison with GP (1)



MOGP1: [1/MMRE, Pred(25), 1/MdMRE, 1/MEMRE, 1/MdEMRE]; MOGP2: [1/MMRE, Pred(25), 1/MdMRE]; MOGP3: [Pred(25), 1/MMRE]; MOGP4: [Pred(25), 1/MdMRE]
GP1: GP_MMRE ; GP2: GP_Pred(25); GP3: GP_MdMRE; GP4: GP_MEMRE; GP5: GP_MdEMRE; GP6:GP_Pred(25)/MMRE; GP7: GP_Pred(25)/MdMRE

Preliminary Empirical Study: Results
Comparison with GP (1)



MOGP1: [1/MMRE, Pred(25), 1/MdMRE, 1/MEMRE, 1/MdEMRE]; MOGP2: [1/MMRE, Pred(25), 1/MdMRE]; MOGP3: [Pred(25), 1/MMRE]; MOGP4: [Pred(25), 1/MdMRE]
GP1: GP_MMRE ; GP2: GP_Pred(25); GP3: GP_MdMRE; GP4: GP_MEMRE; GP5: GP_MdEMRE; GP6:GP_Pred(25)/MMRE; GP7: GP_Pred(25)/MdMRE

Preliminary Empirical Study: Results
Comparison with GP (2)

Table 6. p-values of the Wilcoxon test comparing MOGP and GP (on the test sets)

Dataset	<	GP1	GP2	GP3	GP4	GP5	GP6	GP7
Desharnais	MOGP1	**0.016**	0.778	0.631	**0.004**	**0.024**	0.433	**0.009**
	MOGP2	**0.006**	0.988	0.991	0.616	0.337	0.997	0.658
	MOGP3	**0.005**	0.914	0.908	0.359	0.329	0.899	0.441
	MOGP4	0.088	0.786	0.834	0.108	0.255	0.625	0.293
Miyazaki	MOGP1	**0.003**	0.874	0.417	0.977	0.650	**0.016**	0.701
	MOGP2	**0.003**	0.759	0.445	0.982	0.615	0.081	0.575
	MOGP3	**0.010**	0.996	0.995	0.996	0.993	0.401	0.987
	MOGP4	**0.002**	0.995	0.922	0.993	0.981	0.066	0.978

MOGP1: [1/MMRE, Pred(25), 1/MdMRE, 1/MEMRE, 1/MdEMRE]; MOGP2: [1/MMRE, Pred(25), 1/MdMRE]; MOGP3: [Pred(25), 1/MMRE]; MOGP4: [Pred(25), 1/MdMRE]
GP1: GP_MMRE ; GP2: GP_Pred(25); GP3: GP_MdMRE; GP4: GP_MEMRE; GP5: GP_MdEMRE; GP6:GP_Pred(25)/MMRE; GP7: GP_Pred(25)/MdMRE

Outline

- Background and Motivations
  - Software Development Effort Estimation
  - Effort Estimation with Search-Based Approaches
  - How to Assess Estimation Model Accuracy?
- Empirical Study: Influence of Fitness Function
  - Research Goals
  - GP Setting, Dataset Selection, Validation Method, and Evaluation Criteria
  - Results
- Preliminary Empirical Study: Multi-Objective Genetic Programming
  - Research Goals
  - MOGP Setting, Dataset Selection, Validation Method, and Evaluation Criteria
  - Results
- Conclusions

Conclusions

- GP represents a flexible method that allows project managers to identify their preferred evaluation criterion
  - the choice of the fitness function influences the performance of the models constructed with GP
    - the use of MMRE or MEMRE is not the best choice
      - using them had the effect to degrade a lot of other criteria
    - other accuracy measures are more promising (e.g., Pred(25)/MdMRE)
      - significantly better results than the ones provided by using GP with other fitness functions
      - estimates significantly better than those obtained with CBR
- A preliminary empirical analysis revealed that
  - the best results achieved with MOGP and GP were comparable
  - the choice of the objective vector influences the performance of the models constructed with MOGP

References

- F. Ferrucci, C. Gravino, R. Oliveto, F. Sarro, "Using Evolutionary Based Approaches to Estimate Software Development Effort", in Evolutionary Computation and Optimization Algorithms in Software Engineering: Applications and Techniques, M. Chis, IGI Global, ISBN13: 9781615208098
- F. Ferrucci, C. Gravino, R. Oliveto, F. Sarro, "Genetic Programming for Effort Estimation: an Analysis of the Impact of Different Fitness Functions", in Proceedings of the 2nd International Symposium on Search Based Software Engineering, IEEE Computer Society, pp. 89-98, ISBN: 978-0-7695-4195-2
- F. Ferrucci, C. Gravino, F. Sarro, "How Multi-Objective Genetic Programming is Effective for Software Development Effort Estimation?", SSBSE 2011, to appear
- F. Sarro, "Search-Based Approaches for Software Development Effort Estimation", PROFES 2011 Doctoral Symposium, ACM Inc., pp. 38-43, ISBN: 978-1-4503-0783-3
- F. Ferrucci, C. Gravino, R. Oliveto, F. Sarro "Using Tabu Search to Estimate Software Development Effort", in Proceedings of IWSM/MENSURA 2009. Lecture Notes in Computer Science, Springer, vol. 5891, pp. 307-320, ISBN:978-3-642-05414-3
- F. Ferrucci, C. Gravino, E. Mendes, R. Oliveto, F. Sarro, "Investigating Tabu Search for Web Effort Estimation", in Proceedings of the 36th EUROMICRO Conference on Software Engineering and Advanced Applications (SEAA 2010), IEEE Computer Society, pp.350-357, ISBN: 978-0-7695-4170-9
- A. Corazza, S. Di Martino, F. Ferrucci, C. Gravino, F. Sarro, E. Mendes, "How Effective is Tabu Search to Configure Support Vector Regression for Effort Estimation?" (*Best Paper Award*) , in Proceedings of the 6th International Conference on Predictor Models in Software Engineering (PROMISE 2010), ACM Inc, pp. 1-10, ISBN: 978-1-4503-0404-7

Questions?

Thanks for your attention

Federica Sarro

fsarro@unisa.it
www.dmi.unisa.it/people/**sarro**/www/

University of Salerno