Empir Software Eng (2015) 20:813–843
DOI 10.1007/s10664-014-9300-5

Transfer learning in effort estimation

Ekrem Kocaguneli · Tim Menzies · Emilia Mendes

Published online: 29 March 2014

Abstract When projects lack sufficient local data to make predictions, they try to transfer information from other projects. How can we best support this process? In the field of software engineering, transfer learning has been shown to be effective for defect prediction. This paper checks whether it is possible to build transfer learners for software effort estimation. We use data on 154 projects from 2 sources to investigate transfer learning between different time intervals and 195 projects from 51 sources to provide evidence on the value of transfer learning for traditional cross-company learning problems. We find that the same transfer learning method can be useful for transfer effort estimation results for the cross-company learning problem and the cross-time learning problem. It is misguided to think that: (1) Old data of an organization is irrelevant to current context or (2) data of another organization cannot be used for local solutions. Transfer learning is a promising research direction that transfers relevant cross data between time intervals and domains.

Keywords Transfer learning · Effort estimation · Data mining · k-NN

Communicated by: Martin Shepperd

E. Kocaguneli (✉) · T. Menzies
Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26505, USA
e-mail: kocaguneli@gmail.com

T. Menzies
e-mail: tim@menzies.us

E. Mendes
Department of Software Engineering, Faculty of Computing, Blekinge Institute of Technology, 37179 Karlskrona, Sweden
e-mail: emilia@cs.auckland.ac.nz

1 Introduction

There is a solid body of evidence that data miners can find interesting and useful patterns from *within-company* data; i.e., the data found within one software organization. This data can take many forms including information from mobile phone application store (Harman et al. 2012); natural languages requirements documents (Hayes et al. 2006); logs of software inspections (Menzies et al. 2007; Turhan et al. 2009); or even the power consumption records associated with software (Hindle 2012).

The next challenge after *within-company* learning is how to transfer results across *different companies*. *Transfer learning* divides data into *domains* where each domain is a set of examples plus a probability distribution over some class variable (and this distribution must be learned from the examples) (Pan and Yang 2010). When we jump from domain1 to domain2, the examples and/or the distribution may change and it is the task of the transfer learner to discover how much of domain1 can be suitably applied to domain2.

Transfer learning can be potentially very useful for software engineering (SE). For example, if a project is using some technology for the first time, it can use transfer learning to acquire valuable insights from another project where developers had more experience in that technology. To facilitate this process, many organizations expend much effort to create repositories of software project data. For example, Rodriguez et al. (2012) report 13 active SE repositories that researchers can access; e.g., PROMISE, BUGZILLA, ISBSG, etc.

These repositories are valuable resources for researchers, since they offer benchmark problems for testing techniques. Better still, if other researchers can access the same repositories, then those benchmark problems encourage replication studies where one researcher tries to repeat, improve, or perhaps even refute the results from another.

While these repositories are useful to researchers, it is an open issue if ***repositories of software project data are valuable to industrial software companies***. This issue must be asked for two reasons: Such repositories *may not predict properties for future projects* and they may be *very expensive to build*:

- Zimmermann et al. report over 600 experiments where data from Project1 was used to build a defect predictor for Project2 (Zimmermann et al. 2009). In only 3 % of those experiments did models learned from Project1 adequately predicted defects for Project2. Other researchers offer similar pessimistic results (Menzies et al. 2011, 2012; Posnett et al. 2011, Bettenburg et al. 2012). For example, we have one study where data from six projects is used to predict for defects in a seventh project. The false alarm rate in those predictions was unacceptably high (median values above 65 %) (Turhan et al. 2009).[1]
- These empirical results are troubling since it can be quite costly to create these repositories. For example, Jairus Hihn from the Jet Propulsion Laboratory (Hihn and Habib-agahi 1991) reports that NASA spent two million dollars in the early 1990s to build their own repository of project data from 121 NASA software projects developed in the 1970s and 1980s. If we cannot use such expensive repositories to predict properties in future projects, then all that expense is wasted.

[1] In the literature, this is known as a *negative transfer effect* (Pan and Yang 2010) where transfer learning actually makes matters worse.

It might be argued that, for software engineering, this goal of transfer learning is wrong-headed. After all, software is built by specific people using specific tools for specific tasks on specific platforms. Given that, perhaps all we can ever hope to do is to monitor some current project until some stable pattern emerges for that specific team, tool, task, and platform. But if that was the case, if all conclusions are local and cannot be transferred to other projects, then:

- Managers would have no basis for planning for future software projects.
- We must conclude that these repositories are ***not*** valuable, at least for the purposes of predicting the properties of new projects.

Recent results in one specific area of software engineering suggest that these concerns may be unfounded. In the specific sub-domain of defect prediction, there are two different types of results:

1. Earlier results of transfer learning show that transferring data comes with the cost of reduced performance (Zimmermann et al. 2009; Menzies et al. 2011, 2012; Posnett et al. 2011; Bettenburg et al. 2012);
2. Studies following the earlier results confirm the prior findings of compromised performance through the use of transfer learning in defect prediction, if the data is used in an *as-is* manner. Yet they also add the important finding that filtering the transferred data (e.g., through instance selection) may address the performance issue (Turhan et al. 2009; Ma et al. 2012).

Relevancy filter has been shown to allow effective transfer of data across organizations. A relevancy filter carefully selects some subset of the training data. For example, weights can be applied to control how we prune away irrelevant training data. Such weights can be learned at the level of instances (Turhan et al. 2009; Ma et al. 2012) or data sets (He et al. 2012). Alternatively, partitioning methods can be used to find clusters within the training data that are most relevant to the test data (Posnett et al.2011; Menzies et al. 2011; 2012; Bettenburg et al. 2012).

These promising results from the area of defect prediction motivate our exploration of other areas of software engineering. This paper explores transfer learning for software effort estimation (SEE). Note that the use of *"transfer learning for SEE"* should not be misleading as this paper does not introduce a new set of techniques for SEE. In this research we use a previously introduced cross company (CC) learner (TEAK (Kocaguneli et al. 2012)) on new proprietary data sets and place it in the context of transfer learning. We think it is important to make this distinction, as transfer learning solutions may help important data problems in SE. Previous results show transferring effort estimation results is a challenging task:

- Kitchenham et al. reviewed 7 published transfer studies in effort estimation.[2] They found that in most (4) cases, transferred data generated worse predictors than using within-company information. This is a worrying finding for, e.g., start-up companies since it means they will not be able to make adequate effort estimates until after completing numerous local projects (and, in the meanwhile, they run the risk of projects failing due to inadequate effort allocations).
- In a similar result, Ye et al. report that the tunings to Boehm's COCOMO model have changed radically for new data collected in the period 2000 to 2009 (Yang et al. 2011).

[2]Terminology note: Kitchenham et al. called such transfers "cross-company" learning.

Note that effort estimation is a different problem to defect prediction:

- While defect data sets store information on hundreds to millions of methods,
- Effort data sets are smaller, often they are just a few dozen records.

Since the data sets are so different, we must use different methods to transfer effort estimation data. In prior work (Kocaguneli et al. 2012), we have found that a self-tuning analogy-based effort estimation method called TEAK out-performs other approaches such as linear regression, neural networks, and traditional analogy-based reasoners. Like the above work on defect prediction, TEAK uses relevancy filtering. TEAK explores outwards from each test instance, taking care not to enter "suspicious" regions where consensus breaks down and the training data offers wildly contradictory conclusions. To optimize that process, TEAK first builds a dendogram (a tree of clusters of project data), then computes the variance of the conclusions of the examples in each sub-tree.

This paper uses TEAK as laboratory for studying transfer learning in effort estimation. Using TEAK, we will explore the following research questions.

RQ1: Is transfer learning effective for effort estimation?

We will compare estimates learned from within a single stratification to those transferred across different stratifications.

RQ2: How useful are manual divisions of the data?

TEAK has a choice about where it finds its training data: Either in one stratification or in many. Hence it can assess the relative merits of automated vs manual groups of data. Automated groups are generated by methods like TEAK. The alternative to automatic grouping is manual grouping via *delphi localization*; i.e., the practice where (a) some human expert offers a division of the data then (b) effort models are built from just the data within each such localization.[3]

RQ3: Does transfer learning for effort estimation work across time as well as space?

We find that transfer learning is a unifying framework for two, previously distinct, research themes in software engineering. Turhan discusses the problem of *data set shift* (Turhan 2012) where, within one project, something has changed and the old data for that project no longer applies. This issue of changes within one project is not discussed in the *cross-company learning* literature in SE (Kitchenham et al. 2007; Turhan et al. 2009) since cross-company learning usually focuses on, say, transferring data across two divisions of one company found in two different physical locations. Note that, from the perspective of transfer learning, both problems are the same; i.e., the transfer data from domain1 to domain2 regardless of whether the domains are:

- from the same project at different times (i.e., the data set shift problem);
- or from different projects at similar times in different places (i.e., the cross company learning problem).

[3]Examples of delphi localizations come from Boehm (1981) and Petersen and Wohlin (2009). Boehm divided software projects into one of the "embedded", "semi-detached" or "organic" projects and offered different COCOMO-I effort models for each. Petersen & Wohlin offer a rich set of dimensions for contextualizing projects (processes, product, organization, etc).

This leads to the following conjecture: A transfer learner that works for cross-company learning can also be applied to transferring past data to current projects. To test that conjecture, this paper applies the same transfer learner *without modification* to traditional cross-company learning problems *and* data set shift problems.

We know of many researchers and managers that are so enamored with the fast pace of change in SE that they are willing to throw out conventional wisdom in favor of new ideas, even when those ideas are still immature and not fully investigated. This "everything must be new" mentality often effects our own work. As data mining researchers, we often struggle to find appropriate data for checking our proposed prediction methods. Previously, we have encountered some resistance to the use of seemingly "out-dated" data sets from last century such as COC-81 and NASA-93 for a 21st century paper on effort estimation. Therefore, in this paper, we check if any parts of older data sets have any relevancy on later projects.

Note that, from an economic perspective, this last point is quite important. Collecting data can be very expensive (we saw above that the NASA-93 was the result of a two million dollar project review exercise conducted by NASA). If any organization goes to all the trouble of building such repositories, it seems important and appropriate to understand the life expectancy of that data.

1.1 Contributions and Structure

In summary, the contributions of this study are:

- Building onto prior work, we use transfer learning as a solution to the important problem of local data issues for effort estimation.
- Supporting previous evidence (Lokan and Mendes 2009b; Minku and Yao 2012; Turhan 2012) by empirically showing that effort estimation data can be successfully transfered without performance loss across time and space.
- Evaluating the success of transfer learning on recent proprietary as well as public data sets; hence, providing evidence for practitioners as well as benchmark availability for further research.

This paper extends a prior publication as follows:

- In 2011, Kocguenli et al. offered a limited study on just cross-company learning with 8 data sets (Kocaguneli and Menzies 2011). This paper doubles the number of case studies *as well as* exploring issues of data set shift. That is, that prior publication explored learning across space while here we explore transfer across time *and* space.
- Furthermore, the data used in that prior study is somewhat dated while this paper uses more recent data from organizations building Web-based applications.
- This research uses more evaluation criteria including a recently proposed error measure (Shepperd and MacDonell 2012).
- Lastly, that prior aforementioned work did not recognize the connection of its research to transfer learning. This paper makes that connection via a more extensive literature review.

The rest of this paper is structured as follows. After some background notes, TEAK is evaluated on the proprietary data sets of 8 Web companies from the Tukutuku (Mendes and Mosley 2008) data base (for transfer between domains) as well as publicly available

NASA data sets called Cocomo81[4] and Nasa93[5] (for transfer between time intervals). In the experimentation, each test instance is evaluated in two different scenarios:

- In the first scenario, the test instance is allowed to use only within training data (i.e., restricted to its own domain or time interval).
- In the second scenario, the test instance is allowed to transfer data or use within-company data (i.e., it is allowed to transfer instances from other domains or time intervals). This second scenario study shows that there is a considerable amount of cross data that can be successfully transferred across time and space.

The performance of the test instances in both scenarios are compared according to 8 different error measures (details in Section 3.2) subject to Wilcoxon test (at 95 % confidence). In 6 out of 8 companies, transfer learning resulted in performance that is statistically identical to estimates generated from within-data experiments. In all cases of the transfer learning between time intervals, the within and transferred data performance were statistically the same.

2 Related Work

In this paper, we will refer to transfer learning and software effort estimation as TL and SEE (respectively).

2.1 Transfer Learning

A learning problem can be defined by:

- a specific domain D, which consists of a feature space and a marginal distribution defining this space;
- and a task T, which is the combination of a label space and an objective estimation function.

TL allows for the training and test data to have different domains and tasks (Ma et al. 2012). According to Jialin et al. TL can be formally defined as follows (Pan and Yang 2010): Assuming we have a source domain D_S, a source task T_S, a target domain D_T and a target task T_T; TL tries to improve an estimation method in D_T using the knowledge of D_S and T_S. Note that the assumption in the above definition is that $D_S \neq D_T$ or $T_S \neq T_T$. There are various subgroups of TL, which define the relationship between traditional machine learning methods and various TL settings, e.g., see Table 1 of (Pan and Yang 2010). SEE transfer learning experiments have the same task but different domains, which places them under the category of transductive TL (Arnold et al. 2007).

There are 4 different approaches to TL (Pan and Yang 2010): instance-transfer (or instance-based transfer) (Foster et al. 2010), feature representation transfer (Lee et al. 2007), parameter-transfer (Gao et al. 2008) and relational-knowledge transfer (Mihalkova et al. 2007). The TL approach of the estimation method used in this research corresponds to instance-transfer. The benefits of instance-transfer learning are used in various research

[4] http://goo.gl/WxGXv

[5] http://goo.gl/ioXDy

areas, e.g., Ma et al. use TL for cross-company defect prediction, where they use a weighted Naive Bayes classifier (Ma et al. 2012). Other research areas that benefit from instance-transfer are text classification (Dai et al. 2007), e-mail filtering (Zhang et al. 2007) and image classification (Wu and Dietterich 2004).

2.2 Transfer Learning and SE

TL between different time intervals has been paid very little attention in SEE. In Lokan and Mendes (2009b) Lokan and Mendes found out by using chronological sets of instances that time frame divisions of instances did not affect prediction accuracy. In Lokan and Mendes (2009a), they found out that it is possible to suggest a window size of a time frame of past instances, which can yield performance increase in estimation. They also note that the size of the window frame is data set dependent. Recently Minku and Yao pointed out an important issue with the SEE transfer learning studies (Minku and Yao 2012), which is that SEE is a *"class of online learning tasks"*; they have investigated several scenarios, testing different approaches designed for non-changing as well as changing environments in the context of CC learning. Their findings indicated that CC subsets may be beneficial or detrimental, depending on the moment in time. The algorithm reported in their work can benefit from time dependence and can perform better than WC data when using CC data (Minku and Yao 2012). Our research builds on the prior findings to provide evidence of knowledge transfer through both space and time.

The prior results on the performance of TL (a.k.a. cross-company learning in SEE) are unstable. In their review, Kitchenham et al. (2007) found equal evidence for and against the value of TL in SEE. Out of the 10 studies reviewed by Kitchenham et al., 4 studies favored within data, another 4 studies found that transferring data is not statistically significantly worse than within data, and 2 studies had inconclusive results. In the field of defect prediction, Zimmermann et al. studied the use of transferring data (Zimmermann et al. 2009). Zimmermann et al. found that predictors performed worse when trained on cross-application data than from within-application data. From a total of 622 transfer and within data comparisons, they report that within performed better in 618 cases. Recently Ma et al. defined using data across other domains in the research field of defect prediction as a TL problem (Ma et al. 2012). Ma et al. propose a Naive Bayes variant, so called Transfer Naive Bayes (TNB), so as to use all the appropriate features from the training data. TNB is proposed as an alternative TL method for defect prediction when there are too few training data. According to Ma et al. data transferring problem of defect prediction corresponds to an inductive TL setting; where source and target tasks are the same, yet source and target domains are different. The inductive TL methods are summarized as either instance transfer or feature transfer (Huang et al. 2007). The current literature of TL in defect prediction as well as SEE focuses on instance transfer.

Turhan et al. compared defect predictors learned from transferred or within data. Like Zimmermann et al., they found that transferring *all* data leads to poor estimation method performance (very large false alarm rates). However, after *instance selection* pruned away irrelevant data during TL, they found that the estimators built on transferred data were equivalent to the estimators learned from within data (Turhan et al. 2009). Motivated by Turhan et al. (2009), Kocaguneli et al. (2010) used *instance selection* as a pre-processor for a study of TL in SEE, where test instances are allowed to use only transferred or only within data. In a limited study with three data sets, they found that through instance selection, the performance differences in the predictors trained on transferred or within data were not statistically significant. This limited study was

challenged by Kocaguneli and Menzies in another study that uses 8 different data sets (Kocaguneli and Menzies 2011). The results were identical: performance differences of within and transferred data are not significant.

2.3 Data Set Shift

Turhan offers a formal treatment for this problem of "context change" in software engineering (Turhan 2012). He discusses *data set shift* which appears when training and test joint distributions are different.[6] Turhan investigates different types of data set shift issues as proposed by Storkey (2009)

- Covariate shift: The covariates in test and train sets differ;
- Prior probability shift: Prediction model is found via Bayes Rule, yet the distribution of the dependent variable differs for training and test sets;
- Sample selection bias: The training and test sets are selected from populations with different characteristics, e.g., training set coming from a higher maturity level company is used on a test set of a lower maturity level company;
- Imbalanced data: Certain event (or class) types of interest are rarer compared to other events (or classes);
- Domain shift: The cases where the method of measurement (e.g., performance measures, size measures) changes between training and test conditions.
- Source component shift: Parts of data come from different sources with particular characteristics in varying sizes, e.g., data sets whose instances are collected in different companies or time frames.

Turhan continues his discussion by introducing techniques to handle data set shift problems in software engineering, which are grouped under two main groups: Instance-based techniques and distribution-based techniques. The former group of techniques aim at handling instances (through outlier removal (Turhan et al. 2009), relevancy filtering (Kocaguneli and Menzies 2011) or instance weighting (Zhang and Sheng 2004)), whereas the latter group of techniques aim at regulating the distribution of instances to train and test sets (through stratification (Turhan 2012), cost curves (Jiang et al. 2008) and mixture models (Alpaydin 2010)).

Quoting from Storkey: "Dataset shift and transfer learning are very related." Transfer learning deals with the cases, where there are multiple training scenarios that are partially related and are used to predict in one specific scenario. For example, in the case of our research multiple training scenarios are either data from different companies (transfer of data in space) or data from different time frames (transfer of data in time). Dataset shift is the specific case, where there are only two scenarios (training and test sets) and one of them has no training targets (Storkey 2009). Turhan's mapping of the data set shift types to software engineering is an excellent example of specific transfer learning issues that are becoming imminent.

In the context of SEE, Lokan and Mendes; as well as Minku and Yao investigated the data set shift problem (Lokan and Mendes 2009b; Minku and Yao 2012). The work of

[6]Note that the literature contains numerous synonyms for data set shift including "concept shift" or "concept drift", "changes of classification", "changing environments", "contrast mining in classification learning","fracture points" and "fractures between data". We will use the term as defined in the above text.

Lokan and Mendes posed the question of whether it is better to use all the available past data as the training set or whether it is better to use a subset (window) of past projects. They employed data from the ISBSG database, and their findings showed that using a window can significantly improve estimation accuracy, when compared to using the entire training set. Furthermore, they also point out that the size of the window (i.e., the number of projects to be included as training set) plays an important role in the estimation accuracy and it is data set dependent (Lokan and Mendes 2009b). Minku and Yao's work observes the SEE problem as a class of online learning problems, where the environment is changing (Minku and Yao 2012). Their results show that the estimation accuracy when using CC data is dependent on time, i.e., the CC subsets can be either beneficial or detrimental to estimation accuracy, depending on the moment in time.

In this research, we provide a more general treatment of the transfer learning and propose that our earlier work that we called "cross company learning" (Kocaguneli and Menzies 2011) as well as specific data set issues are in fact particular cases of transfer learning (see Fig. 1). One of the main ideas of this research is that separate problems of cross company and time frame estimation are problems of transfer learning and the data set shift problem is closely related to transfer learning. Hence, instead of proposing different solutions for each specific case of each specific problem, we should be looking for general transfer learning solutions that work for multiple cases in both problem types. Strictly speaking, data set shift and transfer learning problems are not subsets of one another, yet they are closely related. Hence, we believe that CC studies in SE can benefit from both domains (transfer learning and data set shift) considerably.

Data set shift problem differs from the transfer learning problem in the sense that it refers to a non-stationary learning problem (Žliobaitė 2010), where the core assumption is that future involves uncertainty. The main distinction is the fact that the source of a future target instance is not known with certainty, i.e., data cannot be simply decomposed into two disjoint sets and learned with two separate models (Žliobaitė 2010). The idea of transfer learning is to assume the two data sets have different distributions, yet one of them may be used to aid the other; whereas, in the case of data set shift, the changes are detected and the system is adapted to new distributions (instead of the mere use of the other data set).

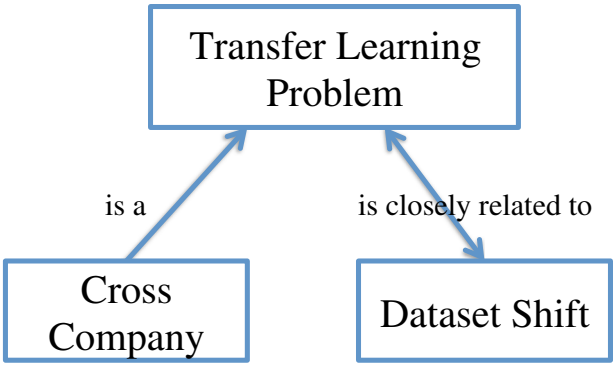


Fig. 1 Prior studies of cross company studies are specific cases of transfer learning, which is very closely related to data set shift problem

3 Methodology

3.1 Dataset

This study uses the Tukutuku data base for TL between domains (Mendes and Mosley 2008). Tukutuku brings together data sets coming from a high number of companies. The version used in this research is composed of a total of 195 projects developed by a total of 51 companies. However, not all the companies in the data set are useful for TL analysis. We eliminated all the companies with less than 5 projects, which yielded 125 projects from 8 companies. The abbreviations used for the 8 companies that were selected and their corresponding number of projects are given in Fig. 2.

The Tukutuku data base is an active project, which is maintained by Emilia Mendes and it includes information collected from Web projects (Mendes et al. 2005). These projects come from a total of 10 different countries around the world (Corazza et al. 2010). Tukutuku data base is characterized by a total of 19 independent variables and a dependent variable. The feature descriptions of the Tukutuku data set are given in (Mendes et al. 2005). The dependent variable is the total effort in person hours.

So as to see the TL performance and selection tendency between time intervals, we used 2 data sets: Cocomo81 and Nasa93. Each of these data sets are divided into 2 subsets of different time periods. The subsets of Cocomo81 are: coc-60-75 and coc-76-rest, where *coc* stands for Cocomo81, *60-75* stands for projects developed from 1960 to 1975 and *76-rest* stands for projects developed from 1976 onwards. It is possible to have different divisions of the data depending on different time frames. We select these particular divisions of time periods so that both subsets span a certain amount of time (e.g., more than a decade) and yet have at least 20 instances. The subsets of nasa93 are: nasa-70-79 and nasa-80-rest. The naming convention is similar to that of Cocomo81 subsets: *nasa* stands for Nasa93 dataset, *70-79* stands for projects developed in the time period of 1970 to 1979 and *80-rest* stands for projects developed from 1980 onwards. The details of these projects are provided in Fig. 3.

Company	# of Projects
tuku1	14
tuku2	20
tuku3	15
tuku4	6
tuku5	13
tuku6	8
tuku7	31
tuku8	18

Fig. 2 The abbreviations for selected companies and the number of projects

Dataset	Features	Size	Description
coc-60-75	17	20	Nasa projects between 1960 and 1975
coc-76-rest	17	43	Nasa projects from 1976 onwards
nasa-70-79	17	39	Nasa projects between 1970 and 1979
nasa-80-rest	17	54	Nasa projects from 1976onwards
		Total : 156	

Fig. 3 Data sets for transfer learning over time period

3.2 Performance Measures

There are multiple performance measures (a.k.a. error measures) used in SEE, which all aim to measure the success of a prediction. In this research, we use a total of 8 performance measures. For example, the absolute error (AE) is the absolute difference between the predicted and the actual values:

$$AE_i = |x_i - \hat{x_i}| \tag{1}$$

(where $x_i, \hat{x_i}$ are the actual and predicted value for test instance i). We use a summary of AE through taking the mean of AE, which is known as Mean AE (MAE).

The Magnitude of Relative Error measure a.k.a. MRE is a very widely used performance measure for selecting the best effort predictor from a number of competing software prediction models(Shepperd and Schofield 1997; Foss et al. 2003). MRE measures the error ratio between the actual effort and the predicted effort and is expressed by the following equation:

$$MRE_i = \frac{|\ x_i - \hat{x_i}\ |}{x_i} = \frac{AE_i}{x_i} \tag{2}$$

A related measure is MER (Magnitude of Error Relative to the estimate (Foss et al. 2003)):

$$MER_i = \frac{|\ x_i - \hat{x_i}\ |}{\hat{x_i}} = \frac{AE_i}{\hat{x_i}} \tag{3}$$

The overall average error of MRE can be derived as the Mean or Median Magnitude of Relative Error measure (MMRE and MdMRE, respectively):

$$MMRE = mean(allMRE_i) \tag{4}$$

$$MdMRE = median(allMRE_i) \tag{5}$$

A common alternative to MMRE is PRED(25), which is defined as the percentage of successful predictions falling within 25 % of the actual values, and can be expressed as follows, where N is the dataset size:

$$PRED(25) = \frac{100}{N} \sum_{i=1}^{N} \begin{cases} 1 \text{ if } MRE_i \leq \frac{25}{100} \\ 0 \text{ otherwise} \end{cases} \tag{6}$$

For example, PRED(25) = 50 % implies that half of the estimates fall within 25 % of the actual values (Shepperd and Schofield 1997).

Other performance measures used here are Mean Balanced Relative Error (MBRE) and the Mean Inverted Balanced Relative Error (MIBRE), both suggested by Foss et al. (2003)

$$MBRE_i = \frac{|\hat{x_i} - x_i|}{min(\hat{x_i}, x_i)} \tag{7}$$

$$MIBRE_i = \frac{|\hat{x_i} - x_i|}{max(\hat{x_i}, x_i)} \tag{8}$$

The above mentioned performance measures are selected due to their wide use in the SEE research. However, none of these error measures are devoid of problems. For instance,

MRE-based error measures have been criticized due to their asymmetry (Foss et al. 2003). This criticism applies to MMRE, MdMRE and Pred(25). The matter of providing an in depth discussion about error measures is out of our scope in this paper. Rather than going into the debate of which error measure is better than the others, we evaluated our results subject to a total of 8 error measures. A recent study by Shepperd and MacDonell provides an excellent discussion of the error measures (Shepperd and MacDonell 2012). In this study Shepperd and MacDonell propose a new unbiased error measure called Standardized Accuracy (SA), which is based on the mean absolute error (MAE). SA's equation is as follows:

$$SA = 1 - \frac{MAE_{P_i}}{\overline{MAE_{P_0}}} \tag{9}$$

MAE_{P_i} is defined to be the MAE of the estimation method P_i. $\overline{MAE_{P_0}}$ is the mean of a large number of (in our case 1000) random guessing. In the random guessing procedure a training instance is randomly chosen with equal probability from the training set (with replacement) and its effort value is used as the estimate of the test instance. SA gives us an idea of how good an estimation method is in comparison to random guessing. Since the term MAE_{P_i} is in the nominator, the higher the SA values, the better an estimation method.

3.3 Instance Selection and Retrieval

One of our goals in this research is to find the selection tendency of test instances, i.e., the percentages of training instances (with respect to the training set size) selected from within or transferred data. This investigation will help us identify what percent of the data across domain or time frame is transferred for a test instance, given the chance that it can select either from within or transferred instances. If test instances were to select mostly from the within domain or within time interval instances (i.e., transferred instances are not much used), then there would not be much need for a study like this; because, within sources would turn out to be the most relevant instances to test instances. However, as we will see in our results section, that is not the case.

Inspecting the selection tendency of an estimation method for test instances makes sense only if the estimation method is a state-of-the-art learner. Because, inspection of the selected instances has its merits, provided that it leads to estimates that are as good as that of the best-of-breed estimation methods. Therefore, we selected to use a variance-based instance selection method that was previously shown to successfully work on TL experiments (across space) on public data sets. In this section we briefly introduce the method used as a TL solution, called TEAK (Kocaguneli et al. 2012). Then we quote prior research to show that TEAK's performance is comparable to or better than various other estimation methods.

3.3.1 ABE0

Analogy-based estimation (ABE) methods generate their estimates by using a data base of past projects. For a *test* project, ABE methods retrieve analogies from a database of past projects. Then the effort values of the retrieved analogies are adapted into an estimate. We use ABE methods in this study since 1) they are widely investigated methods in the literature (Mendes et al. 2003; Chang 1974; Keung 2008; Li et al. 2009; Kadoda et al. 2000; Kocaguneli et al. 2010, 2012), 2) they are particularly helpful for TL studies as they are based on distances between individual project instances.

There is a high number of design options concerning ABE methods such as the distance measure for nearness (Mendes et al. 2003), adaptation of analogy effort values

(Mendes et al. 2003), row processing (Chang 1974; Keung 2008), column processing (Keung 2008; Li et al. 2009) and so on. For example, Keung et al. show that the number of different design options can easily lead to more than 6000 ABE variants (Keung et al. 2012). Here we define ABE0 that is a *baseline* ABE method that combines the methods used in Kadoda et al. (2000), Mendes et al. (2003) and Li et al. (2009):

- Input a database of past projects
- For each test instance, retrieve k similar projects (analogies).
  - For choosing k analogies use a similarity measure.
  - Before calculating similarity, scale independent features to 0-1 interval so that higher numbers do not dominate the similarity measure.
  - Use a feature weighting scheme to reduce the effect of less informative features.
- Adapt the effort values of the k nearest analogies to come up with the effort estimate.

ABE0 uses the Euclidean distance as a similarity measure, detailed in Eq. (10), where w_i corresponds to feature weights applied on independent features. The ABE0 framework does not favor any features over the others, therefore each feature has equal importance in ABE0, i.e., $w_i = 1$. For adaptation, ABE0 takes the median of k projects.

$$Distance = \sqrt{\sum_{i=1}^{n} w_i (x_i - y_i)^2} \tag{10}$$

3.3.2 TEAK

TEAK is a *variance-based* instance selector that discards training data associated with regions of high dependent variable (effort) variance (Kocaguneli et al. 2012). TEAK augments ABE0 with instance selection and an indexing scheme for filtering relevant training examples. In summary, TEAK is a two-pass system:

- **Pass 1** prunes training instances implicated in poor decisions (instance selection): In the first pass, training instances are combined using greedy-agglomerative clustering (GAC), to form an initial cluster tree that we call GAC1; e.g., see Fig. 4. Level zero of GAC1 is formed by leaves, which are the individual project instances. These instances are greedily combined (combine the two closest instances) into tuples to form the nodes of upper levels. Then it prunes the subtrees with high dependent variable (effort) variance.
- **Pass 2** retrieves closest instances to the test instance (instance retrieval): The leaves of the remaining sub-trees are the *survivors* of pass one. They are filtered to pass 2 where they are used to build a second GAC tree (GAC2). GAC2 is generated in a similar fashion to GAC1, then it is traversed by test instances that are moved from root to leaves. When test instances stop at a subtree, the effort value of the closest training instance in the subtree is returned as the estimate.

TEAK is a form of ABE0, so its adaptation method is the same, i.e., take the median of the analogy effort values. A simple visualization of this approach is given in Fig. 5. A very detailed explanation of TEAK can be found in (Kocaguneli et al. 2012).

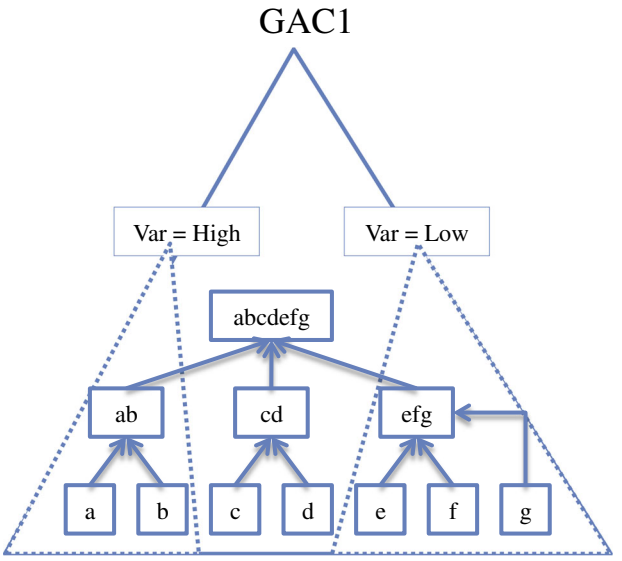


Fig. 4 A sample GAC tree with regions of high and low variance (*dashed triangles*). GAC trees may not always be binary. For example here, leaves are odd numbered, hence node "g" is left *behind*. Such instances are pushed *forward* into the closest node in the higher level. For example, "g" is pushed forward into the "e+f" node to make "e+f+g" node

We use TEAK in this study since it was shown by the leave-one-out experiments of Kocaguneli et al. (2012) that its performance is comparable, or even better, than other commonly-used effort estimators including neural networks (NNet) and linear regression (LR).

That study evaluated different SEE methods using the $win - loss$ calculation of Fig. 6. We first check if two distributions i, j are statistically different according to the Wilcoxon test. In our experimental setting, i, j are arrays of performance measure results coming from two different methods. If they are not statistically different, then they are said to *tie* and

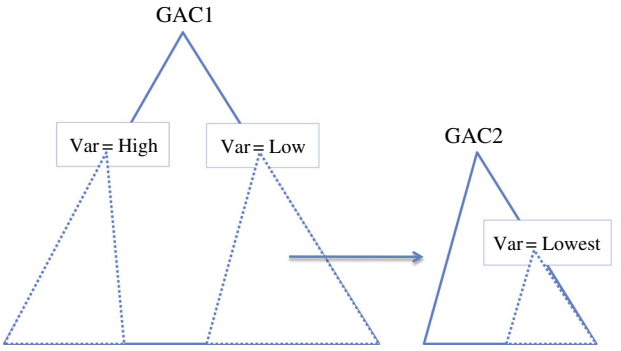


Fig. 5 Execution of TEAK on 2 GAC trees, where the tree on the *left* is GAC1 of Fig. 4 and the one on the *right* is GAC2. The instances in the low variance region of GAC1 are selected to form GAC2. Then test instance traverses GAC2 until no decrease in effort variance is possible. Wherever the test instance stops is retrieved as the subtree to be used for adaptation (var = lowest labeled, dashed triangle of GAC2)

```
win_i = 0, tie_i = 0, loss_i = 0
win_j = 0, tie_j = 0, loss_j = 0
if Wilcoxon(Perf_i, Perf_j) says they are the same then
   tie_i = tie_i + 1;
   tie_j = tie_j + 1;
else
   if Perf_i < Perf_j then
      win_i = win_i + 1
      loss_j = loss_j + 1
   else
      win_j = win_j + 1
      loss_i = loss_i + 1
   end if
end if
```

Fig. 6 Pseudo code for win-tie-loss calculation between methods *i* and *j* w.r.t. performance measures $Perf_i$ and $Perf_j$. If $Perf_i$ and $Perf_j$ are measures like MMRE, MdMRE or MAE, *lower* values are better, whereas for performance measures like Pred(25) *higher* values are better

we increment tie_i and tie_j. On the contrary, if they are different, we update win_i, win_j and $loss_i$, $loss_j$ after a numerical comparison of performance measures. The related pseudo code is given in Fig. 6. In order to reduce any possible bias due to a particular experimental setting, for every experiment 20 runs are made.

In the study of Kocaguneli et al. (2012), TEAK always performed better than the other ABE0 methods, and mostly performed better than neural nets. TEAK's only near-rival was linear regression but (using Fig. 6) TEAK was ranked top nearly twice as much as linear regression.

3.4 Experimentation

The goals of the experiments are set to align with the research questions formed in Section 1. The first goal of the experimentation is aligned with *RQ1* and *RQ2*, where we question the retrieval tendency and the performance. The *retrieval tendency* questions the tendency of a *within* test instance to retrieve *within* or *transferred* data. In other words, given the chance that a test instance had access to *within* and *transferred* data at the same time, what percentage of every subset would be retrieved into *k* analogies used for estimation? Our second goal is aligned with *RQ3* and aims to answer whether TL can enable the use of data from other organizations as well as from other time intervals. For our experimental goals, we employ TEAK as a TL method and compare its performance when trained from just within data versus when trained from a combination of transferred and within data.

Note that besides the above-mentioned experimental goals (i.e., retrieval tendency and performance), one may also question the following case: TEAK uses only CC data and only WC data, then its WC and CC performances are compared. The starting point of this study was in fact that very case (Kocaguneli et al. 2010). In our 2010 study, we solely focused our attention on estimation performance and for the majority of the data sets TEAK's performance (when only CC data is used) was statistically the same to its performance (when only WC data is used). We therefore decided to go one step further and to also focus on the retrieval tendency and let TEAK use WC and CC subsets as it sees fit. This is the approach employed herein.

3.4.1 Performance Comparison

With regard to performance comparison we have two settings: *Within* and *transfer*. In the *within* data setting, only the *within* source is used as the dataset, and a testing strategy of leave-one-out cross-validation (LOOCV) is employed. LOOCV works as follows: Given a *within* dataset of T projects, 1 project at a time is selected as the test and the remaining $T - 1$ projects are used for training, so eventually we have T predictions. The resulting T predictions are then used to compute 8 different performance measures defined in Section 3.2.

Transfer setting uses one instance at a time (since we use LOOCV) from the *within* data as the test set and the *combination of the remaining within instances and all the transferred* data as the training set. In this setting an estimate for each test instance is found via the transfer of the analogies from the training set by TEAK. Ultimately we end up with T predictions adapted from the analogies that are transferred from a training set of transferred and within instances. Finally, the performances under *within* and *transfer* settings are compared. For that purpose, we use both mere performance values as well as win-tie-loss statistics.

3.4.2 Retrieval Tendency

For retrieval tendency experiments we mark every *within* and *transferred* instance in the training set and let the test instance choose analogies from both groups of training instances. Note that retrieved analogies are the unique training instances in the lowest-variance region of GAC2 (see Fig. 5). In this setting our aim is to see what percentage of *within* and *transferred* data would appear among retrieved k analogies, i.e., how much of transferred data is relevant? The retrieval percentage is the average ratio of instances retrieved in analogies to the total size of the training set:

$$Percentage = \frac{Number\ Of\ Retrieved\ Analogies}{Training\ SPACs\ Set\ Size} * 100 \tag{11}$$

4 Results

4.1 Transfer in Space

Comparing the performance of within and transferred data is the first goal of the experimentation. The result of this goal will tell us whether TL can enable instances transfer between domains and time intervals without loss from the performance, or whether the transfer comes at the expense of higher error rates. Due to space constraints and also in order to make the results more readable we equally divided the domain transfer results of the Tukutuku subsets into two figures: Figs. 7 and 8. The time interval transfer results of Cocomo81 and Nasa93 are provided in Figs. 11 and 12, respectively.

Figure 7 shows a uniformity of results. The tie values are very high for 5 out of 8 companies (tuku1, tuku4-to-7), which means that transferred data performance is as good as within data performance. For 1 company, tuku8, within and transferred data performance depends on the error measure: According to all error measures except MMER the within and transferred data performances are very close, whereas for MMRE the within data performance appears to be better. For 2 companies out of 8 (tuku2 and tuku3), the within data performance is dominantly better with a win value of 20. Remember from Section 3.3.2 that

Fig. 7 Performance comparison of within vs transferred data w.r.t. 4 of 8 different performance measures: MAE, MMRE, MdMRE, Pred(25). Win, tie, loss values are w.r.t. within data

Dataset	MAE		
	Win	**Tie**	**Loss**
tuku1	8	12	0
tuku2	20	0	0
tuku3	20	0	0
tuku4	2	18	0
tuku5	0	20	0
tuku6	8	12	0
tuku7	2	18	0
tuku8	1	19	0

Dataset	MMRE		
	Win	**Tie**	**Loss**
tuku1	9	11	0
tuku2	20	0	0
tuku3	20	0	0
tuku4	2	18	0
tuku5	0	19	1
tuku6	5	15	0
tuku7	3	17	0
tuku8	0	20	0

Dataset	MdMRE		
	Win	**Tie**	**Loss**
tuku1	8	12	0
tuku2	20	0	0
tuku3	20	0	0
tuku4	2	18	0
tuku5	0	19	1
tuku6	5	15	0
tuku7	3	17	0
tuku8	0	20	0

Dataset	Pred(25)		
	Win	**Tie**	**Loss**
tuku1	6	10	4
tuku2	18	0	2
tuku3	20	0	0
tuku4	2	18	0
tuku5	1	19	0
tuku6	5	15	0
tuku7	3	17	0
tuku8	0	20	0

Fig. 8 Performance comparison of within vs transferred data w.r.t. 4 of 8 different performance measures: MMER, MBRE, MIBRE, SA. Win, tie, loss values are w.r.t. within data

Dataset	MMER		
	Win	**Tie**	**Loss**
tuku1	0	20	0
tuku2	0	20	0
tuku3	20	0	0
tuku4	0	20	0
tuku5	6	14	0
tuku6	8	12	0
tuku7	6	14	0
tuku8	14	6	0

Dataset	MBRE		
	Win	**Tie**	**Loss**
tuku1	6	14	0
tuku2	20	0	0
tuku3	20	0	0
tuku4	1	19	0
tuku5	0	20	0
tuku6	6	14	0
tuku7	5	15	0
tuku8	2	18	0

Dataset	MIBRE		
	Win	**Tie**	**Loss**
tuku1	2	18	0
tuku2	17	3	0
tuku3	20	0	0
tuku4	1	19	0
tuku5	0	20	0
tuku6	6	14	0
tuku7	4	16	0
tuku8	3	17	0

Dataset	SA		
	Win	**Tie**	**Loss**
tuku1	8	12	0
tuku2	20	0	0
tuku3	20	0	0
tuku4	2	18	0
tuku5	0	20	0
tuku6	8	12	0
tuku7	2	18	0
tuku8	1	19	0

TEAK performs 20 times LOOCV, hence the total of win, tie and loss values for each data set subject to each error measure amounts to 20.

Figure 8 shows the within and transferred data performance comparison for the error measures of MMER, MBRE, MIBRE and SA. The reading of Fig. 8 is exactly the same as of Fig. 7, i.e., it shows the win, tie, and loss values according to 4 error measures. The general pattern we have observed from 4 error measures in Fig. 7 are also visible in Fig. 8. In relation to the error measures MBRE, MIBRE and SA, in 6 data sets (tuku1, tuku2, tuku4-to-7) transferred and within data performances are the same. According to the MMER, within performance is better for 2 data sets: tuku3 and tuku8 (Fig. 9).

The aforementioned results are parallel to the prior results reported by Kocaguneli and Menzies (2011), where they have used 21 public data sets and questioned merely the TL between domains. A summary of their results on 21 public data sets are provided in Fig. 10. As can be seen in Fig. 10, Kocaguneli and Menzies use 4 error measures and identify only 2 data sets (gray highlighted rows) for which within data performance is better than that of transferred data. For $21 - 2 = 19$ data sets, transferred data performance is shown to be statistically significantly the same as that of within.

So as to offer a statistical analysis that has some precedent in the literature, we preferred the win, tie, loss values as reported in Figs. 7 and 8. In Fig. 9 we also report the mean as well as variance of SA values (over 20 runs) for the Tukutuku as well as Cocomo81 and Nasa93 subsets as they provide a comparison w.r.t. random guess (see the denominator of Formula (9) so as to see how random guess is introduced into SA). Note in Fig. 9 that the actual numbers are aligned with the win, tie, loss values reported earlier. Yet, Fig. 9 also shows that some of the data sets are challenging both for within as well as cross company learning, e.g., for tuku3, both within as well as cross learning perform poorly.

4.2 Transfer in Time

Figures 11 and 12 show the performance results of TL in time intervals for Cocomo81 and Nasa93. For Cocomo81, two within sources are defined: 1) projects developed from 1960 to 1975 (called as coc-60-75) and 2) projects developed from 1976 onwards (called as coc-76-rest). Similarly, the subsets of Nasa93 are: 1) projects from 1970 to 1979 (called as nasa-70-79) and 2) projects from 1980 onwards (called as nasa-80-rest). In both Figs. 11 and 12, the tie values are quite high with the smallest tie value of 16. Note that in none of the two figures there is a highlighted row, which means that in none of the time interval instance

	WC		CC	
Subset	**Mean**	**Variance**	**Mean**	**Variance**
tuku1	0.28	0.03	0.20	0.62
tuku2	0.11	0.00	-5.99	5.77
tuku3	-0.08	0.03	-2.54	0.01
tuku4	0.15	0.00	0.52	0.04
tuku5	0.22	0.00	0.20	0.00
tuku6	0.17	0.00	0.17	0.78
tuku7	0.35	0.00	0.31	0.00
tuku8	0.15	0.02	0.11	0.01
coc-60-75	0.26	0	0.25	0.01
coc-76-res	0.23	0	0.23	0.01
nasa-70-79	0.11	0	0.07	0
nasa-80-rest	0.19	0.01	0.2	0.01

Fig. 9 The mean and the variance of SA values over 20 runs for Tukutuku, Cocomo81 and Nasa93 subsets

Dataset	MAE			MMRE			MdMRE			Pred(30)		
	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss
cocomo81e	0	20	0	0	16	4	4	16	0	4	16	0
cocomo81o	0	20	0	2	18	0	2	18	0	2	18	0
cocomo81s	18	2	0	15	5	0	15	5	0	13	5	2
nasa93_center_1	0	20	0	0	20	0	0	20	0	0	20	0
nasa93_center_2	4	16	0	2	18	0	2	18	0	2	18	0
nasa93_center_5	0	20	0	0	12	8	8	12	0	8	11	1
desharnais L1	11	9	0	9	11	0	9	11	0	9	11	0
desharnais L2	0	20	0	0	20	0	0	20	0	0	20	0
desharnais L3	0	20	0	2	18	0	2	18	0	2	18	0
finnish App Type1	0	20	0	0	20	0	0	20	0	0	20	0
finnish App Type 2345	0	20	0	0	17	3	0	17	3	0	17	3
kemerer Hardware1	0	0	20	0	20	0	0	20	0	0	20	0
kemerer Hardware 23456	0	20	0	0	20	0	0	20	0	0	20	0
maxwell App Type1	6	14	0	1	19	0	1	19	0	0	19	1
maxwell App Type2	0	18	2	0	19	1	1	19	0	0	19	1
maxwell App Type3	0	20	0	1	19	0	1	19	0	1	19	0
maxwell Hardware2	0	20	0	0	20	0	0	20	0	0	20	0
maxwell Hardware3	0	20	0	0	20	0	0	20	0	0	20	0
maxwell Hardware5	0	20	0	0	20	0	0	20	0	0	20	0
maxwell Source1	6	14	0	1	19	0	1	19	0	1	19	0
maxwell Source2	0	20	0	0	20	0	0	20	0	0	20	0

Fig. 10 Summary of prior TEAK results (Kocaguneli and Menzies 2011) on 21 data sets, within vs transferred (i.e., win, tie, loss values are w.r.t. within data.). For 19 data sets, transferred and within data performances are the same (note high tie values)

transfer experiments was there a case where we failed to transfer instances between different time intervals. The implications of within and transferred data experiments through instance transfer in time intervals are important for practitioners. TL results on Cocomo81 and Nasa93 subsets show that TL methods, may help companies use aged data sets by identifying which instances are still relevant, hence can be transfered to contemporary estimation tasks. Figures 11 and 12 fundamentally show that decades of time difference can be crossed with the help of instance transfer.

The results of the TL research reported herein on proprietary data sets combined with the prior results of public data sets (Kocaguneli and Menzies 2011) give us a broader picture of the transferred data performance. By using instance transfer methods, such as TEAK between domains, we have:

– 5 out of 8 proprietary data sets (results of this study);
– 19 out of 21 public data sets (results of prior study);

where the performance difference between within and transferred data is not statistically significant. This shows us that from a total of $21 + 8 = 29$ public and proprietary data sets, transferred data performs as well as within data for $5 + 19 = 24$ cases. Also, as for the TL between time intervals, we have 2 data sets subject to 8 error measures ($2 \times 8 = 16$ cases), where transferred data performance is always the same as within data performance.

4.3 Inspecting Selection Tendencies

The second goal of our experimentation is to observe the selection tendencies of the test instances. Figure 13 shows what percentage of instances are selected from within data

Fig. 11 Performance comparison of Cocomo81 subsets for transfer learning in time. Win, tie, loss values are w.r.t. within data

Dataset	MAE		
	Win	**Tie**	**Loss**
coc-60-75	0	20	0
coc-76-rest	0	20	0
Dataset	**MMRE**		
	Win	**Tie**	**Loss**
coc-60-75	0	19	1
coc-76-rest	0	20	0
Dataset	**MdMRE**		
	Win	**Tie**	**Loss**
coc-60-75	0	19	1
coc-76-rest	0	20	0
Dataset	**Pred(25)**		
	Win	**Tie**	**Loss**
coc-60-75	0	19	1
coc-76-rest	0	20	0
Dataset	**MMER**		
	Win	**Tie**	**Loss**
coc-60-75	0	20	0
coc-76-rest	0	20	0
Dataset	**MBRE**		
	Win	**Tie**	**Loss**
coc-60-75	1	19	0
coc-76-rest	0	20	0
Dataset	**MIBRE**		
	Win	**Tie**	**Loss**
coc-60-75	0	19	1
coc-76-rest	0	20	0
Dataset	**SA**		
	Win	**Tie**	**Loss**
coc-60-75	0	20	0
coc-76-rest	0	20	0

sources (diagonal cells) as well as the amount of the transferred data (off diagonal cells) in TL experiments between domains. The first column of Fig. 13 shows the within data sources (8 different companies) as well as their sizes in parenthesis. The second column shows the number of analogies retrieved from GAC2 on average over 20 runs. For each row, the columns *tuku1-to-8* show how the number of analogies in the second column is distributed to each data source. The values outside the parenthesis in each cell of columns *tuku1-to-8* are the number of analogies selected from that data source; the percentage value of that number w.r.t. the second column is given inside the parenthesis. Figures 14 and 15 show the selection tendency of test instances for Cocomo81 and Nasa93 time interval TL experiments. These figures are structured in the same manner as Fig. 13.

The selection tendency values of Tukutuku, Cocomo81 and Nasa93 subsets provide us with suggestions regarding how much data a test instance uses from within and transferred data during the domain and time interval transfer of training instances. From the selection tendency figures, we can identify two findings:

Fig. 12 Performance comparison of Nasa93 subsets for transfer learning in time. Win, tie, loss values are w.r.t. within data

Dataset	MAE		
	Win	**Tie**	**Loss**
nasa-70-79	0	20	0
nasa-80-rest	4	16	0
Dataset	**MMRE**		
	Win	**Tie**	**Loss**
nasa-70-79	0	20	0
nasa-80-rest	4	16	0
Dataset	**MdMRE**		
	Win	**Tie**	**Loss**
nasa-70-79	0	20	0
nasa-80-rest	4	16	0
Dataset	**Pred(25)**		
	Win	**Tie**	**Loss**
nasa-70-79	0	20	0
nasa-80-rest	4	16	0
Dataset	**MMER**		
	Win	**Tie**	**Loss**
nasa-70-79	4	16	0
nasa-80-rest	0	20	0
Dataset	**MBRE**		
	Win	**Tie**	**Loss**
nasa-70-79	2	18	0
nasa-80-rest	2	18	0
Dataset	**MIBRE**		
	Win	**Tie**	**Loss**
nasa-70-79	2	18	0
nasa-80-rest	2	18	0
Dataset	**SA**		
	Win	**Tie**	**Loss**
nasa-70-79	0	20	0
nasa-80-rest	4	16	0

Finding #1: See the second columns of Figs. 13, 14 and 15 that only a very small portion of all the available data is transferred as useful analogies. This finding points to the importance of: a) instance transfer, a.k.a filtering; b) TL methods like TEAK that are capable of transferring relevant analogies between domains and time intervals. The low number of instances selected are also supported by relevant literature: Chang's prototype generators (Chang 1974) replaced training sets of size $T = (514, 150, 66)$ with prototypes of size $N = (34, 14, 6)$ (respectively). That is, prototypes may be as few as $\frac{N}{T} = (7, 9, 9)$ % of the original data. Note that these values are close to how many instances were retrieved in the above results.

Finding #2: When we compare the diagonal and off-diagonal percentages, we see that the values are very close. The second finding bears importance regarding whether or not the TL is worthy of investigation. We see that the amount of data transfered from other domains or time intervals is as much as the instances used from the within data sources. In other words, there is a considerable amount of data that is transfered as relevant across time and space. In addition (recall from the previous subsection) the high amount of transfered data

Dataset	# of Analogies	tuku1	tuku2	tuku3	tuku4	tuku5	tuku6	tuku7	tuku8
tuku1(14)	5.8	0.5(3.6)	1.0(7.1)	0.1(0.7)	0.1(0.7)	0.4(2.9)	0.4(2.9)	1.7(12.1)	1.6(11.4)
tuku2(20)	11.0	1.1(5.5)	1.6(8.0)	0.3(1.5)	0.8(4.0)	0.7(3.5)	0.7(3.5)	2.7(13.5)	3.1(15.5)
tuku3(15)	7.3	0.9(6.0)	1.3(8.7)	0.4(2.7)	0.2(1.3)	0.8(5.3)	0.2(1.3)	1.5(10.0)	1.9(12.7)
tuku4(6)	6.7	0.6(10.0)	1.5(25.0)	0.2(3.3)	0.3(5.0)	0.2(3.3)	0.7(11.7)	1.2(20.0)	2.0(33.3)
tuku5(13)	9.3	2.4(18.5)	1.4(10.8)	0.3(2.3)	0.5(3.8)	0.5(3.8)	0.4(3.1)	1.3(10.0)	2.4(18.5)
tuku6(8)	8.9	0.7(8.8)	1.6(20.0)	0.2(2.5)	0.7(8.8)	0.7(8.8)	0.7(8.8)	1.8(22.5)	2.6(32.5)
tuku7(31)	7.8	1.2(3.9)	1.3(4.2)	0.3(1.0)	0.5(1.6)	0.6(1.9)	0.1(0.3)	1.7(5.5)	2.1(6.8)
tuku8(18)	6.9	1.1(6.1)	1.1(6.1)	0.3(1.7)	0.3(1.7)	0.7(3.9)	0.3(1.7)	1.3(7.2)	1.8(10.0)

Fig. 13 The amount of within and transferred data selected. The first column is the subset names and their sizes in parenthesis. The second column is the average number of retrieved instances in GAC2. The following columns show the number of analogies from each particular subset, in parenthesis the percentage values of these numbers w.r.t. the first column are given

does not come at the expense of higher error rates. This shows us that discarding aged data sets because they are old is a misguided practice. There are relevant instances in old data sets that are relevant to current practices.

To better observe how close within and transferred data percentages are, we plot the percentage values of within and transferred data sources of Tukutuku subsets in Fig. 16a and b, respectively. Figure 16a and b are basically the plots of diagonal and off-diagonal percentage values of Fig. 13. We see from Fig. 16a and b show that the percentages of within and transferred data selection have very close values. To align these percentages, we took the percentile values from 0^{th} percentile to 100^{th} percentile with increments of 10. The resulting percentiles are shown in Fig. 16c. See in Fig. 16c that within and transferred data have similar percentile values, i.e., the selection tendencies are very close to one another. Note that percentage and percentile plots are unnecessary for Figs. 14 and 15, since the closeness of within and transfer data percentages of Cocomo81 and Nasa93 subsets can easily be verified with manual inspection: For Cocomo81 subsets the biggest percentage difference is $6.6 - 4.1 = 2.5\%$; for Nasa93 subsets it is $11.7 - 8.2 = 3.5$ %.

Figure 17 is taken from Kocaguneli and Menzies selection tendency experiments (Kocaguneli and Menzies 2011). In the performance experiments, we have seen the similarity between the results of this research and that of Kocaguneli and Menzies in terms of performance. Comparison of Figs. 16–17 shows that the similarity of results are also valid in terms

Dataset	# of Analogies	coc-60-75	coc-76-rest
coc-60-75 (20)	3.6	0.8 (4.1)	2.8 (6.4)
coc-76-rest (43)	4.5	1.3 (6.6)	3.2 (7.4)

Fig. 14 The amount of within and transferred data from subsets of Cocomo81

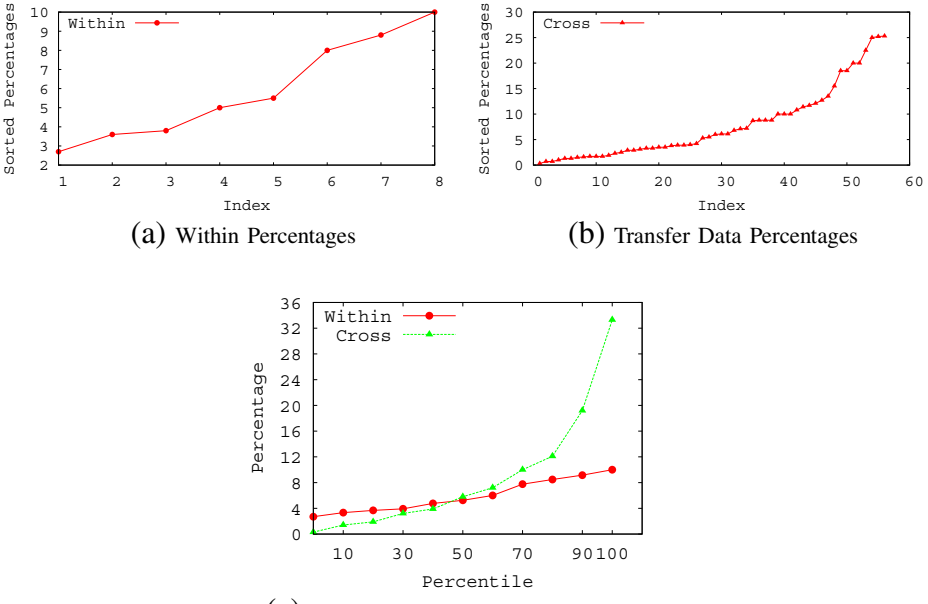
Dataset	# of Analogies	nasa-70-79	nasa-80-rest
nasa -70-79 (39)	7.3	2.6 (6.6)	4.7 (8.7)
nasa 938089 (54)	9.5	3.2 (8.2)	6.3 (11.7)

Fig. 15 The amount within and transferred data selected from the subsets of Nasa93

of the selection tendencies. Similar trends of percentile values (as given shown in Figs. 16 and 17) show that TL between domains of proprietary data sets as well as public data sets have similar selection tendencies.

5 Threats to Validity

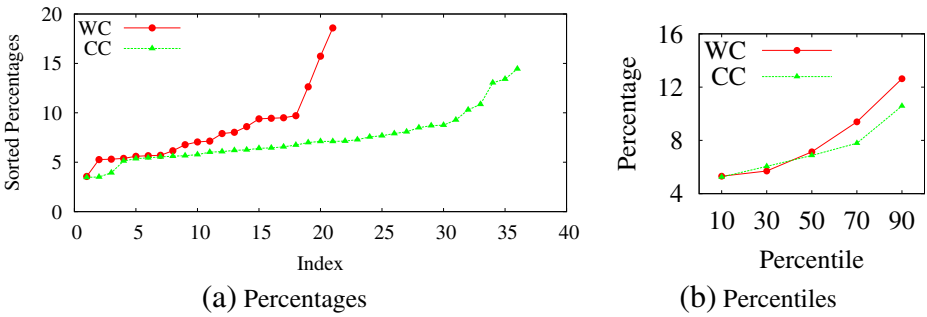
Internal validity questions the extent to which the association between dependent and independent variables holds (Alpaydin 2010). To observe this relationship SEE studies make use of dividing data sets into test and train data. This division is performed in accordance with



(a) Within Percentages (b) Transfer Data Percentages

(c) Within and Tranfer Data Percentiles

Fig. 16 Percentages of instances (a.k.a. analogies) from (**a**) within and (**b**) transferred data. The percentage values come from Fig. 13. Within company percentages are the gray-highlighted diagonal cells, whereas transferred data percentages are the remaining off-diagonal cells. The percentile graph (**c**) shows the percentiles of (**a**) and (**b**)



(a) Percentages (b) Percentiles

Fig. 17 Percentages and percentiles of instances from within and transferred data as given by Kocaguneli and Menzies (2011). Note the similarity of the percentile plot of this figure to that of Fig. 16

various sampling methods such as LOOCV or any N-Way cross validation. However, selecting any particular sampling method is inherently an internal validity threat (Hastie et al. 2008) as different sampling methods are expected to have different bias and variance values. In this research LOOCV is employed, whose alternative is any N-Way cross-validation. In a N-Way cross-validation, data is randomly divided into B bins and each bin is tested on a model learned from the combination of other bins (typical values for B are 3 or 10). Elsewhere (Kocaguneli and Menzies 2012), we show that there is very little difference in the bias and variance values generated for LOOCV and N-way cross-validation. Since two testing strategies have similar bias-variance characteristics for effort datasets, we opted for LOOCV due to the fact that LOOCV is a deterministic procedure that can be exactly repeated by any other researcher with access to a particular data set.

External validity deals with the generality of the results, i.e., whether the results can be generalized outside the specifications of a study (Milicic and Wohlin 2004). This study challenges the validity of prior results of 21 within and transfer data source pairs (Kocaguneli and Menzies 2011) on proprietary data sets coming from 8 Web development companies. The results of this study confirm prior findings, i.e., a total of $21 + 8 = 29$ data sets agree on the reported conclusions, which is way more extensive than other TL studies. Among 10 studies investigated by Kitchenham et al. (2007), 9 of them used a single within-transfer dataset pair, and 1 study used 6 pairs. Therefore, concerning external validity, this research bears higher validity than other similar studies. However, it is important to note that none of the data sets employed here represents a random sample; therefore the results presented herein are much more likely to scale to those companies that manage projects similar to those detailed here.

Construct validity (a.k.a. face validity) observes whether a study measures what it actually intends to measure (Robson 2002). Previous studies have concerned themselves with the construct validity of different performance measures for effort estimation (e.g., Stensrud et al. (2002)). The intention of this paper is neither providing evidence for a particular error measure nor biasing our conclusions due to selection of one particular error measure. Instead, we used 8 different performance measures (which covers a big majority of all the previously used error measures in SEE) aided with win-tie-loss statistics so that our results are able to be benchmarked against other studies.

Note that the selection of the time frame in this study (15 years) is an engineering judgment and bears validity threats concerning the length of the time-frame that can be considered as a set of within-time-period project instances. 15 years may be considered a long

time for the software development, where the characteristic as well as the data distribution may change. Therefore, so as to address this threat, a future study is required to identify how frequently the data distribution changes. The intervals of significant distribution changes may guide the identification of time frame length and the transfer learning methods may use the identified time frames.

6 Future Research

Based on the majority of the companies in the domain TL experiments (depending on the error measure, 5 or 6 companies out of 8) the transferred data performance is the same as the within data performance. In terms of time interval TL, in *all* of the cases, within and transferred data performances were statistically the same. This shows us that TL through methods like TEAK can help transfer instances between domains and time intervals so that transferred data can perform as well as within data.

That said, an interesting fact is that error measures can result in different conclusions. For example, see in Figs. 7 and 8 that transferred data performance for the company tuku8 depends on error measures. This disagreement may cause different companies to make different conclusions depending on the particular error measures they are using. Hence as part of the future work we intend to investigate error measure studies like that of Shepperd and MacDonell (2012), as they bear a significant importance for SEE as well as for the recommendations to practitioner audiences.

Another area of future work would be the exploration of other kinds of transfer learning. For example, this paper has employed *instance-based transfer* where the instances share the same ontology (names of column headers). A more challenging task would be to explore instance-based transfer to take data from different projects collected using different ontologies. For example, (Pan and Yang 2010) described experiments in image processing where data is collected using 2D and 3D sensors, then both projected and clustered within a new synthesized set of dimensions. The key to making that work is to find some (possibly small) number of common reference points to enable the mapping of the raw dimensions onto the synthesized dimensions. Our own experiments in that direction are encouraging, but very preliminary. If other effort estimation researchers want to explore this issue, then we offer them the *Rosetta Stone challenge*:

- The COCOMO Rosetta Stone was created by Reifer et al. (1999) to translate between COCOMO-81 effort estimation model and the revised COCOMO-2000 model (the new model had some new parameters and deleting some old ones).
- The *Rosetta Stone Challenge* is to automatically recreate that manually derived translator from the COCOMO-81 to COCOMO-2000 ontology (or, alternatively, show that some better translator exists).
- Note that COCOMO-81 and COCOMO-2000 data sets are available at the PROMISE web site promisedata.googlecode.com.

In Section 3.4, we noted that one may also question TEAK's performance, when it is allowed to use *"only CC data"*. This case was the question of a 2010 study (Kocaguneli et al. 2010), where we experimented with a limited number of data sets, which may be challenged due to their age. Therefore, we believe that an interesting future direction would be to replicate the *"only CC data set"* scenario with TEAK on newly collected proprietary data sets.

7 Conclusion

When projects lack sufficient local data to make predictions, they can try to transfer information from other projects. In the field of software engineering, transfer learning has been shown to be effective for defect prediction. This paper checked if it was possible to build transfer learners for effort estimation. We explored the following research questions.

RQ1: Is transfer learning effective for effort estimation?

We compared estimates learned from within a single stratification to those transferred across different stratifications. In the majority of the cases, the cross-stratification estimates transferred by TEAK do as well as the within-stratification estimates.

RQ2: How useful are manual divisions of the data?

The selection tendency results of this research show that test instances select equal amounts of instances from within and transferred data sources; i.e., TEAK found no added value in restricting reasoning to just within a delphi localization. We therefore advise that delphi localization should be used with great care, perhaps only after checking if automatic clustering methods can generate better estimates than clustering after delphi localization.

RQ3: Does transfer learning for effort estimation work across time as well as space?

This paper has tested a conjecture from Storkey (2009) "Data set shift" and "cross-company learning" are both very closely related to TL. To explore that conjecture, this paper successfully applied the same transfer learner *without modification* to traditional cross-company learning problems as well as data set shift problems. Hence, we would suggest that it is useful to combine research methods for two previously unrelated research threads in SE: Turhan's "data set shift" and Kitchenham et al.'s "cross-company learning".

Our results also indicate that it may be misguided to think: 1) The data of another organization cannot be used for local solutions and 2) old data of an organization is irrelevant to current context. This is quite an important point as the success of our transfer learners in moving data means that we should not always discard the hard-won lessons of the past. Our history has lessons that are still relevant today.

Acknowledgments The work was partially funded by NSF CCF grant, award number 1302169, and the Qatar/West Virginia University research grant NPRP 09-12-5-2-470.

References

Alpaydin E (2010) Introduction to Machine Learning, 2nd edn. MIT Press

Arnold A, Nallapati R, Cohen W (2007) A comparative study of methods for transductive transfer learning. In: ICDM'07: 17th IEEE international conference on data mining workshops, pp 77 –82

Bettenburg N, Nagappan M, Hassan AE (2012) Think locally, act globally: improving defect and effort prediction models. In MSR'12

Boehm B (1981) Software engineering economics. Prentice hall

Chang C-l (1974) Finding prototypes for nearest classifiers. IEEE Trans Comput C3(11)

Corazza A, Di Martino S, Ferrucci F, Gravino C, Sarro F, Mendes E (2010) How effective is tabu search to configure support vector regression for effort estimation? In: Proceedings of the 6th international conference on predictive models in software engineering

Rodriguez D, Herraiz I, Harrison R (2012) On software engineering repositories and their open problems. In: Proceedings RAISE'12

Dai W, Xue G-R, Yang Q, Yong Y (2007) Transferring naive bayes classifiers for text classification. In: AAAI'07: Proceedings of the 22nd national conference on artificial intelligence, pp 540–545

Foss T, Stensrud E, Kitchenham B, Myrtveit I (2003) A simulation study of the model evaluation criterion mmre. IEEE Trans Softw Eng 29(11):985–995

Foster G, Goutte C, Kuhn R (2010) Discriminative instance weighting for domain adaptation in statistical machine translation. In: EMNLP '10: conference on empirical methods in natural language processing, pp 451–459

Gao J, Fan W, Jiang J, Han J (2008) Knowledge transfer via multiple model local structure mapping. In: International conference on knowledge discovery and data mining. Las Vegas, NV

Harman M, Jia Y, Zhang Y (2012) App store mining and analysis: Msr for app stores. In: MSR, pp 108–111

Hastie T, Tibshirani R, Friedman J (2008) The elements of statistical learning: data mining, inference and prediction, 2nd edn. Springer

Hayes JH, Dekhtyar A, Sundaram SK (2006) Advancing candidate link generation for requirements tracing: the study of methods. IEEE Trans Softw Eng 32(1):4–19

He Z, Shu F, Yang Y, Li M, Wang Q (2012) An investigation on the feasibility of cross-project defect prediction. Autom Softw Eng 19:167–199

Hihn J, Habib-agahi H (1991) Cost estimation of software intensive projects: a survey of current practices. In: 13th international conference on software engineering 1991, pp 276 –287

Hindle A (2012) Green mining: a methodology of relating software change to power consumption. In: Proceedings, MSR'12

Huang J, Smola A, Gretton A, Borgwardt K, Scholkopf B (2007) Correcting sample selection bias by unlabeled data. In: Proceedings of the 19th Annual Conference on Neural Information Processing Systems, pp 601–608

Jiang Y, Cukic B, Menzies T, Bartlow N (2008) Comparing design and code metrics for software quality prediction. In: Proceedings PROMISE 2008, pp 11–18

Kadoda G, Cartwright M, Shepperd M (2000) On configuring a case-based reasoning software project prediction system. UK CBR Workshop, Cambridge, UK, pp 1–10

Keung J (2008) Empirical evaluation of analogy-x for software cost estimation. In: ESEM '08: Proceedings of the second international symposium on empirical software engineering and measurement. ACM, New York, NY, pp 294–296

Keung J, Kocaguneli E, Menzies T (2012) Finding conclusion stability for selecting the best effort predictor in software effort estimation. Automated Software Engineering, pp 1–25. doi:10.1007/s10515-012-0108-5

Kitchenham BA, Mendes E, Travassos GH (2007) Cross versus within-company cost estimation studies: a systematic review. IEEE Trans Softw Eng 33(5):316–329

Kocaguneli E, Gay G, Yang Y, Menzies T, Keung J (2010) When to use data from other projects for effort estimation. In: ASE '10: Proceedings of the international conference on automated software engineering (short paper). New York, NY

Kocaguneli E, Menzies T (2011) How to find relevant data for effort estimation. In: ESEM'11: international symposium on empirical software engineering and measurement

Kocaguneli E, Menzies T (2012) Software effort models should be assessed via leave-one-out validation. Under Review

Kocaguneli E, Menzies T, Bener A, Keung JW (2012) Exploiting the essential assumptions of analogy-based effort estimation. IEEE Trans Softw Eng 38(2):425–438

Lee S-I, Chatalbashev V, Vickrey D, Koller D (2007) Learning a meta-level prior for feature relevance from multiple related tasks. In: ICML '07: Proceedings of the 24th international conference on machine learning, pp 489–496

Li Y, Xie M, Goh T (2009) A study of project selection and feature weighting for analogy based software cost estimation. J Syst Softw 82:241–252

Lokan C, Mendes E (2009a) Applying moving windows to software effort estimation. In: ESEM'09: Proceedings of the 3rd international symposium on empirical software engineering and measurement, pp 111–122

Lokan C, Mendes E (2009b) Using chronological splitting to compare cross- and single-company effort models: further investigation. In: Proceedings of the thirty-second Australasian conference on computer science, vol 91. ACSC '09, pp 47–54

Ma Y, Luo G, Zeng X, Chen A (2012) Transfer learning for cross-company software defect prediction. Inf Softw Technol 54(3):248–256

Mendes E, Mosley N (2008) Bayesian network models for web effort prediction: a comparative study. IEEE Trans Softw Eng 34:723–737
Mendes E, Mosley N, Counsell S (2005) Investigating web size metrics for early web cost estimation. J Syst Softw 77:157–172
Mendes E, Watson ID, Triggs C, Mosley N, Counsell S (2003) A comparative study of cost estimation models for web hypermedia applications. Empir Softw Eng 8(2):163–196
Menzies T, Butcher A, Cok D, Marcus A, Layman L, Shull F, Turhan B, Zimmermann T (2012) Local vs. global lessons for defect prediction and effort estimation. In: IEEE transactions on software engineering, p 1
Menzies T, Butcher A, Marcus A, Zimmermann T, Cok D (2011) Local vs global models for effort estimation and defect prediction. In: IEEE ASE'11. Available from http://menzies.us/pdf/11ase.pdf
Menzies T, Greenwald J, Frank A (2007) Data mining static code attributes to learn defect predictors. In: IEEE transactions on software engineering. Available from http://menzies.us/pdf/06learnPredict.pdf
Mihalkova L, Huynh T, Mooney RJ (2007) Mapping and revising markov logic networks for transfer learning. In: AAAI'07: Proceedings of the 22nd national conference on Artificial intelligence, pp 608–614
Milicic D, Wohlin C (2004) Distribution patterns of effort estimations. In: Euromicro conference series on software engineering and advanced applications, pp 422–429
Minku LL, Yao X (2012) Can cross-company data improve performance in software effort estimation? In: PROMISE '12: Proceedings of the 8th international conference on predictive models in software engineering
Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359
Petersen K, Wohlin C (2009) Context in industrial software engineering research. In: 3rd International Symposium on empirical software engineering and measurement, 2009. ESEM 2009, pp 401 –404
Posnett D, Filkov V, Devanbu P (2011) Ecological inference in empirical software engineering. In: Proceedings of ASE'11
Reifer D, Boehm BW, Chulani S (1999) The Rosetta Stone: Making COCOMO 81 Estimates Work with COCOMO II. Crosstalk. The Journal of Defense Software Engineering, pp 11–15
Robson C (2002) Real world research: a resource for social scientists and practitioner-researchers. Blackwell Publisher Ltd
Shepperd M, MacDonell S (2012) Evaluating prediction systems in software project estimation. Inf Softw Technol 54(8):820–827
Shepperd M, Schofield C (1997) Estimating software project effort using analogies. IEEE Trans Softw Eng 23(11):736–743
Stensrud E, Foss T, Kitchenham B, Myrtveit I (2002) An empirical validation of the relationship between the magnitude of relative error and project size. In: Proceedings of the 8th IEEE symposium on software metrics, pp 3–12
Storkey A (2009) When training and test sets are different: characterizing learning transfer. In: Candela J, Sugiyama M, Schwaighofer A, Lawrence, N (eds) Dataset shift in machine learning. MIT Press, Cambridge, pp 3–28
Turhan B (2012) On the dataset shift problem in software engineering prediction models. Empir Softw Eng 17:62–74
Turhan B, Menzies T, Bener A, Di Stefano J (2009) On the relative value of cross-company and within-company data for defect prediction. Empir Softw Eng 14(5):540–578
Wu P, Dietterich TG (2004) Improving svm accuracy by training on auxiliary data sources. In: Proceedings of the twenty-first international conference on Machine learning, ICML '04. ACM, New York, NY, p 110
Yang Y, Xie L, He Z, Li Q, Nguyen V, Boehm BW, Valerdi R (2011) Local bias and its impacts on the performance of parametric estimation models. In: PROMISE
Zhang H, Sheng S (2004) Learning weighted naive bayes with accurate ranking. In: ICDM '04 4th IEEE international conference on data mining, pp 567–570
Zhang X, Dai W, Xue G-R, Yu Y (2007) Adaptive email spam filtering based on information theory. In: Web information systems engineering WISE 2007, Lecture notes in computer science, vol 4831. Springer Berlin/Heidelberg, pp 159–170
Zimmermann T, Nagappan N, Gall H, Giger E, Murphy B (2009) Cross-project defect prediction: a large scale experiment on data vs. domain vs. process. ESEC/FSE, pp 91–100
Žliobaitė I (2010) Learning under concept drift: an overview. CoRR, arXiv:1010.4784

Ekrem Kocaguneli holds a PhD degree from West Virginia University. He received MSc and BS degrees in Computer Engineering from Bogazici University. His main research interests are software effort estimation, machine learning applications in empirical software engineering and intelligent tools to aid software processes. For further information, his website is: www.kocaguneli.com.

Tim Menzies (Ph.D., UNSW) is a Professor in CS at WVU and the author of over 200 referred publications. In terms of citations, he is one of the top 100 most most cited authors in software engineering (out of 54,000+ researchers, see http://goo.gl/vggy1). At WVU, he has been a lead researcher on projects for NSF, NIJ, DoD, NASA, as well as joint research work with private companies. He teaches data mining and artificial intelligence and programming languages. Prof. Menzies is the co-ounder of the PROMISE conference series (along with Jelber Sayyad) devoted to reproducible experiments in software engineering: see http://promisedata.googlecode.com. He is an associate editor of IEEE Transactions on Software Engineering. the Empirical Software Engineering Journal, and the Automated Software Engineering Journal. For more information, see his web site http://menzies.us or his vita at http://goo.gl/8eNhY or his list of pubs at http://goo.gl/8KPKA.

Dr. Emilia Mendes is Professor in Software Engineering at the Blekinge Institute of Technology (Sweden). She obtained her PhD in Computer Science from the University of Southampton (UK) in 1999. Her research is inter-disciplinary, encompassing four disciplines - Web & Software measurement & metrics, Empirical Software & Web Engineering, IT/Computer Science & Software/Web Engineering education, and Hypertext. To date she has published over 170 refereed publications, which include two books (one edited (2005 - Web Engineering) and one authored (2007 - Cost Estimation Techniques for Web Projects)). For more information, see her home page http://www.bth.se/com/eme.nsf.