1 Replies to Reviewer Comments

1.1 Replies to the Editor Comments

Many thanks for your revised manuscript; we have now collected a complete set of second round reviews.

While the revised version is greatly improved, the standard of presentation and the quality of writing remain below the standards of the journal; and hence we are asking you to revise and resubmit the paper once again.

When revising the paper, please take care to fully address the concerns of the reviewers.

We thank the editor and all reviewers for their helpful comments and such a fast processing of the manuscript. As per requested, we have addressed all issues raised by the reviewers, improving the presentation and quality of writing. The two main issues regarding feature selection and NN-filtering are now discussed in detail in this version of the manuscript. Please find our replies to reviewer comments below.

1.2 Replies to the Reviewer #1 Comments

Reviewer #1: Overall comment:

In my view the authors have done a thorough job of addressing the concerns raised with the previous version of the manuscript. The motivation has been strengthened and the substntial reorganisation and rewriting have made the paper far more coherent. I believe it is a stronger paper as a result.

We thank the reviewer for this positive opinion. We were able to produce a much more stronger version with comments from all reviewers. Please find our replies to your specific comments below.

Specific comments:

- There are still a few challenges in terms of expression: ”using which” in the Abstract, ”Several research” at the beginning of the Introduction, ”benecial” on page 10. I believe a further proof-read would aid the readability of the paper.

Fixed the issues you mention above. We did our best to correct other language specific issues throughout the paper.

- The exclusive use of static code features should be mentioned in the Abstract as this would give the reader an immediate sense of the sort of analyses that follow.

We have edited the abstract accordingly: “... we investigate the applicability of cross-company (CC) data for building localized defect predictors using static code features.”

- For me the paper overdoes the ’this analysis has not been done before’ angle. Certainly it was important to address issues of novelty and motivation, but saying it is novel doesn’t make it so. With all due respect, if a paper appears in a journal such as EMSE then its novelty should be assumed and perhaps described in the Contributions (as is done here) - it should

not need to be pointed out.

Removed corresponding parts from abstract and intro

- There is still a lack of adequate explanation of the basis of similarity - this probably relates more to the comments of Reviewer #2. That is, similarity is said here to be simply a function of distance in n-dimensional code feature space - there is no sense of conceptual similarity. Of course there is no choice in this, as the modules are so masked that any conceptual similarities are hidden. I believe it would be useful to point this out early in the paper.

We have now included a footnote in the second research question: "We should carefully note that we do not make use of any conceptual similarities, since our analysis is based on static code features. As to the issue of conceptual connections within the code, we refer the reader to the concept location and cohesion work of Marcus et al."

- The comment re collecting local data requiring too much time and effort (page 2) is drawn from work on effort estimation. As I pointed out in the prior review, by far the bulk of this work is focused on single whole-project observations - meaning that the associated data set grows very slowly. Collecting within-project module data, as is done here, means that the data set can be acquired far more readily. The comment needs to be qualified in this regard.

In order to address your concern about data collection, we dropped that comment and restated that part as follows:"In the third analysis, we focus on WC models and determine the smallest amount of local data needed for constructing a model. We employ an incremental learning approach to WC data in order determine the number of samples in local repositories for building defect prediction models. "

- The material presented in the second part of page 4 lacks the coherence now evident in the bulk of the paper - it seems a rather disparate and loosley connected discussion.

Removed that part as per requested.

- In my view statements such as 'What may surprise the reader' should not be used - it can be unwise to assume anything that is not founded in evidence.

Removed that statement and re-worded as:"Furthermore, these detectors are widely used despite the fact that they are quite "coarse-grained"..."

- While I can understand why you have assumed no 'knowledge' of the data I still believe that the likelihood of managers downloading alien data from an unknown source and then using this as the basis for decision-making in their new and very expensive project could well be low. Furthermore, the impact of such an approach comes through in the discussion in subsection 5.3, where the authors can only speculate as to why results may have been confounded. Such discussions are only ever able to provide minimal comment minimal regarding the actual features and their influence as really there is no basis for further comment. This does not invalidate the outcomes, of course, but it does mean that there are constaints on the general conceptual insights that can be drawn from them.

Following text is added to introduction: "Nevertheless, there may be some reluctance on the part of programmers and managers to use data collected at other sites. Such caution is to be expected amongst professionals striving to deliver quality products to their clients. In this paper, we suggest that it can be quite beneficial to use data from other sites, particularly when no local data is available. Professional developers should apply our advise cautiously- perhaps via an intial case study on a small part of their system. "

- Re the previous review comment in relation to errors in the form of the references, there are still problems with reference #32 (was #3) - "M,E." should be "Mendes, E."; reference #36 (was #11) - "V.B. ad" should be "V. Basili"; reference #46 (was #19) - inclusion of "url" looks like a LaTeX error. - The spelling of "MacDonell" is still incorrect in the three times it appears in the paper, although it is correct in the list of references.

All these reference related issues are fixed.

1.3 Replies to the Reviewer #2 Comments

Reviewer #2: Abstract
L32-33 English usage is incorrect. Comma in wrong place?

Rewritten as: "Hence, for companies with no local defect data, we recommend a two-phase approach that allows them to employ the defect prediction process instantaneously"

L34 ? the? Defect detection process?

Fixed: "....initiate the defect prediction process..."

Introduction
L44 English ????

Fixed: "Defect prediction studies usually focus on building models..."

P2 L1-5 In the absence? English badly constructed. Spelling L5 Also this is really a well known and obvious point.

Fixed: "When automated tools are not used, manual effort is inevitable to maintain these repositories..."
Also removed the obvious statement.

L10-12 English

Reworded: " Defect prediction literature contains many examples where predictors are learned from within company data. However, there exists no case that attempts to harness cross company data."

L12 I do not believe that these are experiments in the scientific sense they are just analyses: you seem simply to remove data you don't want from existing observational data and analyse it.

Within the text, we now refer to these as analyses.

L17-19 English (s)

Fixed.

***L22 The problem with the first research question is that the current level of knowledge in SE would seem to suggest that WC data (when it exists) is always going to be better and CC is never going to be preferred. I don't see this as a question that we don't already know the answer to. The question is probably: Is CC data ever useful? (ANSWER = only when you don't have WC data.) This needs re-wording.

We agree with the reviewer: it is not a simple matter to determine which is better: WC or CC?

In our paper, we stress that it is a trade-off situation. In all our experiments, WC (local data) does best but it may not be available. CC (external data) can lead to high false alarm rates (and we show that we can tame those rates using nearest neighbor filtering. The resulting detectors are close too, but not quite as good, as using WC data).

What we would add, however, is that the situation is far less defined for effort estimation/ Kitchenham et al. concluded that the value of CC vs WC data for effort estimation is unclear: . . . some organizations would benet from using models derived from cross company benchmarking databases but others would not... (previous draft of the paper discussed Kitchenhan results, however we have removed those as per requested.)

Following your suggestion, we have re-worded this research question and the corresponding answer as follows:
RQ1: "Are CC data ever useful for organizations?"
"CC data are useful in extreme cases such as mission critical projects, where the cost of false alarms can be afforded. Therefore, CC data should be used only when WC data are not available."

***L27 This is the real question and the one we need an answer to: how do we decide on the best way to select the most appropriate error data for any company; this is particularly important when the company has no WC data at all. You say use an NN algorithm (which isn't new) but defining the parameters (and features) to decide what constitutes a suitable NN for any company's data would be. If you are actually saying that the learning algorithm does all this for you then you need to say this within the paper so that the reader knows the limitations of your work.

Thanks for your comment. You are quite correct to strongly protest the lack of discussion on this important point. Previously, we have discussed this issue (about features) in another paper. Notes on those conclusions are now added into this draft (see end of section 3.1). It was a significant oversight that this material was NOT added to previous drafts. We also thank for your comment on NN. As you mention NN does all the work and this is now stated in related parts of the paper.

P3 L1-2 obvious

We agree with the reviewer. however, as it may seem obvious, there is precedent on the contrary in cost estimation literature considering the results reported by Kitchenham et.al.

P3 L8- 9 This seems a very bold statement is it always true that data from site A is useful to site B - this seems difficult to believe - sometimes would be better (or is it just for your data sets) - doesn't ring true to me - clarify.

We agree. Thanks for pointing out this issue. We now include: "...at least for the datasets we analyzed..."

*****8L11 I think it is this filter that we all should know about; it relates directly to the important second research question above. This filter and how it works needs to be emphasised and explained.

The details of the NN filter is now provided in a devoted sub-section 5.1.1. In summary: it is a function of distance in n-dimensional code feature space and does not include any feature subset selection. This filter filters data samples, not data features. Please see new section 5.1.1 for details.

Motivation
P3 L36 I am not sure what you mean by the 'defector' do you mean the 'error detector'?

Fixed as detector.

P4 L49 State what is meant by a defect predictor at this point - do you mean a metric for predicting defects or something else? At some point you need to give an example of what you mean by a defect predictor - better sooner than later.

We have included a footnote in the introduction which seemed more appropriate: "'Throughout the paper, a defect predictor (or predictor) means a binary classification method that categorizes software modules as either defective or defect-free'

P4 L28 This research team (which team - the authors)?

Fixed as: "the authors..."

P5 L7 What are your data miners? The error predictors, the naive Bayes method your algorithm or all of them?

This part is removed as requested by other reviewers. It previously referred to error predictors.

P5 L8/9 It is not clear why the 2 bullet points are related to the preceding statement. The intention of this sentence needs clarifying so that the reader can understand its purpose. I think L7-9 might possibly be irrelevant and detract.

Removed corresponding lines.

L18-19 'per cent' = 'set'? Maybe set is better used twice.

Fixed: "...a remarkably small set of modules are implicated in all faults and that set can change..."

P5 Methodology

P5 L30-31 Experiments? But there is no experimental design and no hypotheses in the statistical sense given so far (no null vs alternative). I can see you persist with this idea of experiments throughout, but it is misleading to confuse a series of apparently straightforward analyses on sub-samples of data taken from sets of observational samples with an experiment.

Within the text, we now refer to these as analyses.

P5 Data, predictors and features are used apparently interchangeably with no explanation. These are presumably static code metrics that software engineers (SEs) would be more familiar with?

Thanks for pointing this out . We now clarified this part by adding this explanation about notation in a footnote at the beginning: "Throughout the paper, the following notation is used: a defect predictor (or predictor) means a binary classification method that categorizes software modules as either defective or defect-free; data refers to MxN matrices of raw measurements of N metrics from M software modules; these N metrics are referred to as features."

P6 L26 Attitude control? - nice if we could do that ? Is this Altitude control?

Fixed: Altitude

P7 L38 predict for the ? 'for' not needed?

Fixed.

P10 L20 ' ....the performance deltas for some treatment...'.

An explanation of what you mean by treatment is still needed (this usage is not the usual statistical one). You don't treat a sample. Do you mean different data subsets that were chosen or something else? The reader may not know. Also delta may not be familiar to an SE. I think you should use performance delta (pd) etc. first and then use pd thereafter. (I don't think you have done this already.) Perhaps you now seem to confuse your own 'experiments' with treatments!

That was a mistake on our side. actually pd = probability of detection. We now introduce this notation in the Performance evaluation section. The term treatment was being used since we explain a tool, i.e. quartile charts, in general terms. Nevertheless we have now reworded that part and removed the terms: treatment and performance delta.

P10 Experiment #1

P10 L37 beneficial???? - I think you should state the alternative (and null) hypothesis within the paragraph not the section title.

Fixed beneficial. Title is updated with the text of the corresponding research question. Added the goal of the analysis: "Our goal is to determine whether using cross company data is beneficial for constructing defect predictors and to identify the conditions under which cross-company data should be preferred to within-company data."

P10 L42 Table 5 does not give an explanation - it is an algorithm without explanation or comments. Do you think that this is sufficient for a SE to follow this is in order to apply your method?

We agree with the reviewer. The term 'explanation' is replaced with 'pseudo code', the table caption is updated accordingly and comments in pseudo-code are added. Furthermore, the notation in the pseudo-code is carefully chosen to be self descriptive. We considered that a pseudo code notation is the best for a SE. We have also updated similar tables, i.e. Table 7 and 8

P12 L26 English usage: ..informs nearly all ..... use of English - examination might inform perhaps ..... but also of ——— English incorrect.

Reworded: "...informs of not only the sources of errors, but also numerous irrelevancies..."

*****L25 What exactly are the extraneous factors: an SE would need to understand what they are. Name some specific ones that are not relevant to your software modules and that were present in the WC data and state why the are extraneous. I can imagine that there will be constructs in most procedural languages that are actually quite similar, whether the language be C, Fortran, or a new version of Cobol!

Re-written:"For example, the defect characteristics of software modules with different complexity levels or sizes may differ (Koru et. al., 2008). In this context, a complicated search algorithm's metrics are irrelevant to the defective behavior of a simple sum *function. Since there are few modules with extreme characteristics (i.e. complexity, size) in a single project, their effect on the overall model are limited. However, when data from multiple companies are combined, the number of these extreme cases, and hence their cumulative effect on the overall model increase significantly. Therefore, factors such as programming language constructs (i.e. object oriented vs. procedural) and project specific requirements (i.e. availability, speed) that have impacts on the module characteristics can be considered as extraneous factors. "*

In fact I am not sure you should even use the word 'cause' (of errors) the metrics are probably just associated with errors and don't cause them; the cause may be the programmer!

Term removed.

L25 - 29 I don't think this is a sentence (there maybe no subject or object) - it therefore is difficult to comprehend its meaning; and, this is a very important explanation that is needed

to justify your poor results here. It needs re-writing.

Re-written, please see our above reply on extraneous factors.

L30 Hence large ..... detection of errors. Do you mean?

Re-written, please see above.

***L25-31 I am not able to follow the argument in this paragraph. This may be because of the poor English. But I can't see how a large data set would necessarily give you spurious results. Poor results might come from a model that is not appropriate for the data or there may not be an effect in the data. If you simply delete all data that does not fit your model and use that which does fit the model you will always get a good model fit and a better prediction will usually follow. This explanation is not good enough for publication. We need to know what is extraneous and why. I think what you are saying is that the modelling procedure cannot sufficiently well distinguish between features or metrics that will give a good result and those that will give a bad result because there are a lot of features in the data! **********So a very important question is how do we know what irrelevant features to remove from the data. Or might you actually be saying that the learning and NN procedure are not very good. You need to answer this.

Re-written. Also, please see our reply to your very last comment about features.

L45 What does (6/7) mean: explain this. This is not good English and is almost impossible to understand. L47 cannot be understood: it needs explanation CC-WC data what's that? S4.3 is badly written and needs re-writing.

Re-written:"....For six projects, the general result holds (i.e. both (pd, pf) increases if defect predictors learned from CC projects are used rather than defect predictors learned from WC projects. See group a in Table **??***).*
For one project, there is no difference in the performances of the defect predictors learned from CC and WC projects (see group b in Table **??***). ..."*

P14 5 Experiment #2
L21 too many limits!

Fixed: 'limits' replaced with 'restricts'

***L32 You need to tell us about this bias - what are you doing to the data? A bias in general is not a good thing to have in your data. It can simply bias the model you use - it's important to tell us what you are doing.

We edited the text accordingly. We do not introduce a bias to the data. Yet, we do introduce a bias to the model by being selective of training data. This is the whole idea behind NN-filtering, we want to bias the model we use. Bias is not a good thing if you have no control over it. But in our case, we know where to bias in order to eliminate noise. The procedure is described in detail in the following paragraphs and also accompanied by a pseudo-code.

L36 Are you saying that you simply use the data which has the same or similar features (metrics) to that which you want to provide the predictions for? Do you do the selection or does the software? You need to answer this.

This is exactly what we do. This selection is done by the software (which corresponds to NN-filtering in our case). The purpose of this whole section is to describe the automatic handling of this selection process by NN-filtering. We have rephrased the corresponding section as follows:

"In this analysis, we try to construct more homogeneous defect datasets from CC data in an automated manner. For this purpose we use a simple filtering method (i.e. NN) which is to be described next. The analysis design is given in Table 7. For this analysis, we use all common features available in NASA projects which is a total of 19 features. These features are marked in NASA Shared column of Table 4."

Then we reserve the following description with the title: "Nearest neighbor filtering."

P15 L28-33 Table 7 gives an algorithm. This is clearly not an experiment you are simply deleting data that you don't want from an observational data set.

As mentioned before, we do not use the term experiment anymore.

L45 Observations can only be made by someone or something. They don't just appear, do they?

Reworded as: "...then we would expect to notice two observations:..."

L46 definite articles missing: the NN and the WC data. It is not really the data but the methods you are comparing aren't you?

Thanks for pointing out. Fixed.

*******P15 L28-L50 Nearest Neighbour! You need to say what features were chosen and are being examined and what the distances are that enable you to say you have a nearest neighbour. Otherwise no one will be able to re-apply your procedure or check it.

As for the features we do not perform any feature selection technique (Please see our reply to your very last comment). We use all available common features in Table 4 *as mentioned in other parts of the paper (i.e. Analysis 1 design, Pseudo-codes for all analyses designs). Since this analysis design is the same except for the NN-filtering part, we did not report the features once more in order not to repeat ourselves. Nevertheless, we now explicitly added the following text to be more clear: "For this analysis, we use all common features available in NASA projects which is a total of 19 features. These features are marked in NASA Shared column of Table 4. "*

Also, as pointed out above, your comments made us realize that we had missed a vital section in our experimental description. Please see our new notes on the complex issue of feature selection, end of section 3.1

As for the distance measure and selection criteria, the text now includes the following explanation available in section 5.1.1: "....We calculate the pairwise Euclidean distances between the validation set and the candidate *training set (i.e. all CC data).....we pick its $k = 10$ nearest neighbors from candidate training set. ..... Using only unique ones, we form the training set....."*

P15 L 28-30 One reason for the difference is likely to be that the data in the CC do not fit the model as well as in the WC. This is not a surprising result as mentioned before providing you select the data to fit the company features .

We certainly agree with the reviewer. Thanks for pointing out this in such a clear way. This comment was really helpful for us to introduce the following issue on extraneous factors: "...Since there are few modules with extreme characteristics (i.e. complexity, size) in a single project, their effect on the overall model are limited. However, when data from multiple companies are combined, the number of these extreme cases, and hence their cumulative effect on the overall model increase significantly...."

L37 .. Third

Fixed.

L42 The problem is we still do not really know how you filtered the features - you have not told us. Or are they simply the features that the WC data would possess? I suspect they are but you need to tell us. How do we replicate this work otherwise.

We think this issue is now clear since we do not perform any feature filtering or selection and we now explicitly mention it in the design part.

L47 - L51 No, surely to out perform one of pd or pf must always be greater than or less than the corresponding pd or pf while the other could be equal or better?

We agree. However, we caution the reader for practice by saying that "Please note that the conjunction of Observation1 and Observation2 is uncommon". Therefore, the results should be used to achieve either higher detection rates or lower false alarms. Unfortunately, there will always be a trade-off and this is clearly observed since these two are mutually exclusive.

P17 L 37 .... are ..... is better

Fixed: "'...is avoided.'

L44 indefinite article missing

Fixed.

L44 When does CC work better than NN give an example from your results.

CC is always worse than NN as clearly stated in the text: "....using NN filtered CC data significantly decreases the false alarms compared to CC data. Yet, we observe that pd's have

also decreased. However, false alarm rates are more dramatically decreased than detection rates.....”

L46 Maybe - is not very scientific - the reasons seem purely speculative, unless there was some statistical or anecdotal evidence for them - your reasons appear without foundation. Unless you can support them as reasons as possible reasons from the data or development context. I would drop them all together. You are dealing with data and 'statistics' and there is always uncertainty and a chance that you will see a result that does not support your hypothesis. The English is weak and ambiguous in this discussion.

First submission of the manuscript did not include these speculations. However, in the previous revision, 'these' possible reasons were asked to be included in the manuscript by other reviewers. Therefore, for the time being we left them in the manuscript.

P18 L35 reveal?

Fixed

**********Experiment #3 But surely it would always be better to use as much data as you possess - otherwise you would be throwing information away?

We agree that more data are good if they contain discriminative info. however, in our case, we show that using more data does not give you anything further. We have analyzed a similar issue on a prior work (ref 32). The important thing is that more data are good only if they contain discriminative information which is not already available in the existing data. Our previous work (ref 32) suggests that static code features reflect a limited information content, hence using more data does not do any better. In general this is supported by the Support Vector Machine research where only the minimum number of 'discriminative' data points are used to construct SVM's. For example, additional data points other than these do not improve the performance of SVM's.

Hypothesis: What is the smallest amount of data needed - might be better? That is if the data can be considered to be representative of the known population of data for a particular company.

We agree. Rephrased the hypothesis as you suggest.

L42 What results? All of them? you need to be more specific.

Fixed. “Analyses #1 and #2 results”. The issue explained in this sentence is applicable to all defect predictors of the first two analyses.

****P19 L29 - 50 It is quite well known that using NN and learning algorithms can lead to worse results when there is a lot of data. This implies (to me at least) that this modelling method does not adequately model the data. In practice we have no control over what data and how much data we get. So we use what we have. Usually the more we have the better the statistical chance there is of finding an effect we are looking for, for a given sample from a known population. We don't normally have the luxury of picking and choosing data to get a better result prior to error detection. Many statisticians would argue that we should never

pick and chose your data without knowing the factors that define the sub-population you are selecting since the results may be biased and not be capable of generalisation to other situations. What we need to know is how do you recognise the most appropriate or similar data to use for learning and NN. And you still haven't told us. The readers need to know.

This section does not include NN and related data selection. We hope that NN issue is now clear as explained in our reply to your previous comments. In this section we take an incremental learning approach to find the smallest number of training examples for defect predictors learned from WC data. For the feature selection issue, please see our reply to your very last comment

P20 L28 offer

Fixed.

L46 English ..... of cases?

Fixed.

L46-50 But surely you have selected 100 observations repeatedly from many more observations using your training/test approach. You do not simply have a single set of 100 observations which you were given. So given any single set of 100 observations we cannot be sure that they will give as good a result as selecting a set of 100 observations from 10000, unless the single set are a random sample from the appropriate population. See my point above about in practice etc. we need to know what the population factors are, i.e. how to choose the NN for a good pd and pf. And to establish what you want we need to be sure that the data collected are a random sample and representative of all possible projects for the company. Will this happen in practice?

***** Really this hypothesis has to do with critical effect size and the sample size is best dealt with using that approach. Yours is an ad hoc approach and I do not think it is likely to be confirmed unless the actual WC data collected is a random sample. You might like to mention critical effect size.

Thanks for pointing out this issue. This is really an important issue with reflections on practical use. We now include this discussion in the text. "However, practitioners should use this approach cautiously. The populations of one hundred examples in our experiments are randomly selected from completed projects with *stratification. Therefore, in practice, any one hundred sample may not necessarily reflect the company characteristics and constructing this initial set may take longer than expected."*

Similarly we added a footnote in the NN section to state the following caveat: "We did not optimize the value of k for each project. We simply used a constant $k = 10$. We consider it as a future work to dynamically set the value of k for a given project."

P21
Section 7 Experiment
L17- 18 English for?

Fixed.

L23 - 25 For clarity these bullets points should be in proper sentence form not abbreviated note form. It does not help the reader as it is.

Fixed: "Then three different types of defect predictors are constructed. First type are defect predictors trained with cross company data (i.e. all 7 NASA projects). Second type of defect predictors are trained with within company data (i.e. random 90% rows of remaining SOFTLAB projects. Finally, the third type are defect predictors trained with nearest neighbor filtered cross company data (i.e. similar *rows from 7 NASA tables)."*

L43 - 44 Does not make sense. Rephrase. L44 - P22 L44 Also needs rewording so that it can be understood. L45-L46 do not seem to connect with L44

Re-written.

P23 L23 with equity - what does this mean? This cannot be the correct word.

Rephrased corresponding part.

So what conclusions should we then draw about your method if whenever pd increases pf increases and pf decreases pd decreases? What does it say about the approach? Why is this important to you - if you can't answer the question I would not keep emphasising it.

******I am still sure that we need to know, in detail, exactly which metrics or features were successfully used in each of your 'experiments'. If they were all used then say so. But in some cases this does not seem to have been the case, since for WC data you have selected certain features or metrics, e.g. your experiment 2 features. An SE needs to assess the worth of those features to their software development and error testing.

This is not specific to our method. It just warns the reader that there is a trade-off between pd and pf (please see ref 5). For issues on feature selection please see our reply to your very last comment.

P23 Section 8 P 24 L4 WC not WV?

Fixed.

L4 - L32 - Still, effort estimation lacks relevance here - it just isn't needed.

Removed.

P25 L32 this effort repeats - I think this could be better worded - it isn't clear what is being repeated here.

Rephrased. "Each member of a review team can inspect 8 to 20 LOC/minute". To remove any confusion we dropped the repeating effort argument.

P26 L23 '..'

Added [21..50] to emphasize that it is an interval.

L45 use of English - only relatively compare - it isn't clear what is meant here drop 'relatively' (I think).

Dropped "relatively"

P27 L2 comma in the wrong place

Fixed.

P27 L47 invisible —- invisibly? Or were invisible?

Fixed: invisibly

P28 L33 our analyses?

Fixed

P 28 L 38 comma? Colon missing? This sentence isn't punctuated correctly?

Fixed.

P29 L11 -18
The point about the data being from culturally diverse development approaches could be expanded, i.e. does this imply that the method is more widely applicable than other methods?

We agree. We can not claim that it is widely applicable than other methods, since we do not analyze other methods. We can at least say that our approach is widely applicable. Expanded this issue by the following argument: "This implies that our approach is widely applicable among different development practices. More precisely, our approach is independent of the processes that yield the final product, at least for the wide range of projects that we have analyzed. "

Conclusions
L26 dramatically - this doesn't mean much? By how much? Can you quantify it using a statistic?

We have now included the median values of pd and pf in the corresponding bullets.

L28 Irrelevancies: how do you know which data are irrelevant? The SE would need to know this or understand how the method decides what is relevant and what is not.

We have added the following: "NN-filtering CC data avoids the high false alarm rates by removing irrelevancies in CC data (i.e. from median value 64 to 32). This removal takes place by automatically selecting similar project data in terms of available static code features and discarding non-similar ones."

L29 No surprise that data from the company will be better than data from another company.

We agree with the reviewer. We would kindly ask the reviewer to consider our previous reply regarding the analysis of WC vs CC data (i.e. Still an open question for cost estimation.)

L42 Is the English correct ? The definite article? company round the globe?

Dropped the term: " ...a completely different company located in another country..."

L 47 - 50 This is not a new idea this is one of the reasons there are global databases. If my memory is not faulty ISBSG had error data from many companies as well as effort estimation data. SEs know this. I think the problem is how do we know which data will suit any given company's needs or how do we know what data items are irrelevant. I think this needs to be addressed in this paper. It doesn't seem difficult to do.

Thank you very much for pointing this out. In light of your other comments and suggestions we believe we have addressed this issue and clarified it in the revised manuscript. So the paper now clearly states that NN filtering directly addresses the automatic selection of relevant data

Comments 1

The authors still do not define the sub-set (or sets) of metrics that are the most successful in predicting errors and which are irrelevant - this would seem to be the most important aspect of the work for a software engineer. I can't imagine that all the features given in the Tables 3 and 4 are used each time - and they do say that there are irrelevancies amongst some of the features in the WC data and they need to be removed.

However, it might be that the authors learning method actually chooses a different sub-set of predictors (features) for each different sample sub-set of data drawn from a complete set of data (and features). (But they don't say this in the paper, as far as I can tell.) However, it is important to tell the reader that this is happening - if indeed it is.

If it is not, then I think the reader would like to know why they are not able to specify the best set of features to use for a given set of data. This seems like a natural question for the reader to want to ask. Also, in this way the SE will have a better idea of what to expect from an application of the approach to their company's data.

For me the problem with the authors' paper is that they don't tell us which sub-set of the metrics are actually being used for any set or sub-set of data they analyse. So we never seem to know which sub-set of features is going to be used for any given sub-set of data. If this is correct, then they should state this clearly so that a SE is not misled. Then as SEs we must assume that we must always collect all possible metrics and let the algorithm decide what is suitable for any given sub-set of data. If I am right about this, then I think an SE might want to know that they are going to put their trust in

We thank the reviewer's efforts and attention on details that were really helpful to improve the quality of our paper. We believe our replies together with this revision make things more clear. Section 5.1.1 is now reserved for NN-filtering. We do not perform feature selection, we use all available common features and the reasons are now explained in a very detailed manner in the new Data section. This part, we believe, fulfills your major concern

about SE expectations of metric subsets. We copy the [start-end] of the related part below, please refer to the text for detailed explanations:

"In all analyses, we used all available features to learn defect predictors. One question is whether it might be better to focus on just some special subset of those features. In previous work, we have found that there is no best single set of features. In fact, this best *set is highly unstable and changes dramatically from data set to data set. That study is summarized, below. The main message of that study is that it is "best" to never assume a* best *set of features. Rather, give them all to a learner and let it work out which ones are best to use or ignore for a particular application.....*

.....

..... the following principles:

- *Do not seek "best" subsets of static code attributes.*
- *Rather, seek instead for learning methods that can combine multiple partial defect indicators like the statistical methods of Naive Bayes.*

Therefore, we do not perform feature subset selection, rather we use all available features in our analyses."

1.4 Replies to the Reviewer #3 Comments

Reviewer #3: The manuscript has considerably improved with regards to readability and clarity. All my earlier review comments have been properly addressed. I think the authors now state much more precisely what exactly the purpose of their research is and how it can help in industrial practice. Therefore, in my opinion, it is absolutely worthwhile to publish the manuscript as-is. Given the rigor the authors have now used in writing up and discussing their results, the manuscript might trigger a quite interesting discussion about the practical value of quality prediction models derived from pure company-external (denoted as "cross-company"), filtered company-external, and company-internal (denoted as "within-company") data.

We thank the reviewer for the positive opinion. This has been possible with the valuable, constructive and insightfull comments of all reviewers..

My only (and last) suggestion relates to the title and keywords: The idea of bridging the gap between using (external) cross-company data and using (internal) within-company data through filtering cross-company data seems to play a prominent role in the authors' suggested method. Nevertheless this element of the method is neither mentioned in the title nor in the list of keywords. Perhaps it might be a good idea to (at least) mention NN-filtering in the list of keywords.

Thanks for the suggestion. It is now added to keywords.

The manuscript still contains a small number of grammar and spelling errors. I assume these will be addressed by the editorial procedures in case the manuscript is accepted. Thus

I don’t mention them here.

We did our best to correct these.