The use of experts in metrics interpretation and analyses

Kari Juul Wedde and Tor Stålhane

Abstract
The main problem with a measurement-based approach to software process improvement is that measurement on a software development process seldom yields many data. Thus, the data can seldom be analysed by statistical methods alone.

Our solution is to combine statistical data analysis with the use of experts. In our context, the experts are the persons that participate in the development process. These persons will, however, not stay experts, except if they get continues training in the shape of feedback from the real data and from the process.

The experts help by providing interpretation of the data, by identifying reasonable models and by identifying improvement opportunities. By stating hypotheses regarding relationships and future observations and then seeing the results, they can update and maintain their expertise. This paper sums up our approach and illustrates it with examples from Norwegian software industry.

1. Introduction
Software Process Improvement (SPI) is about learning to work smarter – continuously changing the processes to get a working environment in which people can do a better job. The underlying assumption is that an improved process will lead to better, faster and cheaper software development. But how can we change these processes and at the same time be sure that the changes are for the better?

Firstly, we have to understand and accept the nature of software development as a creative and innovative process. In spite of all the methods and tools appearing on the marked at a high frequency, software development is in the end a people process. Secondly, systematic and continuous improvement should be based on facts, and these facts come from measurements. In an ever-changing environment, however, it is impossible to collect enough metrics data to draw conclusions based on statistical data analysis alone. We have to include the experts in the interpretation and analyses of metrics data. In this paper we will look into a few questions related to the use of experts:

“Who are the experts”, “When do we need the experts”, “How to use the expert knowledge” and – may be most important - “How can the experts maintain their knowledge - in order to remain experts”.

We will answer these questions based on our experiences from SPIQ (Software Process Improvement for better Quality). SPIQ is a major Norwegian software improvement initiative with the overall goal to increase the competitiveness and profitability of Norwegian software industry by introducing systematic and continuous process improvement.

2. The role of experts in measurement programmes

2.1. Who are the experts
Who are the experts in a software measurement programme? Are they SPI people, data analysis specialists or software engineering best practice specialists? All these specialists can contribute in a measurement programme, but in our opinion the real experts are those who participate in the project - works with the process – from which we collect data. They are the
only ones that have first hand knowledge on how the process works and what influences what, and they are therefore able to interpret the metrics data. Even though all the people that work on the measured process are – in some sense - experts, we do not have the same confidence in the information from each of them. The confidence you can have in expert information depends on the expert’s:

- General knowledge, education and experience
- His role in the project related to the information given
- His ability and opportunity to maintain his expert knowledge [6]

The last point is important. In order to be real experts and to stay in that capacity, they need to continuously test their understanding and knowledge, and they need feedback on the validity when they give information as experts.

It is our experience that this is best done by using feedback sessions. It is, however, important that they receive feedback also when they are not participating in a measurement programme. There are several ways to achieve this, for instance through the PSP (Personal Software Process), participating in reviews – also for cost and time estimates, reliability estimates and so on.

2.2. How to use the experts

Expert knowledge is always important when interpreting the results from data collection and analysis. In our approach, this analysis takes place during the feedback sessions. During these sessions, there are two situations where the participation of experts is crucial for the quality of the results in a measurement program:

- When we have few data – less than five to 10, say.
- When we have many data – enough for a statistical analysis – but when plotting the data, the picture is unclear due to large variations or contradictory data

The experts can give important contributions during data analysis in such cases. The following are the most important points:

- Are the data reasonable? This really boils down to the question: “Do you believe in these data”? This is important for two reasons. Firstly, nobody will act on data they do not trust. Secondly, it will help us to identify problems related to the data collection process. Unfortunately, it may also cause us to throw away extreme data that could have provided new and interesting knowledge about our process.
- Do we have enough data to interpret what we have collected? One of the important questions is whether all the data should be analysed together or split up into subgroups.
- What is the most probable functional relationship between the data? Given that the data should be analysed together, what is their relationship? Will a linear model be most appropriate or do we for instance know about limiting factors, which would make an asymptotic function more appropriate?

3. The SPIQ approach

3.1. The main ideas and solutions

The SPIQ approach to SPI is based on a few vital elements [3]. The first one is that all process improvement should be based on company needs – not on standard models. The company has to set goals according to process status and their marked situation. The second element is facts. Without measuring the effect of improvement actions, it is difficult to know
whether the goals are met or not. The third element is participation. The software development process is a people intensive process. Thus, measurement can not be performed without participation from the people involved. The fourth element is risk management. Including people in data analyses also means to take risks. The last element is learning - both on the organisational and individual level. Maintaining expert knowledge is part of this learning.

The implementation is based on the principles of TQM [5]. The vision of improved quality is a motivating factor and the TQM tools - both the quality tools and the management and planning tools - are part of our approach. Statistical Process Control (SPC) is important in TQM, but not in the SPIQ approach. We have not found SPC to be particularly useful for most software companies [8] [11], and have instead chosen Goal Question Metrics (GQM) [1] as our main method for data collection and analyses - together with Root Cause Analyses (RCA) [9]. Our learning part is based on the ideas in the Experience Factory [2].

3.2. GQM

GQM is a goal-oriented approach to measurement that has gained widespread acceptance. The idea is simple:
1. Set goals pertaining to needs in terms of purpose, perspective and environment.
2. Refine the goals into questions that are tractable.
3. Deduce the metrics and data to be collected in order to answer the questions.
4. Interpret the data, answer the questions and draw conclusions.

Systematic process improvement is based on reliable data combined with knowledge on how the job is actually done. This means that you need experts that know the process, and in a GQM context, the experts on a process are those performing this process. The experts are used in all four of the steps mentioned above, but in the context of this paper, activity number four - also called feedback sessions - is the most important one.

The GQM abstraction sheet is a means for acquiring the information needed for defining the GQM measurement plan. The abstraction sheet has four quadrants – see Figure 1. The upper quadrants show the Questions while the lower half shows the Hypotheses. The hypotheses are important in order to verify the validity of the Questions, and to form models for data analysis. The Baseline Hypotheses quadrant documents the developers’ view of the current status of the measured properties. This means that the team has to use their current knowledge about the process to answer the defined Questions.

The Impact on Baseline Hypotheses quadrant describes the impact of the Variation Factors on the Quality Focus. The developers have to state how they believe that the variation factors will influence the quality focus. This can then later be used to give feedback to the developers in their role as experts.

<table>
<thead>
<tr>
<th>Object</th>
<th>Purpose</th>
<th>Quality Focus</th>
<th>Viewpoint</th>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality Focus:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>etc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline Hypotheses:</td>
<td></td>
<td>Variation Factors:</td>
<td>Question a</td>
<td>Question b</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>etc.</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: GQM abstraction sheet and corresponding GQM tree
3.3. Root Cause Analyses

RCA is one of the most important and versatile of the TQM tools. It consists of three components – data collection, Pareto analysis and the Ishikawa diagram. The data collection helps us to find the facts, the Pareto analysis helps us to focus on the most important problems while the Ishikawa diagram helps us to identify the root causes of the problem. The ideas behind RCA are simple and the topmost level of the approach is as follows:

1. *What is the problem:* identify what we want to improve
2. *Data collection and analyses:* Perform a Pareto analyses on the collected data
3. *Problem analysis:* Use the Ishikawa diagram to identify the *root causes* of the problem.
4. Identify improvement actions

![Figure 2: Pareto diagram](image1)

![Figure 3: Ishikawa diagram](image2)

3.4. The combined method

As you can see from the descriptions of GQM and RCA, they have the same purpose and approximately the same activities. They are using different tools and that is why we have found it useful to combine these two methods. The strength of GQM is that it includes the environment, and the GQM abstraction sheet is a good tool for organising available knowledge. Another strength of GQM is the feedback sessions. It is in the feedback sessions that the experts' contribution has the greatest influence. RCA is not using feedback session for data analysis. Instead they have problem analysis - a brainstorming session with the Ishikawa diagram as the main tool.

<table>
<thead>
<tr>
<th>No</th>
<th>Activity</th>
<th>GQM</th>
<th>RCA</th>
<th>Combined method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Define the improvement strategy</td>
<td>Set goal</td>
<td>Write down a problem statement</td>
<td>Set goal</td>
</tr>
<tr>
<td>2</td>
<td>Define metrics and data collection procedures</td>
<td>Define questions, metrics and hypotheses</td>
<td>Identify problem occurrence data</td>
<td>Define questions, metrics and hypotheses</td>
</tr>
<tr>
<td>3</td>
<td>Data collection</td>
<td>Collect data</td>
<td>Collect data</td>
<td>Collect data</td>
</tr>
<tr>
<td>4</td>
<td>Data analyses</td>
<td>Make plots</td>
<td>Pareto analyses</td>
<td>Make plots</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Feedback sessions</td>
<td>Feedback sessions</td>
<td>Feedback sessions</td>
</tr>
<tr>
<td>5</td>
<td>Problem analyses</td>
<td>Ishikawa diagram</td>
<td>Ishikawa diagram</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Identification of improvement actions</td>
<td>Feedback sessions</td>
<td>Feedback sessions</td>
<td></td>
</tr>
</tbody>
</table>
From Table 1, you can see that the combined method follows GQM for the definition of the measurement programme. The data analyses activity is a combination of GQM and RCA - making plots and performing feedback sessions to interpret the data before a Pareto analysis is conducted on the resulting data set. The problem analysis activity is optional, but if performed, we use the Ishikawa diagram, and feedback sessions are used for the identification of improvement actions.

4. Feedback sessions

Feedback sessions are organised meetings where the people involved in the measurement programme get together regularly in order to interpret and analyse the collected data. How often depends on the amount of data, but once a month is rather normal. Why do we need experts – the people working on the measured project - to perform these tasks? Will not the data themselves give most of the answers, so that a data analyses specialist will be the best person for the job? As said in the introduction, most software projects do not yield enough data for us to rely on statistical methods alone. The data are just indicators that need interpretation, and a data analysis specialist will not have the necessary knowledge of the process to do this.

The list below describes the main activities of the feedback sessions and the role of the experts in each activity.

- **Validate the data.** All data have to be validated, and even though all data have been through validation by the measurement manager at collection time, the experts are more competent to answer the question “Are the data reasonable – do we believe in them?”
- **Look for outliers, trends etc. and seek explanations.** In data analyses, outliers are often excluded from the data set. If we have few data we can not just do that. We have to seek explanations for outliers, trends etc. This is necessary in order to know how to use the data set, which model to use for interpretation and so on. Only people working on the project can provide such explanations.
- **Provide answers to the defined questions.** This is the main task of the feedback sessions. Take the data available, propose an answer to the question and decide on the confidence you have in the data and in the answer. The confidence needed depends on the risk the company is willing to take. Large data sets reduce the risks, but take more time and thus more resources to collect.
- **Identify necessary changes to the measurement plan.** Examples of such changes are new metrics and changed data collection procedures. The role of the experts is to decide if the current metrics are sufficient to answer the questions. A normal reaction during a feedback session is: "In order to answer question one, we also need to know ....". It usually takes two to three feedback session to stabilise a measurement programme.
- **Provide ideas to short term improvement action.** Short-term improvement actions are improvement actions that can be implemented during the lifetime of the measurement plan. Having the people involved in the process to interpret the data will always bring improvement ideas to the surface - “What if we do it this way instead, then this problem will disappear”. Running SPI experiments after the book means that such improvement actions should not be performed, but few companies leave a good idea to the next project just to control the experiment. It requires however, that the data sets in later analyses are split into two, before and after the change. This, in turn, gives fewer data in each set and forces us to rely even more on the experts.
- **Provide ideas to long term improvement actions.** This is not the most important task of the feedback sessions. Good ideas should, however, be documented.
The feedback sessions also play an important role in maintaining the experts’ knowledge. The meetings give the experts feedback on the validity of their information as experts, and thus increase the confidence in the experts’ judgement.

Figure 4 sums up the role of the experts in feedback sessions. The expert combines his knowledge with information from the measurement plan, metrics data and the hypotheses. Based on this combined knowledge, the measurement plan and the hypotheses are revised. Improvement actions are identified, and last but not least, the expert’s own knowledge is revised. An expert that does not get any feedback on his decisions, will not remain an expert for long. In this respect, the hypotheses are of special interest. The hypotheses give the expert’s current view on the measured properties. The collected data will give direct feedback on this view and eventually the need for an update.

Figure 4: The role of the experts in Feedback Sessions

5. Data analyses

5.1. Testing of hypotheses

Data analyses are used in two areas of work in our method – to check the stated theses related to Baseline Hypotheses and Impact on Baseline Hypotheses in the GQM abstraction sheet and to identify improvement opportunities in the development process.

The main problem with testing hypotheses in SPI is the small amount of often seemingly conflicting data. The small data sets give us two problems. One is the problem of being able to accept or reject a stated hypothesis with any decent level of confidence, but even more important is that the results of an analysis will change dramatically if one or a small set of data is removed or included.

The hypotheses are important since they help the developers to check their expertise on the development process and the resulting product. Thus, it is important to be able to give the experts adequate feedback in these matters and with a certain level of confidence.

This can be handled and put into a formalised framework if we are able to attach a confidence rating to each data or set of data. The problem will then be to arrive at a model that maximises both the probability and the confidence. An example is in Figure 5.

The left-hand case in Figure 5 shows two sets of data for the high values – denoted A and B respectively. If we just look at the data, it is not possible to decide whether we should declare group A as outlier and use a linear model or drop group B and for instance use an exponential model instead. Remember that the characteristic “outlier” is not a characteristic of the data, but of a relationship between data and an assumed model [3]. Thus, the confidence we have in each data and the confidence we have in the selected model will be of paramount importance. Only an expert can tell us where to place the confidence.
The right hand case shows a different type of dilemma. One natural conclusion could be that the data set shows a large variance. An expert may, however, possess knowledge that tells us that the two groups of data – C and D – should not be mixed. They can for instance relate to two types of software components. Thus, the solution is to split the data set into two new sets and use a linear regression model for each set.

![Figure 5: Examples of data plots with a confusing message.](image)

### 5.2. Robustness analysis

Robustness analysis is a hybrid technique since it uses a straightforward statistical technique to identify one or more data sets and then uses expert opinion to make a decision. In our work, robustness analysis has only been used as a part of the Pareto analysis but the techniques is general and can - and maybe should - be used as part of any statistical analysis in SPL.

The analysis is simple but tedious to do manually. It can, however, easily be automated. The method can be described as follows:

1. Draw conclusions based on the complete available data set. Call this conclusion C.
2. For a type n robustness analysis, remove n data. If we have a total of N data, this will give us N over n new, reduced data sets. For the sake of brevity we call this number X.
3. Draw your conclusions based each of the data sets generated in step 2. This will give you X new, not necessarily equal, conclusions.

The new set of conclusions can be used in two ways:

- If y of the X conclusions differ from the original conclusion C then we will assign the probability of y/X to the event that the original conclusion is wrong. We group all the data that change the conclusion into the critical data set. This is a straightforward statistical conclusion. If the number y/X is small – less than 0.10, say – we have a standard risk assessment situation. We will have to accept or reject the risk depending on its potential impacts.

- If the number y/X is too high, we have to assess our confidence in the data sets that – if removed – change the original conclusion C. This is where the experts are important. If they indicate that the data sets that change the conclusions are freak events or just plain wrong, we can remove them and keep conclusion C. If the expert claim that the data in the critical data set are just as good and just as relevant as the others, then we should either analyse the results from both Pareto analysis in the following Ishikawa diagram or check our data collection scheme again in order to improve it.
6. Examples

This chapter will give some examples from Norwegian industry - illustrating the role and importance of experts.

6.1. Example 1

The goal for this company was decreased maintenance costs. When they developed their GQM abstraction sheet and their measurement plan, one of their Impact on Baseline Hypotheses was that both the number of errors and the maintenance costs increased with the module’s complexity.

As a result of this, the developers defined a set of characteristics that relate to module complexity. These characteristics should – if successful - be used to focus the testing effort during development and integration. We defined several software characteristics that conventionally are linked to complexity - and thus to the number of errors - such as the number of decisions, the number of input and output parameters and signals and so on, but they all explained only a small part of the variation of number of errors in the components. It was possible to build a model with several parameters that gave acceptable results but the resulting models were complex and thus not particularly useful.

During a feedback session, where these sad data were presented, the developers claimed that the main source of errors was that the design document – written in SDL – became too big. As a result of this, we decided to register the number of pages in the SDL descriptions for each module. The new model, with the number of pages in the design document as the only independent variable, was a success having an adjusted $R^2$ of 0.80. The other, software related factors turned out to be just noise.

6.2. Example 2

The goal for this company was increased reliability. When they developed their GQM abstraction sheet and their measurement plan they also filled in the Impact on Baseline Hypotheses field. One of their hypotheses was that the use of checklists made reviews more efficient – defined as identification of more problems per person-hour spent reviewing.

When the first batch of data had been collected and we should prepare the material for the feedback session, we thought that the difference in efficiency between those reviews that used checklists and those that did not should stick out like a sore thumb. Unfortunately, neither data plots, nor standard statistical testing procedures such as the t-test or ANOVA showed any difference as far as the number of problems was concerned. The only difference was that those reviews where the checklists were used consumed more person-hours.

We presented this rather disappointing result to the developers in the feedback session. The developers were able to provide the solution: The checklists used in the company focused almost only on document layout and not on contents, consistency and so on. No wonder they used more person-hours.

6.3. Example 3

The goal for the company under consideration here was straightforward – they wanted to reduce the number of errors inserted into the code during development [7]. They started out with a GQM process in order to decide on the data to collect and how to collect them. This gave them a data collection schema and two sets of error categories – one set related to the coding process and one related to the project’s environment.

When they had collected data from two projects they performed a Pareto analysis and a robustness analysis on the data. The robustness analysis indicated that the risk of using the
data as is was acceptable – less than 20%. The two most important error causes identified by the Pareto analysis were “Incomplete requirements” and “Incorrect use of language features”. These were then analysed further by using the Ishikawa diagram.

Given enough time and resources, we could instead have kept on collecting data in order to find out what characterises a project with incomplete requirements and then repeat the process of data definition, collection and analysis but probably not even the most adherent purists would take matters that far. Anyhow – the use of experts supported by the Ishikawa diagram enabled us to identify a set of important root causes, which could be implemented in the organisation.

7. Conclusions

To conclude, we will answer the questions stated in the introduction related to the role of experts in the interpretation and analyses of metrics data.

The first question was “Who are the experts?” In the context of SPI measurement programmes, the experts are those who do the real work – perform the process – and thus have first hand knowledge on how the process works and what influences what.

The second question was “When do we need the experts?” In SPI, the experts are needed in defining the improvement strategy - problems to be solved and opportunities to take advantage of. They are needed to define the metrics and the data collection procedures, to interpret and analyse the metrics data and to identify and implement improvement actions. In the SPIQ approach, the interpretation and analyses is performed in feedback sessions, and as our examples have illustrated, feedback sessions without those experts, have minimal value.

The third question was “How to use the expert knowledge?” The expert knowledge is mainly used in the feedback session and in brainstorming sessions such as the use of the Ishikawa diagram. This way the experts can obtain common experiences by converting individual tacit knowledge into explicit shared knowledge [10].

And last but not least “How can the experts maintain their knowledge – so that they remain experts?” The only way experts can maintain their expertise is when they are able to continuously test their understanding and knowledge and get feedback on the validity when they give information as experts. In the SPIQ approach – and in GQM – this is in general ensured by participation in feedback sessions and especially by the use of hypotheses.

8. References