LETTER FROM THE PRESIDENT

The case for simplicity

The president of ENBIS, Fabrizio Ruggeri, shows, using examples from his own work, that we cannot always treat complex problems with simple methods

In my previous letter I mentioned that complex problems often cannot be treated with simple methods, although Occam's razor should guide us in our work: *'One should not increase, beyond what is necessary, the number of entities required to explain anything.'*

To argue in favour of this statement, I would just like to mention some works in which I have been involved in recent years. One of the latest cases involved noisy measurements from an atomic force microscopy (AFM) experiment. AFM is a type of scanned proximity probe microscopy that can measure the adhesion strength between two materials at the nano-newton scale. In AFM, a cantilever beam is adjusted until it bonds with the surface of a sample, and then the force required to separate the beam and the sample is measured from the beam deflection. Beam vibration can be caused by factors such as thermal energy of the surrounding air, or the footsteps of someone outside the laboratory. The vibration of a beam acts as noise on the deflection signal. This noise must be removed for the data to be useful. The model we used was based on wavelets, a sophisticated mathematical method that allows any function in a given set (namely, the space of square integrable functions on the real line) to be represented as a mixture of scaled and shifted versions of one or two functions. Therefore, the signal is considered as a function and represented as a mixture of wavelets, whose weights (or coefficients) are the object of the statistical analysis. It is important to check if the coefficients are significant (that is non null in this context) and statistical methods (Bayesian ones in our case) and decision analysis are used to estimate them. By properly setting many coefficients equal to zero and removing the corresponding wavelet functions from the mixture, we obtain a de-noised signal.

We also used wavelets to consider input/output operations in a computer disk and to model the number of packets transferred over a given period. Evaluation of the performance of storage systems is a key aspect of the design and implementation of computers with heavy I/O workloads. There are difficulties and expenses involved in obtaining actual measurements of disk usage, so disk performance simulators are usually fed with synthetic traces. Unfortunately, the generation of realistic disk traces is a difficult and unsolved task. As in many areas of computing and telecommunications, the time processes of disk usage exhibit dependencies that span over long ranges. Furthermore, the series have bursty* behaviours that cannot usually be captured by commonly-used times series methods or by exponential, or Poisson, arrival times. Fractional Brownian motion and fractional Gaussian noise are unrealistic for various reasons. Multiplicative cascade models have been considered in the literature as a way of capturing the sporadic behaviour of the series. Such models have multi-fractal properties, providing a rich structure that is able to capture the behaviour of series of disk usage. A cascade model on the coefficients of a Haar wavelet allows simulation of a multifractal process.

In another work on telecommunications, we used Brownian motion to model a cumulative traffic network and to compute the loss probability. This is the probability of losing traffic data and it represents a measure of the quality of the service provided by the system.

Other models we considered were dynamic linear models, in which costs of different activities in the construction of industrial plants change over time depending on some evolving variables. Costs are forecast before bidding for the plant construction, using both past data and experts' opinions, as prescribed by the Bayesian paradigm.

We have used Bayesian belief networks to model human and organisational factors leading to accidents in the maritime industry. Finally, we have devoted most of our recent work to describing failures in repairable systems, like gas escapes and subway doors. The model used in this case was a non-homogeneous Poisson process.

You can find more details on these works on my web page: www.mi.imati.cnr.it/~fabrizio.

The methods I have illustrated are typically considered in the talks and the workshops organised the ENBIS Special Interest Group (SIG) on 'Stochastic Modelling'. ENBIS has SIGs and they are the places for networking (one of the motivating reasons for the existence of ENBIS).

There are lights and shadows in the activities of the SIGs and our goal is to increase the number of people being active in them. Interested people are warmly invited to attend not only the annual ENBIS conference in Wroclaw (Poland) from 18-20 September, but also to attend the workshops organised by the SIGs and the meetings of the SIGs when yearly activities are planned and issues relevant for the SIGs are discussed.

* You may not find 'bursty' in your dictionary but it is used in the literature about these time processes to mean that every once in a while in a usual time series there is a big jump: a burst.

A Variation Reduction Algorithm

Stefan Steiner and Jock MacKay propose a novel solution to a problem that occurs in a wide variety of engineering, scientific or management processes

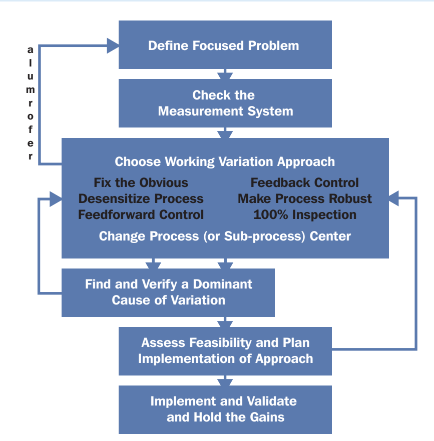


Figure 1: **A variation reduction algorithm.**

Reducing variation in outputs is a key part of process improvement. For mass-produced components, reducing variation can reduce cost, improve function, and increase customer satisfaction all at the same time. Excess variation can have dire consequences: the component may have to be scrapped and reworked; more inspections; more customer returns; impairment of function; reduced reliability; and durability. Examples of variation reduction, in a wide variety of engineering, scientific or management processes, are:

- Reducing rework due to brake rotor imbalance;
- Improving the reliability of a personal digital assistant;
- Increasing average yield of a chemical process;
- Reducing variation in camshaft lobe geometry; and
- Reducing the average number of errors made in the translation of technical documents.

Variation reduction is best done step-by-step. In Figure 1, we propose an algorithm for high- to medium-volume manufacturing processes. We designed it to identify low-cost changes such as improvements to the control plan or new process settings.

In most applications of the algorithm, we search for a dominant cause: a variable input that is responsible for much of the variation in the output. If we apply the Pareto principle (80/20 rule) to causes of variation, we expect that only a few will be dominant. They may be 'special' or 'common' in the language of Statistical Process Control.

We illustrate the algorithm with a case study. An iron foundry produced veined brake rotors for machining elsewhere. The machining plant inspected 100 per cent of the rotors for balance and welded a weight into the veins if the imbalance was too severe. We call a rotor needing added weight 'a balance reject'. The reject rate, initially around 25 per cent, jumped to 50 per cent coincident with a change from a four-cavity to a six-cavity core mould to increase productivity in the foundry. The cores create the veins when the rotor is cast.

The foundry was convinced that the change to the six-cavity mould did not increase the balance rejects. A dimensional analysis of the six-cavity mould and core-making process had shown all characteristics well within specification. The increased reject-rate could not be explained by any other changes. Each party blamed the other. The foundry formed a problem-solving team to return the reject rate to its historical level.

To determine imbalance, the machining plant measured the centre of gravity (a distance and direction from the rotor centre) that was then translated into a weight (gm) and orientation needed to balance the rotor. A balance reject was any rotor needing weight greater than 15g. The team selected balance weight as the output because they knew that if they could reduce the weight, they could eliminate the rework, regardless of the orientation.

The team selected 300 rotors from the previous week's production (Figure 2). The team set its goal as 'at least 75 per cent of

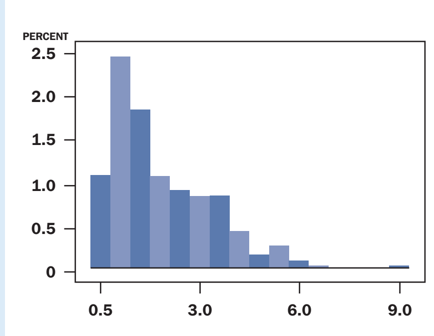


Figure 2: **Balance Weights in the Baseline Investigation**

the rotors requiring a weight of less than 15g'.

We have seen many failures in such exercises because a team tried to reduce variation without an adequate measurement system, so in the second stage of the algorithm, we ensure that the measurement system is not itself a dominant cause of variation. In the case study, the team selected three rotors with initial measured weights of 3, 10 and 32g, and then measured them six times each on three separate days. There was little operator effect since the gauges were automated. The measurement system was highly repeatable, relative to the variation seen in the baseline study.

At the next stage of the algorithm, we consider how to reduce the output variation. Approaches are:

- Fix the obvious: Use knowledge of a dominant cause to implement an obvious solution;
- Desensitise the process: Change the process settings to reduce the sensitivity of the output to changes in a dominant cause; and
- Implement feed forward control: predict the output using measured values of a dominant cause; adjust the process to reduce variation.

Approaches not requiring the identification of a dominant cause are:

- Implement feedback control: Predict the output using current and past output values and adjust the process to reduce variation;
- Make the process robust: Change the process settings to reduce variation;
- Use 100 per cent inspection: Use an inspection scheme to select units with less variation in the output; and
- Move the process centre closer to the target.

To choose a working approach, we consider the nature of the problem and process, our current state of knowledge, and for each of the seven approaches, the:

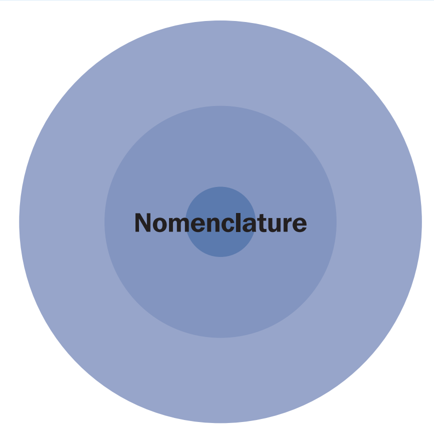


Figure 3: Concentration Diagram of Weight Locations

- Knowledge required to implement the approach;
- Likelihood and cost of obtaining this process knowledge;
- Likelihood of successful implementation; and
- Probable cost of implementation

In the brake rotor case, the team first considered the non-cause based approaches. They ruled out 100 per cent inspection since that was current and costly. They eliminated feedback control since there is no strong pattern in the variation over time, so they did not know how to adjust the process. Robustness or move the process centre were possibilities but, without more process knowledge, were unlikely to succeed. The team decided to proceed with a search for the dominant cause.

We recommend the method of elimination: partition the causes of variation into families and use new or available data to rule out all but one as the dominant cause. We can use elimination recursively to narrow down the potential dominant causes.

The brake rotor team first recorded the location of the welded rework weight for the 140 rejects from the baseline investigation, on a concentration diagram (Figure 3). There was a non-symmetric pattern of balance weight locations. Since the machining process is rotationally symmetric and the casting process is oriented, the team had eliminated all causes in the machining operation.

The team next compared 30 rejects and 30 balanced brake rotors. They measured 26 foundry-determined characteristics on each and identified two, thickness variation and core position (offset), that were substantially different for balanced and unbalanced rotors.

The team wanted to verify that core thickness and position were substantive causes and the six-cavity mould was not. They used two levels for each input and a 23 factorial design.

From the interaction plot in Figure 4, the team concluded that low thickness variation using the four-cavity mould produced the optimal results. Thus, the dominant cause lay in the core moulding process.

In general, after we verify a particular input as the dominant cause, we consider the feasibility of the cause-based approaches. If we rule out the cause-based approaches, we have three options:

- Reformulate the problem in terms of the dominant cause;
- Reconsider the non-caused based approaches; and
- Search for a more specific dominant cause.

If we decide to reformulate the problem, we restart the algorithm with the goal of reducing variation in the dominant cause. We sometimes reformulate a problem several times. Eventually, we must select one of the variation reduction approaches.

Having chosen a variation reduction approach, in the next stage of the algorithm, we look in detail at the feasibility of the selected working approach. We:

- Examine the process to see if it meets the conditions for the approach to be effective;
- Determine what further knowledge is required;
- Plan and conduct investigations to acquire the knowledge;
- Determine the solution, i.e. how the process will be changed;
- Estimate the benefits and costs of the proposed change; and
- Look for possible negative side effects

If the working approach is feasible, we validate and implement the solution. Otherwise, we must reconsider the other variation reduction approaches.

In the brake rotor case, the team made the 'obvious fix' and recommended that the foundry go back to the original four-cavity core mould. The reject rate dropped to its historic levels. The team had met the project goal. Alternatively the team could have considered trying to compensate for the variation in core thickness. In theory this could be accomplished either by making a one time change to the process (desensitisation) or measuring core thickness variation for each rotor and adjusting the process as needed for each rotor to compensate for the measured thickness variation (feed forward control). Both these suggestions were rejected in favour of the more direct obvious solution.

A new core-making process was already available in the plant but not in use. The team knew that this cold box process was stable dimensionally and they expected much less thickness variation with this process.

In the final stages of the algorithm, we reassess the baseline performance after the process change to ensure that the project goal has been met. We also examine other process outputs to check for negative side effects. Finally we implement and lock the change into the process or its control plan. We recommend monitoring the process output and auditing the process change until we are certain that the solution is effective and permanent.

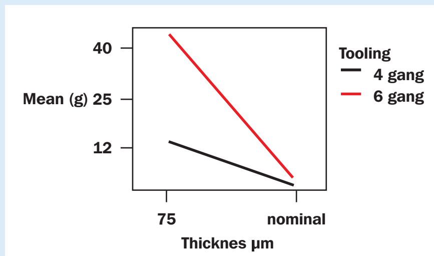


Figure 4: Interaction Plot from Verification Experiment

With the implementation of the cold box method, the process was greatly improved. The rate of balance rejects dropped to 0.2 per cent, a large reduction from the 50 per cent at the start of the project. The machining plant eliminated the expensive rework stations and scrapped the few remaining balance rejects in the new process.

The algorithm works well in manufacturing. Its unique features are the search for a dominant cause using the method of elimination and the demarcation of the variation reduction approaches. To be successful, the algorithm should be embedded in a global improvement system such as Six Sigma.

Stefan Steiner and Jock MacKay,
Department of Statistics and Actuarial Science, University of Waterloo,
Waterloo, Ontario, Canada
shsteiner@uwaterloo.ca
rjmackay@uwaterloo.ca

STATISTICS UNRAVELLED

Beware of serial correlation

John Logsdon again urges you to understand the structure of your data

In the last article, we discussed the importance of the structure in any data analysis. In this issue, I focus on a structure not often recognised that would fail the simple test I gave last time, unless you thought about it. The structure is serial correlation, where successive measurements are made on the same item. It is almost always analysed incorrectly.

One of my first statistical jobs was to analyse some corrosion data. There was a large number of test coupons of different types and compositions oxidised at a variety of temperatures. Each coupon was weighed periodically and the increased weight with time represented the metal lost. If a single time step showed enhanced oxidation, we would expect the subsequent growth to be lower. This is an example of serial correlation: each observation depends on the previous observation. In this instance, the correlation is negative.

Most scientists use standard statistical programs to analyse such data. This will generally give a sensible central prediction but the confidence intervals will be completely misleading.

Consider a simple model of the form

$y = Ax + B + \varepsilon$

where ε represents the errors in y where the trend is linear. The errors are completely independent: for example the error in the third observed value depends in no way on the second observed value.

We simulate such an experiment. Generate 100 data sets based on a given model, such as

$y = x + \varepsilon$ so A = 1 and B = 0

Each data set has x = 1, 2,…,50. The values of y are calculated from the formula, adding random numbers sampled from a normal distribution with mean = 0 and sd = 0.2.

Use standard regression to fit the model to the first ten points only. Estimate values of slope and intercept (A and B) from the data in each set and use this fitted model to predict y values for all 50 points. Each prediction line also has its own prediction error and we calculate the root mean square of all these errors. This is the naïve prediction error. We also calculate the standard deviation of the 100 sets of actual y values. This is the actual prediction error.

The source for this experiment may be found at http://www.enbis.org/SCW4.R.

Using R, we load all the commands by:

```
source('SCW4.R')
run.basic()
```

will generate the model above. Fig. 1 compares the naïve predictions from a standard model in red with the standard deviations calculated at each step from actual values for each of the family of curves, shown in green. They are quite similar.

FIG. 1: Basic linear case uncorrelated

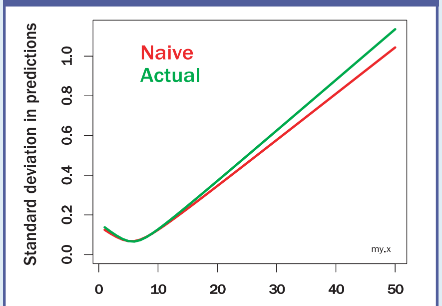
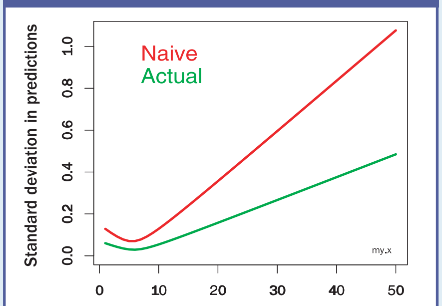


FIG. 2: Correlation = –0.701 sd = 0.203



We compare this result with a calculation where there is serial correlation, as in Fig. 2, generated by:

```
run.simple.corr()
```

This, by default, assumes a correlation of -0.7 between successive values. The naïve red line is nowhere near the green lines calculated from the actual data. On the other hand, setting the correlation, for example, to +0.7 will show that the naïve line underestimates the envelope.

```
run.simple.corr(corr=0.7).
```

Setting the correlation to zero will show that all three lines are again similar.

This experiment shows that wherever the correlation is non-zero, the predictions from a standard statistical package are wrong.

Imagine the implications of ignoring the true confidence intervals. As well as in corrosion, serial correlation occurs in things like wear so that, if the correlation is positive, failures could occur well before the expected date. Or, if the correlation is negative, excessive provision may be put aside for failures that do not materialise.

We can extend this by generating numbers that represent a random walk as well as being serially correlated. In a random walk, each point is randomly selected from a normal distribution for which the mean is the value of the previous point. We can add a random walk to a general trend to create a drifting sequence of data values.

Two further functions: run.random.walk() and run.random.drift() generate such examples. The naïve lines almost always underestimate the envelope unless the correlation is large and negative. This is because the errors accumulate; the standard statistical procedures do not allow for this. Fig. 3 shows a typical result. The corrosion example is a random walk with non-linear drift.

So how should you model such data? In the first instance, it is important to consider the science. Is there, for example, some naturally occurring feedback mechanism? A moment's search on the Internet will reveal a vast amount of literature and software on the analysis of repeated measurements. I will not attempt this here: it can be a complex undertaking requiring skilled application of a statistics package.

These days we have no such excuse for not modelling serial correlation; tomorrow we shall have even more power and so expect to tackle even more complex problems.

FIG. 3: Random Walk: Correlation = 0.706 sd = 0.203