Measurement-Type “Calibration” of Expert Estimates Improves Their Accuracy and Their Usability: Pavement Engineering Case Study

Edgar Daniel Rodriguez Velasquez[1,2]
Carlos M. Chang Albitres[2], and Vladik Kreinovich[3]
[1]Civil Engineering, Univ. de Piura in Peru, edgar.rodriguez@udep.pe
Departments of [2]Civil Engineering and [3]Computer Science
University of Texas at El Paso
cchangalbitres2@utep.edu, vladik@utep.edu

1. Experts Are Often Used for Estimation

- Sometimes, experts are used because no measuring instruments can replace these experts.
- For example, in dermatology, estimates of a skilled expert are more accurate results than of any algorithm.
- This is one of the main reasons why,
  - in spite of numerous expert systems,
  - human doctors are still needed and still valued.
- In other cases, in principle, we can use automatic systems, but experts are still much cheaper to use.
- An example of such situation is pavement engineering.
- In principle, we can use an expensive automatic vision-based system to gauge the condition of the pavement.
- However, it is much cheaper – and faster – to use human raters.

2. Expert Estimates Are Often Very Imprecise

- Humans rarely have a skill of accurately evaluating the values of different quantities.
- For example, it is well known that humans drastically overestimate small probabilities.
- Correspondingly, underestimate the probabilities which are close to 1.
- Since most people's estimates are very inaccurate, it is difficult to find good expert estimators.
- It is well known that there is a high competition to get into medical schools.
- Even in pavement engineering, finding a good rater is difficult.

3. It Is Difficult to Find Good Experts: Example from Pavement Engineering

- According to a current standard, the condition of a pavement is evaluated by using a special index.
- This Pavement Condition Index (PCI) combines different possible pavement faults.
- To gauge the accuracy of a rater candidate,
  - many locations across the US
  - use criteria developed by the Metropolitan Transportation Commission (MTC) of California.
- A crucial part of the rater certification is a field survey exam.
- In this exam, a rater evaluates 24 test sites that have been previously evaluated by expert raters.

4. Pavement Engineering (cont-d)

- Candidate's PCI values are then compared with the PCI values of the expert rater.
- The expert's values are taken as the ground truth (GT).
- To certify, the rater must satisfy the following two criteria:
  - at least for 50% of the evaluated sites, the difference should not exceed 8 points, and
  - at least for 88% of the evaluated sites, the difference should not exceed 18 points.
- MTC provided a sample of 18 typical candidates.
- Out of these candidates, only 5 (28%) satisfy both criteria and thus, pass the exam and can be used as raters.

5. Problems

- What can we do to increase the number of available experts?
- And for those who have been selected as experts, can we improve the accuracy of their estimates?

6. Measuring Instruments Are Also Sometimes Not Very Accurate

- We are interested in situations when expert serve, in effect, as measuring instruments.
- Measuring instruments are usually much more accurate then human experts.
- Still, they are sometimes not very accurate.
- Even when they are originally reasonably accurate, in time, their accuracy decreases.
- When the measuring instrument becomes not very accurate, we do not necessarily throw it away.
- For example, before we step on the scales, they already show 10 pounds.
- We do not necessarily throw away these scales: instead, we adjust the starting point.

7. Calibration (cont-d)

- When a household device for measuring blood pressure starts producing weird results,
  - the manufacturers do not advise the customers to throw it away and to buy a new one,
  - they advise the customers to come to a doctor's office and to calibrate the customer's instrument.
- In general, calibration is a routine procedure for measuring instruments; we measure the same quantities:
  - by using our measuring instruments – resulting in the values $x_1, \ldots, x_n$, and
  - by using a much more accurate ("standard") measuring instrument – resulting in the values $s_1, \ldots, s_n$.
- In many cases – like in the above scales example – the main problem is the bias.

8. Calibration (cont-d)

- We compensate for the bias by subtracting the estimated value.
- The resulting corrected values $x_i + b$ are closer to the ground truth s_i.
- A reasonable way to estimate the bias is to use the Least Squares method: $\sum_{i=1}^{n} ((x_i + b) - s_i)^2 \to \min$.
- In some cases,
  - there is also a relative systematic error,
  - when each value is under- or over-estimated by a certain percentage.

9. Calibration (cont-d)

- To compensate for this under- and over-estimation, we need to multiply by an appropriate constant; e.g.:
  - if all the values are overestimated by 10%,
  - then each ground truth value s_i is replaced by the biased value $s_i + 0.1 \cdot s_i = 1.1 \cdot s_i$.
- To compensate for this relative bias, we thus need to multiply all the measurement results by $1/1.1$.
- In general, we need to replace the original measurement results x_i by corrected values $a \cdot x_i$ for some a.
- In general, to compensate for both absolute and relative biases, we replace x_i with $a \cdot x_i + b$.

10. Calibration (cont-d)

- The values a and b can be found by the Least Squares method: $\sum\limits_{i=1}^{n}((a\cdot x_i+b)-s_i)^2\to \min$.
- After that:
  - instead of using the original measurement result x produced by the measuring instrument,
  - we calibrate it into a more accurate value

$$x'=a\cdot x+b.$$

- In addition to such a linear calibration, it is sometimes beneficial to use non-linear calibration.
- Sometimes, a quadratic or cubic calibration is used – which leads to more accurate measurement results.
- In many practical situations, it is also beneficial to use fractional-linear re-scaling $x'=\dfrac{a\cdot x+b}{1+c\cdot x}$.

11. Our Idea: Let Us Calibrate Experts

- A natural idea is that since experts serve as measuring instruments, we can similarly calibrate the experts.
- Namely, instead of using the original expert estimates:
  - we first re-scale the original expert estimates in accordance with the appropriate calibration function,
  - and then we use these re-scaled values instead of the original expert estimates.
- As a result – just like for measuring instruments – we will hopefully get more accurate estimates.
- In some situations,
  - when for some experts, their original estimates were not very accurate,
  - we may end up with re-scaled estimates of acceptable quality, so we can use them.

12. Such Calibration is Indeed Helpful

- A good example of the efficiency of such calibration is expert's estimations of small probabilities.
- According to Kahnemann and Tversky, these estimates e_i are way off.
- However, the values $e'_i = a \cdot \sin^2(b \cdot e_i)$ are much more accurate.
- Namely, for $p_i < 20\%$, the worst-case difference $|p_i - e_i|$ is 8.6%.
- This is more than 40% of the original probability value.
- The worst-case difference $|p_i - e'_i|$ is 0.7%.
- This is 3.5% of the original probability value, and is, thus, an order of magnitude more accurate.

13. We Applied Our Idea to Pavement Engineering

- We started with the 18 rater candidates from the original MTC sample.
- In the original test, only five of these candidates passed the exam: rater candidates R6, R8, R9, R14, and R15.
- Originally, we compare this rater's ratings r_i with the 24 corresponding ground truth values s_i.
- Instead, we first found the values a and b that minimize the sum of the squares $\sum_{i=1}^{24}((a \cdot r_i + b) - s_i)^2$.
- Then used the re-scaled values $r'_i = a \cdot r_i + b$ to compare with the ground truth.

14. As a Result, More Experts Are Selected

- Based on the re-scaled ratings, four more candidates passed the test: candidates R1, R3, R5, and R11.
- This means that these four folks can now be used for rating pavement conditions; of course:
  - instead of using their original ratings r_i,
  - we first re-scale them to $r_i' = a \cdot r_i + b$ for this rater's a and b.
- As a result, we can accept 9 raters.
- Thus, the acceptance rate is now no longer $5/18 \approx 28\%$, it is $9/18 = 50\%$.

15. For Most Originally Selected Experts, Re-Scaling Leads to More Accurate Estimates

- After re-scaling, one of the originally accepted candidates – R9 – no longer fits.
- For this rater, we use his original ratings.
- For the remaining four originally selected raters, re-scaling improves the accuracy of their estimates:
  - for R6, the mean square rating error decreases from 11.21 points to 10.01 points – a decrease of 9.9%;
  - for R8, the mean square rating error decreases from 10.00 points to 8.66 points – a decrease of 6.4%;
  - for R14, the mean square rating error decreases from 8.62 to 6.95 points – a decrease of 19.4%; and
  - for R15, the mean square rating error decreases from 6.47 points to 6.21 points – a decrease of 4.0%.

16. First Auxiliary Result: Why 50%?

- In the MTC procedure,
  - as the first threshold,
  - we consider the accuracy with which we should have at least 50% of the measurements.
- In other words, we compare the median of the empirical distribution with some threshold.
- But why 50%? Why not select a value corresponding to, say, 40% or 60%?
- The only explanation that MTC provides is that selecting 50% leads to empirically the best results.
- But why? Here is our explanation.
- We want to find a parameter describing how distribution of expert's approximation errors.

17. Why 50% (cont-d)

- This may be the standard deviation, this may be some other appropriate parameter.
- We want the relative accuracy with which we determine this parameters to be as good as possible.
- We estimate this parameter based on a frequency f that corresponds to some probability p.
- It is known that, after n observations, $f - p$ is approximately normally distributed, with 0 mean and

$$\sigma[p] = \sqrt{\frac{p \cdot (1-p)}{n}}.$$

18. Why 50% (cont-d)

- We can measure the relative accuracy both:
  - with respect to the probability p of the original event and
  - with respect to the probability $1-p$ of the opposite event.
- We want both relative accuracies to be as small as possible.
- The relative accuracy with which we can find the desired probability p is equal to

$$\frac{\sigma[p]}{p}=\sqrt{\frac{1-p}{n\cdot p}}=\sqrt{\frac{1}{n}\cdot\left(\frac{1}{p}-1\right)}.$$

19. Why 50% (cont-d)

- Similarly, the relative accuracy with which we can find the probability $1-p$ is equal to

$$\frac{\sigma[p]}{1-p}=\sqrt{\frac{p}{n\cdot(1-p)}}=\sqrt{\frac{1}{n}\cdot\left(\frac{1}{1-p}-1\right)}.$$

- We need to make sure that the largest of these two values is as small as possible.

- One can check that the largest of these two values is

$$\sqrt{\frac{1}{n}\cdot\left(\max\left(\frac{1}{p},\frac{1}{1-p}\right)-1\right)}=$$

$$\sqrt{\frac{1}{n}\cdot\left(\frac{1}{\min(p,1-p)}-1\right)}.$$

- This expression is a decreasing function of $\min(p,1-p)$.

20. Why 50% (cont-d)

- Thus, for the relative standard deviation to be as small as possible, $\min(p, 1-p)$ must be as large as possible.
- This expression grows from 0 to 0.5 when p increases from 0 to 0.5, then decreases to 0.
- Thus, its maximum is attained when $p = 0.5$ – and this is exactly what MTC recommends.
- Thus, we have a theoretical explanation for this empirically successful recommendation.

21. Why 88%

- There are many different independent reasons why an expert estimate may differ from the actual value, so:
  - the expert uncertainty can be represented as
  - a sum of a large number of small independent random variables.
- It is known that, under reasonable condition, the distribution of such a sum is close to normal.
- This result is known as the Central Limit Theorem.
- Thus, we can safely assume that the distribution of expert uncertainty is normal.

22. Why 88% (cont-d)

- For a normal distribution with 0 mean,
  - if the probability for the value to be within ± 8 is 50%,
  - then the probability for the value to be within ± 18 is indeed close to 88%.
- This explains the second part of the MTC test.
- In both cases, our explanations seem to be simple and natural.
- We would not be surprised if it turns out that,
  - when selecting the corresponding numbers,
  - the authors of the MTC test were inspired not only by the empirical evidence,
  - but also by similar simple theoretical ideas.

23. Acknowledgments

This work was supported in part by the US National Science Foundation grant HRD-1242122 (Cyber-ShARE).