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

1st Edgar Daniel Rodriguez Velasquez
Department of Civil Engineering
Universidad de Piura in Peru (UDEP)
Piura, Peru
edgar.rodriguez@udep.pe and
Department of Civil Engineering
University of Texas at El Paso
El Paso, Texas, USA
edrodriguezvelasquez@miners.utep.edu

2nd Carlos M. Chang Albitres
Department of Civil Engineering
University of Texas at El Paso
El Paso, Texas, USA
cchangalbitres2@utep.edu

3rd Vladik Kreinovich
Department of Computer Science
University of Texas at El Paso
El Paso, Texas, USA
vladik@utep.edu

***Abstract*—In many applications areas, including pavement engineering, experts are used to estimate the values of the corresponding quantities. Expert estimates are often imprecise. As a result, it is difficult to find experts whose estimates will be sufficiently accurate, and for the selected experts, the accuracy is often barely within the desired accuracy. A similar situations sometimes happens with measuring instruments, but usually, if a measuring instrument stops being accurate, we do not dismiss it right away, we first try to re-calibrate it – and this re-calibration often makes it more accurate. We propose to do the same for experts – calibrate their estimates. On the example of pavement engineering, we show that this calibration enables us to select more qualified experts, and make estimates of the current experts more accurate.**

***Index Terms*—expert estimates, calibration, pavement engineering**

I. Introduction

Expert are often used for estimation. In many real-life problems, experts are used to estimate the values of different quantities.

Sometimes, experts are used because no measuring instruments has yet been invented to replace these experts.

For example, in medicine, while many measurements are possible, in some areas (e.g., in dermatology), an estimate of a skilled expert still leads to more accurate results than any known 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, where, in principle, we can use an expensive automatic vision-based system to gauge the condition of the pavement, but it is much cheaper – and faster – to use human raters.

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

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 – and, correspondingly, underestimate the probabilities which are close to 1; see, e.g., [2] and references therein.

As a result, it is difficult to find good experts. 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, but even in pavement engineering, finding a good rater is difficult.

It is difficult to find good experts: example from pavement engineering. According to a current standard [1], the condition of a pavement is evaluated by using a special Pavement Condition Index (PCI), a numerical characteristic that combines different possible pavement faults.

To gauge the accuracy of a rate candidate, many locations across the US use criteria developed by the Metropolitan Transportation Commission (MTC) of California [13]. A crucial part of the rater certification is a field survey exam, in which a rater evaluates 24 test sites that have been previously evaluated by expert raters.

Candidate's PCI values are then compared with the PCI values of the expert rater – which 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 between the rater's estimate and the ground truth should not exceed 8 points, and
- at least for 88% of the evaluated sites, the difference between the rater's estimate and the ground truth 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.

Problems.

- What can we do to increase the number of available experts?
- And for those who have been selected as experts – and whose accuracy is barely tolerable – can we improve the accuracy of their estimates?

II. Our Main Idea: Let Us Calibrate Experts the Same Way We Calibrate Measuring Instruments

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, but still, they are sometimes not very accurate – and even when they are originally reasonably accurate, in time, their accuracy decreases.

When a measuring instrument is not very accurate, we do not throw it away, we calibrate it. When the measuring instrument becomes not very accurate, we do not necessarily throw it away.

For example, when we try to use the scales to find our weight, and 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.

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 – instead, they advise the customers to come to a doctor's office and to calibrate the customer's instrument by using the doctor's more accurate instrument as the ground truth.

In general, calibration is a routine procedure for measuring instruments; see, e.g., [14]. In this procedure, 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 $s_1, \ldots, s_n$.

Then, we find the values a and b for which the re-scaled values $a \cdot x_i + b$ are the closest to the ground truth s_i.

This is usually done by using the Least Squares method, when we find the values a and b for which the sum of the squares of the differences attains the smallest possible value [14], [15]:

$$\sum_{i=1}^{n} ((a \cdot x_i + b) - s_i)^2.$$

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. For example, in many practical situations, it is beneficial to use fractional-linear re-scaling

$$x' = \frac{a \cdot x + b}{1 + c \cdot x};$$

see, e.g., [3]–[5], [10]–[12].

Our idea: let us calibrate experts. A natural idea is, instead of dismissing inaccurate potential experts, calibrate them – and even for current experts, we can calibrate them and thus, in principle, improve their accuracy.

Such calibration is indeed helpful. A good example of the efficiency of such calibration is expert's estimations of small probabilities. As we have mentioned earlier, these estimates e_i are way off, they are very different from the actual probabilities p_i [2]. However, it turns out that if we apply an appropriate non-linear transformation, and use the values $e'_i = a \cdot \sin^2(b \cdot e_i)$ instead of the original estimates e_i, we get much more accurate fit; see, e.g., [6]–[9]. Namely, for probability below 20%:

- the worst-case difference between the original estimates e_i and the actual probabilities was 8.6% – more than 40% of the original probability value – while
- the worst-case difference between the re-scaled estimates e'_i and the probabilities p_i is 0.7% – 3.5% of the original probability value, an order of magnitude more accurate.

III. Results of Applying Our Idea to Pavement Engineering: More Experts are Selected, and Their Estimates Are More Accurate

What we did. We started with the same 18 rater candidates. In the original test, only five of these candidates passed the exam: rater candidates R6, R8, R9, R14, and R15.

For each rater, instead of directly comparing this rater's ratings r_i with the 24 corresponding ground truth values g_i, we first found the values a and b that minimize the sum of the squares

$$\sum_{i=1}^{24} ((a \cdot r_i + b) - r_i)^2,$$

and then used the re-scaled values $r'_i = a \cdot r_i + b$ to compare with the ground truth.

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 – provided that instead of using their original ratings r_i, we first re-scale them to $r'_i = a \cdot r_i + b$, where the coefficients a and b have been determined for each of these raters.

As a result, we can accept 9 raters. Thus, the acceptance rate is now no longer $5/18 \approx 28\%$, it is $9/18 = 50\%$.

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, which means that for this rater, we cannot re-scale, we have to use his original ratings.

For the remaining four originally selected raters, re-scaling improves the accuracy of their estimates:

- for rater R6, the mean square rating error decreases from 11.21 points to 10.01 points – a decrease of 9.9%;
- for rater R8, the mean square rating error decreases from 10.00 points to 8.66 points – a decrease of 6.4%;
- for rater R14, the mean square rating error decreases from 8.62 points to 6.95 points – an impressive decrease of 19.4%; and
- for rater R15, the mean square rating error decreases from 6.47 points to 6.21 points – a decrease of 4.0%.

IV. AUXILIARY RESULTS: WHY 50%? WHY 88%?

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 (corresponding to 50%) of the empirical distribution with some threshold. But why 50%? Why not select a value corresponding to, say, 40% or 60% and compare this value with the appropriate threshold?

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. 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 to-be-determined probability p. It is known (see, e.g., [15]) that, after n observations, the difference $f-p$ between the observed frequency f and the actual (unknown) probability p is approximately normally distributed, with 0 means and standard deviation

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

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

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

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

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

To get the most accurate estimate of the desired parameters, 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 equal to

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

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

This expression is a decreasing function of $\min(p,1-p)$. Thus, for the relative standard deviation to be as small as possible, the expression $\min(p,1-p)$ 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 as p continues to grow. 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.

Why 88%. There are many different independent reasons why an expert estimate may differ from the actual value. As a result, the expert uncertainty can be represented as a sum of a large number of small independent random variables.

It is known – see, e.g., [15] – 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. 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.

REFERENCES

[1] ASTM International, *Standard Practice for Roads and Parking Lots Pavement Condition Index Surveys*, International Standard D6433-18, 2018.

[2] D. Kahneman, *Thinking, Fast and Slow*, Farrar, Straus, and Giroux, New York, 2011.

[3] V. G. Knorring, Y. Ya. Kreinovich, and V. D. Mazin, "Measurement Information: Scales and Conversions", *Measurement Techniques*, 2002, Vol. 45. No. 2, pp. 113–115 (Russian version February 2002, pp. 3-4).

[4] I. N. Krotkov, V. Kreinovich, and V. D. Mazin, "Methodology of designing measuring systems, using fractionally linear transformations," *Measuring Systems. Theory and Applications*, Novosibirsk Electrical Engineering Institute, Novosibirsk, Russia, 1986, pp. 5–14 (in Russian).

[5] I. N. Krotkov, V. Kreinovich, and V. D. Mazin, "General form of measurement transformations which admit the computational methods of metrological analysis of measuring-testing and measuring-computing systems," *Izmeritelnaya Tekhnika*, 1987, No. 10, pp. 8–10 (in Russian); English translation: *Measurement Techniques*, 1987, Vol. 30, No. 10, pp. 936–939.

[6] J. Lorkowski and V. Kreinovich, "Fuzzy Logic Ideas Can Help in Explaining Kahneman and Tversky's Empirical Decision Weights", *Proceedings of the 4th World Conference on Soft Computing*, Berkeley, California, May 25–27, 2014, pp. 285–289.

[7] J. Lorkowski and V. Kreinovich, "Granularity Helps Explain Seemingly Irrational Features of Human Decision Making", In: W. Pedrycz and S.-M. Chen (eds.), *Granular Computing and Decision-Making: Interactive and Iterative Approaches*, Springer Verlag, Cham, Switzerland, 2015, pp. 1–31.

[8] J. Lorkowski and V. Kreinovich, "Fuzzy Logic Ideas Can Help in Explaining Kahneman and Tversky's Empirical Decision Weights", In: L. Zadeh et al. (Eds.), *Recent Developments and New Direction in Soft-Computing Foundations and Applications*, Springer Verlag, 2016, pp. 89–98.

[9] J. Lorkowski and V. Kreinovich, *Bounded Rationality in Decision Making Under Uncertainty: Towards Optimal Granularity*, Springer Verlag, Cham, Switzerland, 2018.

[10] V. D. Mazin and V. Kreinovich, *An important property of fractional-linear transformation functions,* Leningrad Polytechnical Institute and National Research Institute for Scientific and Technical Information (VINITI), 1988, 13 pp. (in Russian).

[11] V. D. Mazin and V. Kreinovich, "A Universal Sensor Model", *Proceedings of the 12th International Conference Sensor'2005*, Nuremberg, Germany, May 10–12, 2005, pp. 317–322.

[12] H. T. Nguyen, V. Kreinovich, C. Baral, and V. D. Mazin, "Group-Theoretic Approach as a General Framework for Sensors, Neural Networks, Fuzzy Control, and Genetic Boolean Networks", *Proceedings of the 10th IMEKO TC7 International Symposium on Advances of Measurement Science*, St. Petersburg, Russia, June 30 – July 2, 2004, Vol. 1, pp. 65–70.

[13] Metropolitan Transportation Commission (MTC), *MTC Rater Certification Exam*, Streetsaver Academy, San Francisco, California, 2018.

[14] S. G. Rabinovich, *Measurement Errors and Uncertainty: Theory and Practice*, Springer Verlag, New York, 2005.

[15] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.