appearing in Experts and Consensus in Social Science - Critical Perspectives from Economics, Sociology, Politics, and Philosophy. Editors: Carlo Martini and Marcel Boumans, Series title: Ethical Economy - Studies in Economic Ethics and Philosophy, Springer

Validating Expert Judgment with the Classical Model

Roger M. Cooke
Resources for the Future,
Dept. Mathematics, TU Delft

April 30, 2013

Abstract: The classical model derives performance based weights for combining expert judgments, based on calibration or seed variables from the experts' field. Since publication of the TU Delft expert judgment database in 2005, various authors have attempted to use this data base for cross validation, splitting the seed variables into training sets and test sets. These attempts are reviewed. Many pitfalls and biases in cross validation efforts are identified and explained. A proposal for performing cross validation, based largely on the work of Eggstaff, Mazzuchi and Sarkani (2013) is formulated and illustrated with data from recent expert judgment studies.

1. Introduction

In 2008, Cooke and Goossens (2008) published the TU Delft data base comprised of 45 studies in which experts assessed calibration, or "seed" variables; variables for which true values are known post hoc. The Classical Model[1] (Cooke 1991) was used to derive Performance Weight (PW) and Equal Weight (EW) combinations (Decision Makers, DM). For an explanation of performance weighting and performance measures see (Cooke 1991, 2008, Flandoli 2010 or Eggstaff et al 2013). Suffice to say that calibration or statistical accuracy is measured as the p-value of the "null hypothesis" that an expert's probability statements are statistically accurate (we want not to reject the null hypothesis) and informativeness is measured as Shannon relative information with respect to a uniform or loguniform background measure. An expert's combined score is the product of his/her p-value and informativeness, and satisfies an asymptotic scoring rule constraint. This entails that an expert is weighted only if his/her p-value is above a threshold, which is chosen so as to optimize the combined score of the DM. In global weighting the informativeness score is averaged over all calibration variables, and the same weights are applied to all variables; with item specific weighting the informativeness for each item is used and the weights differ from item to item. All results reported here, except those in Cooke (2008a) use only global weights.

The TU Database is unique in providing expert assessments of variables in their fields whose true values are known post hoc. Researchers have used this data base to explore new models and to study whether performance on the calibration variables predicts performance on the variables of interest. In a few studies, variables of interest were later observed, enabling out-of-sample validation. In most cases the variables of interest are not observable on timescales relevant for the decision problem. Therefore, various forms of cross validation have been suggested. Clemen (2008) proposed a Remove-One-At-a-Time (ROAT) method according to which the calibration variables were removed one at a time and predicted by the model initialized on the remaining calibration variables. The predictions, though originating from different decision makers, were pooled and compared with the equal weight decision maker. On the 14 studies selected for this

[1] So called because of an analogy with classical hypothesis testing.

exercise, Clemen found that PW outperformed EW on 9, which was not statistically significant. Cooke (2008a) noted that this procedure is biased against PW since each calibration variable is predicted by a decision maker in which experts who assessed that particular item badly are up-weighted. It is commonly observed that removing one calibration variable can influence an individual expert's statistical likelihood by a factor 3 or more, a feature explained by the fact that statistical accuracy is a very fast function.

Variations on the ROAT approach have been performed by other researchers. Lin and Cheng (2008) examined 28 of the 45 studies and found PW significantly out performing EW, although PW's out-of-sample performance was degraded. Lin and Cheng (2009) used ROAT on 40 studies finding no significant difference between PW and EW[2]. Lin and Huang (2012) used ROAT with the Brier score in a regression based study of the effects of aggregation method, dependence, number of experts and seed variables and overconfidence on the Brier score (defined as 1 minus the quadratic scoring rule).

Other researchers have undertaken cross validation without ROAT. Cooke (2008a) looked at half-half splits in 13 studies with at least 14 calibration variables. Flandoli et al (2010) examined five datasets, choosing 30% of the number of calibration variables as the size of the test set, provided this number was at least 8, otherwise the test set was 8. They recoded the classical model in R, but did not implement item weights or the log uniform background measure. They randomly drew 500 partitions into training and test sets of the fixed sizes. The most extensive study of this kind is Eggstaff et al (2013), which initializes the global weights model on *all* non empty subsets of seed variables and in each case predicts the complementary subset, again using only global weights. Studies with large numbers of seed variables were split into separate studies to prevent combinatoric explosion. In total 62 expert judgment studies were analysed.

Studies differ in expert subject matter, in numbers and training of experts, in the methods of recruitment and methods of elicitation. For this reason, a numerical representation of out-of-sample validity at the study level would be desirable. For each study, Eggstaff et al (2013) average the combined scores of PW and EW for each number K of variables in the training set, for K = 1 to N – 1, where N is the number of seed variables. The same experts, the same calibration variables, and the same information background measures apply for all training set choices within one study. However the statistical power of the test set goes down as the training set size increases, there are many more studies for values of K near N/2, and these studies have overlapping training sets. With this in mind the PW and EW combined scores are averaged for each size K, for K = 1..N–1. To aggregate these up the study level we may either average the score differences (PW – EW) or take the geometric mean (geomean) of the ratios PW/EW.

Whereas the difference of scores inherits the scores' dimension (meters minus meters is meters), the ratio of scores is dimensionless (meters divided by meters is an absolute number). In aggregating ratios of positive numbers we must take the geometric mean, or geomean[3]. The ratio

[2] There large differences between the in-sample values in these two papers, and those found in the original studies.

[3] To see this suppose on two comparisons the scores were (PW=4, EW=1) and (PW=1, EW=4) The performance is identical, but the average of ratios is 1/2(4+1/4) =2.125. The Geomean is $(4\times1/4)^{1/2}=1$. Eggstaff et al report only the average scores for each size of the training sets, so we consider the ratios of averages. Since the average is always greater or equal to the geomean, the numerator and denominator in these comparisons would both be smaller if we

of PW and EW can be compared across training set sizes and across studies. The geomean of the ratios of combined scores of all comparisons per study are plotted in Figure 1. In 45 of the 62 studies (73%) the geomean of combined score ratios PW / EW was greater than unity. When PW's combined score exceeded that of EW, it tended to exceed by a greater amount than when EW's combined score exceeded that of PW. The best eyeball assessment is to compare the mass of lines above and below the baseline of 1. The geomean of the geomeans for each study was 2.46. Summarizing, PW outperforms EW in out of sample cross validation on more than two thirds of the studies, and the combined score of PW is more than twice that of EW.

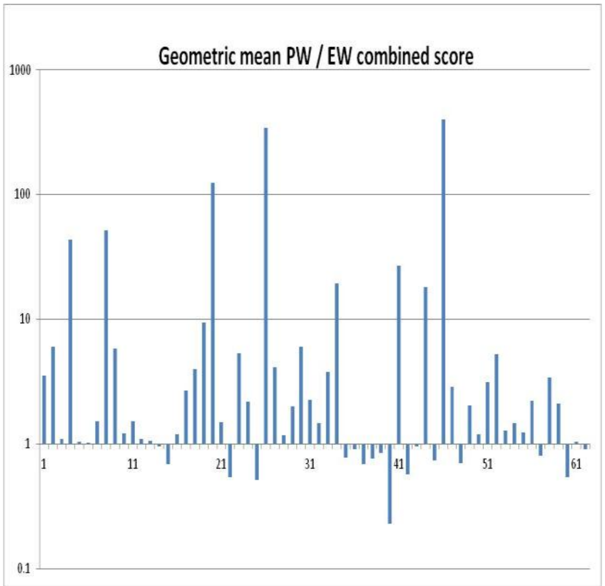


Figure 1: 62 studies, per study: geomeans of comparisons of PW/EW combined score ratios.

The accuracy of a DM in terms of proximity of the median to the true value is not directly related to the scoring variables of statistical accuracy and informativeness. Eggstaff et al (2013) report an accuracy advantage of PW over EW comparable to the differences in combined scores; however that feature is not pursued in this paper.

This paper addresses cross validation. First, the issue of scoring rules for individual variables is dealt with, followed by a demonstration of the bias in Remove-One-at-a-Time (ROAT) cross validation. Realistic expectations for cross validation are developed in section 3. Comparisons based on 5 or fewer seed variables would require a large number of independent and statistically identical studies to detect significant differences; 10 seed variables provide a more powerful basis for comparison. Section 4 analyses 13 studies reported since 2012, of which 4 involved more than 10 seed variables. Based on a suggestion of Eggstaff et al. (2013) initializing the model on all subsets of one or two seed variables, and evaluating on all seed variables attests to the superiority of Performance Weighting (PW) over Equal Weighting (EW), or when that is not attested, enables understanding in terms of the number of experts and their individual performance.

took the geomeans of combined scores of each separate K-tuple of training variables. It's impossible to say if there is an overall effect of this choice.

The cross validation exercises reported here were performed with the software system EXCALIBUR, which has been extensively tested in over 20 years of use. It is freely available at http://risk2.ewi.tudelft.nl/oursoftware/6-excalibur. For cross validation of the Asian Carp data, the MATLAB code used in Eggstaff et al (2013) was graciously provided by the authors.

2. Scoring rules for individual variables

Scoring rules were originally introduced as a tool for elicitation. An expert gives a mass or density function for an uncertain quantity which is later observed, and a scoring rule assigns a number to the assessment-plus-realization. Strictly proper scoring rules are such that an expert achieves his maximal expected score by and only by giving the assessment which correspond to his/her true belief. The classical model uses asymptotically strictly proper scoring rules based on sets of assessments and sets of realizations. Many authors have suggested using strictly proper scoring rules for individual variables, and summing the scores over a set of variables, an idea strongly discouraged in Cooke (1991). A simple example tees up the issue. Suppose an expert assess the probability of Heads with a coin of unknown composition as 1/2. On each toss with the coin, the score is the same for Heads and Tails. If these individual scores are added, then the sum score after 100 tosses is also independent of the actual sequence of outcomes; 50 Heads and 50 Tails gets the same score as 100 Heads. Table 1 compares the quadratic score (positively sensed, on [-1,1]) averaged over 1000 predictions of rain of two experts.

Table 1: Two experts assessing next day probability of rain on 1000 days

Probability of Rain next day:		%5	15%	25%	35%	45%	55%	65%	75%	85%	95%	Totals
expert 1	assessed	100	100	100	100	100	100	100	100	100	100	1000
	realized	5	15	25	35	45	55	65	75	85	95	500
expert 2	assessed	100	100	100	100	100	100	100	100	100	100	1000
	realized	0	0	0	0	0	100	100	100	100	100	500
Quadratic score expert 1 = 0.665; Quadratic score expert 2 = 0.835												

Both experts are equally informative in the sense that they both attribute 5 % probability to one hundred next days, etc. Expert 1 is statistically perfectly accurate, expert 2 is massively inaccurate, yet expert 2 scores higher than expert 1. The reason is that such rules decompose as the sum of a "calibration" and "resolution" terms (De Groot and Fienberg 1986). Resolution measures the expert's ability to separate the variables into statistically distinct subsets, regardless whether the distributions assigned to the subsets correspond to the expert's assessments. High resolution overwhelms bad statistical accuracy in the above example.

2. ROAT Bias

To understand the ROAT bias, suppose two experts state the probability of heads. Let $P_1(Heads) = 0.8$ and $P_2(Heads) = 0.2$ be the probability of heads for experts *1* and *2*. Suppose that the decision maker's probability is a weighted combination of the experts' probabilities, $P_{dm} = wP_1 + (1-w)P_2$, where the weight of each expert, given observed data, is proportional to the likelihood of each expert's distribution, given the data[4]. After observing *n* Heads and *m* Tails, the experts' likelihood ratio is

$$\frac{0.8^{n} \times 0.2^{m}}{0.2^{n} \times 0.8^{m}} = 0.8^{n-m} \times 0.2^{m-n} \qquad (1)$$

If $m = n$, then the weight ratio is 1, and $w = 1/2$. If $m = n - 1$ then the weight ratio is *4* and $w = 4/5$ Thus if we remove *one* Tail, re-initialize our model and predict the Tail which was removed, we find that the predicted probability of Heads is $P_{dm}(Heads) = (4/5) \times 0.8 + (1/5) \times 0.2. = 0.68$. Removing one Tail, strongly tilts the model toward expert 1, and our prediction probability for heads is *0.68*. At the same time we evaluate this model on the Tail which we removed, hence the likelihood for this model on this observation is *0.32*. The same holds, mutatis mutandis, when we remove a Head. Suppose we observe $n = m$; then the PW model would use $w = ½$. If we truly validated out of sample with $n = m$ fresh observations, the PW likelihood would be 0.5^{2n} , but the ROAT value would be 0.32^{2n}. ROAT punishes PW relative to EW by a factor $(0.32/0.5)^{2n}$. The classical model is more complex than this simple probabilistic model, but the same behavior can be observed in simple artificial examples[5].

ROAT sampling is NOT out-of sample validation. Each seed variable is removed one at a time, the model is re-initialized on the remaining seed variables and used to predict the removed variable. Each prediction is made by a DIFFERENT model. To appreciate how much these models may differ, Table 2 gives the 23 different performance weights for the eight experts that arise as the 23 seed variables in the Eudisp case are removed one at a time. Evidently, the differences between ROAT and true out-of-sample validation can be substantial. The weight for expert 4 varies from 1 to 0.4; that of expert 5 from 0 to 0.59. This is a consequence of the well know volatility of the calibration score, which is observed on every robustness analysis of every study. The recalculation of weights when item j is removed tends to give more weight to experts who assessed item j badly, and hence this volatility is put to work AGAINST the PW model.

Table 2: Weights for ROAT in Eudisp

Removed variable	Expert							
	1	2	3	4	5	6	7	8
None	0	0	0	0.7683	0.2317	0	0	0
1	0	0	0	0.8086	0.1914	0	0	0
2	0	0	0	0.7928	0.2072	0	0	0
3	0	0	0	0.8071	0.1929	0	0	0
4	0	0	0	0.7303	0	0.2697		
5	0	0	0	0.4094	0.5906	0	0	0
6	0	0	0	0.7022	0.2978	0	0	0

[4] Such likelihood weights are not proper scoring rules, and do not account for informativeness, nonetheless there is a strong analogy with the classical model, as the driving term in that model is the likelihood of the hypothesis that an expert is well-calibrated.

[5] Take expert 1 (2) with 5, 50 and 95 percentiles equal to 0, 4, 8, (2, 6, 10). Take 3 realizations = 0.5, 3 realizations = 9.5, 2 realizations = 4 and 2 realizations=6. PW and EW coincide and hence should be identical on out of sample validation; however, the score of PW under ROAT is a factor 105 lower than that of EW.

7	0	0	0	0.6777	0.3223	0	0	0
8	0	0	0	0 .7928	0.2072	0	0	0
9	0	0	0	0.806	0.194	0	0	0
10	0	0	0	0.8645	0.1355	0	0	0
11	0	0	0	0.7003	0.1638	0	0	0.1359
12	0	0	0	0.7042	0.1632	0	0	0.1325
13	0	0	0	0.7659	0.2341	0	0	0
14	0	0	0	0.6996	0.1654	0	0	0.135
15	0	0	0	0.6287	0.1637	0	0.07593	0.1317
16	0	0	0	0.704	0.296	0	0	0
17	0	0	0	0.6996	0.1655	0	0	0.1349
18	0	0	0	0.6286	0.1638	0	0.07588	0.1317
19	0	0	0	0.704	0.296	0	0	0
20	0	0	0	0.6499	0.1537	0	0.07101	0.1254
21	0	0	0	0.5016	0.1307	0	0	0.3677
22	0	0	0	1	0	0	0	0
23	0	0	0	0.4094	0.5906	0	0	0

3. Cross validation without ROAT: what to expect

Absent observation of variables of interest, one option for some form of cross validation splits the calibration variables into a training set used to initialize the model and a test set used to assess performance. Cooke (2008) applied this method to 13 studies having at least 16 seed variables. In 20 of the 26 cases the PW outperformed Equal Weights (EW). The probability of seeing 20 or more "successes" on 26 trials (77%), if the probability of success were 0.5, is 0.001247. In this exercise both global and item weights were used, according to which performed best on the training set. Cross validation with item weights is possible with EXCALIBUR, but it is extremely time consuming. A large exercise enabling the choice between global and item weights would require recoding the model.

Intuitively, if we select experts who are 'statistically more accurate than average' in-sample, it is implausible that they should consistently be statistically less accurate out of sample. At worst they might be no better than randomly chosen experts out of sample (reversion to the mean). Further it is plausible that averaging a large set of experts will be less informative than averaging a small subset. How many similar studies we need to detect these effects depends on the size of the effects, the number of seed variables in each study and the number of studies. This section performs some indicative calculations.

We consider p = (0.05, 0.45, 0.45, 0.05) as the interquantile interval probabilities, and let s(N) = (s_1(N),s_2(N),s_3(N),s_4(N)) be the sample distribution based on N independent samples from p. The likelihood ratio test statistic

$$2N\ I(s(N)\mid p)\ =\ 2N\ \Sigma_{i=1..4}\ s(N)\ \ln(s_i(N)/p_i) \qquad (2)$$

is asymptotically chi square distributed with 3 degrees of freedom if s consists of independent samples from p. If F_3 is the cdf of the chi square distribution with 3 df, then $1-F_3(2N\ I(s(N) \mid p))$ is the p-value of s(N), and it is uniformly distributed on the interval [0, 1]. For small N the distribution is not uniform. Figure 1 shows the p-value distributions for s(5), s(10), and s(20) based on independent samples from p. Note that s(5) is concentrated in the middle of the [0, 1] interval, its 17th percentile is 0.394 and not 0.17 (the value for a uniform variable).

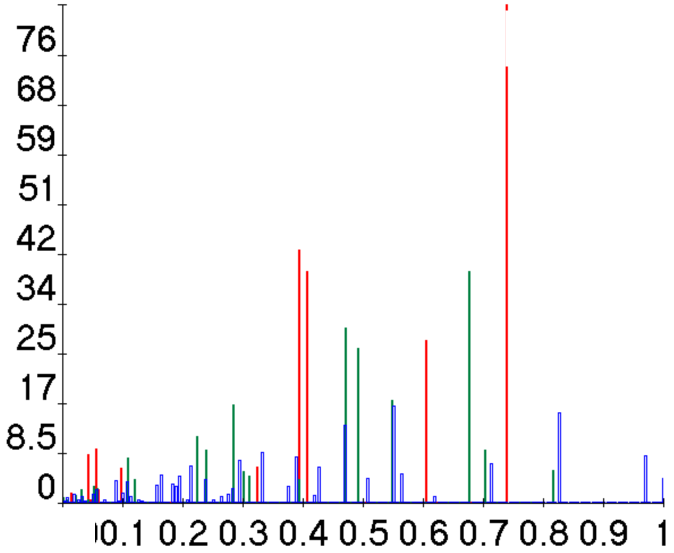


Figure 2: P-values for s(5) (red), s(10) (green) and s(20) (blue) sampled independently from p.

If the distribution s is not sampled from p then the mass functions of the three cases in Figure 1 shift toward zero, however the shift is much slower for smaller N. Suppose the samples were actually sampled from the distribution p** = (0.3, 0.2, 0.2, 0.3). This would be the sampling distribution of an expert who has only a 40% chance of catching the realizations in his 90% central confidence band, corresponding to severe overconfidence. Figure 2 shows the mass function of p-values for p**(5), p**(10) and p**(20).

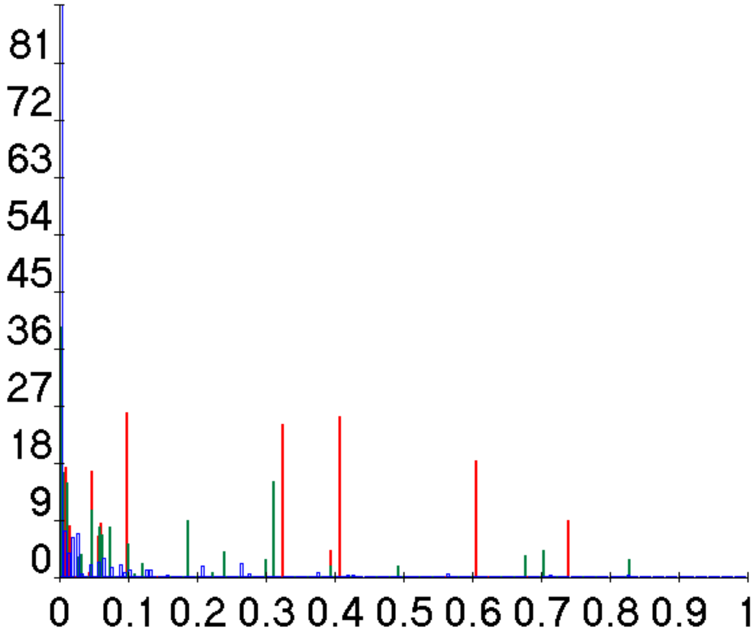


Figure 3: P-values for p**(5) (red), p**(10) (green) and p**(20) (blue) sampled independently from p.

The p-value of p**(5) has a 36% chance of falling below the standard rejection threshold 0.05; for p**(10) that chance is 54%. If two experts generate sample distributions from p and p**, then on 5 seed variables there is a 24% chance that the p-value of p** will be *greater* than that of p; on 10 seed variables that chance drops to 14%.

Suppose we have three types of assessors, the first type's interquantile hits are sampled independently from p. The second type's hits are sampled independently from p**; the third are sampled independently from p* = (0.15, 0.3. 0.3, 0.15). Type p* shows overconfidence, though not as severe as p**. The first type is perfectly calibrated, and his p-value is asymptotically uniform as the number of seed variables goes to infinity. The second type (p**) has p-values distributed as in Figure 2, for 5, 10 and 20 seed variables. When an assessor of each type states a 5 percentile, for example, the probability that the realization falls beneath that 5 percentile is 30% for p** and 15% for p*. The mean and standard deviations of the p-values on 5 seed variables of these three types are shown in Table 3, as computed by simulation.

Table 3: Mean and standard deviation of p-values on 5 seed variables for three types of DM

5 seed vbls	mean	σ
P	0.50	0.23
p*	0.41	0.25
p**	0.23	0.23

Figure 3 plots the mean of p-values for p, p* and p** as a function of the number of seed variables.

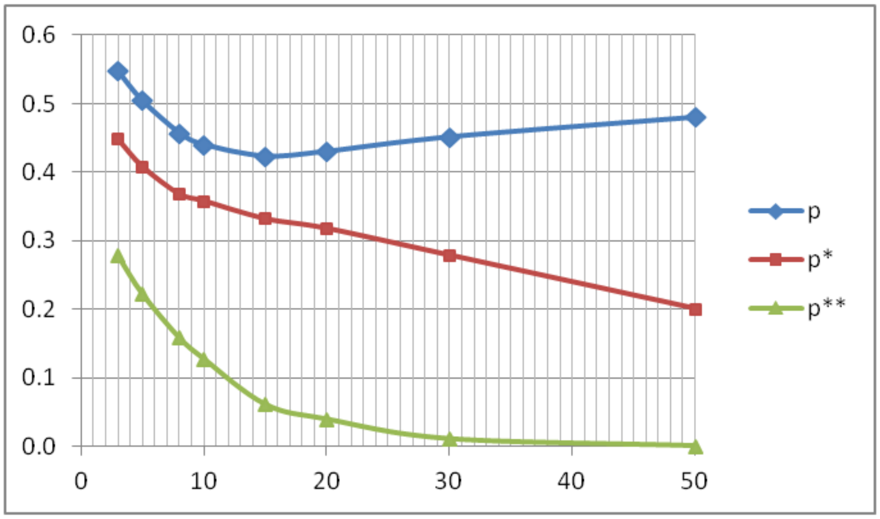


Figure 3: Mean p-values of p, p* and p** as function of number of seed variables.

Note that the mean for p starts above the asymptotic value of 0.5, then dips down to 0.42 after 15 seed variables before climbing back to 0.5. The expected value of the severely overconfident assessor doesn't drop below 0.05 until 17 seed variables. On 5 seed variables the distribution with severe overconfidence and a location bias, p*** = (0.5,0.3,0.1, 0.1), will have and expected p-value of 0.06.

We may think of p and p* as representing high scoring assessors, while p** and p*** are representative of low scoring assessors. The following thumb rules apply: assuming that the information scores are equal, we cannot statistically distinguish between high scoring experts (p and p*) on 50 seed variables (the highest number on any study was 55). High scoring assessors can be statistically distinguished from severe overconfidence (p**) on 20 seed variables[6], while severe overconfidence plus location bias can be statistically distinguished from high scoring assessors on 10 variables.

We now consider T independent studies with 5 seed variables, where on ALL studies the DM is p, p* or p**. How many studies would we need to decide which type of DM we have? An approximate answer is derived by noting that the mean of averaging p-values over T independent studies is the mean value in Table 3, and the standard deviation is approximated by $\sigma/\sqrt{T}$. Figure 4 shows the 5 percentile of p's sample average (dotted), the mean, 5 and 95 percentiles of p*'s sample average (dashed) and the mean and 95 percentile of p**'s sample average (solid); each is a function of the number of studies T. The 5 and 95 percentiles of p and p** cross at T = 8. That means that if we average p-values over 8 studies, each based on 5 seed variables, we can be 95% certain that the average of p**'s p-values will be in the critical region for p; and conversely if we average p's p-values, we can be 95% certain of being in the critical region for p**. Put a bit loosely, if two DM's had interquantile probabilities corresponding to calibration scores 0.5 and

[6] This is a thumb rule, the 5% lower confidence bound for p is 0.05, for p* this bound depends on the number of seed variables, since the expected p-value of p* goes to zero as the number of seed variables goes to infinity. For 5, 10 and 20 seed variables the 5% lower bounds are 0.02, 0.01 and 0.005 respectively.

0.23 on 5 seed variables, we could distinguish them statistically on 8 identical and independent studies.

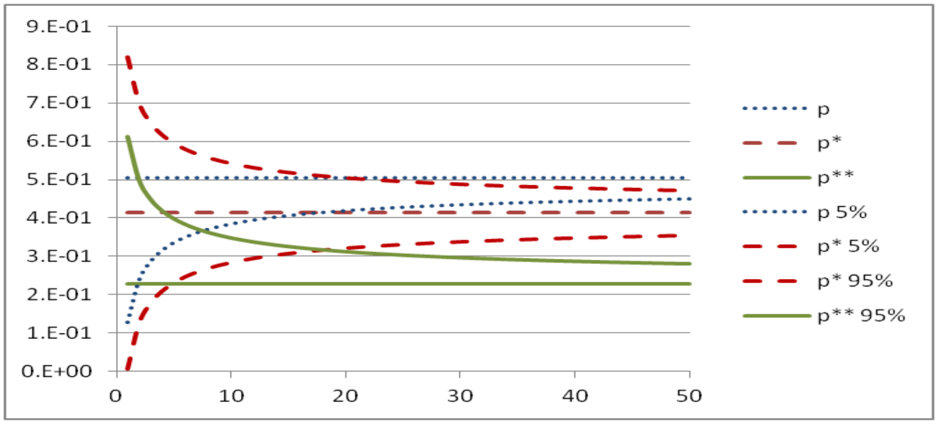


Figure 4: Number of studies with 5 seed variables for distinguishing DMs

Note that slopes of the percentiles go to zero, making the point of intersection unstable to small perturbations. Table 4 gives the number of studies need to distinguish these DMs

Table 4: Number of studies to distinguish DMs with 5 seed variables

Number of studies to distinguish pairs of DMs		
p vs p*	perfect vs overconf (p vs p*)	75
p vs p**	perf. vs severe overconf (p vs p**)	8
p* vs p**	overconf vs severe overconf (p* vs p**)	19

Whereas perfect calibration can be distinguished from severe overconfidence based on 8 studies with 5 seeds, distinguishing perfect calibration (p) from mere overconfidence (p*) on 5 seeds requires 75 studies.

Table 5 combines the information of Tables 3 and 4, but for studies involving 10 seed variables. Overconfidence and severe overconfidence can now be distinguished on 11 studies, instead of 19. Curiously, the mean of the p-values for p has dropped (one can see this in Figure 3) with 10 seeds the distribution is still far from uniform. As a result, distinguishing perfect calibration (p) and overconfidence (p*) actually requires more studies (90 instead of 75).

Table 5: p-values of DMs based on 10 seeds, and number of studies to distinguish DMs

10 seed vbls	mean	Std	Number of studies to distinguish pairs of DMs		
P	0.44	0.21	p vs p*	perfect vs overconf	90
p*	0.36	0.25	p vs p**	perfect vs severe overconf	5
p**	0.13	0.20	p* vs p**	overconf vs severe overconf	11

These results circumscribe what we can realistically expect from cross validation with appropriate caveats:

- The DM's of individual studies are not all the same,

- Calibration dominates the weighing, but information is also important
- The number and value of the experts in a study is also important. If there are several high scoring experts in a study, then fluctuating scores of experts will tend to cancel, keeping PW's performance high, but with one or no high scorers, extra instability can be expected, causing PW to perform poorly.

Most studies use 10 seed variables; if we split the seed variables into training at test sets of 5, then we may need in the order of 20 studies to distinguish the type of differences between p* and p**, but we need many more to distinguish p and p*, which is the range within which PW and EW calibration scores typically lie.

In contemplating cross validation, the first question is, what questions do we want to answer? On could formulate the following:

- Do the DM's calibration and information scores in-sample predict those out-of-sample?
- Is PW better than EW out of sample?

When a cross validation study initializes the performance based DM on K of the N calibration variables, the following issues arise: (1) If K is close to N, then the number of out-of-sample predictions, N-K, is small, statistically unpowerful, and predictions are subject to the ROAT bias. (2) If K is small, then the power of the calibration score is lowered, thereby reducing the ability to distinguish high and low statistical likelihood. (3) A straddling bias may arise when training and testing sets are complementary halves of the seed variables. The intuition behind this is as follows: two independent random numbers X and Y become negatively correlated if we conditionalize on their sum. When the sum is fixed, one variable can get larger only at the expense of the smaller.

Straddling bias

The straddling effect can be observed theoretically as follows: repeatedly draw independently 10 realizations from the distribution p*, divide them into disjoint sets of five, and denote by V_1, V_2 the distribution of p-values in the first and second sets. Since V_1, V_2 are independent, their correlation ρ_{12} is zero. The partial correlation given the p-value of the whole set S is defined as

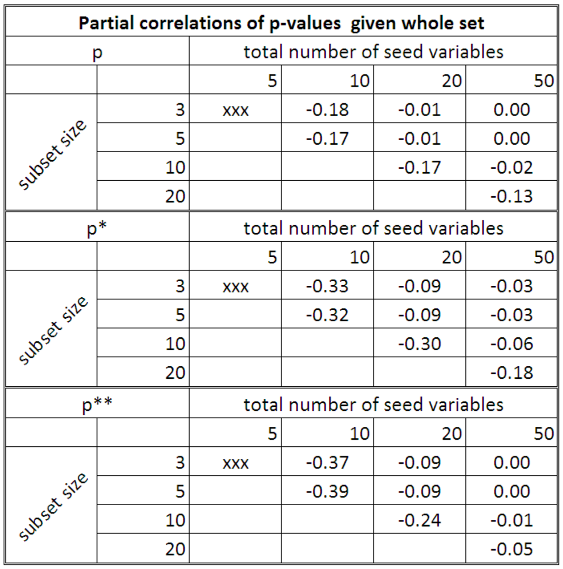
$$\rho_{12\,|\,S} = \frac{\rho_{12} - \rho_{1s}\rho_{2s}}{(\,(1-\rho_{1s}^{2})\,(1-\rho_{2s}^{2}))^{1/2}} = \frac{-\rho_{1s}^{2}}{1-\rho_{1s}^{2}}\,.$$

since $\rho_{1s\,=\,}\rho_{2s}$; $\rho_{12} = 0$. If S represents the “reduced power p-value” of all 10 variables[7], and if this negative partial correlation is strong, then the p values of the first and second sets of 5 will tend to straddle the p-value of the whole set. This combined with the fact that the performance weight scores have higher variance than equal weight scores will tend to cause performance

[7] This would hold after the power p-values on samples of size 10 has been reduced to the power of sample size 5. We find that for assessor p (perfect calibration) V_1 and V_2 straddle the reduce power 10 sample p-value with probability 0.24. For assessors p* and p** straddling occurs with probability 0.32 and 0.54 respectively. Note that for p all p-values have approximately the same expectation (namely ½). For p* and p* the expected p-value is decreasing in sample size, which is why we have to equalize power to see the straddle effect.

weights to outperform on only one of the two subsets, giving an overall 50% chance that PW exceeds EW, assuming the informativeness scores are roughly equal. Note that if the size of disjoint sets decreases, then ρ_{1s} decreases as well; if the size is greater than one half, then ρ_{12} will be positive. In either case the partial correlation will move towards zero. The effect of a straddling bias in real datasets has not yet been studied, but it is a potential problem.

Table 6: Partial correlations



Partial correlations of p-values given whole set					
p		total number of seed variables			
		5	10	20	50
subset size	3	xxx	-0.18	-0.01	0.00
	5		-0.17	-0.01	0.00
	10			-0.17	-0.02
	20				-0.13
p*		total number of seed variables			
		5	10	20	50
subset size	3	xxx	-0.33	-0.09	-0.03
	5		-0.32	-0.09	-0.03
	10			-0.30	-0.06
	20				-0.18
p**		total number of seed variables			
		5	10	20	50
subset size	3	xxx	-0.37	-0.09	0.00
	5		-0.39	-0.09	0.00
	10			-0.24	-0.01
	20				-0.05

Table 6 shows the partial correlations for p, p* and p** for seed variable sets of size 3, 5, 10, 20 and 50. Notice that the partial correlations are rather weak for the perfectly calibrated assessor whose interquantile probabilities are p. However, for p* and p** the effect is sizeable, even for disjoint subsets of size 3 in a set of 10 seed variables.

Using the data of Eggstaff et al (2013), Table 7 breaks the out-of-sample geomean of mean-score ratios PW/EW and arithmean of mean-score differences PW-EW into the number of variables in the training set and in the test set. For each training set size K (dot-shaded), we collect all comparisons with K in the training set, and consider the geomean of their out-of-sample score ratios and the arithmean of their out-of-sample score differences. This procedure aggregates over test sets of different sizes since the 62 studies differ in total number of seed variables. Similarly we aggregate over all test sets of size K (plane-shaded), thus aggregating over different sizes of training sets. There are 62 studies in total, and 62 with training sets of size one, and also 62 with test sets of size 1. The 35 studies with at lest 11 seed variables have a training set of size 10, and these same 35 studies also have a test set of size 10. Displaying the data in this way, we see that the geomean "likes" small training sets and large test sets. The arithmean "likes" larger training sets up to size 9. Large test sets have greater statistical power, which tends to drive down both the PW and EW scores, and also drives down their difference. This explains arithmean's decreasing behaviour in test size.

Table 7: Geomean and Arithmean as function of training size and test size, using data of Eggstaff et al (2012). The number of studies with training set size 16 and greater are too small to draw conclusions.

#train	geomean	# test	geomean	# studies	#train	arithmean	# test	arithmean
1	3.14	1	1.20	62	1	0.031	1	0.216
2	5.96	2	1.35	62	2	0.059	2	0.211
3	4.66	3	1.59	62	3	0.087	3	0.180
4	3.52	4	1.88	62	4	0.114	4	0.139
5	2.83	5	2.20	61	5	0.133	5	0.121
6	2.24	6	2.66	59	6	0.142	6	0.101
7	1.88	7	3.06	59	7	0.174	7	0.067
8	1.60	8	3.84	55	8	0.220	8	0.043
9	1.41	9	2.84	53	9	0.197	9	0.022
10	1.23	10	3.02	35	10	0.031	10	0.017
11	1.13	11	3.32	32	11	0.017	11	0.014
12	1.06	12	4.45	25	12	0.005	12	0.018
13	1.08	13	4.03	20	13	0.029	13	0.011
14	0.96	14	1.65	15	14	0.002	14	0.044
15	1.05	15	2.25	10	15	0.033	15	0.042
16	1.09	16	1.79	7	16	0.040	16	0.030
17	0.98	17	1.22	6	17	0.001	17	0.016
18	0.81	18	1.28	2	18	-0.102	18	-0.001
19	0.96	19	0.95	1	19	-0.015	19	-0.013
20	0.95	20	0.93	1	20	-0.015	20	-0.019
21	1.04	21	0.93	1	21	0.012	21	-0.018

Figure 5 compares the results of aggregating up to the study level by taking the geomean of the mean-score ratios (left panel) and the arithmetic mean of the mean-score differences (right panel), where “mean-scores” refers to combined scores averaged over training sets of the same size, per study. The left panel of Figure 5 was already presented in Figure 1. Since the studies are indexed from small to large numbers of seed variables, we readily note that a larger number of seed variables lowers the PW and EW scores and also the score differences. A similar effect was noted in Table 7. Figure 5 highlights the differences between geometric versus arithmetic aggregation, but the superiority of PW over EW is evident from either perspective.

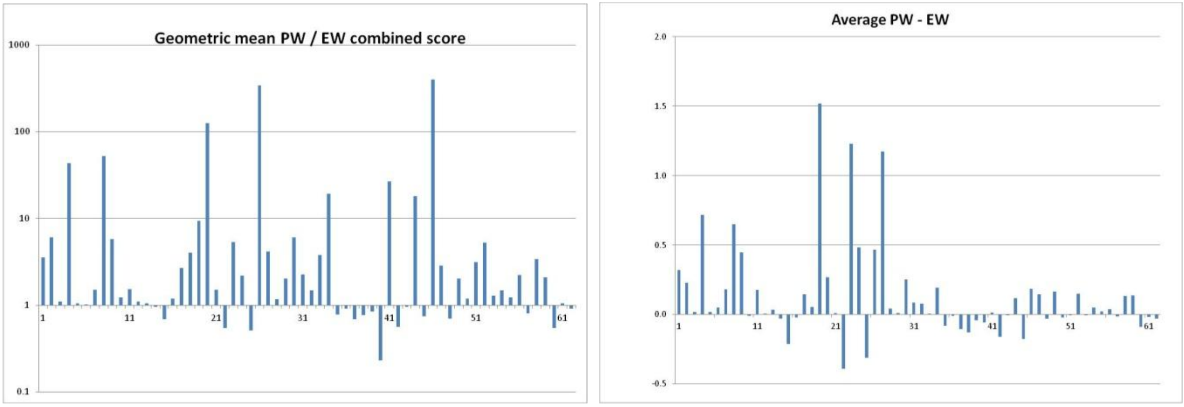


Figure 5: Geomean (left) of PW and EW score ratios and arithmetic mean (right) of PW and EW score differences, for each of 62 studies analysed in Eggstaff et al (2013). If a study had N seed variables, the PW and EW scores were

averaged over training sets of size K, K = 1 … N-1 and aggregated with either geo- or arithmetic means to determine an out-of-sample performance indicator per study .

The "small training set" cross validation has the advantages of avoiding the ROAT and the straddle bias. Eggstaff et al (2013) noted that with small training set the actual scores did not predict the scores on the larger set of calibration variables, but the superiority of performance weighting against equal weighting was attested. If this finding is corroborated, then a smaller number of calibration variables would be defensible, if a smaller number achieved adequate coverage of the problem domain. Based on the results of section 3, the current recommendation is that the test set should comprise at least 8, preferably 9 variables.

4. Cross Validation of Recent Data

The four studies discussed here are recent and have not yet been published. Figures 6 through 9 show the results of initializing the PW model on all subsets of size K=1 and K=2, and using these to predict all calibration variables. Using all seeds instead of the out of sample seeds is done to enable uniform comparison with all-sample results. The difference between all-sample and out of sample is modest for small K values. The curve is the contour of calibration × information that corresponds to the all-sample EW. The all-sample PW and EW are indicated by solid and outlined stars respectively. Geomeans are given in the caption to each figure for the ratio of the PW combined score based on one and two seed variable initializations and all-sample EW. The geomean of all geomeans in the cases analysed here is 2.57

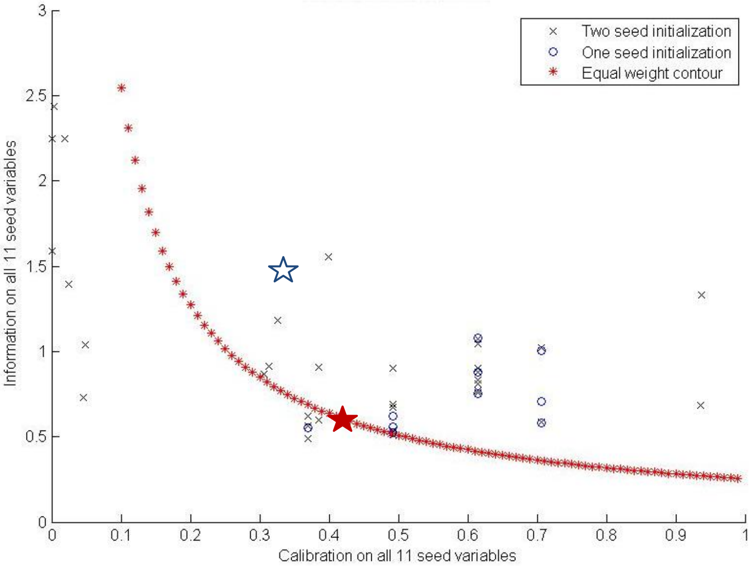


Figure 6 Ice sheet, 11 seed variables: geomean PW/EW one seed = 1.48, two seeds = 0.59. PW > EW on 9 of 11 one seed initializations, and on 34 of 55 two seed initializations, overall on 43/66 = 65% of these initializations.

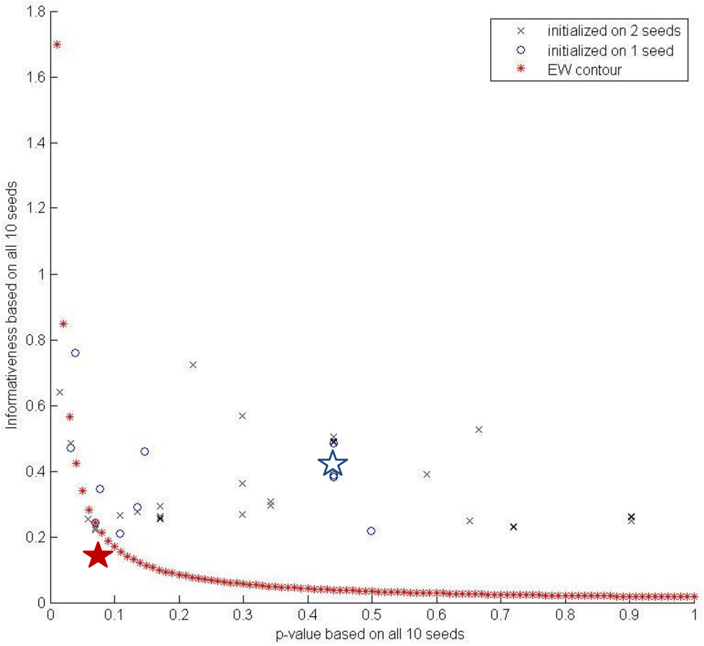


Figure 7: Obesity, 10 seed variables: geomean PW/EW one seed = 3.46, two seeds = 6.19. PW > EW on 9 of 10 one seed initializations, and on 39 of 45 two seed initializations, overall on 48/55 = 87% of these initializations.

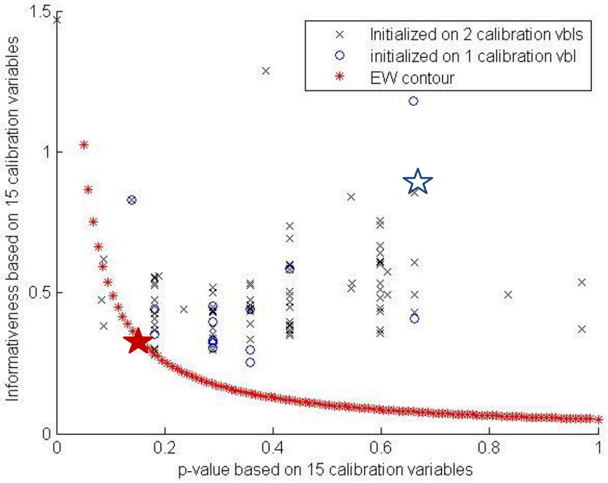


Figure 8: Asian carp, 15 seed variables: : geomean PW/EW one seed = 2.64, two seeds = 3.22. PW > EW on 15 of 15 one seed initializations, and on 101 of 105 two seed initializations, overall on 116 / 126 = 92% of these initializations.

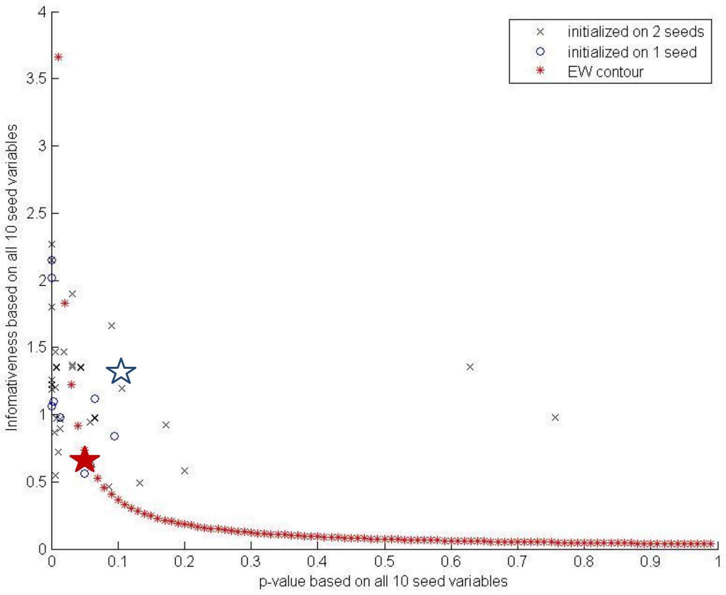


Figure 9: Fistula, 10 seed variables; geomean PW/EW for one seed = 34.8, for two seeds 0.34. PW > EW on 2 of 10 one seed initializations, and on 19 of 45 two seed initializations, overall on 21 / 55 = 38% of these initializations.

In the Fistula case there is no discernible pattern for PW to outperform EW out of sample, for the others there is. In addition to numbers of experts and seed variables, out of sample performance depends on the experts themselves. In the Fistula case, there were 8 experts, all of whom scored poorly. However, on one or two seed variables, one expert may have achieved a high score and gathered dominant weight, only to degrade on the entire set of seed variables.

The Asian carp study with 15 seed variables allows us to illustrate aggregation while keeping at least 9 variables in the training set. Table 9 shows results for each size of the training set from 1 to 6. “Arithmean” means that the experts’ combined score results were averaged over all tests with the same size of training set; similarly “geomean” indicates that the geometric mean was taken. Column 9 considers the ratio of arithmeans, column 10 takes the ratio of geomeans, and column 11 considers the difference of arithmeans.

Table 9: Asian Carp cross validation

1. Nr in Training set	2. Nr of training sets	3. PW median	4. PW Arithmean	5. PW geomean	6. EW median	7. EW Arithmean	8. EW geomean	9. Arith(PW) / Arith(EW)	10. Geo(PW) / Geo(EW)	11. Arith(PW) - Arith(EW)
1	15	0.1134	0.1260	0.1173	0.0688	0.0582	0.0554	2.1648	2.1158	0.0678
2	105	0.1470	0.1953	0.1491	0.0561	0.0670	0.0613	2.9170	2.4311	0.1284
3	455	0.1679	0.2019	0.1636	0.0755	0.0764	0.0678	2.6438	2.4136	0.1255
4	1365	0.2001	0.2406	0.1695	0.0957	0.0864	0.0747	2.7852	2.2676	0.1542
5	3003	0.2285	0.2626	0.1962	0.0864	0.0971	0.0822	2.7059	2.3858	0.1656
6	5005	0.2382	0.2808	0.1970	0.1114	0.1084	0.0902	2.5918	2.1834	0.1725

Note that all of the PW and EW scores increase in the training set size, reflecting the diminishing power of the test set. Note that the difference between the PW arithmean and geomean is greater than this difference with EW. The geomean is always less than or equal to the arithmean, and the difference becomes greater as the (positive) numbers are more variable. Indeed, if one of the aggregated non-negative numbers is zero the geomean is zero, no matter how large the other numbers are. This tendency of the geomean to be driven by the smallest of highly variable non-negative numbers cautions against uncritical use of the geomean. The geomean of a normal variable with mean 1 and standard deviation 0.8, truncated at 0.001 is 0.42, while its mean is 1.04. With this in mind, the ratio of geomeans (column 10) would punish PW for its greater variability. The difference of arithmeans (column 11) is affected by the decrease in statistical power. The ratio of arithmeans (column 9) avoids both these issues and is the current favorite. The geomean of column 9 is 2.62, which is the geomean of column 4 divided by the geomean of column 7. Figure 8 compares the PW and EW scores ratios and differences, for each of the 9949 training sets of size 1..6.

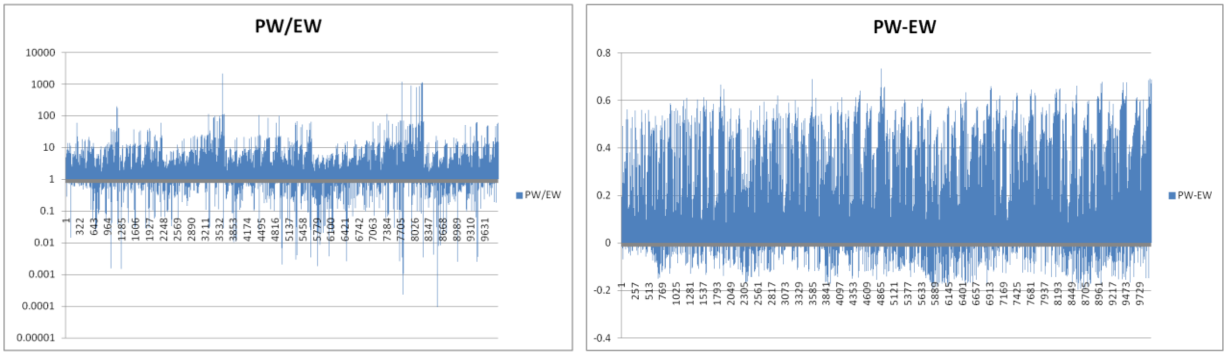


Figure 10 Ratios and differences of PW, EW scores for Asian Carps, based on 1 to six training variables out of 15 calibration variables.

Conclusion

Cross validation is more complex that appears at first sight. It is easy to see that single variable scores like the Brier or quadratic score are inappropriate, although they are still misused for this purpose. Less evident but still readily demonstrable is the bias in the ROAT method. Cross validation without ROAT is more challenging. Considering the straddle bias and the power loss relative to the expected size of the PW - EW differences, and the number of studies, the training set should be less than half of the total number of seed variables, and should leave at least 8, preferably 9, seed variables in the test set.

To derive a single out-of-sample validation number for each study, the following procedure reflects the best current insight:

I. Average the PW and EW combined scores (calibration × information) scores over training sets of size K which leave at least 9 variables in the test set. Call these averages PW(K), EW(K).
II. For each K, compute the ratio PW(K) / EW(K) and the difference PW(K) – EW(K).
III. Take the geomean over K of these ratios: $\prod_{K=1..K^*} [PW(K)/EW(K)]^{1/K^*}$ (preferred) and the arithmean of the differences: $(1/K^*) \sum_{K=1\ldots K^*} PW(K) - EW(K)$ (for comparison); where $N - K^* = 9$, and N is the number of seed variables. If N = 10, use $K^* = 2$.

The geomean is preferred in (III) as the ratio PW(K)/EW(K) is dimensionless, and PW(K) – EW(K) is affected by the statistical power of the test set. Throughout all these comparisons, the experts, and seed variables are the same, and the information scores can be meaningfully compared. Taking the geomean over all comparisons with the same training set size is not recommend, as (a) the number of comparisons can be very large, (b) the geomean is strongly influenced by the minimum value, and (c) the PW has greater variability than EW.

Future work could be profitably directed to performing cross validation with item weights and performing cross validation on all studies in the TU Delft base including recent studies. Of course, the studies are designed to pick up the large differences in expert performance, and are fit for that purpose. However we parse the Eggstaff data, PW seems to outperform EW convincingly. Since these are out of sample results, it is not surprising that the difference in performance is less than in sample. Further, these studies were not designed to optimally enable cross validation. As we better understand how cross validation should be done, we may modify the study designs.

References

Clemen, R.T (2008)" Comment on Cooke's classical method" Reliability Engineering & System Safety, Volume 93, Issue 5, May 2008, Pages 760-765

Cooke, R.M. (2008) Special issue on expert judgment, Editor's Introduction Reliability Engineering & System Safety, 93, Available online 12 March 2007, Volume 93, Issue 5, May 2008, Pages 655-656.

Cooke, R.M., (2008a) Response to Comments, Special issue on expert judgment Reliability Engineering & System Safety, 93, 775-777, Available online 12 March 2007. Volume 93, Issue 5, May 2008.

Cooke, R.M. (2011) "Pitfalls of ROAT Cross Validation Comment on Effects of Overconfidence and Dependence on Aggregated Probability Judgments" appearing Journal of Modelling in Management,

Cooke, R.M., ElSaadany, S., Xinzheng Huang, X. (2008) On the Performance of Social Network and Likelihood Based Expert Weighting Schemes, Special issue on expert judgment Reliability Engineering & System Safety, 93, 745-756, Available online 12 March 2007, Volume 93, Issue 5, May 2008.

Cooke, R.M. (1991) Experts in Uncertainty, Oxford University Press.

De Groot, M. and Fienberg, SA. (1986) Comparing probability forecasters: basic binary concepts and multivariate extensions, in P. Goel and A. Zellner (eds) Bayesian Inference and Decision Techniques, Elsevier, New York, 1986,

Eggstaff,J.W., Mazzuchi,T.A. Sarkani, S. (2013) The Effect of the Number of Seed Variables on the Performance of Cooke's Classical Model, in preparation

Flandoli, F. Giorgi W.P. Aspinall, W. and Neri A (2010). " Comparing the performance of different expert elicitation models using a cross-validation technique" appearing in Reliability engineering and System Safety.

Lin, Shi-Woei, and Bier, V.M. (2008) "A Study of Expert Overconfidence" Reliability Engineering & System Safety, 93, 775-777, Available online 12 March 2007. Volume 93, Issue 5.

Lin, Shi-Woei, Cheng, Chih-Hsing (2008) "Can Cooke's Model Sift Out Better Experts and Produce Well-Calibrated Aggregated Probabilities?" Department of Business Administration, Yuan Ze University, Chung-Li, Taiwan Proceedings of the 2008 IEEE IEEM

Lin, Shi-Woei, Cheng, Chih-Hsing (2009) "The reliability of aggregated probability judgments obtained through Cooke's classical model", Journal of Modelling in Management, Vol. 4 Iss: 2, pp.149 – 161

Shi-Woei Lin, Ssu-Wei Huang, (2012) "Effects of overconfidence and dependence on aggregated probability judgments", Journal of Modelling in Management, Vol. 7 Iss: 1, pp.6 – 22.