Averaging Quantiles ?

Roger M.Cooke
Resources for the Future, Univ Strathclyde
Jan. 17, 2015

Abstract The proposal to combine expert distributions by averaging quantiles is examined. We show that this is equivalent to taking the harmonic mean of the experts' densities. Based on 31 recent professional expert judgment studies, the informativeness of averaging quantiles is on a par with performance based weighting but statistical accuracy is significantly degraded. In over half of the studies, the hypothesis that averaging quantiles produced statistically accurate probability statements would be rejected at the 5% level with rejection at the 0.1% level for a quarter of the studies. Analysis of a recent large cluster of applications at the World Health Organization suggests an explanation of poor performance and a way forward. Informativeness and statistical accuracy are negatively correlated for most experts, but are positively correlated among the statistically most accurate experts. When applicable, weighting only the statistically most accurate experts promises better results.

Introduction

Lichtendahl et al (2013) suggest that averaging experts' quantiles (AvQ) might give a better decision maker than an equal weight, or "averaging probabilities" (AvP) combination of their distribution functions. They note that AvQ is "sharper" than AvP. Flandoli et al (2011) also used this technique in their analysis of the Classical Model (CM). Averaging quantiles is easier to compute than averaging distributions, and is frequently employed by unwary practitioners. This note first shows that the density of AvQ is the harmonic mean of the densities of the combined distributions, and illustrates the effects on a simple example. The performances of AvQ, ASP and Performance Weighted (PW) combinations are then compared on the 31 professional expert judgment studies since 2006 based on CM.

Analysis

Let *F* and *G* be CDFs from experts 1 and 2, with densities *f, g*. Let *AvQ* , *avq* denote respectively the CDF and density of the result of averaging the quantiles of *F, G*. Then

$$AvQ^{-1}(r) = \tfrac{1}{2}\left(F^{-1}(r) + G^{-1}(r) \right). \qquad (1)$$

A good intuitive interpretation (Andrea Bevilacqua, personal communication) notes that *AvQ* takes the average of the experts' median values and a confidence interval whose width is the average of the experts' confidence intervals. The position of the median within the confidence interval depends on the distributions.

To gain further insight into eq (1), take derivatives of both sides:

$$1/avq(AvQ^{-1}(r) = ½\,(1/f\,(F^{-1}(r)) + 1/g^{-1}(G^{-1}(r))), \qquad (2)$$

$$avq(AvQ^{-1}(r) \;=\; \frac{2}{(1/f\,(F^{-1}(r)) + 1/g^{-1}(G^{-1}(r)))}\;. \qquad (3)$$

Eq. (3) says that *avq* is the harmonic mean of *f* and *g*, evaluated at points corresponding to the r-th quantile of each distribution. The harmonic mean of *n* numbers strongly favors the smallest of these numbers: the harmonic mean of 0.01 and 0.99 is 0.0198. To appreciate this fact, consider a flexible and tractable class of distributions on the unit interval:

$$F(x) = 1 - a^{-\frac{x^b}{1-x^b}}\,;\; F^{-1}(r) = \left(-\frac{\ln(1-r)}{\ln(a)}\cdot\left(1 - \frac{\ln(1-r)}{\ln(a)}\right)^{-1}\right)^{\frac{1}{b}} \qquad a > 1;\, b > 0 \quad (4)$$

Figure 1 shows two expert distributions from this class, *F* and *G*, and also shows AvQ, AvP and the geometric mean of *F*,*G*.

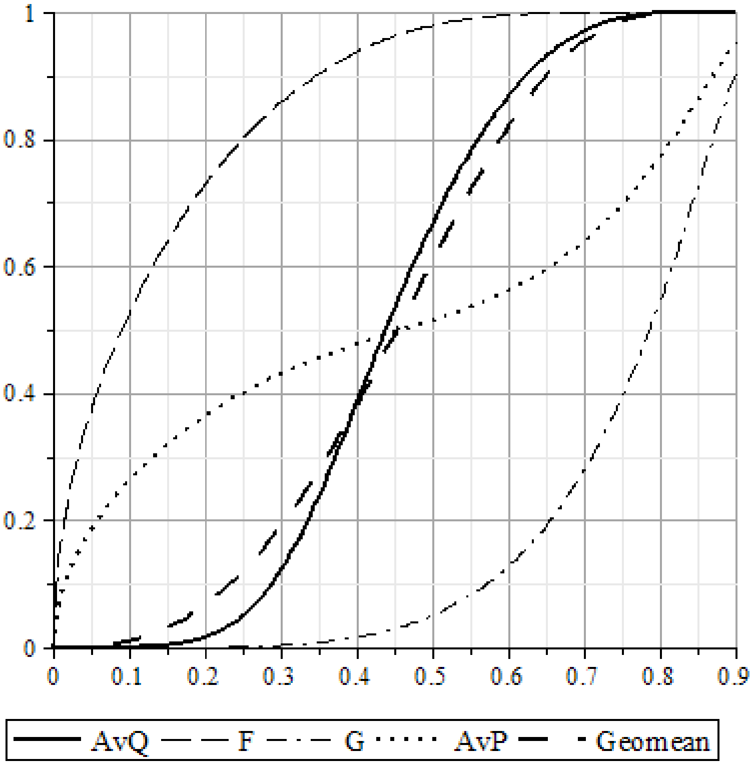


Figure 1:F(a=5, b=0.5), G(a=5, b=5), AvQ = Average Quantiles, AvP = Average Probabilities

The slope of *AvQ* is close to the smaller of the slopes of *F* and *G*; causing *AvQ(x)* to grow slowly for small *x* and decay quickly for large *x*. Table 1 shows that, despite the large disagreement between experts *F* and *G*, the *AvQ* combination has a *10-90%* confidence interval whose width is the average of that of the experts. *AvP* in contrast has a much wider confidence interval. Note that AvQ is more concentrated than the Geomean. This corresponds to the fact that the harmonic mean of distinct numbers is less than the geometric mean. The MAPLE scripts for these computations are included as an appendix.

Table 1: 10- and 90-percentiles of the distributions in Figure 1

Distribution	10%-tile	50%-tile	90%-tile	CI
F	0.004	0.09	0.346	0.343
G	0.572	0.79	0.899	0.327
Average Quantiles	0.288	0.44	0.623	0.335
Average Probbilities	0.101	0.47	0.901	0.800

The higher concentration of the *AvQ* combination would be very valuable if statistical accuracy were achieved. Statistical accuracy can be addressed with real experts assessing real variables from their fields for which true values are known post hoc.

Performance on real expert data

Using the *31* professional contracted expert judgment studies performed since 2006, it is possible to compare *AvQ,* and performance weighting (*PW*) as done in CM (Cooke, 1991) . In these studies, panels of 4 to 21 experts assessed between 7 and 17 calibration variables from their fields for which the true values were known post hoc. These studies were performed after the 2006 publication of the TU Delft expert judgment base of 45 studies (Cooke and Goossens 2008), and will also be made available to researchers.

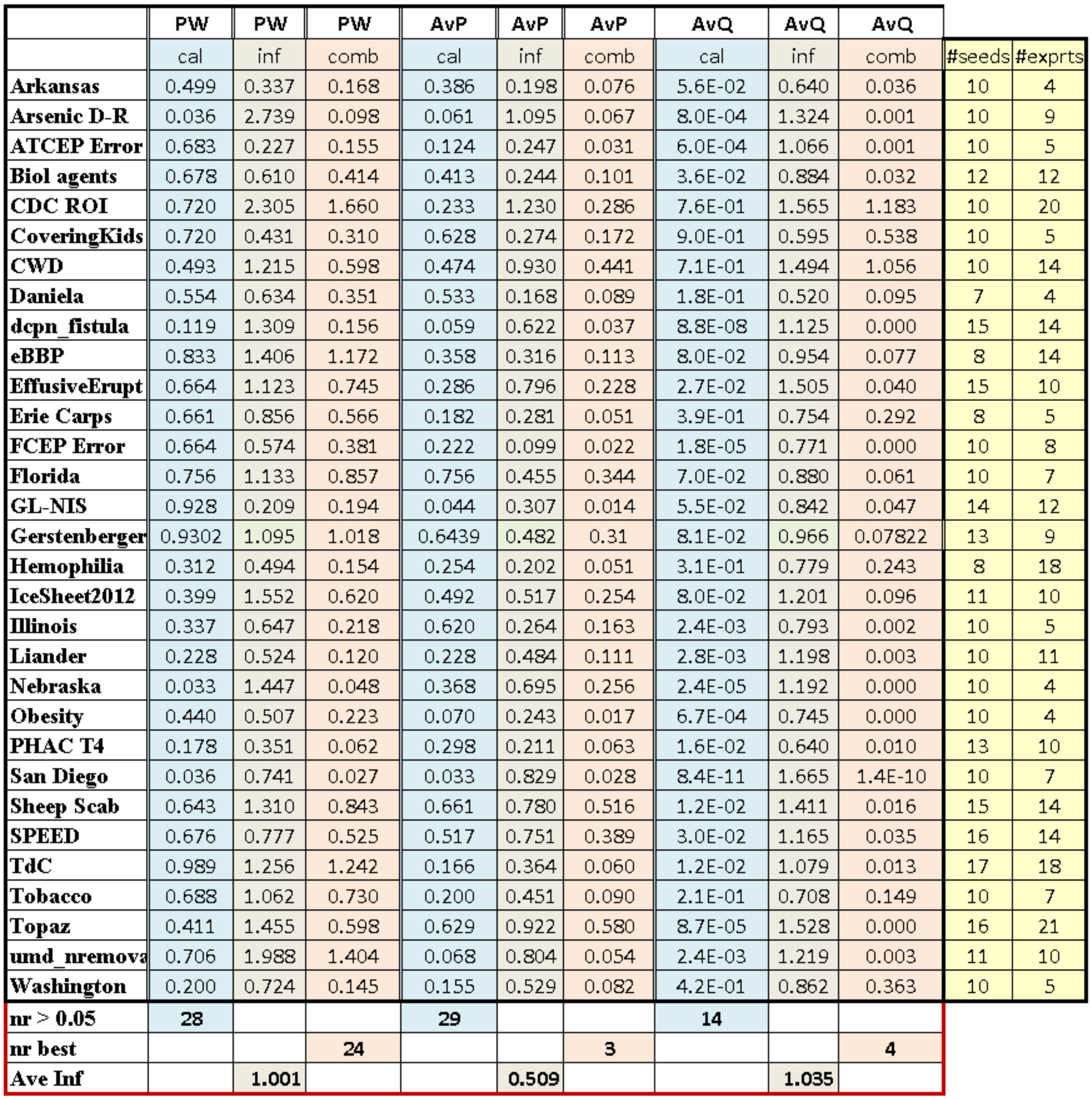
Whereas the pre 2006 studies contain several from the dawn of structured expert judgment, the recent studies were much better resourced, executed and documented. They were contracted and overseen by a variety of organizations including the Robert Wood Johnson Foundation, US EPA, US NOAA, Public Health Agency of Canada, PrioNet (Canada), Sanguin, British Government, European Community, NUMO (Japan), and Bristol University (UK). Prior to release of the post 2006 studies, the data underlying the results reported here can be obtained on request from the author.

In performing this comparison, the global weights combination was used and experts who assessed less than the full set of seed variables were excluded. This causes the *PW* and solutions used here to differ slightly from the solutions that will be published with the full datasets. The integrity of the present comparison is not affected; it was done to facilitate checks by third parties.

The performance of *AvQ*, *AvP* and *PW* are compared with regard to statistical accuracy (measured as the p-value at which one would falsely reject the hypotheses that the probability assessments were statistically accurate), information (measured as Shannon relative information with respect to a user supplied background measure) and a combined score (the product of the former two). Shannon relative information is used because it is scale invariant, tail insensitive, slow and familiar. The combined score satisfies a long run proper scoring rule constraint, and involves choosing an optimal statistical accuracy threshold beneath which experts are unweighted. Details on these scoring measures are found in Wittmann et al (2014). Out of sample validation is treated in Cooke et al (2014).

Table 1 gives the results. *AvQ* is the best of the three in 4 of the 31 cases, its informativeness is slightly higher than that of *PW*, and substantially higher than *AvP*. The statistical accuracy of *AvQ* is substantially below that of AvP and *PW*. In 8 cases its p-value is below 0.001, and in 17 cases (53%) the hypothesis that *AvQ* is statistically accurate would be rejected at the 5% level. Graphical interpretation of Table 1 is found in Figure 2.

Table 2



	PW	PW	PW	AvP	AvP	AvP	AvQ	AvQ	AvQ		
	cal	inf	comb	cal	inf	comb	cal	inf	comb	#seeds	#exprts
Arkansas	0.499	0.337	0.168	0.386	0.198	0.076	5.6E-02	0.640	0.036	10	4
Arsenic D-R	0.036	2.739	0.098	0.061	1.095	0.067	8.0E-04	1.324	0.001	10	9
ATCEP Error	0.683	0.227	0.155	0.124	0.247	0.031	6.0E-04	1.066	0.001	10	5
Biol agents	0.678	0.610	0.414	0.413	0.244	0.101	3.6E-02	0.884	0.032	12	12
CDC ROI	0.720	2.305	1.660	0.233	1.230	0.286	7.6E-01	1.565	1.183	10	20
CoveringKids	0.720	0.431	0.310	0.628	0.274	0.172	9.0E-01	0.595	0.538	10	5
CWD	0.493	1.215	0.598	0.474	0.930	0.441	7.1E-01	1.494	1.056	10	14
Daniela	0.554	0.634	0.351	0.533	0.168	0.089	1.8E-01	0.520	0.095	7	4
dcpn_fistula	0.119	1.309	0.156	0.059	0.622	0.037	8.8E-08	1.125	0.000	15	14
eBBP	0.833	1.406	1.172	0.358	0.316	0.113	8.0E-02	0.954	0.077	8	14
EffusiveErupt	0.664	1.123	0.745	0.286	0.796	0.228	2.7E-02	1.505	0.040	15	10
Erie Carps	0.661	0.856	0.566	0.182	0.281	0.051	3.9E-01	0.754	0.292	8	5
FCEP Error	0.664	0.574	0.381	0.222	0.099	0.022	1.8E-05	0.771	0.000	10	8
Florida	0.756	1.133	0.857	0.756	0.455	0.344	7.0E-02	0.880	0.061	10	7
GL-NIS	0.928	0.209	0.194	0.044	0.307	0.014	5.5E-02	0.842	0.047	14	12
Gerstenberger	0.9302	1.095	1.018	0.6439	0.482	0.31	8.1E-02	0.966	0.07822	13	9
Hemophilia	0.312	0.494	0.154	0.254	0.202	0.051	3.1E-01	0.779	0.243	8	18
IceSheet2012	0.399	1.552	0.620	0.492	0.517	0.254	8.0E-02	1.201	0.096	11	10
Illinois	0.337	0.647	0.218	0.620	0.264	0.163	2.4E-03	0.793	0.002	10	5
Liander	0.228	0.524	0.120	0.228	0.484	0.111	2.8E-03	1.198	0.003	10	11
Nebraska	0.033	1.447	0.048	0.368	0.695	0.256	2.4E-05	1.192	0.000	10	4
Obesity	0.440	0.507	0.223	0.070	0.243	0.017	6.7E-04	0.745	0.000	10	4
PHAC T4	0.178	0.351	0.062	0.298	0.211	0.063	1.6E-02	0.640	0.010	13	10
San Diego	0.036	0.741	0.027	0.033	0.829	0.028	8.4E-11	1.665	1.4E-10	10	7
Sheep Scab	0.643	1.310	0.843	0.661	0.780	0.516	1.2E-02	1.411	0.016	15	14
SPEED	0.676	0.777	0.525	0.517	0.751	0.389	3.0E-02	1.165	0.035	16	14
TdC	0.989	1.256	1.242	0.166	0.364	0.060	1.2E-02	1.079	0.013	17	18
Tobacco	0.688	1.062	0.730	0.200	0.451	0.090	2.1E-01	0.708	0.149	10	7
Topaz	0.411	1.455	0.598	0.629	0.922	0.580	8.7E-05	1.528	0.000	16	21
umd_nremova	0.706	1.988	1.404	0.068	0.804	0.054	2.4E-03	1.219	0.003	11	10
Washington	0.200	0.724	0.145	0.155	0.529	0.082	4.2E-01	0.862	0.363	10	5
nr > 0.05	28			29			14				
nr best			24			3			4		
Ave Inf		1.001			0.509			1.035			

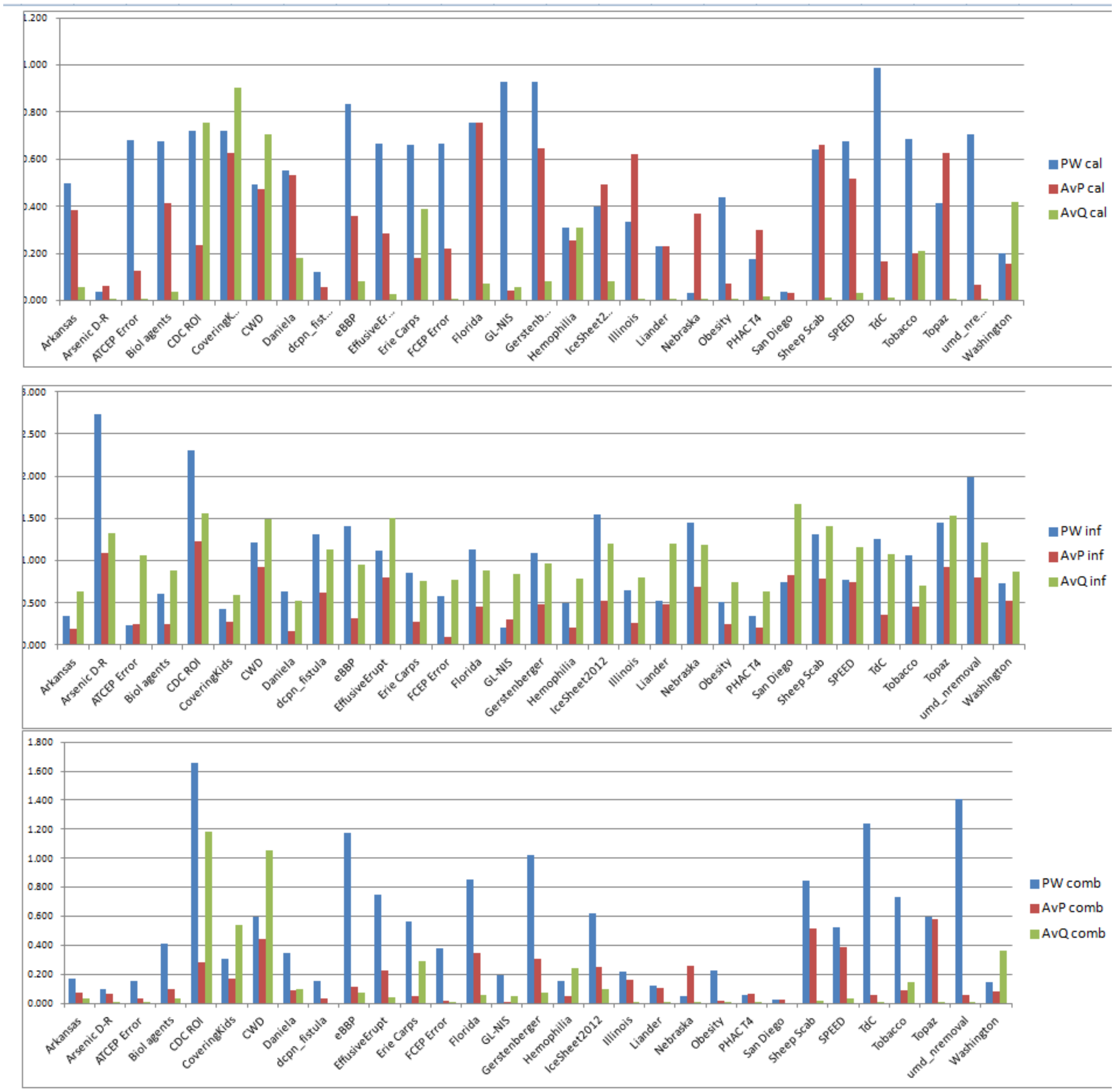


Figure 2: Statistical accuracy, informativeness and combined scores

This data provides evidence on how performance is affected by the number of experts and number of calibration variables. Focusing on statistical accuracy, Figure 6 graphs the number of calibration variables and number of experts against the statistical accuracy scores, for *AvQ*, *AvP*, and *PW*. It appears that *AvQ* degrades as the number of calibration variables increases. There is a 39% chance that a randomly selected study has more than 10 calibration variables. None of the 8 studies with *AvQ* statistical accuracy above 0.1 have more than 10 calibration variables. The statistical power of the measure of statistical accuracy increases with the number of calibration variables and this would tend to suppress statistical accuracy scores of all experts and combinations alike. However, no such effect is observed for *AvP* or *PW*. The number of experts does not have a marked effect on any of the combinations.

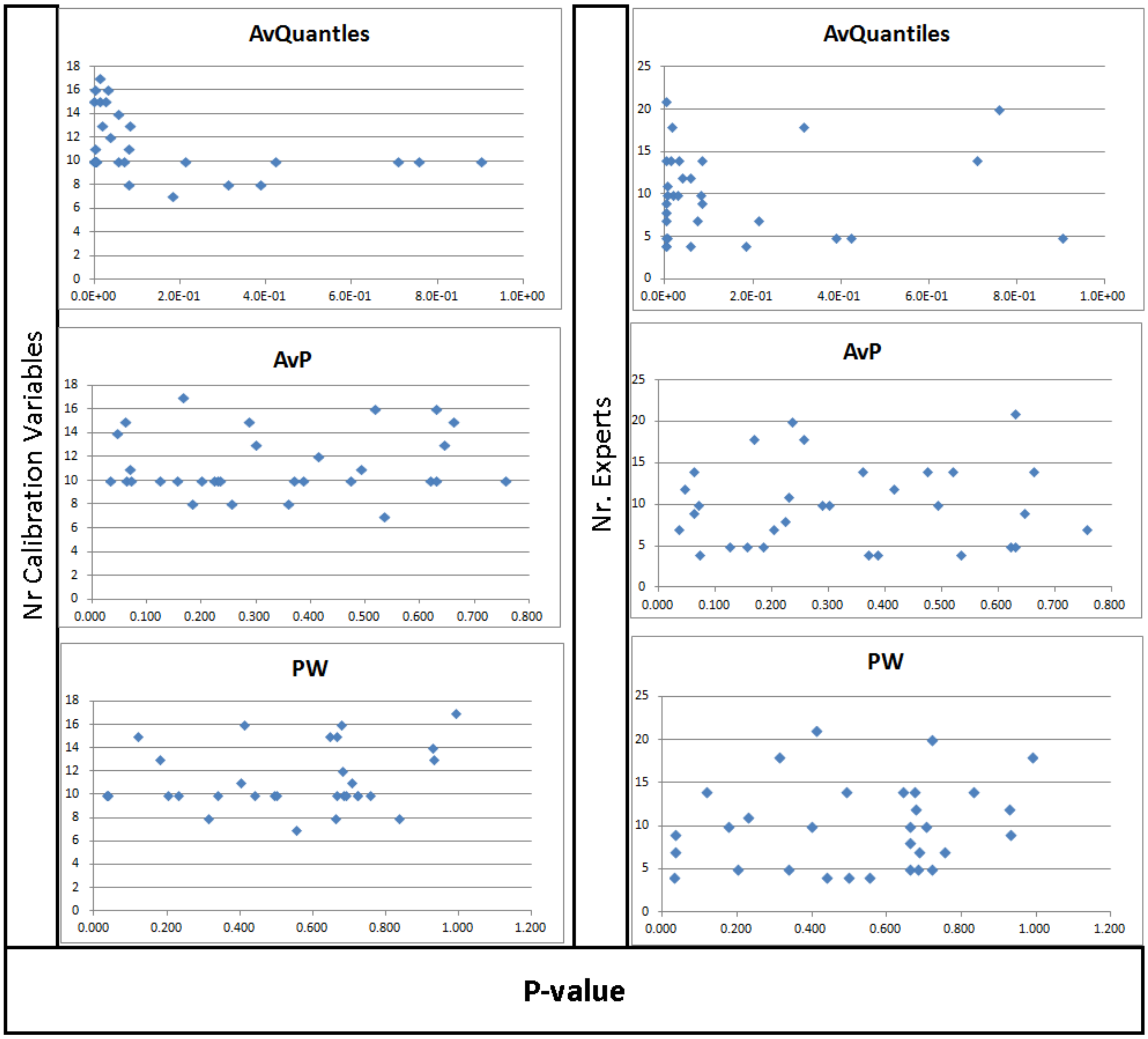


Figure 4: P-values plotted against number of calibration variables and number of experts

Discussion

It would be desirable to have a simple algorithm of combining experts which improves the informativeness relative to AvP without degrading the statistical accuracy. From the data analyzed above it appears that AvQ does improve informativeness, but sacrifices statistical accuracy. Some hints on why this happens, and how the situation might be improved can be gleaned from a very large study recently completed by the World Health Organization (WHO). This study involved 72 epidemiologists and health professionals over the whole world assessing relative frequencies of infection pathways for various pathogens in various regions of the world. In total, 134 distinct panels were formed involving overlapping sets of 6 to 30 experts each. Full documentation is in preparation, but expert results are already available. The experts assessed similar, though not identical calibration variables. Although informativeness scores are relative to a background measure on a support that is computed per panel, because of the variables' similarity, it is reasonable to compare informativeness and statistical accuracy scores of all experts. Figure 5 shows a pronounced negative correlation between log statistical accuracy and informativeness. There is tendency for the more informative experts to be less accurate statistically. Moreover, the statistically accurate experts are in the minority. Only 4 of the 72 experts returned statistical accuracy scores above the traditional 5% rejection threshold.

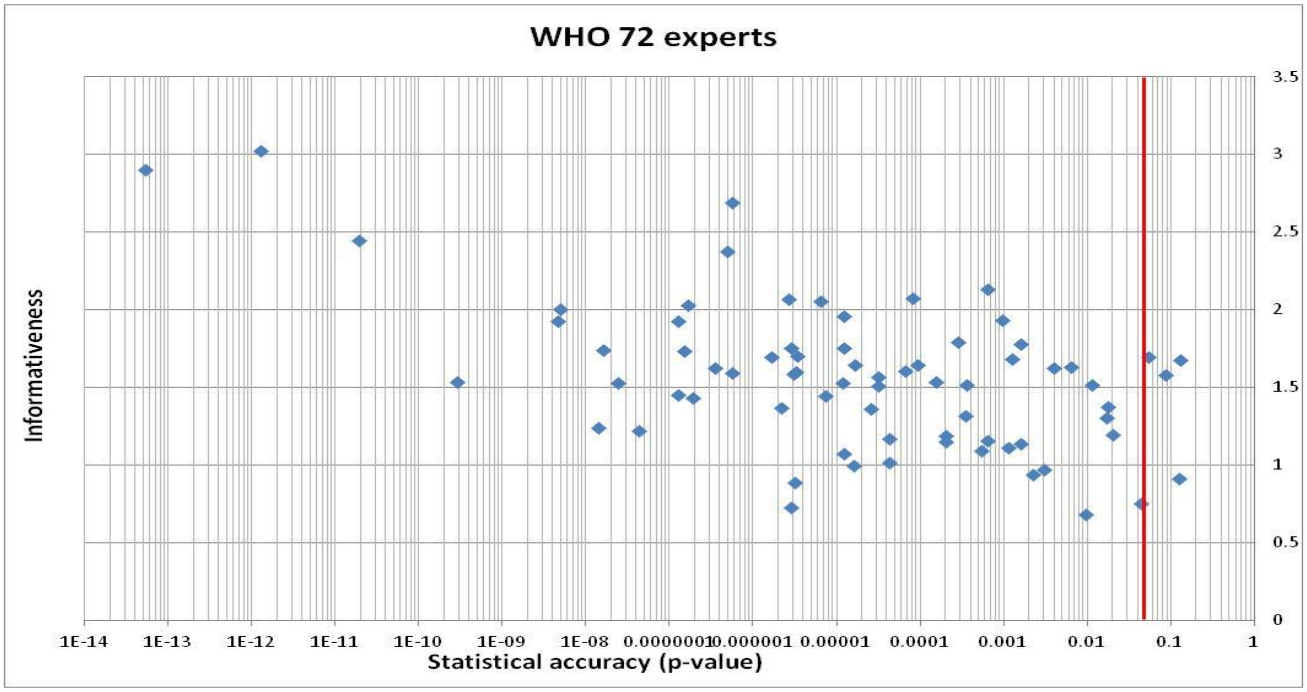


Figure 5: Informative accuracy and informativeness for 72 experts in the WHO study.

In a typical panel where the statistically inaccurate experts outnumber the accurate experts, AvQ will exhibit confidence bands more resembling the inaccurate experts. Indeed, any combination rule which up-weights informative experts without regard for their statistical accuracy is likely to suffer the same fate.

Figure 5 also suggests that the negative correlation between informativeness and statistical accuracy is strongest among the more inaccurate experts. This suggestion is confirmed in Figure 6, which plots the running rank correlation between informativeness and statistical accuracy for experts with accuracy above α, where α runs from worst to best. For the highest α this correlation is trivially 1, for the lowest α it is equal to the rank correlation in the entire sample of 72 experts.

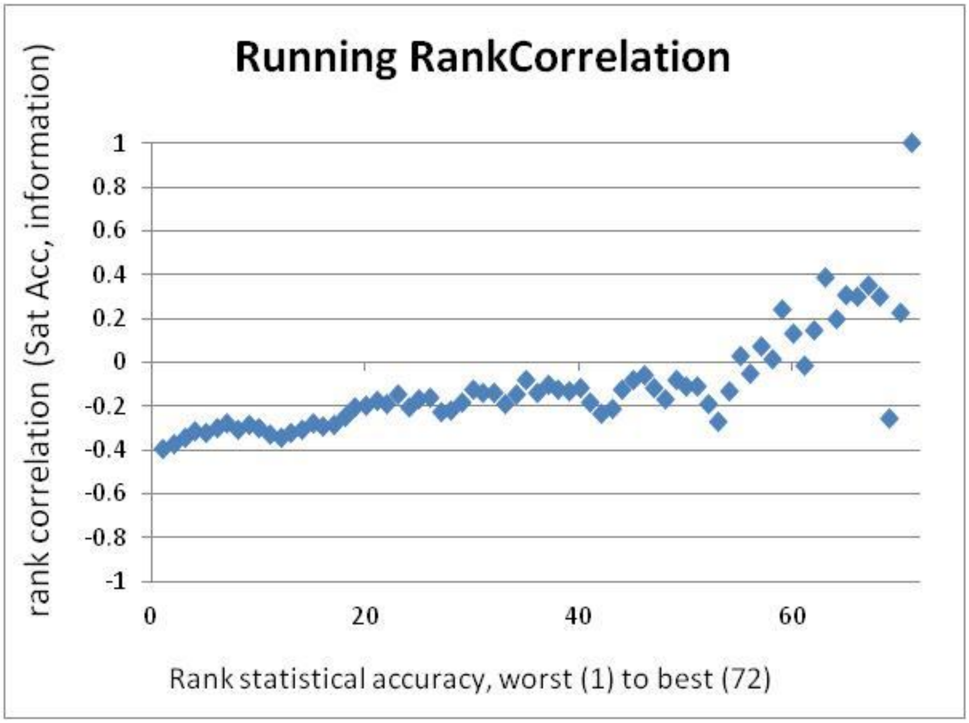


Figure 6 Running rank correlation of informativeness and statistical accuracy in the WHO expert data

The rank correlation becomes effectively zero among the 20% most accurate experts, and it is actually positive among the top 10%. This suggests that a combination rule which is restricted to, say, the top 10% most accurate experts stands a good chance of improving informativeness without losing statistical accuracy. It is interesting to remark in this regard, that the classical model usually concentrates the weight among 1 to 3 most accurate experts.

Three caveats apply. First, in many panels there are no statistically accurate experts. The most accurate expert may score well below the 5% threshold. Second, any such combination rule requires the assessment of calibration variables. Finding good calibration variables from the experts' field is time consuming and requires that the analyst dive deeply into the subject matter. The main appeal of AvQ and AvP - that they do not require calibration variables - would be lost. The classical model involves choosing a p-value threshold for selecting the experts to be weighted, based on optimizing the performance of the combination. It remains to be seen whether an "a priori" threshold can deliver comparable or better performance. Finally, these suggestions are based on only one, albeit large, data set. These caveats can be addressed by analyzing the extensive expert data already in the public domain (Cooke and Goossens 2008).

Conclusion

Based on this data analysis, AvQ does not deliver statistically accurate combinations in panels of real experts. As yet, combination rules that do not take expert performance into account, have been unable to deliver both informativeness and statistical accuracy. Recourse to calibration variables may be unavoidable. The classical model has remained unchanged for 25 odd years and has generated a wealth of data which provides a test bed for new ideas. The classical model provides a baseline for informative and statistically accurate combinations; it most unlikely that it can't be improved upon or surpassed.

References

Lichtendahl, Jr.,K. C., Grushka-Cockayne, Y., Winkler, R. L., (2013) Is It Better to Average Probabilities or Quantiles? MANAGEMENT SCIENCE Vol. 59, No. 7, July 2013, pp. 1594–1611 ISSN 0025-1909 (print) . ISSN 1526-5501 (online) http://dx.doi.org/10.1287/mnsc.1120.1667 ©2013 INFORMS

Cooke, R.M., Wittmann, M.E., Lodge, D.M., Rothlisberger, J.D., Rutherford E.S., Zhang, H. and Mason, D.M. (2014) "Out-of-Sample Validation for Structured Expert Judgment of Asian Carp Establishment in Lake Erie", Integrated Environmental Assessment and Management, open access. DOI: 10.1002/ieam.1559

Cooke, Roger M., Goossens, L.H.J. (2008) TU Delft Expert Judgment Data Base, Special issue on expert judgment Reliability Engineering & System Safety, 93, 657-674, Available online 12 March 2007, Issue 5, May 2008.http://onlinelibrary.wiley.com/doi/10.1002/ieam.1559/abstract

Cooke R., (1991) Experts in Uncertainty; Opinion and Subjective Probability in Science, Oxford University Press; NYork Oxford, 321 pages; ISBN 0-19-506465-8

Wittmann, M.E., Cooke, R.M., Rothlisberger, J.D., Rutherford, E. S., Zhang, H., Mason, D., Lodge, D.M. (2014): Structured expert judgment to forecast species invasions: Bighead and silver carp in Lake Erie, Conservation Biology . DOI: 10.1111/cobi.12369 http://onlinelibrary.wiley.com/doi/10.1111/cobi.12369/full

Flandoli, F., Giorgi, E., Aspinall W. P., and Neri, A., (2011) Comparison of a nexpert elicitation model with the Classical Model, equal weights and single experts, using a cross-validation technique. Reliability Engineering and System Safety, 96, 1292-1310. doi:10.1016/j.ress.2011.05.012.

APPENDIX

$>$ $with(student) : with(Student[Calculus1]) : reset :$

$>$ $F := (a, b, x) \rightarrow 1 - a^{-\frac{x^b}{1-x^b}} :$

$>$ $Finv := (a, b, r) \rightarrow \left(-\frac{\ln(1-r)}{\ln(a)} \cdot \left(1 - \frac{\ln(1-r)}{\ln(a)} \right)^{-1} \right)^{\frac{1}{b}} :$

$>$ $f(a, b, x) := -a^{-\frac{x^b}{1-x^b}} \left(-\frac{x^b\, b}{x\,(1-x^b)} - \frac{(x^b)^2\, b}{(1-x^b)^2\, x} \right) \ln(a) :$

$>$ $aa := 5 : b1 := .5 : b2 := 5 : b3 := .4 :$

$>$ $H :=$**proc**$(x) : unassign('r') : assign(solve(\{\mathrm{Finv}(aa, b1, r) + Finv(aa, b2, r) = 2 \cdot x, r \geq 0, r \leq 1\})_1) :$ **return** r; **end proc**:

$>$ $FF :=$**proc**(x) :**return** $F(aa, b1, x)$:**end proc**:

$>$ $GG :=$**proc**(x) : **return** $F(aa, b2, x)$:**end proc**:

$>$ $geom :=$**proc**(x) : **return** $\frac{int\left((f(aa, b1, z) \cdot f(aa, b2, z))^{.5}, z = 0..\mathrm{x} \right)}{int\left((f(aa, b1, z) \cdot f(aa, b2, z))^{.5}, z = 0..1 \right)}$: **end proc**:

$>$ $avp :=$**proc**(x) : **return** $\frac{(F(aa, b1, x) + F(aa, b2, x))}{2}$: **end proc**:

```
> plot( [H, FF, GG, avp, geom], 0 ..0.9, legend = [ "AvQ", "F", "G", "AvP",
      "Geomean"], linestyle = [solid, dash, dashdot, dot, spacedot], color
      = [black, black, black, black, black]) :
```