revised

Supplementary Online Material Structured Expert Judgment for Asian Carp with Out-of-Sample Validation

This supplementary online material contains details on the expert scoring measures, and on the expert data analysis for the Asian Carp study.

1. Performance Measures and Combination: the classical model

There are two generic, quantitative measures of expert performance, *calibration* and *information*. Loosely, calibration measures the statistical likelihood that a set of experimental results correspond, in a statistical sense, with an expert's assessments. More precisely, it is the p-value at which we would falsely reject the hypothesis that an expert's probability statements were accurate. In this study the 5, 50 and 95 percentiles, or quantiles, were elicited from each expert for each of the continuous variables. Hence, each expert divides the range of possible outcomes of each variable into 4 intervals: less than or equal to the 5% value, greater than the 5% value and less than or equal to the 50% value, etc. The probabilities for these intervals are expressed as a vector

$p = (p_1, p_2, p_3, p_4) = (0.05, 0.45, 0.45, 0.05)$.

Calibration
If N quantities are assessed, each expert may be regarded as a statistical hypothesis, namely that each realization falls in one of the four inter-quantile intervals with probability vector *p*. Suppose we have realizations $x_1, \ldots x_N$ of these quantities. We may then form the sample distribution of the expert's inter quantile intervals as:

$s_1(e) = \#\{ i \mid x_i \leq 5\% \text{ quantile}\}/N$
$s_2(e) = \#\{ i \mid 5\% \text{ quantile} < x_i \leq 50\% \text{ quantile}\}/N$
$s_3(e) = \#\{ i \mid 50\% \text{ quantile} < x_i \leq 95\% \text{ quantile}\}/N$
$s_4(e) = \#\{ i \mid 95\% \text{ quantile} < x_i \}/N$
$s(e) = (s_1, \ldots s_4)$

Note that the sample distribution depends on the expert *e*. If the realizations are indeed drawn independently from a distribution with quantiles as stated by the expert then the quantity

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

is asymptotically distributed as a chi-square variable with 3 degrees of freedom. This is the likelihood ratio statistic, and I(s | p) is the relative information of distribution *s* with respect to *p*. Extracting the leading term of the logarithm yields the familiar chi-square test statistic for goodness of fit. There are advantages in using the form in (1) (Cooke 1991).

If after a few realizations the expert were to see that all realization fell outside his 90% central confidence intervals, he might conclude that these intervals were too narrow and might broaden them on subsequent assessments. This means that for this expert the uncertainty distributions are *not* independent, and he learns from the realizations. Expert learning is not a goal of an expert

judgment study. Rather, the problem owner wants experts who do not need to learn from the elicitation. Independence is not an assumption about the expert's distribution but a desideratum of the problem owner. Hence the decision maker scores expert *e* as the statistical likelihood of the hypothesis

H_e: "the inter quantile interval containing the true value for each variable is drawn independently from probability vector p."

A simple test for this hypothesis uses the test statistic (1), and the likelihood, or p-value, or **calibration score** of this hypothesis, is:

$$Cal(e) = \text{p-value}(e) = Prob\{2NI(s(e)\ p) \geq r \mid H_e\}$$

where *r* is the value of (1) based on the observed values $x_1, \ldots x_N$. It is the probability under hypothesis H_e that a deviation at least as great as *r* should be observed on *N* realizations if H_e were true. Calibration scores are absolute and can be compared across studies. However it is appropriate to equalize the power of the different hypothesis tests by equalizing the effective number of realizations. To compare scores on two data sets with *N* and *N' realizations*, we simply use the minimum of *N* and *N'* in (1), without changing the sample distribution *s*.
The calibration score uses the language of simple hypothesis testing to measure the degree to which the data supports the hypothesis that the expert's probabilities are accurate. Low scores, near zero, mean that it is unlikely that the expert's probabilities are correct. High scores, near 1, indicate good support.

Information
The second scoring variable is information. Loosely, the information in a distribution is the degree to which the distribution is concentrated. Information cannot be measured absolutely, but only with respect to a background measure. Being concentrated or "spread out" is measured relative to some other distribution.

Measuring information requires associating a density with each assessment of each expert. To do this, we use the unique density that complies with the experts' quantiles and is minimally informative with respect to the background measure. This density can easily be found with the method of Lagrange multipliers. For a uniform background measure, the density is constant between the assessed quantiles. The background measure is not elicited from experts as indeed it must be the same for all experts; instead it is chosen by the analyst.

The uniform and log-uniform background measures require an *intrinsic range* on which these measures are concentrated. The classical model implements the so-called *k*% overshoot rule: for each item we consider the smallest interval *I = [L, U]* containing all the assessed quantiles of all experts and the realization, if known. This interval is extended to

$$I^* = [L^*, U^*];\ L^* = L - k(U-L)/100;\ \ U^* = U + k(U-L)/100.$$

The value of k is chosen by the analyst. A large value of *k* tends to make all experts look quite informative, and tends to suppress the relative differences in information scores. The default

value $k = 10$ is used here. *The* **information score** of expert *e* on assessments for uncertain quantities 1…N is

$$Inf(e) = \text{Average Relative information w.r.t. Background} = (1/N)\ \Sigma_{i=1..N} I(f_{e,i} \mid g_i)$$

where g_i is the background density for variable i and $f_{e,i}$ is expert e's density for item i. This is proportional to the relative information of the expert's joint distribution given the background, under the assumption that the variables are independent. As with calibration, the independence assumption reflects a desideratum of the decision maker and not an elicited feature of the expert's joint distribution. The information score does not depend on the realizations. An expert can give himself a high information score by choosing his quantiles very close together. The information score of *e* depends on the intrinsic range and on the assessments of the other experts. Hence, information scores cannot be compared across studies.

The above information score is chosen because it is familiar, tail insensitive, scale invariant and "slow". The latter property means that relative information is a slow function; large changes in the expert assessments produce only modest changes in the information score. This contrasts with the likelihood function in the calibration score, which is a very "fast" function. This causes the product of calibration and information to be driven by the calibration score.

Combination: Decision Maker
The **combined score** of expert *e* will serve as an (unnormalized) weight for *e*:

$$w_\alpha(e) = Cal(e) \times Inf(e) \times \mathbb{1}_\alpha(Cal(e) \geq \alpha), \qquad (2)$$

where $\mathbb{1}_\alpha(Cal(e)\alpha) = 1$ if $Cal(e) \geq \alpha$, and is zero otherwise. The combined score thus depends on α; if $Cal(e)$ falls below cut-off level α, expert *e* is unweighted. The presence of a cut-off level is imposed by the requirement that the combined score be an asymptotically strictly proper scoring rule. That is, an expert maximizes his/her long run expected score by and only by ensuring that his probabilities $p = (0.05, 0.45, 0.45, 0.05)$ correspond to his true beliefs (Cooke, 1991). α is similar to a significance level in simple hypothesis testing, but its origin is to measure 'goodness' and not to reject hypotheses.

A combination of expert assessments is called a "decision maker" (DM). All decision makers discussed here are examples of linear pooling; the classical model is essentially a method for deriving weights in a linear pool. "Good expertise" corresponds to good calibration (high statistical likelihood, high p-value) and high information. Weights that reward good expertise and pass these virtues on to the decision maker are desired.

The reward aspect of weights is very important. We could simply solve the following optimization problem: find a set of weights such that the linear pool under these weights maximizes the product of calibration and information. Solving this problem on real data, one finds that the weights do not generally reflect the performance of the individual experts. As an expert's influence on the decision maker should not appear haphazard, and "gaming" the system with assessments tilted to achieve a desired outcome should be discouraged, we must impose a strictly scoring rule constraint on the weighting scheme.

The scoring rule constraint requires the term $\mathbb{1}_\alpha(Cal(e) \geq \alpha)$ in eq (2), but does not indicate what value of α we should choose. Therefore, we choose α to maximize the combined score of the resulting decision maker. Let $DM_\alpha(i)$ be the result of linear pooling for any item *i* with weights proportional to (2):

$$DM_\alpha(i) = \sum_{e=1,..E} w_\alpha(e)\, f_{e,i} \;/\; \sum_{e=1,..E} w_\alpha(e) \tag{3}$$

The *optimized global weight DM* is DM_{α^*} where α* maximizes

$$\textit{calibration score}(DM_{a^*}) \times \textit{information score}(DM_{\alpha^*}). \tag{4}$$

This weight is termed global as the information score is based on all the assessed calibration items.
A variation on this scheme allows a different set of weights to be used for each item. This is accomplished by using information scores for each item rather than the average information score:

$$w_\alpha(e,i) = 1_\alpha(\textit{calibration score}) \times \textit{calibration score}(e) \times I(f_{e,i} \mid g_i) \tag{5}$$

For each α we define the *Item weight* DM_α for item *i* as

$$IDM_\alpha(i) = \sum_{e=1,..E} w_\alpha(e,i)\, f_{e,i} \;/\; \sum_{e=1,..E} w_\alpha(e,i) \tag{6}$$

The *optimized item weight DM* is IDM_{α^*} where α* maximizes

$$\textit{calibration score}(IDM_a) \times \textit{information score}(IDM_\alpha). \tag{7}$$

The non-optimized versions of the global and item weight DM's are obtained by setting $\alpha = 0$.

In this study the global and item weighting schemes are identical.

2. Details of the Asian Carp study

The Briefing booklet sent to the experts is available here:
https://www.dropbox.com/s/8ingqnc8389znwl/Briefing%20Booklet%20Erie%20Carps_Final.pdf

Table 1. List of participating experts. In the Results section, experts are not listed in the same order as in this table

Name	Title, affiliation and expertise
Duane C. Chapman, MSc	*Research Fish Biologist, United States Geological Survey, River Studies: Invasive Carp Research Program. Chapman is affiliated with the Asian Carp Working Group, Asian Carp Rapid Response Team, Mississippi River Basin*

	Panel on Aquatic Nuisance Species and the American Fisheries Society.
Joseph V. DePinto, Ph.D.	*Senior Scientist, Limnotech. A former professor of environmental engineering, DePinto conducts aquatic ecosystem structure and functioning research, and designs education and management programs, with emphases on the Great Lakes region.*
Tomas O. Höök, Ph.D.	*Assistant Professor of Fisheries and Aquatic Sciences, Purdue University, Department of Forestry and Natural Resources. Focuses on fish and fisheries ecology in the Laurentian Great Lakes.*
Timothy B. Johnson, Ph.D.	*Research Scientist, Ontario Ministry of Natural Resources, Great Lakes Fisheries Division. Johnson's expertise is in bioenergetics models, specifically for Lake Erie, and has studied the biology of invasive round goby.*
Roger L. Knight	*Lake Erie Fisheries Program Administrator, Ohio Department of Natural Resources, Division of Wildlife. Serves on the Lake Erie Committee and the Council of Lake Committees (Great Lakes Fisheries Commission)*
Stuart A. Ludsin, Ph.D.	*Associated Professor, The Ohio State University Department of Evolution, Ecology and Organismal Biology. Ludsin's expertise is on mechanisms that regulate fish population and community structure and dynamics, food web interactions and natural resource management.*
Charles P. Madenjian, Ph.D.	*Research Fishery Biologist, United States Geological Survey, Western Basin Ecosystems Branch, Lake Michigan Section. Madenjian is a quantitative fisheries biologist and has focused on fish bioenergetics modeling in the Great Lakes.*
Peter Meisenheimer	*Executive Director, Ontario Commercial Fisheries Association. A biologist who represents commercial fisheries in Ontario, currently a member of the Canadian Committee of Advisors of the Great Lakes Fishery Commission and Chair of the Ontario Species at Risk Public Advisory Committee.*
Mark A. Pegg, Ph.D.	*Associate Professor, School of Natural Resources at the University of Nebraska Lincoln. Pegg specializes in fisheries management, the impacts of aquatic nuisance species including Asian carps, and restoration ecology.*
Kevin Reid	*Ph.D. candidate, University of Guelph, and Assessment Manager and Fisheries Biologist-Technical Advisor Ontario Commercial Fisheries Association.*
Brian J. Shuter, Ph.D.	*Professor, Department of Ecology and Evolutionary Biology, University of Toronto and Research Scientist Aquatic Research & Development Section Ontario Ministry of Natural Resources. Shuter focuses on food web dynamics, population ecology and growth/production models for fish and zooplankton.*

Table 2. Calibration variable descriptions and acronym identifiers for each variable that are used in Figure 1 and Table 5.

Calibration Variables	
1	Biomass of walleye in Lake Erie in 2011 (metric tons km^2) (WY11)
2	Biomass of round goby in Central Basin Lake Erie 2011 (metric tons km^2) (RG11)

3	Biomass of rainbow smelt in Lake Erie in 2011 (metric tons km^2) (RS11)
4	Biomass of gizzard shad in Lake Erie in 2011 (metric tons km^2) (GS11)
5	% of fish in diet of smallmouth bass (age 2+) (Central Basin) (SMBa11)
6	% of fish in diet of white bass (yearling) (Central Basin) (WBy11)
7	% of fish in diet of white bass (age 2+) (Central Basin) (WBa11)
8	% of fish in diet of yellow perch (yearling) (Central Basin) (YPy11)
9	% of fish in diet of yellow perch (age 2+) (Central Basin) (YPa11)
10	% of rainbow smelt in diet of walleye (yearling) (Central Basin) (RS_WYy11)
11	% of rainbow smelt in diet of walleye (age 2+) (Central Basin) (RS_WYa11)
12	% of round goby in diet of walleye (yearling) (Central Basin) (RG_WYy11)
13	% of round goby in diet of smallmouth bass (age 2+) (Central Basin) (RG_SMBa11)
14	% of round goby in diet of yellow perch (age 2+) (Central Basin) (RG_YPa11)
15	kg of Asian carps captured in Marseilles and Dresden pools (CAWS) in 2012 (pool12)

Table 3 shows the results of the individual expert and equally-weighted decision maker (EW DM) scores with discrepancy. The EW DM had a calibration score of 0.31, indicating that we would not reject the hypothesis that EW's probability assessments were accurate. Experts' calibration scores were high for a panel of this size, and varied from 2E-6 to 0.53, with 9 of the experts scoring above 0.05. Note that all experts assessed all 15 calibration variables, with the exception of expert 8 who assessed only 11, thus reducing the effective number of items to 11 (see discussion following eq (1)). The number of calibration variables corresponds to the power of the statistical test used to calculate each expert's p-value; using 11 instead of 15 effective calibration items has the effect of raising the calibration scores of the other experts. For example, with 15 effective items, expert 4's p-value would decrease from 0.7606 to 0.661. Without equalizing the effective number of calibration variables, expert 8 would enjoy an advantage relative to the others.

The EW DM is less informative than any of the experts individually (Table 3; Figure 1). The column "unnormalized weight" (column 6, Table 3), is the product of the numbers in columns 2 and 4 and is the combined score used in performance based weighting. Columns 7 and 8 show, respectively, the relative information between each expert and the EW DM, which is a sort of 'average expert'. We see that these scores typically lie in the interval [0.9, 3.4] for all variables. These information scores have no absolute meaning, as they depend on the given set of expert assessments. However, they can be meaningfully compared to the changes in the resulting DM

caused by removing experts and calibration items one at a time. This is discussed in the robustness analysis section below (and see Tables 5,6).

Table 3. Expert and equally-weighted decision maker scores with discrepancy. The first column gives the labels of experts – including the decision-makers. The second column shows the calibration scores for all experts. The third and fourth columns indicate the average information on all variables and on the calibration variables, respectively. Discrepancy is shown in columns 7 and 8 as the relative information of each expert with respect to the equal weight combination on all variables (7) and calibration variables (8).

Expert and EW DM scores with Discrepancy							
Expert	**P-Value**	**Mean rel. Info, ALL**	**Mean rel. Info, Calibr vbls**	**# Assessed Calibr vbls**	**UnNormalized wgt**	**Rel.Inf to EW DM, All Vbls**	**Rel.Inf to EW DM, calibr. Vbls**
1	0.1815	1.409	0.6121	15	0.1111	0.8712	0.4807
2	0.1227	0.6903	0.6648	15	0.08159	0.6083	0.5314
3	0.005634	3.789	1.47	15	0.008283	3.169	1.095
4	0.7606	3.812	0.8562	15	0.6513	3.252	0.4717
5	0.666	2.16	0.84	15	0.5595	1.641	0.5993
6	1.93E-06	1.494	1.381	15	2 .66E-06	1.375	1.298
7	0.05946	1.852	1.158	15	0.06883	1.176	0.8126
8	0.615	4.348	1.086	11	0.6678	3.42	0.5294
9	0.5276	2.56	1.288	15	0.6797	1.654	0.8251
10	0.2587	2.617	0.8282	15	0.2142	2.08	0.4858
11	0.5276	2.53	0.8071	15	0.4258	1.678	0.5083
EW DM	0.3126	0.5748	0.2943	15	0.09197	0	0

Robustness analysis

The effect of removing an expert other than expert 4 is zero (Table 5), illustrating how robust the model is to experts. Removing expert 4 induces a relative information score (with respect to the original DM) of 1.872, which is at the lower end of the numbers in column 7 of Table 3. This indicates that the loss of expert 4 induces a change that overall is smaller than the differences among the experts themselves. The perturbed DM in this case still exhibits good performance.

For robustness on items, the loss of any single calibration variable has virtually no effect on the DM, weight 1 always goes to expert 4 (Table 6). The statistical accuracy (p-value) and informativeness scores of the perturbed DMs do change somewhat, but the changes are small.

Overall, the informativeness and statistical accuracy of the optimized performance based DM were very good, and this conclusion is quite robust against loss of a single expert and loss of a single calibration variable.

Table 4. Robustness on experts, showing the result of removing experts one at a time and re-computing the DM

Robustness on Experts for Optimized PW DM					
Expert removed	Mean rel. Info, ALL	Mean rel. Info, Calibr vbls	P-Value	Rel.Info wrt Orig. DM, All	Rel.Info wrt Orig. DM, calibr vbls
1	3.801	0.8029	0.7606	0	0
2	3.668	0.8299	0.7606	0	0
3	3.794	0.8477	0.7606	0	0
4	2.138	0.8319	0.666	1.872	0.7984
5	3.806	0.8433	0.7606	0	0
6	3.789	0.737	0.7606	0	0
7	3.812	0.8562	0.7606	0	0
8	3.801	0.8562	0.661	0	0
9	3.769	0.8562	0.7606	0	0
10	3.798	0.8562	0.7606	0	0
11	3.806	0.8265	0.7606	0	0
None	3.812	0.8562	0.7606	0	0

Table 5. Robustness on items, showing the result of removing experts one at a time and re-computing the DM (does not apply for equal weights)

revised

Robustness on Calibr vbls for Optimized PW DM					
Item removed	Mean rel. Info, ALL	Mean rel. Info, Calibr vbls	P-Value	Rel.Info wrt Orig. DM, All	Rel.Info wrt Orig. DM, calibr vbls
WY11	0.8555	0.8555	0.659	0	0
RG11	0.7995	0.7995	0.659	0	0
RS11	0.8797	0.8797	0.659	0	0
GS11	0.8563	0.8563	0.659	0	0
SMBa11	0.7905	0.7905	0.659	0	0
WBy11	0.8667	0.8667	0.3992	0	0
WBa11	0.8941	0.8941	0.659	0	0
YPy11	0.783	0.783	0.659	0	0
YPa11	0.8706	0.8706	0.659	0	0
RS_WYy11	0.8586	0.8586	0.659	0	0
RS_WYa11	0.8584	0.8584	0.659	0	0
RG_WYy11	0.8857	0.8857	0.659	0	0
RG_SMBa11	0.8811	0.8811	0.659	0	0
RG_YPa1	0.8741	0.8741	0.659	0	0
pool12	0.8894	0.8894	0.659	0	0
None	0.8562	0.8562	0.661		

Out of sample Validation

Variations on the Remove-One-At a -Time (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[1]. 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.

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

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[2]. The ratio 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 2. 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.

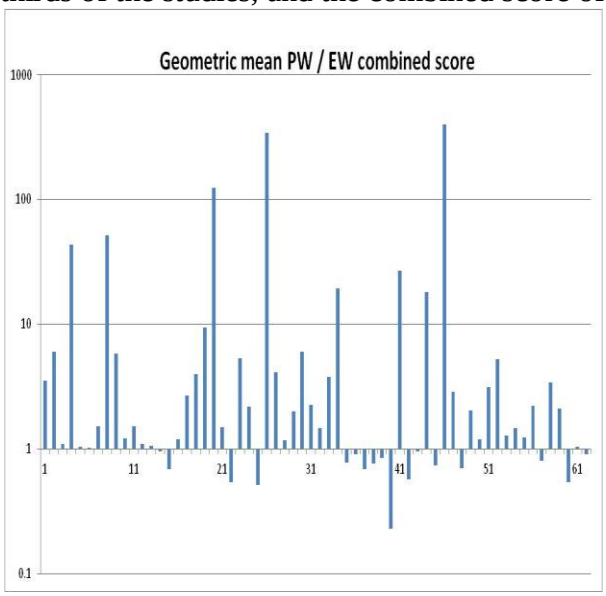

Figure 2: 62 studies, per study: geomeans of comparisons of PW/EW combined score ratios.

[2] 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 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.

Figure 3 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 3 was already presented in Figure 2. 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. Figure 3 highlights the differences between geometric versus arithmetic aggregation, but the superiority of PW over EW is evident from either perspective.

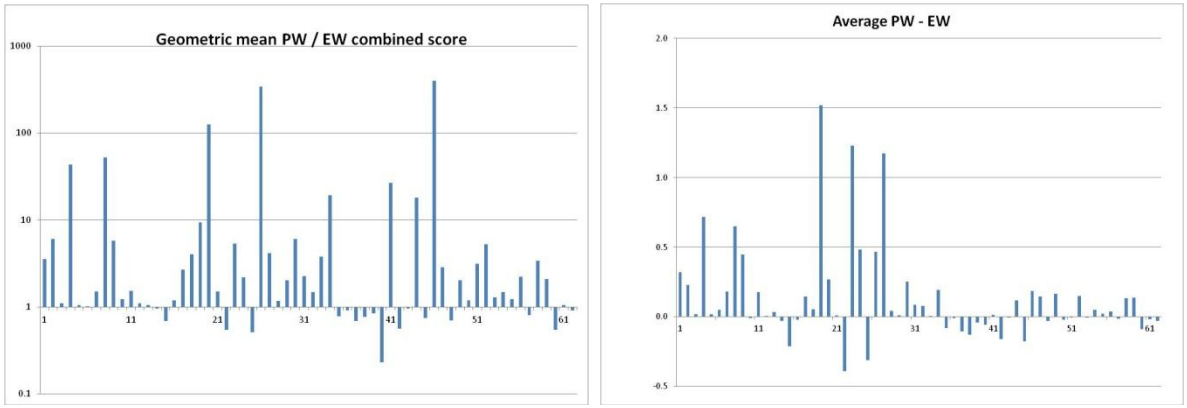

Figure 5: Geomean (left) of PW and EW score ratios and arithmetic mean (right) of PW and EW score differences, for each of 62 studies analysed in Eggstaff et al (2013). If a study had N seed variables, the PW and EW scores were averaged over training sets of size K, K = 1 … N-1 and aggregated with either geo- or arithmetic means to determine an out-of-sample performance indicator per study .

The 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 here.