INSEAD

The Business School
for the World®

Faculty & Research
Working Paper

Combining Interval Forecasts

Anil GABA
Ilia TSETLIN
Robert L. WINKLER

2014/58/DSC

Combining Interval Forecasts

Anil Gaba*

Ilia Tsetlin**

Robert L. Winkler***

October 1, 2014

* The Orpar Chaired Professor of Risk Management, Academic Director, Center for Decision Making and Risk Analysis, Professor of Decision Sciences at INSEAD, 1 Ayer Rajah Avenue Singapore 138676. Email: anil.gaba@insead.edu

** Associate Professor of Decision Sciences at INSEAD, 1 Ayer Rajah Avenue Singapore 138676. Email: ilia.tsetlin@insead.edu

*** James B Duke Professor of Business Administration, Professor of Statistical Science at Fuqua School of Business, Duke University, Durham, NC 27708-0120 USA.
Email: rwinkler@duke.edu

Abstract

In combining forecasts, a simple average of the forecasts performs well, often better than more sophisticated methods. In a prescriptive spirit, we consider some other parsimonious, easy-to-use heuristics for combining interval forecasts and compare their performance with the benchmark provided by the simple average, using real-life data sets consisting of forecasts made by professionals in their domain of expertise. The relative performance of the heuristics is influenced by the degree of overconfidence in the experts' intervals. With a moderate to high degree of overconfidence, two of the heuristics outperform the simple average, with the best creating wider combined intervals and locating the intervals better in terms of the accuracy of the midpoints of the intervals. If there is not much overconfidence, the median and simple average perform best. The results provide some easy-to-use alternatives to the simple average, with an indication of when each might be preferable.

Key words: Interval Forecasts; Combining Expert Forecasts; Heuristics; Overconfidence.

1. Introduction

Decisions are usually made in the face of some uncertainty, and subjective assessments of uncertainty about unknown quantities are needed, as well-defined, easy-to-model data-generating processes often do not exist in real life settings. Even if models are used, subjective judgments are frequently needed in the modeling process and also to adjust the output from the models to account for unreliability of some modeling assumptions. Subjective assessments of uncertainty are obtained from someone with expertise concerning the unknown quantities of interest and are often expressed in terms of interval forecasts, which are easily understood by both the experts and the decision makers. For example, a financial analyst might provide a 90% interval forecast for the price of oil three months or twelve months ahead. This interval, also called a 90% forecast interval or predictive interval, implies that the analyst assigns a 0.90 probability that the interval will include the realized value.

Suppose that interval forecasts for an unknown quantity are obtained from more than one expert in order to obtain more information. How should we aggregate such forecasts? Is there any benefit in combining the interval forecasts? What are some reasonable measures for evaluating intervals from individual experts and combined intervals? These are the types of questions that we explore in this paper.

There is an extensive literature on combining probability distributions; for reviews, see Genest and Zidek (1986), Cooke (1991), Clemen and Winkler (2007), and Ranjan and Gneiting (2010). Clemen (1989) and Armstrong (2001) review the broader area of combining forecasts, which was later popularized by Surowiecki (2004), who coined the phrase "the wisdom of crowds." Two overriding messages from this literature are that combining forecasts can be beneficial and that simple combining methods are more robust, easier to use, and often perform better than more complex methods.

Suppose that we obtain 90% intervals for an unknown quantity from *k* different experts. In combining these intervals, there are several issues that we might take into consideration. One important issue is the possibility of overconfidence in the individual assessments, which leads to intervals being overly narrow in retrospect (Soll and Klayman 2004). Another issue is the tendency of experts to have positively dependent forecast errors, which greatly reduces the gains from increasing the number of forecasts being combined (Clemen and Winkler 1985). We could work with a detailed model that attempts to take account of such issues. However, such models can require the estimation of many parameters and can be quite sensitive to these estimates as well as to various underlying assumptions. Instead, consistent with the above-noted implications of previous work on combining forecasts, we consider some parsimonious, easy-to-use heuristics for combining interval forecasts.

Two main desiderata of interest in interval forecasts, as with other forecasts involving probability, are calibration and sharpness. For example, we would like a series of 90% interval forecasts to be well-calibrated in the sense that they capture the realized value close to 90% of the time. Intervals that show overconfidence, for example, are not well-calibrated. In addition to calibration, the sharpness of the interval forecasts is very important. If two experts are similar in terms of calibration but one systematically provides narrower (i.e., sharper) intervals, then that expert's intervals are more informative and therefore more valuable.

To evaluate the performance of the heuristics, we use a scoring rule that trades off calibration and sharpness. In addition to the overall measure of performance given by this scoring rule, we also consider some partial measures of performance: the relative frequency (the rate at which the intervals capture the realized value, which provides information about calibration), the mean absolute error of the midpoints (a measure of how "well-located" the interval is), and the average width of the combined intervals (a measure of sharpness).

To study the heuristics, we use two data sets of real-life forecasts. One data set is from a study at a major brokerage and investment firm where analysts provided 90% interval forecasts for various quantities in the financial markets. The other data set is based on forecasts of GDP growth and inflation as part of the Federal Reserve Bank of Philadelphia's Survey of Professional Forecasters. These two data sets are interesting in two respects. First, the forecasts were all provided by professionals in their domain of expertise. Second, these two data sets together provide an attractive contrast in terms of degrees of overconfidence in the individual forecasts, ranging from a very high degree of overconfidence exhibited by the analysts at the brokerage firm to a milder but still relevant degree of overconfidence in the SPF forecasters' GDP growth forecasts to virtually no overconfidence in their inflation forecasts.

Our results show that, as expected, consulting multiple experts and combining their interval forecasts leads to improved forecasts and the gains generally increase at a decreasing rate with each additional forecast. Most of the benefit from aggregation accrues from the first five to ten forecasts. The relative performance of our heuristics depends on the degree of overconfidence in the individual forecasts. If there is overconfidence in individual interval forecasts, as in the cases of the analysts and the GDP growth forecasts, the best overall performance is shown by a heuristic that combines probability averaging of endpoints with simple averaging of midpoints, followed by a heuristic with only the probability averaging of endpoints. On the other hand, if there is no overconfidence in the individual forecasts, as with the inflation forecasts, the order of relative performance is reversed; a heuristic that takes medians of the endpoints performs the best, followed by the simple averaging of endpoints.

The capture rate of the combined intervals, their average width, and the mean absolute error of their midpoints shed further light on the nature of the heuristics and their relative performance in the different cases. We show how these measures interact with each other and relate to the overall performance of the heuristics on the scoring rules in addition to how they relate to the degree of overconfidence in the individual intervals, which is an important pointer to which heuristics should be more promising to use.

Our analysis, which involves forecasts made by professionals in their domains of expertise, has important prescriptive implications for combining interval forecasts. We show that considerable improvements over the individual forecasts from the experts can be obtained from some simple, easy-to-use combining heuristics and identify factors that favor particular heuristics. The simple average has a long track record of excellent performance in combining forecasts, and we show that with one exception, the other heuristics can hold their own against the benchmark of the simple average. Since interval forecasts can be understood easily by decision makers, can be made more easily and quickly than forecasts of complete probability distributions, and can be aggregated more easily and quickly with these heuristics than through the development of detailed models, such heuristics deserve further study. This could lead to more frequent consideration of uncertainty and the wisdom of multiple experts in important real-world situations.

In §2, we formally define our heuristics and the measures we use to evaluate their performance. We describe our data sets and study the performance of the individual intervals and the heuristics in §3, followed by a summary and discussion in §4.

2. Heuristics and Evaluation Measures

There are many different ways that interval forecasts can be combined. We consider five heuristics that are very easy to apply and seem reasonable based on the past research on combining of probability distributions. Those heuristics are presented in §2.1, followed in §2.2 by measures that we use to summarize and evaluate the individual interval forecasts and the combined interval forecasts created by the heuristics.

2.1 Heuristics for Combining Interval Forecasts

Suppose that we have k $100(1-\alpha)\%$ forecast intervals $[L_i, U_i]$, $i = 1,\ldots,k,$ for a random variable $\tilde{x},$ and let $[L_1^*, U_1^*],\ldots,[L_5^*, U_5^*]$ denote the $100(1-\alpha)\%$ combined forecast intervals for $\tilde{x}$ obtained from the k individual intervals via the 5 heuristics considered in this study. In this subsection we describe those five heuristics.

H1. Simple Averaging: $L_1^* = \dfrac{1}{k}\sum_{i=1}^{k} L_i,\ U_1^* = \dfrac{1}{k}\sum_{i=1}^{k} U_i.$ This heuristic just takes a simple average of the endpoints of the intervals. In combining point forecasts or probability forecasts, just as in summarizing data, simple averages are often used, not only because of their simplicity but also because of their good performance and robustness in many real-life settings. If we assume that the individual interval forecasts are symmetric in terms of probability, so that $F_i(L_i) = 1 - F_i(U_i) = \alpha/2,\ i = 1,\ldots,k,$ where F_i represents forecaster *i*'s cdf for $\tilde{x}$, then H1 corresponds to averaging quantiles, which has been shown to perform well (Lichtendahl et al. 2013).

H2. Median: $L_2^* = Median(L_1,\ldots,L_k)$, $U_2^* = Median(U_1,\ldots,U_k)$. The median is another measure commonly used in summarizing data, and it is less sensitive to extreme values than the mean. Hora et al. (2013) study the combination of probability distributions via the median cdf and show that it has many desirable properties.

H3. Enveloping: $L_3^* = Min(L_1,\ldots,L_k)$, $U_3^* = Max(U_1,\ldots,U_k)$. This heuristic is consistent with the idea that each expert's window on the world is only a partial view and that the aggregate view should envelop all of these views so that no forecasts, however extreme, are discarded or discounted. Unless the *k* individual intervals are identical, enveloping will yield a wider interval than the other heuristics. Also, by yielding wide intervals, it is one way to overcome overconfidence that might be exhibited by the individual forecasts.

H4. Probability Averaging: L_4^* and U_4^* satisfy $\frac{1}{k}\sum_{i=1}^{k} F_i(L_4^*) = \alpha/2$ and $\frac{1}{k}\sum_{i=1}^{k} F_i(U_4^*) = 1-\alpha/2$, where the individual interval forecasts are assumed to be based on normal cdfs with $F_i(L_i) = 1 - F_i(U_i)$ $= \alpha/2$, $i = 1,\ldots,k$. In words, L_4^* is the value at which the average of the cumulative probabilities $F_1,\ldots,F_k$ is $\alpha/2$, and U_4^* is the value at which the average of the cumulative probabilities is $1-(\alpha/2)$. As with H1, H4 involves simple averaging, but this averaging is of probabilities instead of quantiles. This heuristic is also in the spirit of yielding a wider interval but in a more tempered manner than H3, which simply considers the most extreme endpoints.

H5. Simple Averaging of Midpoints and Probability Averaging of Endpoints: The midpoint of the combined interval from H5 is $M_5^* = \frac{1}{k}\sum_{i=1}^{k} M_i$, where $M_i = (L_i + U_i)/2$ is the midpoint of the i^{th} interval. The width of the combined interval is $W_5^* = W_4^* = U_4^* - L_4^*$. H5 combines aspects of H1 and H4. The midpoint, M_5^*, is the same as the midpoint M_1^* of the H1 interval, and the width is the same as the width of the H4 interval. Thus, $L_5^* = M_1^* - 0.5W_4^*$ and $U_5^* = M_1^* + 0.5W_4^*$. The H5 interval can be viewed as taking the H1 interval and adjusting its width (borrowing the width of the H4 interval), or as taking the H4 interval and shifting it up or down so that its midpoint is the same as that of the H1 interval.

2.2 Measures for Summarizing and Evaluating Interval Forecasts

Before the value of $\tilde{x}$ is known, an interval forecast for $\tilde{x}$ (including a combined interval forecast) can be summarized by statistics such as its midpoint and width. If *k* intervals are to be combined, summary measures such as the mean and standard deviation of the *k* midpoints and of the *k* widths can tell us something about the k intervals. The midpoint and width of a combined interval forecast can be compared to the summary statistics of the set of intervals being combined, and we can compare the summary measures for interval forecasts from the five heuristics in §2.1.

Of course, the key question of interest for a forecast is how well it performs in view of the realized value x of $\tilde{x}$. For measuring this kind of performance, a value for just one or a few forecasts is not very helpful. Instead, we compute average values of the measures over a series of forecasts. These averages can be taken over a series of forecasts for different quantities (e.g., high temperatures on different days) from an individual forecaster or a particular combining heuristic. Alternatively, they can be taken over a series of forecasts given by different forecasters or different heuristics for the same quantity (e.g., high temperature on a given day) or over both different quantities and different forecasters. In evaluating the individual forecasters and the five heuristics, we focus primarily on the following measures.

Average *Q*-score: A $100(1-\alpha)\%$ interval forecast $[L,U]$ can be evaluated by the *Q*-score, a strictly proper quantile scoring rule (Jose and Winkler 2009):

$$Q(L,U,x) = 2g - (\alpha/2)(U-L) - (L-x)^{+} - (x-U)^{+},$$

where $w^{+} = \max\{w, 0\}$, a higher score is better, and g is a constant that can be chosen to scale the score as desired. Without loss of generality, we set $g = 0$, in which case the score is always negative and a less negative score is better. In our empirical studies, we work with 90% interval forecasts, which correspond to $\alpha = 0.10$.

The $-(\alpha/2)(U-L)$ term represents a penalty associated with the width of the interval, and the last two terms represent penalties imposed if x falls below or above the interval. The tradeoff is that a wider interval is given a higher penalty for width but avoids or reduces the penalty for not “capturing” the realized value in the interval. For a series of interval forecasts for different quantities, forecasters, or heuristics, we use the average *Q*-score, $\bar{Q}$, as an overall measure of the accuracy of the interval forecasts, taking into account both sharpness and calibration. Good interval forecasts are sharp in the sense of having narrow intervals and well-calibrated in the sense of having a “capture rate” close to the $100(1-\alpha)\%$ indicated by the forecast.

Relative Frequency: Over a series of interval forecasts, the relative frequency (*RF*) of times the interval captures the realized value is an informative statistic. With the 90% interval forecasts we consider, the forecasts are perfectly calibrated if $RF = 0.90$, or 90%. $RF < 90\%$ indicates overconfidence: intervals too narrow to capture the expected percentage of realized values in the intervals. Similarly, $RF > 90\%$ indicates underconfidence.

Average Interval Width: It can be helpful to compare the average interval width $\bar{W}$ over a series of interval forecasts. This measure indicates the degree of uncertainty about a variable being forecast that is implied by the intervals and can help explain any overconfidence or underconfidence revealed by *RF*.

Mean Absolute Error: This refers to the absolute error of the midpoint of the interval as a point forecast, the difference between the midpoint and the realized value of $\tilde{x}$: $AE = |M - x|$. The mean absolute error (*MAE*) is the average absolute error over a series of forecasts. This is a common measure of accuracy for point forecasts, and it indicates how "well-located" the intervals are, thereby providing different information than $\bar{W}$.

In addition to these measures, we consider some other measures as well. For example, the number of realized values below lower bounds and the number above upper bounds can tell us whether the values not captured by the intervals are divided roughly equally between being above and being below the intervals, or whether there is a strong asymmetry (e.g., a tendency to systematically underforecast or overforecast $\tilde{x}$).

3. Empirical Studies

We analyzed data from two data sets of forecasts. The first data set is from a study we designed and conducted with analysts at a brokerage and investment group. The analysts made interval forecasts involving stock exchange indices and oil and gold prices. The second data set is based on experts' probability forecasts of GDP growth and inflation for the Survey of Professional Forecasters (SPF). In both data sets, the forecasters were experts who followed the variables of interest on a regular basis. Relevant details concerning these data sets and summary statistics related to the analysts' and SPF experts' forecasts are presented in §3.1, and our analysis of the performance of the heuristics is presented and discussed in §3.2.

3.1 Data Description and Summary Statistics

Data from Analysts

Fifty-nine analysts at a major international brokerage and investment group, CLSA, participated in a study in which they provided interval forecasts. CLSA is headquartered in Hong Kong, with more than 1,350 professionals located in 15 major Asian cities and in other major financial centers such as London, New York, and Sydney (see https://www.clsa.com/home.php). The median age of the participants, all with university or advanced degrees, was 36 years, and the median years of service with CLSA (not including experience at other financial houses) was five years. The analysts in our sample were based in New York, Tokyo, and Hong Kong. They participated in an online survey where they provided 90% interval forecasts one month, two months, and three months ahead for some of the financial quantities that were of most interest to and were continuously tracked by them: the price of Oil in

US$/barrel from the Brent EUCRBRDT Index (Oil1 to Oil3), the price of Gold in US$/oz from the Bloomberg GOLDS COMDTY Index (Gold1 to Gold3), the Dow Jones Industrial Average Index (DJ1 to DJ3), the Nikkei NKY Index (NK1 to NK3), and the Hang Seng HIS Index (HS1 to HS3).

SPF Data

In addition to the analysts' data, we also analyzed the forecasts of annual GDP growth (1992-2009) and annual inflation (1992-2010) reported by experts surveyed in the first through fourth quarters (Q1-Q4) of each year as part of the Federal Reserve Bank of Philadelphia's Survey of Professional Forecasters. Panelists reported probabilities that percentage changes in U.S. real GDP and U.S. inflation for the full year would fall in each of ten predetermined bins. The experts did not all participate in the survey every year. Our data involved 25-50 experts per year for both GDP and inflation forecasts. The data are part of the data set downloaded from http://www.phil.frb.org/econ/spf and analyzed by Lichtendahl et al. (2013). We used data starting from 1992 to have the same ten bins over the period of our analysis. Since the panelists were not asked directly about interval forecasts, we approximated the SPF panelists' continuous predictive distributions by fitting piecewise-linear cdfs to their reported discrete distributions, as in Lichtendahl et. al. (2013). We then took the 0.05 and 0.95 quantiles of the fitted cdfs to construct 90% prediction intervals.

Summary Statistics

Summary statistics involving the analysts' forecasts are presented in Table 1 for the three lead times (e.g., DJ1 to DJ3) for all five quantities of interest. The overall measure of "goodness" of the forecasts, the average *Q*-score, cannot be compared for different quantities, because it depends on the scaling of and the degree of uncertainty about $\tilde{x}$, but *Q*-scores for different lead times for the same variable are comparable. As anticipated, $\overline{Q}$ tends to be better (less negative) for shorter lead times, with the only violations occurring in going from HS2 to HS3 and Gold2 to Gold3. The *MAE* of the interval midpoints is also better for shorter lead times except for the same violations as $\overline{Q}$.

The average interval widths in Table 1 decrease with shorter lead time for all quantities, which is not surprising. The relative frequencies, however, are not consistent in their shifts as lead time decreases, tending to decrease from lead time 1 to 2 and then increase from 2 to 3. Overall, *RF* is always below 90%, indicating overconfidence, and often exhibits an extremely high degree of overconfidence, getting as low as 10.17% for NK2. For each quantity except Gold, the shortest lead time has the least overconfident intervals, suggesting that the increases in interval widths observed for longer lead times are not sufficient.

Finally, if we confine our attention to those forecasts for which the interval does not capture the realized value, the realized values are predominantly below or predominantly above the intervals, never a roughly even mix between the two sides. The number of forecasts for which *x* is below the interval and the number with *x* above the interval for a given quantity and a given lead time, shown in the last column of Table 1, demonstrates this asymmetric tendency. For example, HS2 and Gold1 both had very low values of *RF*, 13.6% for HS2 and 18.6% for Gold1. HS2 had 51 misses, with *x* below the interval for 49 forecasts and above for 2 forecasts, whereas for Gold1's 48 misses, *x* was below for 1 forecast and above for 47. This indicates high correlations among the forecast errors from the different analysts.

Tables 2 and 3 give the same set of statistics for the SPF forecasts of GDP and inflation, respectively, with year-by-year results shown only for Q1 to save space because the year-by-year results for Q2-Q4 are very similar in nature. Averages across the years are shown for Q1-Q4 in order to look at lead time effects. Note that unlike the analysts' forecasts, with Oil1 indicating the shortest lead time and Oil3 the longest for forecasts of oil prices, Q1 represents the longest lead time and Q4 the shortest for the SPF forecasts. The analysts made all of their forecasts at the same time for values of the quantities one, two, and three months ahead. The SPF forecasts are all for the same full-year values of GDP and inflation, with the longest lead times corresponding to Q1 forecasts, made in the first quarter of the year, and the shortest corresponding to Q4 forecasts.

For inflation, $\bar{Q}$ and *MAE* improve for shorter lead times. For GDP, the same thing is true for *MAE*, but $\bar{Q}$ is slightly worse in going from Q2 to Q3 and Q3 to Q4. Also similar to the results for the analysts, the average width decreases with shorter lead times for both variables.

The biggest difference between the analysts' forecasts and the SPF forecasts involves *RF*. The GDP forecasts exhibit overconfidence (average *RF* near 70% for Q1-Q3, declining to 58.6 for Q4), but it is much less severe than for the analysts (average *RF* of 37.4%). And for inflation, the forecasts are not overconfident, hovering around 90% for all four lead times. The misses for a given quantity tend to cluster on one side of the interval for both GDP and inflation, as for the analysts, although that evidence is weaker for the inflation forecasts because the high *RF* values mean that there are not that many misses in the first place.

Some differences between the characteristics of the analysts' forecasts and the SPF forecasts might be expected. The analysts had little or no experience at making formal probability forecasts or interval forecasts, and there was not much chance to learn from experience. Many of the SPF experts, on the other hand, had previous experience at making probability forecasts through the SPF program itself and perhaps otherwise. Also, as noted earlier, the elicitation methods were different, with the analysts directly providing their 90% intervals and the SPF experts providing probabilities for predetermined bins, from which their 90% intervals were constructed. A similarity that is important in our view is that both the analysts and SPF forecasters can be considered substantive experts with respect to the quantities they were forecasting and may often make some types of forecasts of those quantities, whether formal or not, as part of their normal occupations.

3.2 Performance of the Heuristics

In this section, we show the performance of the different heuristics for combining intervals in the three cases of the analysts, GDP, and inflation. In each of these cases, for each forecasting variable, we formed random subgroups of $k = 2,\ldots,20$ forecasters, with 10,000 simulations for each *k*, and created combined intervals using heuristics H1 to H5. We then computed the four performance measures for the combined intervals: average *Q*-score $(\bar{Q})$, relative frequency (*RF*), average interval width $(\bar{W})$, and mean absolute error of the interval midpoint (*MAE*).

For reasons of space and brevity, we present results aggregated across the five forecasting variables and three time horizons for the analysts and across the different years with Q1 only for GDP and inflation; similar graphs for the Q2-Q4 SPF forecasts look virtually the same. Before combining intervals for the analysts, we rescaled their forecasts in terms of return, where return = (forecast/actual value at the time of forecast) –1, in order to place all the forecast intervals on the same scale. The realized values were similarly rescaled to yield realized returns. No such rescaling was necessary for GDP and inflation, as those variables of interest were the same from year to year.

Figure 1 shows the average *Q*-score for combined intervals from heuristics H1 to H5 as a function of *k* for the analysts, GDP, and inflation. The relative performance of heuristics in terms of *Q*-score is similar for analysts and GDP, but the order is different for inflation. For the analysts and GDP, the *Q*-score increases at a decreasing rate with *k* for all of the heuristics except H3, under which it increases rapidly and is the best *Q*-score among all heuristics for the first few values of *k* (up to *k* = 4 for analysts and *k* = 3 for GDP) but thereafter decreases. Overall, putting aside H3, H5 appears to perform the best, followed by H4, H1, and then H2.

For inflation, H2 seems to perform the best followed by H1, then H4 and H5 which are mostly overlapping, and finally H3 which performs the worst. Also, the gain in $\bar{Q}$ from a higher *k* is minimal under H2 and H1, shows no discernible increase under H4 and H5, and is negative under H3. These differences in the overall relative performance of heuristics between analysts and GDP on one hand and inflation on the other can be attributed to the nature of the heuristics combined with the degree of overconfidence in the underlying interval assessments of the forecasters. The overall average relative frequency for the 90% intervals is 37.4% for analysts, 67.4% for GDP(Q1), and 88.7% for inflation(Q1), implying a high degree of overconfidence for the analysts, a somewhat lesser but still substantial degree of overconfidence with GDP, and no overconfidence to speak of with inflation.

How factors such as this difference in degree of overconfidence translate into the overall relative performance $\bar{Q}$ of the heuristics is better understood by looking at the nature of the heuristics through partial measures of their performance. Figures 2, 3, and 4 show the relative frequency, average width, and mean absolute error of the interval midpoint, respectively, of the combined intervals under the five heuristics as a function of *k* for the analysts, GDP, and inflation. These three partial measures of performance interact with each other, leading to the overall performance measure, the *Q*-score.

The width is directly related to the *Q*-score because, as noted when the *Q*-score was defined in §2.2, it includes a term $-(\alpha/2)(U-L)=-(\alpha/2)W$ that is proportional to *W*. The width also relates to the penalties incurred when *x* is not captured by the interval, represented by the terms $-(L-x)^+$ and $-(x-U)^+$. The wider the interval, the farther *x* has to be from the midpoint before one of these two penalties is activated, and the smaller the penalty is for a given *x* when it is activated. Thus, *W* affects both the penalty for width and the penalty for *x* being outside the interval, with the former penalty increasing and the latter penalty potentially decreasing as *W* increases.

The relative frequency affects the *Q*-score because the higher *RF* is, the less often the penalty for *x* being outside the interval comes into play. The wider the interval, the higher *RF* will be, all other things equal, so a higher *RF* is related to the good and bad implications for the *Q*-score of a larger *W*. For a 90% interval, the optimal balance between these good and bad implications occurs when $RF=90\%$.

The *MAE* is also connected to all of these measures. The *Q*-score for a 90% interval can be expressed as $Q(AE,W)=-0.05W-(AE-0.5W)^+$. Thus, all other things equal, the penalty associated with the second term of $Q(AE,W)$ is more likely to kick in when *AE* is larger. Similarly, a larger *AE*, which indicates that *x* is likely to be more distant from the center of the interval, makes it more likely that the interval will not capture *x*. Of course, higher values of *AE* imply a higher *MAE*, so we can expect a higher *MAE* to be associated with a lower *RF*.

Note in Figure 2 that the relative frequencies for *k* =1 (37.4% for the analysts, 67.4% for Q1 GDP, and 88.7% for Q1 inflation) correspond to the degrees of overconfidence for the three cases, with a lower *RF* implying greater overconfidence. In all three cases, *RF* increases with *k* under all of the heuristics except for H1 and H2. Under H1 for the analysts, *RF* goes up slightly and is then roughly constant with *k*, whereas it declines under H2. For GDP, *RF* is roughly constant for both H1 and H2, declining slightly under H1. *RF* is uniformly highest under H3 compared to the other four heuristics, which is not surprising at all since H3 creates the widest combined interval, encompassing the combined intervals from the other heuristics. H3 moves quickly from overconfidence to underconfidence as *k* increases. After H3, the next highest relative frequencies are for H5 and H4, but note that the *RF* for H5 is slightly higher or equal to that of H4, even though both heuristics create the same interval width. This has to do with better placement (lower *MAE*) of the combined interval under H5, as discussed below. H5 and

H4 both become less overconfident as *k* increases in all three cases, but only for inflation do they reach the level of being underconfident. The relative frequencies of H1 and H2 are the lowest in all cases.

Figure 3 on interval widths is similar for all three cases. H1, which is simple averaging of the lower and upper bounds, must lead to an average width that is constant over *k*. The average widths increase with *k* under all other heuristics except H2, which creates marginally smaller widths after *k* = 2. Further, the average width is highest under H3, followed by H4 and H5, which have the same width. As noted earlier, H3 must create the largest widths as it encompasses combined intervals from all other heuristics. H4 and H5 lead to wider intervals than H1 and H2 but not as wide as H3.

Finally, looking at Figure 4, one can see that *MAE* declines with *k* under all heuristics except for H3, where it actually increases. The *MAE* is best under H1 and H5 (by definition, it is the same under H1 and H5), followed closely by H2, then H4, and the worst is H3. The midpoint of H1and H5 by definition is equivalent to simple averaging of the midpoints of individual intervals, and that improves the *MAE* of the midpoint of the combined interval, consistent with idea of accuracy of point forecasts improving well with simple averaging. H3, on the other hand, is just focused on the extreme lower and upper bounds of the individual intervals in a group. Hence, its midpoint will be quite volatile and prone to high values of *MAE* with increasing *k*. H4 is similar in this respect to H3, but in a more controlled and less volatile way. Hence, it does much better than H3 in terms of *MAE* and worse than H1/H5 and H2 but not too much worse if the individual intervals do not exhibit strong overconfidence.

Combining the observations from Figures 2, 3, and 4 discussed above sheds light on the overall performance of heuristics as shown by Figure 1 on *Q*-scores. H1 can substantially improve the placement of the combined interval by improving the accuracy of the midpoint. However, any meaningful degree of overconfidence in the individual intervals is not corrected for, even with a high *k*. Any increase in *Q*-score with *k* seems to be due simply to the better accuracy of the midpoint, as is the case for the analysts and GDP. On the other hand, if there is no overconfidence in the individual intervals, as in the case of inflation, H1 performs very well, as the combined interval gets better placed as *k* increases.

H2 also increases the accuracy of the midpoint with *k*, but at the same time it creates marginally smaller interval widths than H1, thus performing even worse than H1 in accounting for overconfidence. This is seen in Figure 1, where the *Q*-score under H2 is almost flat over *k*, the worst of all heuristics for analysts and mostly so for GDP. In contrast, the *Q*-score for H2 does the best of the heuristics for inflation, where overconfidence is not an issue, and the marginally narrower intervals make it slightly better than H1.

H3 is an extreme brute-force heuristic that can do well for very small values of *k* when there is substantial overconfidence. For example, with the analysts and GDP, the average relative frequency of the individual intervals (for *k* = 1) is well below 90% for the 90% intervals. As *k* increases, H3's *RF* rather quickly goes to 90% and beyond due to the rapidly increasing width of the combined interval, resulting in H3 creating very underconfident forecasts. At the same time, the *MAE* of H3 increases with *k*. Hence, as *k* increases, the *Q*-score of H3 first does well for the first few values of *k* due to the needed correction for overconfidence but thereafter begins to decline quickly as this correction becomes overdone in terms of excessive width. For inflation, on the other hand, the individual intervals are on average already well calibrated. Thus, H3 creates excessive width in the combined interval even for small values of *k*, which leads to a *Q*-score that declines with *k* from the start and performs by far the worst among all heuristics.

H4 provides some of the benefits of H3 when there is overconfidence, but in a more tempered way, without creating excessive width by just the extreme values and without declining in performance beyond very small *k*. This balancing works reasonably well when there is overconfidence, but when the individual intervals are not overconfident, the increases in width provided by H4 dampen gains in performance. This is consistent with results showing that combining by averaging probabilities, as in H4, results in combined forecasts that move in the direction of less overconfidence or more underconfidence as compared with the individual forecasts (Hora 2004, Ranjan and Gneiting 2010, Lichtendahl et al. 2013). Thus, the combined forecasts are less overconfident than overconfident individual forecasts and are underconfident if the individual forecasts are not overconfident. Finally, the shifts in the endpoints are such that the midpoint for H4 does not perform as well as H1, H2, and H5 in terms of *MAE*, which in turn hurts its *Q*-score.

H5 combines the beneficial characteristics of H1 and H4, whereby it improves the accuracy of the midpoint and at the same time accounts for overconfidence by increasing the width beyond just the average width of the individual intervals but not to the extreme shown by H3. This is reflected by the *Q*-score of H5 with the analysts and GDP, where it performs the best. In terms of the location of H5's intervals, its borrowing of H1's midpoint makes it the leader, along with H1, in terms of *MAE*. It can also be seen in the relative frequencies, which show that H5's forecasts reduce overconfidence more effectively than all of the heuristics except the too-extreme H3. However, if there is not much underlying overconfidence, as in the case of inflation, the higher width compared to that under H1 and H2 makes it underperform H1 and H2. On the whole, though, unless we are reasonably confident that any overconfidence is minimal, H5 might be the safest heuristic to use.

What are the implications of these results for the choice of *k*, the number of experts to consult when obtaining forecasts? Looking at Figure 1, we see that for the analysts and GDP, most of the gains in average *Q*-score for H5 and H4, the best-performing heuristics in those cases, are attained by $k = 5$, with much smaller gains from $k = 5$ to $k = 10$ and beyond. For inflation, with no overconfidence, H2 and H1 perform best, and most of their gains in *Q*-score are realized by $k = 3 - 5$. This suggests that if we are quite unsure about the degree of overconfidence to expect in a given situation, any *k* from 5 to 10 might be a good choice. Others studying the aggregation of forecasts have come to similar conclusions about *k* (Armstrong 2001, Hora 2004, Budescu and Chen 2014, Mannes et al. 2014).

4. Summary and Discussion

The best (or only) source of information regarding some uncertain quantities of interest is often expert judgment, which is frequently elicited in terms of interval forecasts or a few quantiles instead of the more complex assessment of entire probability distributions. Multiple experts are typically consulted and their forecasts are combined. Extensive work on combining forecasts shows that simple combining methods are more robust, easier to use, and often perform better than more complex methods. Moreover, decision makers may be less inclined to use the more complex methods because they tend to involve detailed modeling and the estimation of various model parameters.

Taking all of these things into consideration, we investigated the performance of some parsimonious, easy-to-use heuristics for combining interval forecasts, using real-life data sets consisting of forecasts made by professionals in their domain of expertise. Our results show that, as with point forecasts, consulting multiple experts and aggregating their interval forecasts with these heuristics yields improved forecasts, and we are able to identify factors that favor particular heuristics.

In our empirical studies, the relative performance of the heuristics is influenced by the presence of overconfidence in the individual intervals. If there is a high degree of overconfidence in these intervals, then H3, H4, and H5 do well against the benchmark heuristic H1 involving simple averaging of the endpoints. This happens because H3, H4, and H5 create wider combined intervals, a needed correction for overconfidence. However, if this is overdone, as is the case with H3 beyond very small values of *k*, then the combined interval can be underconfident and the performance of the heuristic suffers. Between H4 and H5, H5 has a slight advantage. In addition to creating a needed adjustment of the width of the combined interval, H5 locates the interval better in the sense of its midpoint being a more accurate point forecast, essentially combining the merits of H4 with those of H1. This is what we observe with the analysts and GDP, for which the individual intervals are clearly overconfident. On the other hand, if there is not much overconfidence in the individual intervals, as with inflation, then H1 and H2, which do not create wider combined intervals and at the same time have more accurate midpoints, perform better.

When we are dealing with decisions in real time, of course, we want to combine individual forecasts before we see the realizations. Can data on interval forecasts from the same set of experts on past realizations of the same quantities be helpful? Absent such data, how can we predict the calibration (particularly the degree of overconfidence) of the individual intervals?

Correcting for overconfidence can be tricky. The degree of overconfidence can vary depending on the expert, the quantity, and the way the judgments are assessed (Klayman et al. 1999). In our study, the degree of overconfidence exhibited by the interval forecasts for individual experts varied from situation to situation, ranging from a very high degree of overconfidence to no overconfidence. Even within the same domain, the SPF forecasters were overconfident for GDP growth but not for inflation. Thus, previous performance from similar forecasts or from "test questions" designed to help analysts calibrate experts might not be reliable.

Some studies have found that assessment procedures other than direct assessment can reduce overconfidence in interval forecasts. Soll and Klayman (2004) found that overconfidence could be reduced by assessing the two endpoints separately (e.g., asking for the .05 quantile and then the .95 quantile instead of asking for the 90% interval), and asking in addition for the median provides further reduction. Budescu and Du (2007) expanded on this approach, assessing multiple quantiles, fitting a full predictive distribution to those quantiles, and determining an interval forecast from the distribution. Haran et al. (2010) assessed probabilities for preselected bins chosen to cover the entire range of feasible outcomes and determined an interval forecast from the resulting probability distribution. Jain, et al. (2013) took a different approach, that of unpacking the time horizon; before considering forecasts for the horizon of primary interest, the assessors were asked to make forecasts for the same quantity at intermediate time horizons. All of the modifications in these studies led to reduced overconfidence.

None of the above studies looked into implications for combining the individual interval forecasts, but our empirical studies relate to their assessment procedures. The analysts provided interval forecasts directly, which has been shown to lead to greater overconfidence than other procedures. However, they also worked with unpacked time horizons, which reduced overconfidence in Jain et al. (2013). The SPF forecasters, on the other hand, assessed probabilities for preselected bins as in Haran et al. (2010). Predictive distributions were then fit to these probabilities, just as Budescu and Du (2007) fit distributions to assessed quantiles, and interval forecasts were determined from these predictive distributions. Thus, we might expect the way the SPF forecasts were assessed and combined to lead to lower overconfidence than exhibited by the analysts, which is what happened.

Current views on overconfidence as a cognitive bias run the gamut. Glaser, et al. (2013, p. 405), after stating that "Overconfidence is often regarded as one of the most prevalent judgment biases," note that there is a debate about whether overconfidence might be "an ecological and statistical illusion that just seems to exist but is not real." They quote O'Hagan et al. (2006, p. 82): "Juslin et al. (2000) argue vehemently that, when all potential artifacts associated with calibration tasks and their analysis are taken into account, there is little evidence of any meaningful cognitive overconfidence bias." However, interval forecasts that capture fewer realizations than expected, often by a considerable margin, represent a phenomenon that is extremely common, whether it is caused by a cognitive bias or not. Overall, as mentioned earlier, unless we are fairly confident that any overconfidence is minimal, using H5 with 5 to 10 experts to combine interval forecasts might be a reasonable approach.

Another issue noted in §1 is the tendency of experts to have positively dependent forecast errors. This tendency is indicated in our empirical studies by the last columns of Tables 1-3, as noted in §3.1. The primary impact of such dependence is to reduce the potential improvement in performance from additional experts. If the experts in our empirical studies had been less dependent, the gains from all of our heuristics could have been even greater. More diversity can help in reducing dependence, where lack of diversity is reflected by experts who have similar training, read the same literature, subscribe to the same theories, use the same models, have access to the same data sets, and so on.

We have tried to use simple heuristics involving few or no assumptions. H1, H2, and H3 are best in this sense, requiring no assumptions. H4 and H5 use a normal distribution assumption to determine the endpoints of the combined interval forecast, but that doesn't require any assumptions about overconfidence or other cognitive factors. What other relatively simple heuristics could be considered? Overconfidence is associated with narrower intervals, so we might try to correct for overconfidence by simply increasing the width of all of the experts' intervals by a fixed factor, leaving the midpoints unchanged. However, this approach is rather ad hoc, and the multiplying factor would be hard to select because overconfidence has been shown to vary not only across individuals, but across assessment procedures, quantities being assessed, and lead time for a given individual, as noted above.

What about weighted averages, which are used quite often? The idea of giving "better experts" higher weight is intuitively appealing, but it is not as simple as the heuristics considered here because it requires estimation of the weights, whether subjectively or via models or data on past performance. Rowe (1992, p. 161) notes that "A considerable number of studies have examined the relative worth of various weighting schemes, and have generally found there to be little advantage (if any) in using differential over equal weighting."

Trimmed averages are a special case of weighted averages that have received some attention to combine forecasts (Yaniv 1997, Jose et al. 2014). The estimation of weights amounts to deciding which forecasts to trim (omit), giving them a weight of zero while giving the remaining forecasts equal weights. The usual trimming of extreme values can be advantageous with underconfident forecasts, while overconfident forecasts call for trimming some of the least extreme values (Jose et al. 2014). Thus, choices between these two types of trimming (exterior and interior trimming), including how much to trim, make this approach comparable in difficulty to choosing the multiplicative factor when multiplying the interval widths by a constant.

Simple averages are widely used in combining forecasts. They offer simplicity, good performance without making any restrictive assumptions, and robustness. We used the simple average and four other easy-to-use heuristics also satisfying the principal of parsimony to combine interval forecasts. With no overconfidence in the individual forecasts, the simple average performed well, as expected, as did the median. When there was overconfidence in the individual intervals, two of the other heuristics stacked up quite well against the benchmark provided by the simple average. This provides practitioners a few convenient combining methods as possible alternatives to the simple average, with an indication of when each might be preferable.

Acknowledgments

We are extremely grateful to CLSA for arranging for their analysts to participate in our study, and to Casey Lichtendahl and Yael Grushka-Cockayne for sharing their "cleaned" data set of forecasts from the Federal Reserve Bank of Philadelphia's Survey of Professional Forecasters. We are also extremely grateful to Zhi Chen for his excellent assistance in the data analysis. The Centre for Decision Making and Risk Analysis (CDMRA) at INSEAD provided financial support for this project.

References

Armstrong, JS (2001) Combining forecasts. Armstrong JS, ed. *Principles of Forecasting: A Handbook for Researchers and Practitioners* (Kluwer Academic, Norwell, MA), 417-439.

Budescu D, Chen E (2014) Identifying expertise to extract the wisdom of crowds. *Management Sci.* Forthcoming.

Budescu DV, Du, N (2007) Coherence and consistency of investors' probability judgments. *Management Sci.* 52(11):1731-1744.

Clemen, RT (1989) Combining forecasts: A review and annotated bibliography. *Internat. J. Forecasting* 5(4):559-583.

Clemen RT, Winkler RL (1985) Limits for the precision and value of information from dependent sources. *Oper. Res.* 33(2):427-442.

Clemen RT, Winkler RL (2007) Aggregating probability distributions. Edwards W, Miles RF, von Winterfeldt D, eds. *Advances in Decision Analysis* (Cambridge University Press, Cambridge, UK), 154-176.

Cooke RM (1991) *Experts in Uncertainty: Opinion and Subjective Probability in Science* (Oxford University Press, Oxford, UK).

Genest C, Zidek JV (1986) Combining probability distributions: A critique and an annotated bibliography. *Statist. Sci.* 1(1):114-135.

Glaser M, Langer T, Weber M (2013) True overconfidence in interval estimates: Evidence based on a new measure of miscalibration. *J. Behavioral Decision Making* 26(5):405-417.

Haran U, Moore, DA, Morewedge CK (2010) A simple remedy for overprecision in judgment. *Judgment and Decision Making* 5(7):467-476.

Hora, SC (2004) Probability judgments for continuous quantities: Linear combinations and calibration.

Management Sci. 50(5):597-604.

Hora SC, Fransen BR, Hawkins N, Susel I (2013) Median aggregation of distribution functions. *Decision Anal.* 10(4):279-291.

Jain K, Mukherjee K, Bearden JN, Gaba A (2013) Unpacking the future: A nudge toward wider subjective confidence intervals. *Management Sci.* 59(9):1970-1987.

Jose VRR, Grushka-Cockayne Y, Lichtendahl KC Jr (2014) Trimmed opinion pools and the crowd's calibration problem. *Management Sci.* 60(2):463-475.

Jose VRR, Winkler RL (2009) Evaluating quantile assessments. *Oper. Res.* 57(5):1287-1297.

Juslin P, Winman A, Olsson H. (2000) Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psych. Review* 107(2):384-396.

Klayman J, Soll JB, González-Vallejo C, Barlas S (1999) Overconfidence: it depends on how, what, and whom you ask. *Organ. Behav. Human Decision Processes* 79(3):216-247.

Lichtendahl KC Jr, Grushka-Cockayne Y, Winkler RL (2013) Is it better to average probabilities or quantiles? *Management Sci.* 59(7):594-1611.

Mannes, AE, Soll, JB, Larrick, RP (2014) The wisdom of select crowds. *J. Personality Social Psych.* Forthcoming.

O'Hagan AO, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T (2006) *Uncertain Judgments: Eliciting Experts' Probabilities* (Wiley, Chichester, UK).

Ranjan R, Gneiting T (2010) Combining probability forecasts. *J. Roy. Statist. Soc. B* 72(1):71-91.

Rowe G (1992) Perspectives on expertise in the aggregation of judgments. Wright G, Bolger F, eds. *Expertise and Decision Support* (Plenum, New York).

Soll JB, Klayman J (2004) Overconfidence in interval estimates. *J. Experimental Psychology: Learning, Memory, and Cognition* 30(2):299-314.

Surowiecki J (2004) *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations* (Doubleday, New York).

Yaniv I (1997) Weighting and trimming: Heuristics for aggregating judgments under uncertainty. *Organ. Behav. Human Decision Processes* 69(3):237-249.

Quantity	Average *Q*-Score	Average Width	Relative Frequency	MAE of Midpoints	Misses < L , Misses > H
DJ1	-127.62	1155.15	59.32%	349.69	0, 24
DJ2	-498.55	1539.96	33.90%	937.53	38, 1
DJ3	-757.25	1847.86	23.73%	1296.32	44, 1
HS1	-292.51	2147.97	59.32%	709.49	22, 2
HS2	-1282.25	2725.85	13.56%	2229.63	49, 2
HS3	-919.17	3275.47	38.98%	1820.81	32, 4
NK1	-107.10	1296.08	86.44%	249.87	7, 1
NK2	-991.37	1739.39	10.17%	1622.05	52, 1
NK3	-1464.31	2115.61	13.56%	2271.71	51, 0
Gold1	-42.38	125.67	18.64%	84.31	1, 47
Gold2	-56.89	175.48	23.73%	110.61	1, 44
Gold3	-36.57	206.31	67.80%	91.49	4, 15
Oil1	-1.39	12.19	57.63%	4.00	0, 25
Oil2	-4.96	16.69	25.42%	9.88	44, 0
Oil3	-7.52	20.53	28.81%	13.73	42, 0
Average			37.40%		

Table 1. Summary Statistics for Analysts.

Year	Number of Forecasters	Average *Q*-Score	Average Width	Relative Frequency	MAE of Midpoints	Misses < L , Misses > H
1992	36	-0.5371	3.25	38.89%	1.61	0, 22
1993	31	-0.1525	2.83	96.77%	0.39	0, 1
1994	27	-0.1860	2.93	81.48%	0.85	0, 5
1995	25	-0.2029	2.72	92.00%	0.57	2, 0
1996	35	-0.8216	3.03	22.86%	2.08	0, 27
1997	33	-0.7670	2.91	18.18%	2.03	0, 27
1998	29	-0.7432	2.62	13.79%	1.83	0, 25
1999	30	-0.6788	3.43	26.67%	2.19	0, 22
2000	33	-0.1670	3.03	90.91%	0.73	0, 3
2001	30	-0.2335	3.47	80.00%	0.81	6, 0
2002	30	-0.2199	3.19	90.00%	0.76	1, 2
2003	33	-0.1755	3.49	96.97%	0.531	0, 1
2004	27	-0.2388	3.85	85.19%	0.86	4, 0
2005	32	-0.1914	2.84	90.63%	0.48	3, 0
2006	49	-0.1823	2.99	91.84%	0.58	4, 0
2007	46	-0.1667	2.88	84.78%	0.70	7, 0
2008	43	-0.6161	3.15	37.21%	1.75	27, 0
2009	39	-0.5509	3.52	74.36%	1.94	10, 0
Q1 Avg.	33.8	-0.3795	3.12	67.36%	1.15	
Q2 Avg.	36.1	-0.2573	2.86	74.64%	0.90	
Q3 Avg.	34.9	-0.2580	2.43	70.30%	0.83	
Q4 Avg.	34.8	-0.3030	1.79	58.59%	0.77	
Average	34.9	-0.2994	2.55	67.72%	0.91	

Table 2. Summary Statistics for GDP: Year by Year for Q1 and Overall Averages for Q1-Q4.

Year for Q1	**Number of Forecasters**	**Average *Q*-Score**	**Average Width**	**Relative Frequency**	**MAE of Midpoints**	**Misses < L , Misses > H**
1992	36	-0.1454	2.77	94.44%	0.68	2, 0
1993	30	-0.1937	2.58	86.67%	0.82	4, 0
1994	26	-0.1259	2.52	100.00%	0.58	0, 0
1995	26	-0.2414	2.38	80.77%	0.83	5, 0
1996	33	-0.1438	2.62	90.91%	0.57	3, 0
1997	33	-0.1692	2.56	81.82%	0.71	6, 0
1998	29	-0.2770	2.27	72.41%	0.96	8, 0
1999	30	-0.1421	2.85	100.00%	0.43	0, 0
2000	33	-0.1368	2.53	93.94%	0.42	0, 2
2001	30	-0.1396	2.62	96.67%	0.46	0, 1
2002	30	-0.1534	2.70	96.67%	0.50	1, 0
2003	32	-0.1490	2.82	93.75%	0.45	0, 2
2004	26	-0.2198	2.96	80.77%	1.07	0, 5
2005	32	-0.3021	2.50	59.38%	0.96	0, 13
2006	50	-0.1597	2.81	92.00%	0.59	0, 4
2007	46	-0.1391	2.58	84.78%	0.58	0, 7
2008	45	-0.1860	2.95	91.11%	0.61	3, 1
2009	39	-0.1676	3.07	92.31%	0.55	1, 2
2010	38	-0.1493	2.81	97.37%	0.56	1, 0
Q1 Avg.	33.9	-0.1758	2.68	88.72%	0.65	
Q2 Avg.	36.5	-0.1552	2.55	91.10%	0.58	
Q3 Avg.	34.1	-0.1356	2.27	92.15%	0.46	
Q4 Avg.	35.1	-0.1228	1.81	90.34%	0.39	
Average	34.9	-0.1474	2.33	90.58%	0.52	

Table 3. Summary Statistics for Inflation: Year by Year for Q1 and Overall Averages for Q1-Q4.

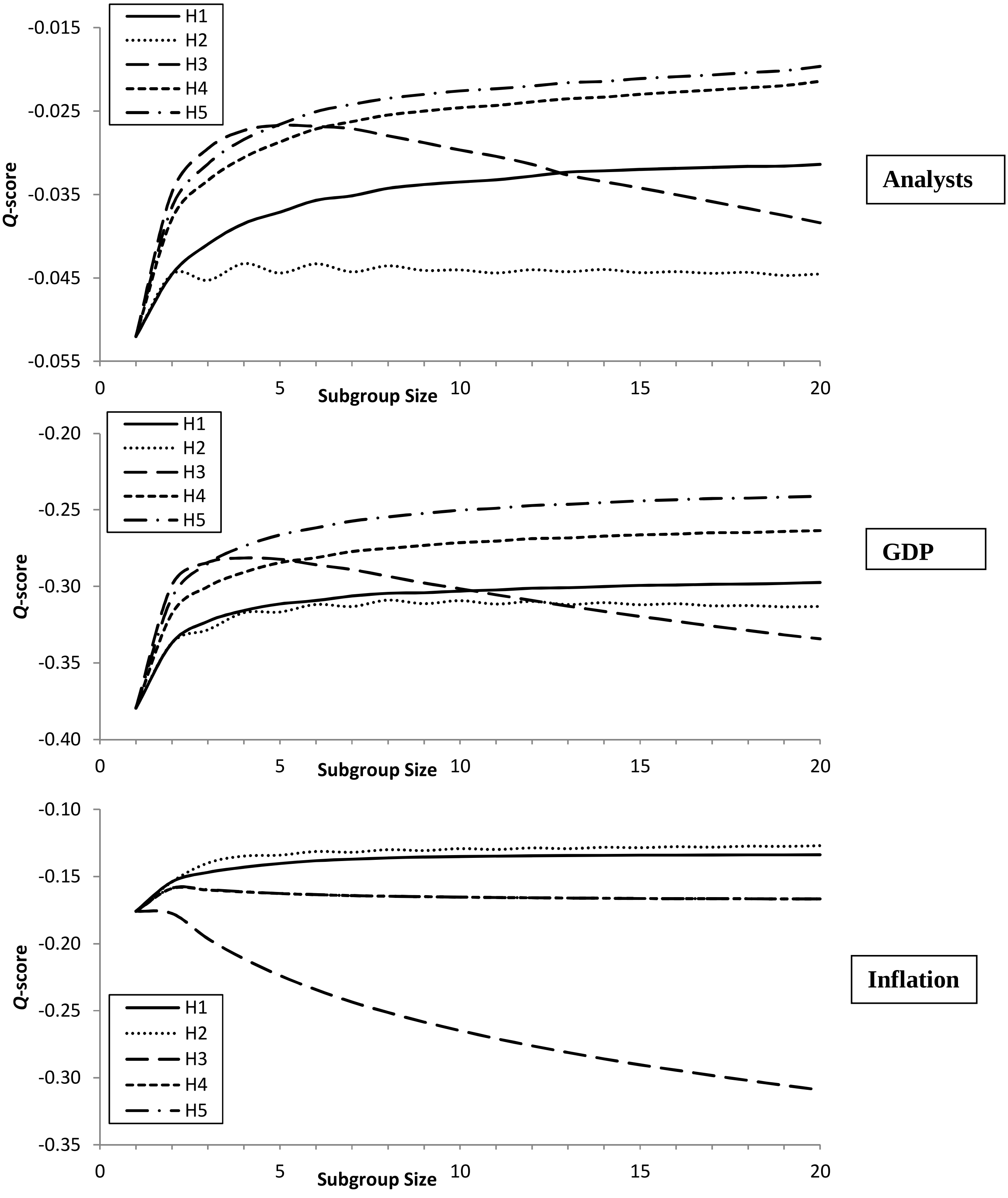

Figure 1. Average *Q* Scores of Combined Intervals under Different Heuristics (H1 to H5) as a Function of Subgroup Size (*k*) for Analysts and for Q1 with GDP and Inflation.

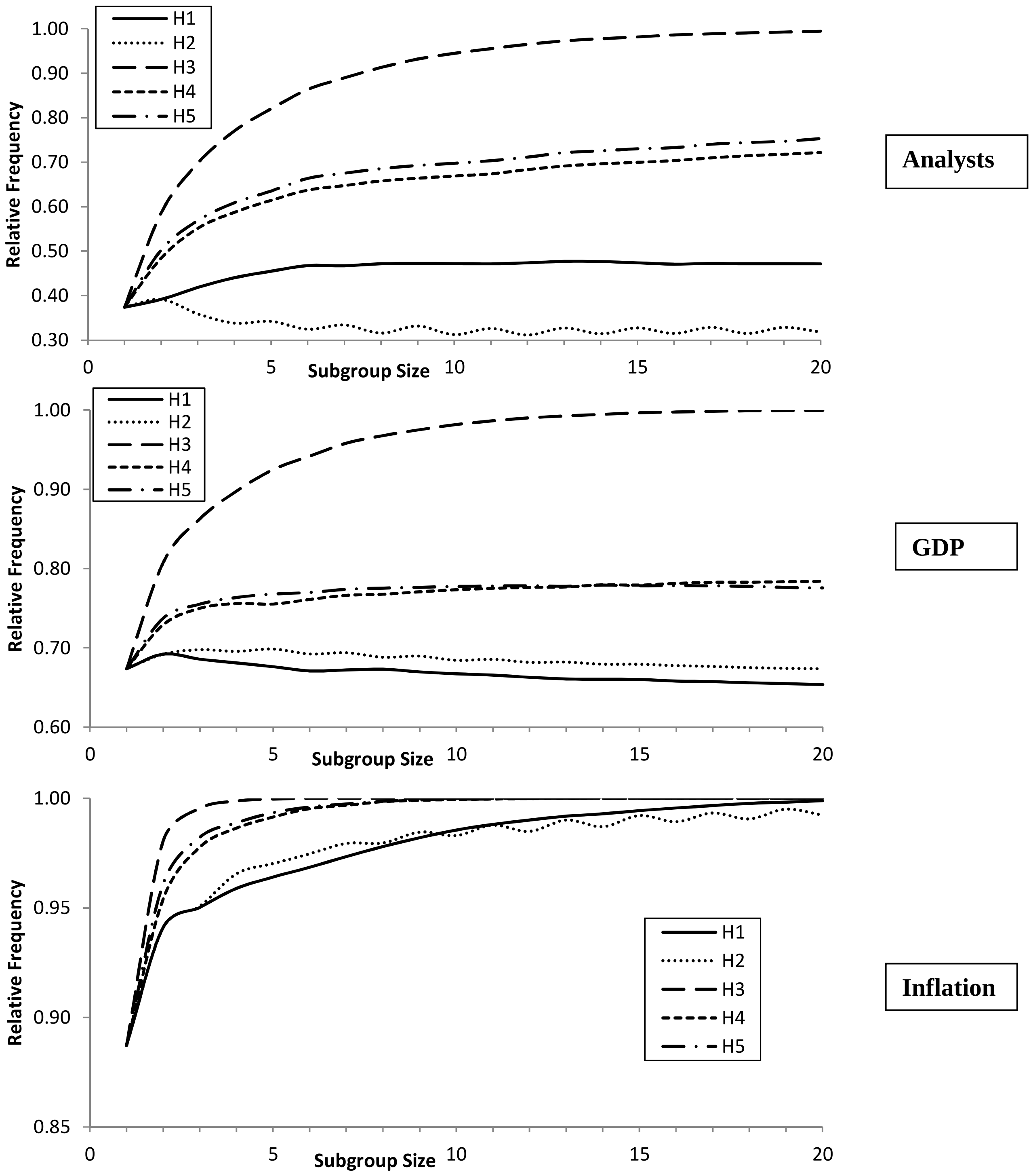

Figure 2. Relative Frequencies of Combined Intervals under Different Heuristics (H1 to H5) as a Function of Subgroup Size (*k*) for Analysts and for Q1 with GDP and Inflation.

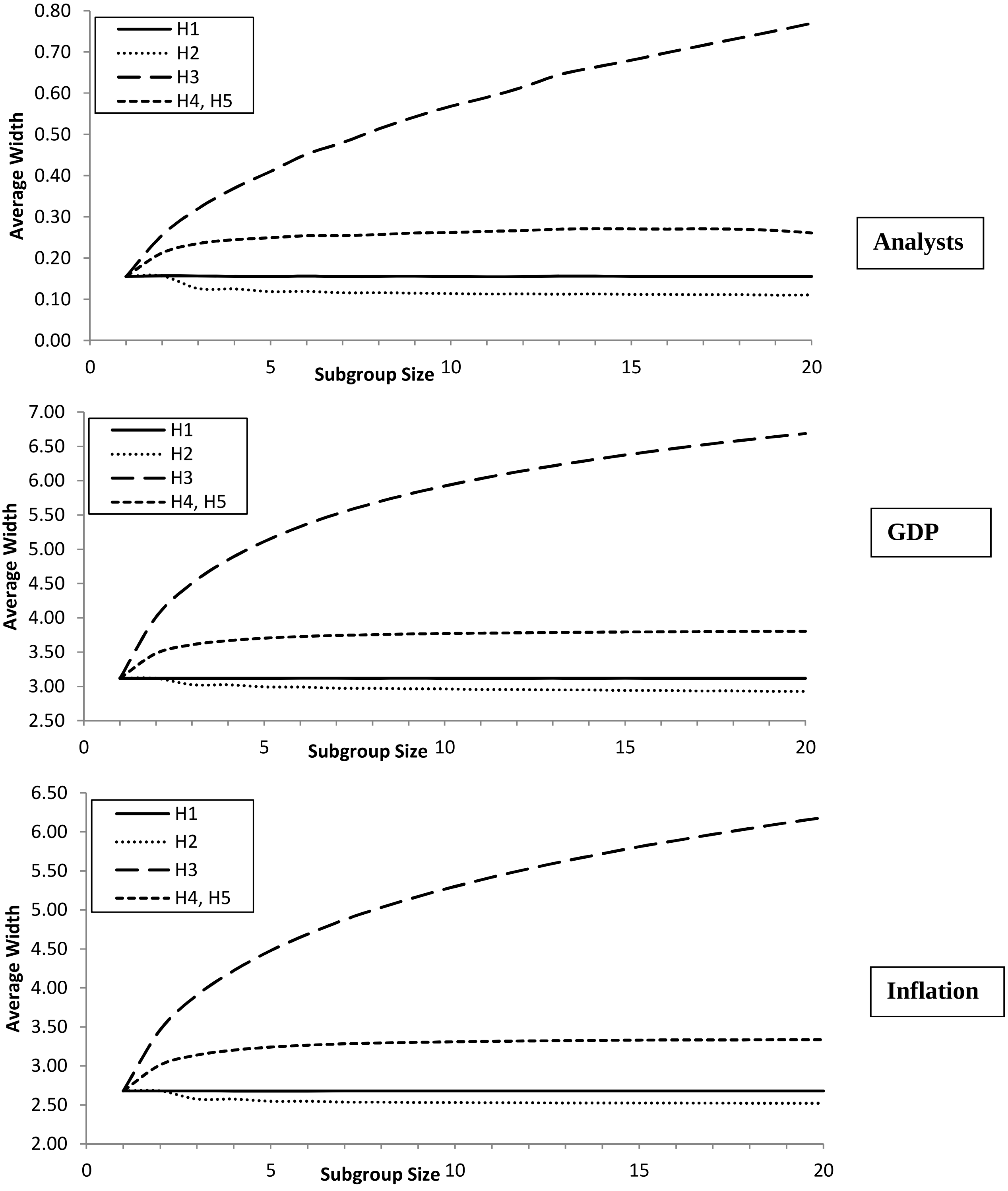

Figure 3. Average Widths of Combined Intervals under Different Heuristics (H1 to H5) as a Function of Subgroup Size (*k*) for Analysts and for Q1 with GDP and Inflation.

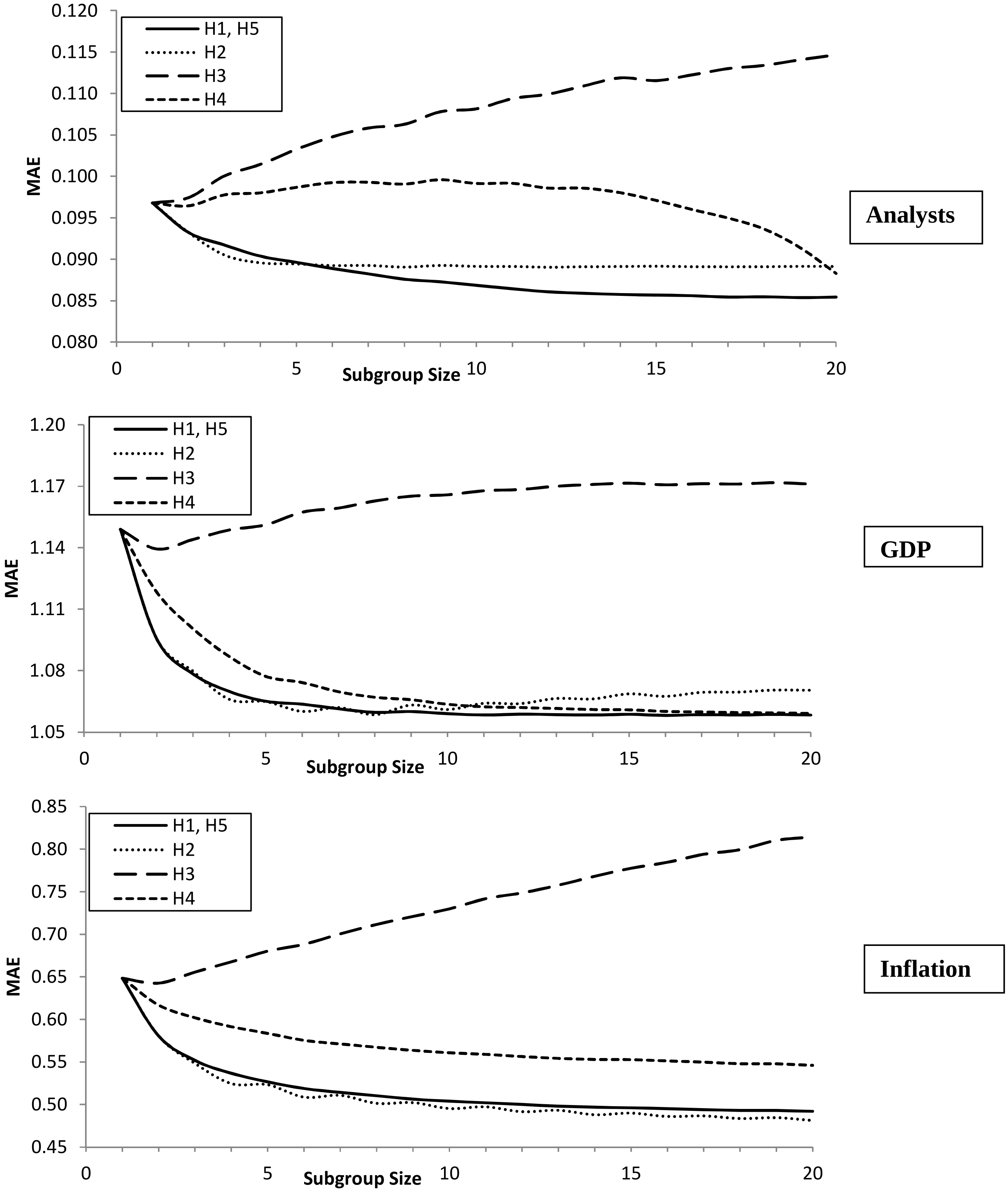

Figure 4. Mean Absolute Error of Midpoints of Combined Intervals under Different Heuristics (H1 to H5) as a Function of Subgroup Size (*k*) for Analysts and for Q1 with GDP and Inflation.

Europe Campus
Boulevard de Constance
77305 Fontainebleau Cedex, France
Tel: +33 (0)1 60 72 40 00
Fax: +33 (0)1 60 74 55 00/01

Asia Campus
1 Ayer Rajah Avenue, Singapore 138676
Tel: +65 67 99 53 88
Fax: +65 67 99 53 99

Abu Dhabi Campus
Muroor Road - Street No 4
P.O. Box 48049
Abu Dhabi, United Arab Emirates
Tel: +971 2 651 5200
Fax: +971 2 443 9461

www.insead.edu

INSEAD

The Business School
for the World®

Printed by INSEAD