Judgment and Decision Making, Vol. 14, No. 5, September 2019, pp. 617–619

A response to Mandel's (2019) commentary on Stastny and Lehner (2018)

Paul Lehner* Bradley Stastny†

Abstract

Stastny and Lehner (2018) compared the accuracy of forecasts in an intelligence community prediction market to comparable forecasts in analysis reports prepared by groups of professional intelligence analysts. To obtain quantitative probabilities from the analysis reports experienced analysts were asked to read the reports and state what probability they thought the reports implied for each forecast question. These were called *imputed probabilities*. Stastny and Lehner found that the prediction market was more accurate than the imputed probabilities and concluded that this was evidence that the prediction market was more accurate than the analysis reports. In a commentary, Mandel (2019) took exception to this interpretation. In a re-analysis of the data, Mandel found a very strong correlation between readers' personal and imputed probabilities. From this Mandel builds a case that the imputed probabilities are little more than a reflection of the readers' personal views; that they do not fairly reflect the contents of the analysis reports; and therefore, any accuracy results are spurious. This paper argues two points. First, the high correlation between imputed and personal probabilities was not evidence of substantial imputation bias. Rather it was the natural by-product of the fact that the imputed and personal probabilities were both forecasts of the same events. An additional analysis shows a much lower level of imputation bias that is consistent with the original results and interpretation. Second, the focus of Stastny and Lehner (2018) was on the reports as understood by readers. In this context, even if there was substantial imputation bias it would not invalidate accuracy results; it would instead provide a possible causal explanation of those results.

Keywords: forecasts, prediction market, imputed probability

1 Introduction

Stastny and Lehner (2018 [S&L2018]) compared the accuracy of forecasts in an intelligence community prediction market (ICPM) to comparable forecasts in analysis reports prepared by groups of professional intelligence analysts. To obtain quantitative probabilities from the analysis reports, experienced analysts were asked to read the reports and state what probability they thought the reports implied for each forecast question. These were called *imputed probabilities*. S&L2018 found that the prediction market was more accurate than the imputed probabilities and concluded that this was evidence that the prediction market was more accurate than the intelligence reports. In a commentary, Mandel (2019, [M2019]) strongly disagreed with this interpretation. In a re-analysis of the data M2019 found a strong correlation between readers' personal and imputed probabilities. From this, and some other results, M2019 builds a case that the imputed probabilities are largely a biased reflection of the readers' personal views; and that they do not fairly reflect the contents of the analysis reports at all.

*The MITRE Corporation. Email: plehner@mitre.org.

†Work completed while at The MITRE Corporation, currently at Google.

This paper examines the two primary concerns raised in M2019. First whether the high correlation between personal and imputed probabilities is evidence of substantial imputation bias. Second, how the existence of imputation bias, at any level, impacts that validity of the accuracy results.

2 Measuring imputation bias

Imagine that one is given the task of personally forecasting the average monthly temperature in Toronto over the next year; and then given the secondary task of imputing a forecast by reading and copying the average forecasted temperature found on a weather web site. Two things will occur. First, the personal and imputed forecasts will be highly correlated. It is after all hotter in the summer and colder in the winter. Second, the imputed forecasts will be completely unbiased and independent of the personal forecasts. After all, the reader is just copying a posted forecast. In this case the strong correlation between personal and imputed forecasts has nothing to do with imputation bias.

In general, personal and imputed probabilities will be correlated because they are both forecasts of the same events. To correctly measure imputation bias, it is necessary to parse out the covariation attributable to the common events.

S&L2018 measured imputation bias by examining the *relative* extent to which personal and imputed probabili-

TABLE 1: Correlations between personal and imputed probabilities across readers. (N in parentheses, italics indicates difference between within-reader and between-reader correlation is significant at least at p<.05.)

	Imputed				
Personal	Reader 1	Reader 2	Reader 3	Reader 4	Reader 5
Reader 1	0.435 (65)	0.170 (51)	0.485 (15)	*0.024* (35)	0.511 (20)
Reader 2	0.370 (51)	0.466 (64)	0.320 (27)	*−0.404* (22)	0.379 (21)
Reader 3	0.265 (15)	0.218 (27)	0.455 (27)		
Reader 4	0.229 (35)	0.343 (22)		0.154 (41)	0.455 (23)
Reader 5	0.634 (20)	0.523 (21)		*−0.033* (23)	0.733 (26)

ties tracked within and across readers. A non-parametric test was developed that compared within and across reader probabilities separately for each forecast question. A statistically significant effect was found but was characterized by S&L2018 as readers "... did a pretty good job of putting aside their personal views."

Another way to measure imputation bias, which is much closer to the approach in M2019, is to compare the correlation of personal and imputed probabilities within and across readers. The correlation between one reader's imputed probabilities with another reader's personal probabilities cannot be the result of imputation bias. So, it provides a baseline against which to compare within-reader personal and imputed correlations.

Table 1 shows these correlations. To interpret this table, consider the first data row. For the 65 forecasts from Reader 1, the correlation between that reader's personal and imputed probabilities was .435. This is statistically significantly greater than the correlation between Reader 1 personal and Reader 4 imputed probabilities (.024, N=35), but less than the correlation between Reader 1 personal and Reader 5 imputed probabilities (.511, N=20).

Across the five readers there are 11 instances where the within-reader correlation was higher than the across-reader correlations and 5 instances where the across reader correlations were higher. The weighted average within-reader correlation was 0.43 and across-reader correlation was 0.26. Very roughly this suggests that about 12% of the variance in imputed probabilities can be attributed to imputation bias. This is consistent with the original non-parametric test results.

3 Understanding the relevance of imputation bias

In the view expressed in M2019 the existence of imputation bias implies that the imputed probability method is not a valid approach to assigning a numerical probability forecast to an analysis report; and therefore, is not a valid approach to measuring the forecasting accuracy of analysis reports. M2019 correctly point out that S&L2018 did not provide any evidence that imputations are reliable, much less that they are a valid measure of what is written in the report.

The problem with this argument, at least as paraphrased above, is the word "valid". It presupposes that for each forecast there is a "true" or "correct" intended probability implied by the analysis report and that imputed probabilities are measures of that intended probability.

S&L2018 did not presume the existence of a correct probability interpretation. There is no reason for example to presume that the authors of the analysis reports had numerical probabilities in mind when they wrote their reports. If a report states "fair chance" then the authors may have meant nothing more than 'fair chance'; and if the authors did have a quantitative probability in mind when they wrote "fair chance ..." it's not clear why the accuracy of the report should be judged by what the authors did not write. S&L2018 measured the accuracy of the reports as those reports were understood by readers.

In this view, even substantial imputation bias would not invalidate comparative accuracy results but would instead provide a potential causal explanation of those results. If imputation bias were as substantial as M2019 suggested, then the relative accuracy of the ICPM over analysis reports might be attributed entirely to imputation bias. As it turns out, the imputation bias is not nearly as substantial as M2019 suggests and a reasonable explanation may just be a standard crowd wisdom explanation – mathematically aggregating independent forecasts tends to be more accurate than forecasts generated through consensus collaboration.

4 Discussion

Forecasting research that examines the accuracy of human analysts generally requires that analysts express forecasts as numerical probabilities (e.g. Tetlock and Gardner, 2015). But numerical probabilities may not be analysts' natural mode of thinking or communicating. And even if they do

have numerical probabilities in mind, they may intentionally not include them in their products.[1] The unique contribution of S&L2018 was the ability to estimate the accuracy of the forecasts that analysts actually publish.

S&L2018 found that the ICPM yielded forecasts that were more accurate than analysis reports as those reports were interpreted by readers. This may have been due to in part reader (mis) interpretation; in part to the fact that crowd wisdom methods tend to be more accurate than consensus methods; or in part to some other factors. In any case it does suggest that there is substantial opportunity to substantially improve the forecast quality of analytic products.

Regarding interpretation issues, there are a variety of steps analysts can take to reduce interpretation errors. Numerical probability forecasts would help, but as noted above analysts prefer not to include them in their reports. However, in preparing their reports analysts could use the imputed probability approach in S&L2018 to assess the likely interpretations of their products; and would then have a basis for determining whether they need to tighten the language in a report to better ensure their intended reader interpretations.

Regarding crowd wisdom methods, this research does provide some additional credence to the claim that forecasts from crowd wisdom methods tend to be more accurate than forecasts that result from traditional consensus forming methods. However, rather than focus on the particular crowd wisdom method examined in S&L2018 analysts should attend to a growing research literature demonstrating various ways to improve forecasting accuracy (e.g., Tetlock & Gardner 2015, Mandel & Barnes 2014). There are a variety of approaches to incorporating alternative forecasting methods into analytic reasoning. And importantly, as shown by Mandel and Barnes (2014), it is very feasible to track and measure forecasting accuracy. If analysts so choose, they can adopt an evidence-based approach to improving their analytic forecasting tradecraft. S&L2018 was intended to contribute to such a venture.

[1] Anecdotally, analysts have told us that readers tend to interpret quantitative probabilities as being the result of a formal statistical analysis. While they are mindful of the vagaries associated with verbal certainties, they believe that including numerical certainties would make their reports subject to even more misinterpretation.

References

Mandel, D. (2019). Too soon to tell if the US intelligence community prediction market is more accurate than intelligence reports: Commentary on Stastny and Lehner (2018). *Judgment and Decision Making, 14*(3), 288–292.

Mandel, D., & Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *PNAS, 111*(30) 10984–10989.

Stastny, B., & Lehner, P. (2018). Comparative evaluation of the forecast accuracy of analysis reports and a prediction market. *Judgment and Decision Making, 13*(2), 202–211.

Tetlock, P., & Gardner, D. (2015) *Superforecasting: The art and science of prediction*. Crown: New York, NY, USA.