Selecting the Optimal Number of Crowd Workers for Forecasting Tasks

Completed Research Full Paper

Arthur Carvalho
Farmer School of Business
Miami University
arthur.carvalho@miamioh.edu

Majid Karimi
College of Business Administration
California State University San Marcos
mkarimi@csusm.edu

Abstract

Crowdsourcing platforms enable requesters to elicit information from thousands of workers worldwide. A question that arises when outsourcing a task to a crowd is how many crowd workers a requester should hire. Focusing on forecasting tasks, we provide a methodological way of answering that question by formally describing how the number of crowd workers relates to the expected forecast error. Specifically, we discuss how an error curve associates a number of hired crowd workers with a fixed and a variable error. We then suggest an optimality concept that enables requesters to select an optimal number of crowd workers based on the tolerable variable error. In an illustrative study on Amazon Mechanical Turk, we demonstrate how highly prescriptive our approach is. In particular, we show how requesters can perform *ex-ante* and *ex-post* analyses by determining the optimal number of workers before running a task and by determining the fixed error after running the task and collecting forecasts.

Keywords

Crowdsourcing, forecasting, forecast evaluation, forecast aggregation.

Introduction

Crowdsourcing platforms give employers (requesters) access to a global population of (crowd) workers at virtually any time. The process of outsourcing a task to a crowd can happen in many different ways. We consider in this paper the crowdsourcing of well-defined, short tasks — sometimes called *microtasks* — that can be completed individually online and do not require highly specialized skills (Gadiraju et al., 2015). Several different microtasks have been successfully outsourced to crowds, e.g., the analysis of sentiments behind textual data (Borromeo and Toyama, 2015) and image collection (Yan et al., 2009).

Clearly, there are crucial factors related to crowd composition that may lead to successful crowdsourcing practices, e.g., the diversity (Hong *et al.*, 2020), skills (Budescu and Chen, 2015), and size (Robert and Romero, 2015) of the crowd. It is the latter factor that we focus on in this paper. In particular, we study the impact of crowd sizes on collective forecasting accuracy. The forecasting tasks we consider involve eliciting numerical values from individuals in a crowd to predict the outcome of an unknown event. We assume that a realized outcome (*ground-truth* value) will eventually be observed and can be used to determine *forecast errors*. Our task formulation is generic enough to capture several realistic problems, e.g., forecasting financial asset prices, demand for a product, event attendance, sports outcomes, etc.

We argue that there are two major challenges when determining the optimal number of crowd forecasters. First, traditional statistical techniques to determine sample sizes, such as power analysis (Cohen, 2013), may not be appropriate in crowdsourcing contexts as they do not consider forecast errors. Our approach differs because we rely on results that link the number of hired crowd workers to an expected forecast error. Specifically, our techniques rely on an *error curve* in which each number of hired crowd workers is linked to a *fixed* and a *variable error*, where the variable error decreases with more workers.

The second challenge we face is that "optimality" may have multiple definitions. For example, an optimal number of crowd workers can mean the maximum number a budget can pay for or the minimum number

that produces a requester's acceptable expected forecast error. Focusing on forecasting tasks, we define our optimality concept based on that latter idea, i.e., an optimal solution is based on a requester's tolerable variable error. Having established a formal optimality criterion, we derive the formulae required to determine the optimal number of crowd workers given a tolerable variable error and vice versa. To demonstrate the prescriptive power of our solution, we ran a demonstrative study on the crowdsourcing platform Amazon Mechanical Turk where we asked crowd workers to predict the outcome of a basketball game. When discussing that study, we highlight how a requester can use our approach *ex-ante* to define the optimal number of workers. We also suggest a bootstrapping algorithm to be used in an *ex-post* analysis to estimate the error curve once ground-truth values are available.

Research Background and Related Work

The idea of crowdsourcing forecasts gained prominence with the advent of research on *superforecasters* (Tetlock and Gardner, 2016). In short, under some conditions, it has been suggested that some individuals are consistently more accurate than the general public or experts in forecasting tasks as well as that non-experts can become incredibly accurate forecasters with appropriate training (Mellers et al., 2015). Another academic pillar supporting the crowdsourcing of forecasts is called the *wisdom of crowds* (Surowiecki, 2005). This research stream suggests that the elicitation and proper aggregation of forecasts produced by a crowd results in a more accurate forecast in expectation than any individual forecast.

Research on forecast aggregation has been well-studied by the decision analysis community, and it predates the research work on superforecasters and the wisdom of crowds. For example, the core idea that lower correlations among individual forecast errors result in greater expected benefits from combining forecasts was first developed in the 1960s (Bates and Granger, 1969). In the 1980s, hundreds of related research had already been published (Winkler *et al.*, 1989). Our formulae in this paper do not consider correlations of forecast errors because that would require knowledge of the historical accuracy of crowd forecasters — a crucial piece of information that is not available on most generic crowdsourcing platforms, such as Amazon Mechanical Turk. Our work instead relies on an *error curve*, i.e., a formal association between the number of crowd forecasters and the *expected* forecast error. In that regard, Lamberson and Page (2012) and Davis-Stober et al. (2015) proved a more generic result concerning the shape of error curves than what we rely on in this paper by allowing a non-zero covariance among forecast errors.

Carvalho et al. (2016) computationally derived the error curve and then approximated the underlying function using piecewise (segmented) linear regression analysis. The authors further suggested the intersection of segments as the optimal number of crowd workers a requester should hire. Unfortunately, that approach requires estimating the error curve first, i.e., knowing ground-truth values. As such, it cannot be applied *ex-ante* to determine the optimal number of crowd workers unless one must assume that forecasts reported on future forecasting tasks will resemble previously reported forecasts. That assumption allows a previously estimated error curve to be reused. Alternatively, our method relies only on the existence of the error curve but not necessarily on estimates of its parameters.

Given an error curve and under our proposed optimality concept, we show that the marginal benefits of hiring one extra crowd worker in order to decrease forecast errors quickly diminish. This result shows the power of *small crowds*. Related to that topic, there has been a flurry of recent experimental and theoretical work demonstrating that there is value in building small groups from a crowd. For example, Vercammen et al. (2019) found that a group of six crowd workers has a collective IQ one standard deviation above the mean for the general population. Callison-Burch (2009) showed that combining the judgments of five non-experts is enough to achieve the equivalent quality of experts in translation tasks. Goldstein et al. (2014) found that smaller, smarter crowds can be identified in advance and that they beat the wisdom of the larger crowd. Overall, our work validates the above stream of research through formal mathematical modeling.

Mathematical Model

We consider a scenario where a *requester* elicits a *point forecast* from a group of *crowd workers*. It is well-known in the decision analysis (Makridakis and Winkler, 1983) and machine learning (Dietterich, 2000) fields that, under certain conditions, aggregating forecasts from independent experts/models tends to improve the expected accuracy of the final forecast. A natural instance of our studied setting is when a newsvendor requests forecasts concerning the future demand for a product to properly build inventory

(Carvalho and Karimi, 2021). Other examples include predicting stock prices and estimating the number of future students accepting offers sent by a university. Formally, given $n \in \mathbb{Z}^+$ crowd workers, each expert $i \in \{1, \dots, n\}$ reports a forecast $f_i \in \mathbb{R}$. We assume the reported forecasts are independent and identically distributed (i.i.d.). We further assume that a true (ground-truth) value $t \in \mathbb{R}$ will materialize in the future, e.g., a true stock price or product demand will eventually be observed. In practical applications, this latter condition is required to measure forecast accuracy and issue outcome-contingent payments. Although we focus on single events, we note that our model and results naturally extend to multiple independent events.

Error Curve

Having established the basic notation we use throughout the paper, we now move to understand what happens with the expected forecast error when aggregating forecasts from different numbers of crowd workers. To do so, we rely on some traditional results in statistical inference. For example, to estimate errors, we rely on the mean square error (MSE), a robust measure that considers the variance and the bias of an estimator. Given a point forecast $f \in \mathbb{R}$ and the realized value $t \in \mathbb{R}$, the MSE is defined as follows:

$$\text{MSE}(f, t) = (f - t)^2 \tag{1}$$

Naturally, the lower the MSE, the more accurate a forecast is. To estimate the error when aggregating forecasts for a random group of g crowd workers, for $g \in \{1, 2, \dots, n\}$, we consider all combinations of crowd workers $C_{n,g} = \frac{n!}{(n-g)! \times g!}$ in groups of size g. For each group, we perform forecast aggregation by calculating the simple, equal-weight average of the reported forecasts, which we denote by $\bar{f}^{(g,x)}$, for $x \in \{1, 2, \dots, C_{n,g}\}$. That is, $\bar{f}^{(g,x)}$ is the average forecast from a group of size g indexed by x. We focus on equal-weight averages because they tend to perform well in practice and, at the same time, there are limits to the accuracy gains from combining forecasts optimally beyond simple averages (Genre et al., 2013). That said, we define $\bar{f}^{(g,\cdot)}$ as the average of the aggregated forecasts for groups of size g:

$$\bar{f}^{(g,\cdot)} = \frac{\sum_{x=1}^{C_{n,g}} \bar{f}^{(g,x)}}{C_{n,g}}$$

The corresponding variance is defined as:

$$\text{var}^{(g,\cdot)} = \frac{\sum_{x=1}^{C_{n,g}} \left(\bar{f}^{(g,x)} - \bar{f}^{(g,\cdot)}\right)^2}{C_{n,g}}$$

After aggregating the reported forecasts for each possible group of size g, one can then estimate the underlying expected error produced by groups of that size using MSE as follows:

$$\epsilon_g = \frac{\sum_{x=1}^{C_{n,g}} \left(\bar{f}^{(g,x)} - t\right)^2}{C_{n,g}}$$

Repeating the above procedure for all group sizes $g \in \{1,2, \dots, n\}$ produces what we call the *error curve*, i.e., a relationship between the number of crowd workers in a group (g) and the expected error for the aggregate forecast (ϵ_g).

Characterizing the Error Curve

We next introduce a well-known result in decision analysis that formalizes the relationship between the number of crowd workers (group size) and the expected error concerning the aggregate forecast. In other words, we formally define the shape of the error curve.

Proposition 1. $\epsilon_g = \alpha + \beta \times g^{-1}$, for $\alpha, \beta \geq 0$ and $g \in \{1, 2, \dots, n\}$.

Proof. Given the bias-variance decomposition of the mean squared error (Geman et al., 1992), we have that:

$$\epsilon_g = \frac{\sum_{x=1}^{C_{n,g}} \left(\bar{f}^{(g,x)} - t\right)^2}{C_{n,g}} = \text{bias}\left(\bar{f}^{(g,x)} - t\right)^2 + \text{var}^{(g,\cdot)}$$

Recall that any $\bar{f}^{(g,x)}$ is the average of g independent and identically distributed random variables. Moreover, $\text{var}^{(g,\cdot)}$ is the variance of the sampling distribution, which means that $\text{var}^{(g,\cdot)} = \sigma^2/g$, where σ is the standard deviation for the whole population of forecasts. Therefore, the above equation becomes:

$$\epsilon_g = \text{bias}\left(\bar{f}^{(g,x)} - t\right)^2 + \frac{\sigma^2}{g}$$

We can now define $\alpha = \text{bias}\left(\bar{f}^{(g,\cdot)}, t\right)^2$ and $\beta = \sigma^2$, thus completing the proof. □

The above proposition shows that the expected forecast error decreases monotonically when increasing the number of aggregated forecasts. This result captures what was empirically found by Makridakis and Winkler (1983), namely that forecasting accuracy improves while the variability of the accuracy among different combinations decreases as the number of forecasts in the average increases. Moreover, the first theorem in the work by Lamberson and Page (2012) reduces to Proposition 1 when the covariance among the errors of the reported forecasts is zero. An interesting aspect of Proposition 1 is that it allows for crowds to be wrong. That is, $\lim_{g\to\infty} \epsilon_g = \alpha$. In other words, α is a fixed error independent of the crowd size, whereas β/g is a variable error that is affected by the crowd size.

Optimal Number of Crowd Workers

Given the theoretical result in Proposition 1, a natural question to ask is how to determine an "optimal" number of crowd workers for a given forecasting task. We note that aggregating forecasts from the whole population of n crowd workers results in the lowest expected error. However, it might not be feasible to hire that many crowd workers in practice. Thus, one must first define what "optimal" means. Recall that according to Proposition 1, there are two major components contributing to the error ϵ_g, namely a fixed error α and a variable error β/g. One can reduce, in expectation, the variable error by increasing the number of crowd workers reporting forecasts. Given the above observation, we next suggest a new optimality concept based on how much variable error the requester is willing to accept. We later illustrate how this optimality concept can help the requester define *ex-ante* the optimal number of crowd workers for a forecasting task. In this way, our ideas constitute a decision support system regarding the number of crowd workers to be hired. We start by defining the error reduction sequence, i.e., the reduction in the variable error when the requester adds one extra crowd worker. For the sake of illustration, consider what happens with the error g when moving from $g = 1$ crowd worker to a group of $g = 2$ crowd workers:

$$s_{1,2} = \epsilon_1 - \epsilon_2 = \alpha + \beta - \alpha - \frac{\beta}{2} = \frac{\beta}{2}$$

Now, from $g = 2$ to $g = 3$, we have that:

$$s_{2,3} = \epsilon_2 - \epsilon_3 = \alpha + \frac{\beta}{2} - \alpha - \frac{\beta}{3} = \frac{\beta}{6}$$

More generally, moving from $g = x$ to $g = x + 1$ causes an error reduction of:

$$s_{x,x+1} = \epsilon_x - \epsilon_{x+1} = \frac{\beta}{x \times (x+1)}$$

Since the β term is fixed, we can therefore ignore it and focus only on the associated fraction. The error reduction sequence $\left(s_{1,2}, s_{2,3}, s_{3,4}, \cdots, s_{n-1,n}\right)$ is then defined as $\left(\frac{1}{2}, \frac{1}{6}, \frac{1}{12}, \cdots, \frac{1}{(n-1)\times n}\right)$. Focusing first on a population of infinite size ($n = \infty$), we show next that the series $\left(s_{1,2}, s_{2,3}, s_{3,4}, \ldots\right)$ is convergent.

Proposition 2. $\sum_{x=1}^{\infty} s_{x,x+1} = 1$.

Proof. After expanding the above summation, we have:

$$\sum_{x=1}^{\infty} s_{x,x+1} = \frac{1}{2} + \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \cdots + \left(\frac{1}{y} - \frac{1}{y+1}\right) + \cdots$$

for $y \in \mathbb{Z}^+$. After rearranging the above terms, we obtain that:

$$\sum_{x=1}^{\infty} s_{x,x+1} = 1 + \left(\frac{1}{3} - \frac{1}{3}\right) + \left(\frac{1}{4} - \frac{1}{4}\right) + \cdots + \left(\frac{1}{y+1} - \frac{1}{y+1}\right) + \cdots = 1 \qquad \square$$

Given the above results, consider next *partial sums* defined as $S_y = \sum_{x=1}^{y-1} s_{x,x+1}$. The corollary below characterizes S_y.

Corollary 1. $S_y = \sum_{x=1}^{y-1} s_{x,x+1} = 1 - 1/y$.

Proof. After expanding the above summation, we have:

$$\sum_{x=1}^{y-1} s_{x,x+1} = \frac{1}{2} + \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \cdots + \left(\frac{1}{y-1} - \frac{1}{y}\right)$$

After rearranging the above terms, we obtain that:

$$\sum_{x=1}^{y-1} s_{x,x+1} = 1 + \left(\frac{1}{3} - \frac{1}{3}\right) + \left(\frac{1}{4} - \frac{1}{4}\right) + \cdots + \left(\frac{1}{y-1} - \frac{1}{y-1}\right) - \frac{1}{y} = 1 - \frac{1}{y} \qquad \square$$

We note that the above corollary enables a requester to choose the ideal number of crowd workers based on the acceptable amount of variable error via partial sums of the error reduction sequence. We are now in a position to formally define our optimality concept.

Definition 1 (γ-optimality). We say that a number of crowd workers $y \in \mathbb{Z}^+$ is γ-optimal, for $0 \leq \gamma < 1$, when aggregating forecasts from y crowd workers leads to a reduction of at least γ of the total variable error.

For example, say that a requester is willing to accept at most 25% of the total variable error. This means that the total reduction in variable error must be at least 75%. This can be achieved by hiring $y = 4$ crowd workers since $s_{1,2} + s_{2,3} + s_{3,4} = 1/2 + 1/6 + 1/12 = 0.75$. That is, we say that hiring four crowd workers is 0.75-optimal. More generally, one can find the ideal number of crowd workers y for a desired reduction in variable error γ by solving the following optimization problem:

$$\begin{aligned} \min \quad & y \\ s.t. \quad & \gamma \leq S_y \\ & y \in Z^+ \end{aligned} \qquad (2)$$

In words, the optimization problem in (2) leverages Corollary 1 by finding the smallest positive integer y such that the associated partial sum S_y is greater than or equal to the desired variable error reduction γ. That is, the variable error reduction associated with y is at least the desired value γ. Figure 1 shows the variable error reduction for different numbers of crowd workers. An interesting point to note is that there is not too much benefit — accuracy-wise — beyond a certain number of crowd workers. For example, relying on only three crowd workers causes the total variable error to decrease by 66.6%. This number goes all the way to 90% with ten crowd workers and 99% with 100 crowd workers. This result highlights the wisdom of small crowds we previously discussed in Section 2.

The above results rely on a population of infinite size. In particular, the following adjustment must be made to the optimization program in (2) for a population of finite size n:

$$\begin{aligned} \min \quad & y \\ s.t. \quad & \gamma \leq \frac{S_y}{S_n} \\ & y \in \{1, 2, \ldots, n\} \end{aligned} \qquad (3)$$

In words, the series $s_{1,2} + s_{2,3} + s_{3,4} + \cdots + s_{n-1,n}$ no longer conveniently sum to one. Instead, it equates to S_n, which now represents the greatest variable error reduction. The first constraint in (3) takes that fact into account when normalizing the variable error reduction.

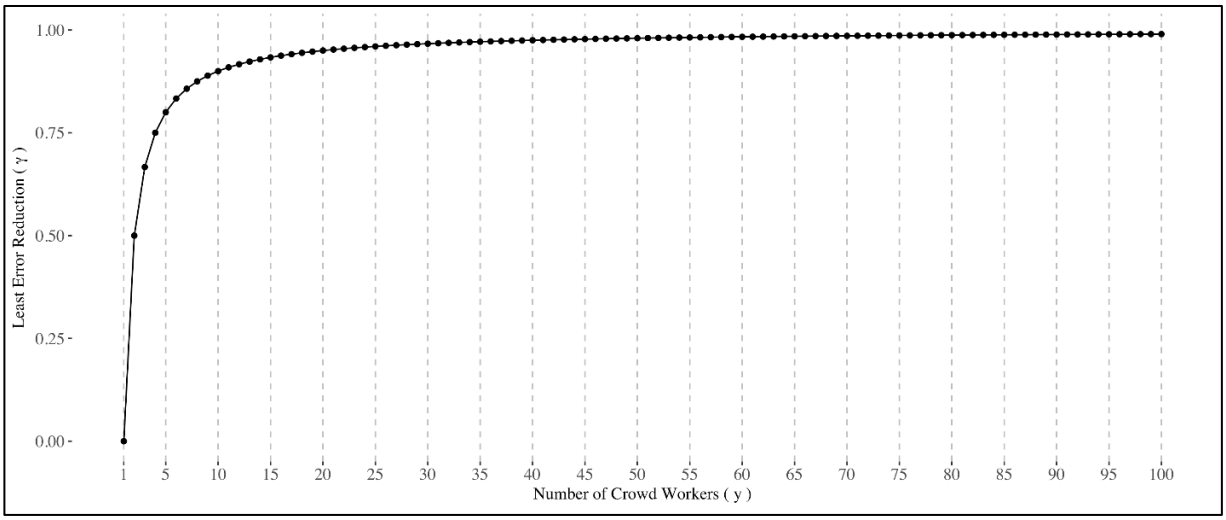

Figure 1. Variable Error Reduction (γ) for Different Numbers of Crowd Workers (y).

Illustrative Study

The optimization problems in (2) and (3) provide a highly prescriptive way of determining the optimal number of crowd workers for a certain forecasting task. Moreover, the proposed solution is not dependent on any unknown term that defines the error curve, such as α and β in Proposition 1. Nonetheless, having access to ground-truth values, a requester might be interested in estimating the fixed error (α) as well as the coefficient of the variable error (β). That can provide useful information when, for example, future forecasting tasks and crowd composition are expected to reflect previous ones. By means of an illustrative study, we highlight in this section how our theoretical findings can be used both in an *ex-ante* way to determine the optimal number of crowd workers to hire as well as in an *ex-post* way to determine the error curve and the associated limitations when using crowd workers in a given task.

Amazon Mechanical Turk

Our study concerns the crowdsourcing of forecasts. Specifically, we ran an experiment similar to the work by Carvalho et al. (2016). We hired a total of 50 crowd workers from the crowdsourcing platform Amazon Mechanical Turk (MTurk) to predict the outcome of an NBA game. Data produced by MTurkers have been widely used in information systems (Jia et al., 2017), accounting (Farrell et al., 2017), and operations management (Lee et al., 2018), among other business research communities. MTurk is particularly suitable for our purposes since the assumptions behind our model and results hold true there. For example, detailed information about crowd workers, including expertise and experience with forecasting tasks, may not be immediately available, meaning that our i.i.d. assumption is appropriate when randomly selecting a group of MTurkers to produce forecasts. Moreover, the performance on previous tasks, including forecasting tasks, is not publicly available, meaning that a requester has little to no relevant information to discriminate MTurkers based on historical forecast accuracy. That is a strong indicator in favor of using simple averages when aggregating forecasts — as assumed by our model — as opposed to more elaborate weighted averages.

To elicit forecasts, we adapted the interface used by Carvalho et al. (2016). After explaining the task and underlying interface, MTurkers could assign probability values to different outcomes by simply moving a slider from one side to another (see Figure 2). Since the underlying task has only two outcomes, a single probability value can represent a crowd worker's forecast, as required by our model. The slider's initial value was set at the 0.5/0.5 prediction. The crowd worker could assign more probability mass to a team's victory by moving the slider closer to the team's name. Whenever moving the slider, the crowd worker could see the amount of money made under the two different outcomes as well as the associated probability distribution (forecast). The quadratic (Brier) scoring rule was used to define payments (Brier et al., 1950).

This forecasting task has a natural realized outcome in that the requester can simply observe the result of the game; if, say, the home team is the winner of the game, then the realized outcome associated with that team is equal to 1. Otherwise, it is set to 0. Ground-truth data is required when estimating the error curve in an *ex-post* analysis. The forecasts used in our analysis are then the probability values associated with the home team winning the game.

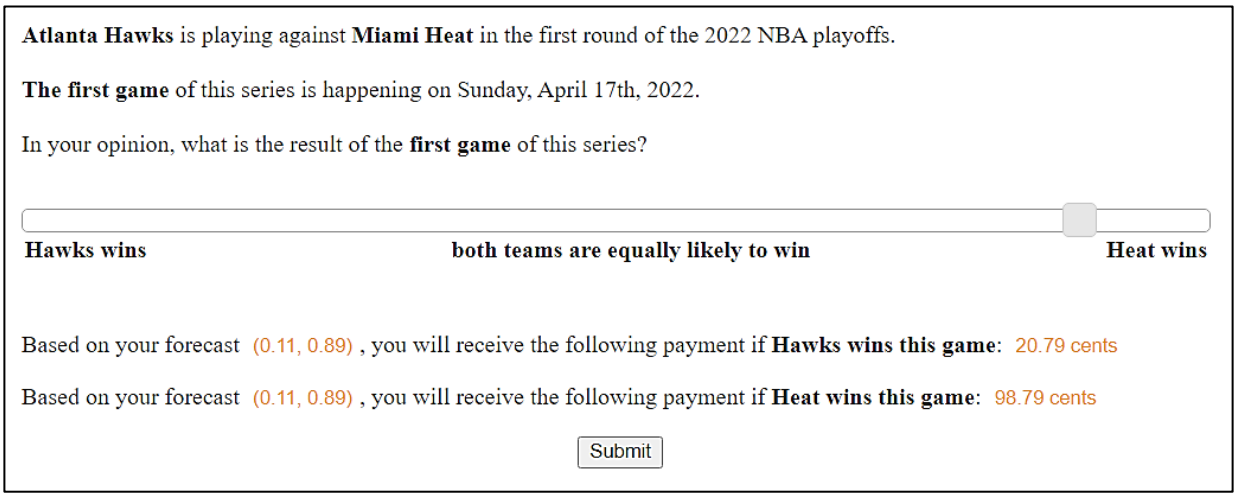

Figure 2. Interface Used for Data Collection.

Ex-Ante Analysis

Having described a forecasting task in the previous subsection, a question that arises is: how many crowd workers should a requester hire? Traditional statistical approaches based on power analysis (Cohen, 2013) are not suitable since they do not take forecast evaluation into account. Alternatively, the requester can explicitly rely on expected accuracy results by following the ideas we previously described. For example, say the requester has the budget to hire five crowd workers. Corollary 1 then suggests a reduction of 80% in the variable error, i.e., hiring five workers is 0.8-optimal. If the requester needs further accuracy guarantees, e.g., a 0.96-optimal solution, then solving the optimization problem in (2) implies that the requester needs to hire 25 crowd workers.

The above analysis assumes an infinite population of crowd workers. In the case of MTurk, this is a reasonable assumption for our purposes, given its massive size. In particular, it has been estimated that there are between 100,000 and 500,000 MTurkers (Difallah et al., 2018). Adjusting the above analysis by solving the optimization problem in (3) when $n = 100{,}000$, we still obtain that hiring five and 25 crowd workers are approximately 0.8- and 0.96-optimal.

Note how the *ex-ante* analysis is universal, i.e., it works for any similar forecasting task since it does not rely on the requester knowing the values of α and β to determine the error curve. We next discuss an approach to estimate those parameters in an *ex-post* analysis.

Ex-Post Analysis

Having elicited forecasts and obtained ground-truth values, the requester may now determine how good the crowd workers are in an *ex-post* analysis. In particular, the ground-truth values help to define the error curve characterized by Proposition 1. The error curve, in turn, determines the expected performance of crowd workers in future, similar forecasting tasks by defining the fixed error. To unequivocally determine the error curve, the requester needs access to the entire population of crowd workers and their forecasts, which is unlikely to be feasible in practice. Given a sample of $m << n$ crowd workers and their reported forecasts, the requester can estimate the error curve by following a bootstrapping approach. For example, to estimate the error when aggregating g forecasts, the requester can sample g crowd workers, aggregate their forecasts by using simple, equal-weight averaging, and measure the MSE of the aggregate forecast.

Clearly, the resulting error estimate regards a single sample. One can generalize this approach by creating a total of ℵ bootstrap resamples and, after repeating the above procedure ℵ times, obtain ℵ error estimates for a group of size g. The final forecast error estimate for that group is then the average of those ℵ error estimates. Algorithm 1 formalizes the above approach.

for all $g \in \{1,2,\dots,m\}$ **do**

  for all $j \in \{1,2,\dots,\aleph\}$ **do**

    θ = sample g from the m available forecasts

    $average = \text{mean}(\theta)$

    $error[\text{j}] = \text{MSE}(average, t)$

  end for

  $\hat{\epsilon}_g = \frac{\sum_{j=1}^{\aleph} error[\text{j}]}{\aleph}$

end for

Algorithm 1. Bootstrapping Procedure

Using Algorithm 1, we estimate the error curve for the 50 crowd workers in our data set using a total of $\aleph = 1{,}000{,}000$ bootstrap resamples. Thereafter, given the estimated errors for different numbers of crowd workers, we fit the equation we derive in Proposition 1 to our data using the Gauss-Newton nonlinear least-squares method. This results in the values $\hat{\alpha} = 0.05894$, $\hat{\beta} = 0.0884$ (residual sum of squares less than 5×10^{-8}), and the estimated error-curve equation being equal to $\hat{\epsilon}_g = 0.05894 + 0.0884/g$. Figure 3 shows the data and fitted curve. From it, one can see that the estimated fixed error produced is approximately 0.059. That is, the total error is still above zero no matter how many crowd workers are used, thus showing that crowds are not necessarily perfect forecasters.

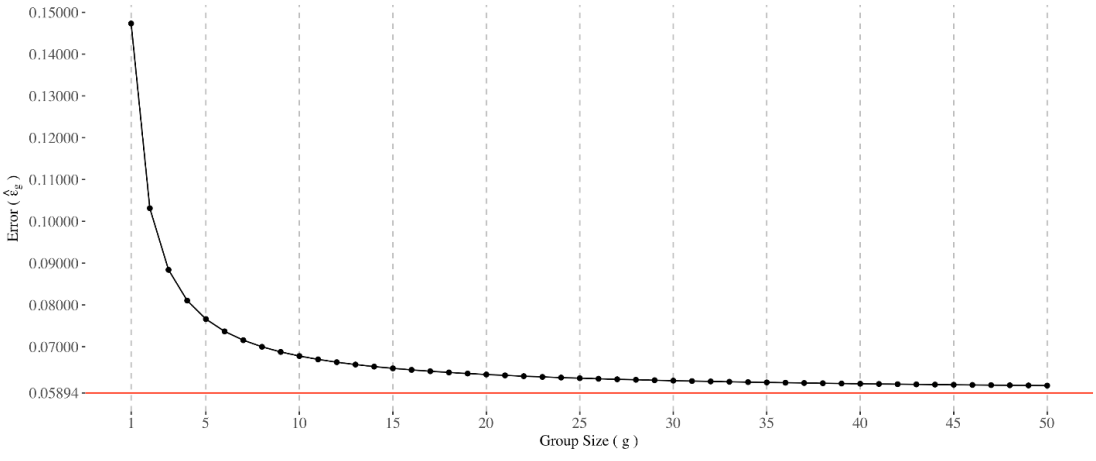

Figure 3. Estimated Error Curve.

Conclusion

A crucial problem when a requester elicits information from a crowd is to determine how many crowd workers are needed. In this paper, we proposed a highly prescriptive approach that solves that problem in forecasting tasks. We started by formally quantifying the impact of the number of crowd workers on the expected forecast error, resulting in an artifact called the error curve. In particular, we connected each group size to a fixed and a variable error. We then suggested a method to determine the optimal number of crowd workers given a tolerable variable error. Our formulae also allow a requester to estimate the variable error given a number of crowd workers. That feature is important because the requester might have budget

constraints and, thus, be able to hire only a limited number of crowd workers in practice. Our approach provides some formal guarantees on the variable error associated with that number.

We argue that the assumptions behind our model and methods are rather natural. For example, it is expected that forecasting tasks have a realized outcome (ground truth), which is required for the construction of an error curve. Furthermore, our results also rely on the aggregation of forecasts using simple, equal-weight averages. That is desirable in most crowdsourcing tasks because equally weighting the forecasts reported by all crowd workers reduces the risk of a highly inaccurate aggregate forecast since that approach requires no estimation of weights or parameters and, consequently, it is robust by not being sensitive to misestimation (Winkler, 2015). Moreover, decisions on weighting should be based on what is known about the background and expertise of the crowd workers whose forecasts are being aggregated. That makes weighting infeasible in crowdsourcing platforms such as MTurk, which has a massive population of crowd workers and limited information about them. These points, together with the fact that many crowdsourcing platforms do not provide a formal communication channel between workers, make the assumption that crowd workers' forecasts are independent and identically distributed acceptable.

Besides finding the ideal number of crowd workers for a desirable variable error and vice versa, we also discussed a bootstrapping algorithm to fully estimate the error curve. That algorithm can be helpful when the requester acquires ground-truth data and believes that crowd workers are likely to perform similarly in future forecasting tasks. In such cases, the estimated error curve enables the requester to determine fixed errors, i.e., how wrong the crowd will be in future tasks regardless of the number of hired workers. In an illustrative study on Amazon Mechanical Turk, we highlighted how our ideas could be used in an *ex-ante* and in an *ex-post* sense, i.e., to determine the optimal number of workers before running a crowdsourced task and to estimate the error curve after running the task.

In terms of future work, a natural extension of our ideas is to create methods to determine the optimal number of crowd workers beyond forecasting tasks. A challenge that can occur in those settings is that ground-truth values might not be available. For example, sentiment analysis or image labeling tasks can be subjective. Consequently, the idea of creating a universal error curve is not valid. On the other hand, there are some evaluation techniques a requester can rely on which do not depend on ground-truth values. For example, in tasks involving multiple-choice questions, one can ask crowd workers to endorse the answer most likely to be true and to predict the empirical distribution of the endorsed answers. Thereafter, the crowd workers are evaluated by the accuracy of their predictions — i.e., how well they match the empirical frequency — as well as how surprisingly common their answers are. That evaluation approach is the core behind the seminal Bayesian Truth Serum method (Prelec, 2004). That said, it is worth it investigating an optimal criterion for multiple-choice tasks alongside how to define an optimal number of crowd workers under the Bayesian Truth Serum method. Finally, it might be fruitful to investigate other forecast aggregation techniques beyond averaging. For example, there are statistical methods that aggregate forecasts by simulating a consensus-reaching process among the forecasters (Carvalho and Larson, 2013; DeGroot, 1974). Under such techniques, a new error curve might result in fixed errors lower than that derived in Proposition 1. Overall, we believe the above research directions will lead to a more effective and efficient information aggregation from crowds.

REFERENCES

Bates, J. M., and Granger, C. W. J. 1969. "The Combination of Forecasts," *Journal of the Operational Research Society* (20:4), pp. 451–468.

Borromeo, R. M., and Toyama, M. 2015. "Automatic vs. Crowdsourced Sentiment Analysis," in *Proceedings of the 19th International Database Engineering & Applications Symposium*, pp. 90–95.

Brier, G. W. 1950. "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review* (78:1), pp. 1–3.

Budescu, D. V., and Chen, E. 2015. "Identifying Expertise to Extract the Wisdom of Crowds," *Management Science* (61:2), pp. 267–280.

Callison-Burch, C. 2009. "Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 286–295.

Carvalho, A., Dimitrov, S., and Larson, K. 2016. "How Many Crowdsourced Workers Should a Requester Hire?," *Annals of Mathematics and Artificial Intelligence* (78:1), pp. 45–72.

Carvalho, A., and Karimi, M. 2021. "Aligning the Interests of Newsvendors and Forecasters through Blockchain-Based Smart Contracts and Proper Scoring Rules," Decision Support Systems (151).

Carvalho, A., and Larson, K. 2013. "A Consensual Linear Opinion Pool," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pp. 2518–2524.

Cohen, J. 2013. *Statistical Power Analysis for the Behavioral Sciences,* New York: Routledge.

Davis-Stober, C. P., Budescu, D. V., Broomell, S. B., and Dana, J. 2015. "The Composition of Optimally Wise Crowds," *Decision Analysis* (12:3), pp. 130–143.

DeGroot, M. H. 1974. "Reaching a Consensus," *Journal of the American Statistical Association* (69:345), pp. 118–121.

Dietterich, T. G. 2000. "Ensemble Methods in Machine Learning," *Multiple Classifier Systems*, pp. 1–15.

Difallah, D., Filatova, E., and Ipeirotis, P. 2018. "Demographics and Dynamics of Mechanical Turk Workers," in *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, pp. 135–143.

Farrell, A. M., Grenier, J. H., and Leiby, J. 2017. "Scoundrels or Stars? Theory and Evidence on the Quality of Workers in Online Labor Markets," *Accounting Review* (92:1), pp. 93–114.

Gadiraju, U., Demartini, G., Kawase, R., and Dietze, S. 2015. "Human Beyond the Machine: Challenges and Opportunities of Microtask Crowdsourcing," *IEEE Intelligent Systems* (30:4), pp. 81–85.

Geman, S., Bienenstock, E., and Doursat, R. 1992. "Neural Networks and the Bias/Variance Dilemma," *Neural Computation* (4:1), pp. 1–58.

Genre, V., Kenny, G., Meyler, A., and Timmermann, A. 2013. "Combining Expert Forecasts: Can Anything Beat the Simple Average?," *International Journal of Forecasting* (29:1), pp. 108–121.

Goldstein, D. G., McAfee, R. P., and Suri, S. 2014. "The Wisdom of Smaller, Smarter Crowds," in *Proceedings of the 15th ACM Conference on Economics and Computation*, pp. 471–488.

Hong, H., Ye, Q., Du, Q., Wang, G. A., and Fan, W. 2020. "Crowd Characteristics and Crowd Wisdom: Evidence from an Online Investment Community," *Journal of the Association for Information Science and Technology* (71:4), pp. 423-435

Jaspersen, J. G. 2022. "Convex Combinations in Judgment Aggregation," *European Journal of Operational Research* (299:2), pp. 780-794.

Jia, R., Steelman, Z. R., and Reich, B. H. 2017. "Using Mechanical Turk Data in IS Research: Risks, Rewards, and Recommendations," *Communications of the Association for Information Systems* (41), pp. 301–318.

Lamberson, P. J., and Page, S. E. 2012. "Optimal Forecasting Groups," *Management Science* (58:4), pp. 805–810.

Lee, Y. S., Seo, Y. W., and Enno, S. 2018. "Running Behavioral Operations Experiments Using Amazon's Mechanical Turk," *Production and Operations Management* (27:5), pp. 973–989.

Makridakis, S., and Winkler, R. L. 1983. "Averages of Forecasts: Some Empirical Results," *Management Science* (29:9), pp. 987–996.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., and Tetlock, P. 2015. "Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions," *Perspectives on Psychological Science* (10:3), pp. 267–281.

Prelec, D. 2004. "A Bayesian Truth Serum for Subjective Data," *Science* (306:5695), pp. 462–466.

Robert, L. P., and Romero, D. M. 2015. "Crowd Size, Diversity and Performance," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1379–1382.

Surowiecki, J. 2005. *The Wisdom of Crowds*, New York: Anchor

Tetlock, P. E., and Gardner, D. 2016. *Superforecasting: The Art and Science of Prediction*. New York: Random House.

Vercammen, A., Ji, Y., and Burgman, M. 2019. "The Collective Intelligence of Random Small Crowds: A Partial Replication of Kosinski et al.(2012)," *Judgment and Decision Making* (14:1), pp. 91-98.

Winkler, R. L. 2015. "Equal Versus Differential Weighting in Combining Forecasts," *Risk Analysis* (35:1), pp. 16–18.

Winkler, R. L., Grushka-Cockayne, Y., Lichtendahl Jr, K. C., and Jose, V. R. R. 2019. "Probability Forecasts and Their Combination: A Research Perspective," *Decision Analysis* (16:4), pp. 239-260.

Yan, T., Marzilli, M., Holmes, R., Ganesan, D., and Corner, M. 2009. "mCrowd: A Platform for Mobile Crowdsourcing," in *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, pp. 347–348.