Retrospective Analysis of Technology Forecasting: In-Scope Extension

Contract Number: HQ0034-11-C-0016

13 AUGUST 2012

Final Report

Submitted to:
Dr. Melissa Flagg
Director, Technical Intelligence
OSD AT&L/OASD(R&E)

Submitted by:
Carie Mullins
Senior Engineer
The Tauri Group, LLC
6363 Walker Lane
Suite 600
Alexandria, Virginia 22310
Phone: (724) 449-6177
www.taurigroup.com

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE **13 AUG 2012**	2. REPORT TYPE **N/A**	3. DATES COVERED

4. TITLE AND SUBTITLE **Retrospective Analysis of Technology Forecasting: In-scope Extension**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **The Tauri Group 6363 Walker Lane Suite 600 Alexandria VA 22310**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**UU**	**78**	

EXECUTIVE SUMMARY

The purpose of this study is to obtain a larger data set of verified technological forecasts than was obtained during the previous effort (at least 1,000), which will allow us to achieve the sample size necessary to identify predictive trends and causal relationships associated with forecast accuracy if they exist. A previous study of 310 verified forecasts revealed that technology forecasts developed using quantitative methods were more accurate than were other methods, forecasts about autonomous systems and computers were more accurate than were forecasts about other technology area tags, and that while forecasts tended to be neither optimistic nor pessimistic, there was not a strong correlation between the nine attributes we studied and the accuracy of a forecast. It is the purpose of this study to reevaluate and add to those findings using a larger sample size.

The Assistant Secretary of Defense for Research and Engineering (ASDR&E) is focused on developing tools and techniques to improve technological forecasting to guide future technology development. In support of these efforts The Tauri Group conducted an analysis of technology forecasting methods. The analysis will inform current and future efforts to improve forecasting, support automated methods of forecasting and, and establish a performance baseline against which new forecasting methods and tools can be compared.

Methodology and research results

The Tauri Group conducted a broad survey and analysis of retrospective forecasts from academia, industry, government and other sources. Forecast documents were thoroughly reviewed and 2,279 forecasts were extracted from those documents. Of these forecasts, 2,092 were found to be timely, specific, complete and relevant enough to be further verified and assessed for accuracy. These "assessable" forecasts were extracted from 300 forecast documents. Data collection metrics are summarized in figure ES-1.

Collected data	
Number of forecast documents	300
Assessable forecasts	2,092
Verified forecasts	1,055
Number of verification documents	2,016

FIGURE ES-1. COLLECTED DOCUMENTS AND FORECASTS

The forecasts were categorized using nine objective attributes including forecasting methodology, technology area tag, and timeframe, as seen in figures ES-2 and ES-3. We were able to verify 1,058 forecasts to establish if the predicted event occurred and if so when. The sample of 1,055 "verified" forecasts were used to provide descriptive statistics of our data and statistical analysis of the general population of forecasts.

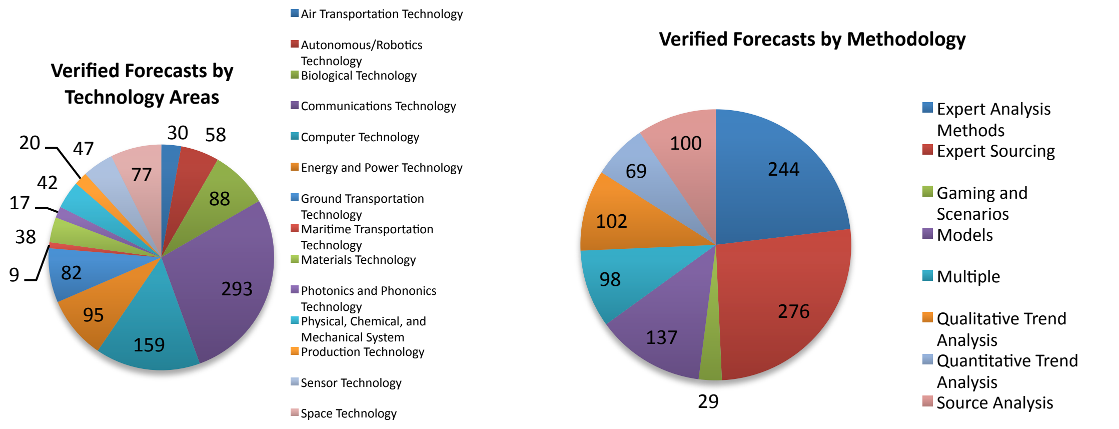


FIGURE ES-2. NUMBER OF FORECASTS BY TECHNOLOGY AREA TAG AND METHDOLOGY

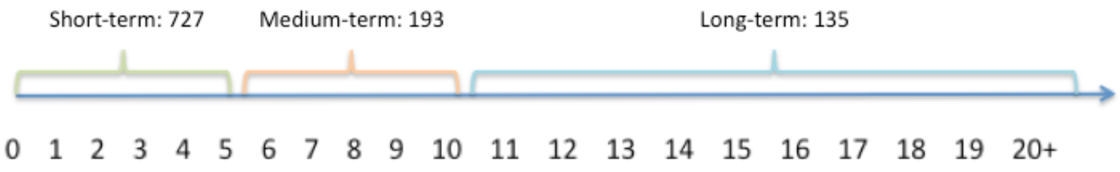


FIGURE ES-3. NUMBER OF FORECASTS BY TIME FRAME

Key Findings

Verified forecasts were analyzed descriptively and statistically. Forecast accuracy was analyzed against the nine attributes to identify statistical differences and trends within the sample and greater population. The key statistical and language findings are shown in table ES-2.

TABLE ES-2. KEY STATISTICAL AND LANGUAGE FINDINGS

In general, forecasts provide more accurate predictions than uninformed guesses. Six of the eight methodologies statistically are more accurate than a theoretical probability of success (random guess). Although qualitative trend analysis and gaming and scenarios methods have observed accuracies better than a random guess, at a 95% confidence interval there is no statistical evidence that these methods would perform better than a guess.
Forecasts based on numeric trends are more accurate than forecasts based on opinion. Forecasts generated from quantitative trend analyses have statistically higher success rates than do forecasts generated from other methodologies.
Forecasts are more likely to overestimate the event date. This is a change from our previous study, which indicated that there was a balance between pessimistic and optimistic forecasts.
Short -term forecasts are more accurate than medium- and long-term forecasts.
A predictive model of forecast accuracy could not be developed. Forecast accuracy appears to be influenced by a random component or some other attribute not captured in the study.
Forecasts that clearly describe timeframe, technology, predicted event, and associated performance metrics are more informative.

1.0 INTRODUCTION

1.1 BACKGROUND

As DoD’s Chief Technology Office, the Assistant Secretary of Defense for Research and Engineering (ASDR&E) focuses on developing tools and techniques that will provide an advantage for making investment decisions for technology development. Technology development forecasts, which include forecasts for emergence of new technologies, the evolution of existing technologies, or the migration of technologies to new application areas, provide an important tool for informing investment decisions. Understanding which forecasting methods provide the best forecasts for different types of technologies can impact ASDR&E’s technology development decisions and development of decision tools. The Retrospective Analysis of Technology Forecasting project was initiated to provide this key understanding to ASDR&E.

1.2 OBJECTIVES

Study Goals

- Provide a rigorous statistical evaluation of science and technology forecasting methods using a large sample size (at least 1,000 verified forecasts)
- Recommend a standard language for expressing forecasts in future studies

The purpose of this study is to obtain a larger data set of verified technological forecasts than was obtained during the previous effort (at least 1,000), which will allow us to achieve the sample size necessary to identify predictive trends and causal relationships associated with forecast accuracy if they exist. The results provide a baseline for forecast accuracy, as well as insight into the elements of a forecast that contribute to its accuracy. The analysis will be used to inform current and future efforts to improve forecasting, support fielding of more automated methods of forecasting, and enable comparison of new forecasting methods. A significant output of this project is a recommended standard language for expressing forecasts in future studies, which can be used to improve the utility of future forecasts. A standard lexicon based on analyst efforts to interpret forecasts is included in appendix B.

2.0 METHODOLOGY AND RESEARCH DESIGN

The Tauri Group conducted a broad survey and analysis of literature from academia, industry, and open sources that generated 2,279 forecasts from 300 documents. The analysis categorized the forecasts using nine attributes that were described in the previous version of the study, including forecast methodology, technology area tags, and time frame. From this data, we conducted a statistical analysis to evaluate each forecast method, identify which method(s) provided the best results given a combination of technology area tags and time frames, and identify which attributes were the most significant for a given forecasting methodology. We also created a dictionary of unambiguous language to enable clear and concise interpretations of forecasts. The intent of this dictionary is to help users read and understand future forecasts.

The study was conducted in four phases:

1. ***Collect forecast documents.*** We collected forecast documents from academia, industry, government, and other sectors. Throughout the collection process, we reviewed forecast

documents to determine if certain geographic regions, technology area tags, or publication types were underrepresented. We then initiated targeted document searches to improve representation across regions, technologies, and publication sources.

2. ***Extract technology forecasts from documents and record attributes in database.*** During this phase, we extracted technology forecasts that were embedded in forecast documents and added them to the Quickbase database that was used during the previous effort. We conducted additional research to determine whether the forecasted events had occurred, and, if so, when they occurred.
3. ***Verify whether forecasted events occurred.*** During this phase, we conducted research to verify whether forecasted events had occurred and, if so, when they occurred.
4. ***Assess forecast accuracy and identify key findings.*** During the original study, we developed several statistical tests to determine how accurate forecasts were and to identify attributes associated with forecast accuracy. These tests were repeated on the larger data set.

The methodology applied to this study is consistent with that of the original study. The only differences were the means used to collect and verify forecasts, as well as the use of deep web research and broad sourcing. The study methodology is summarized in figure 1.

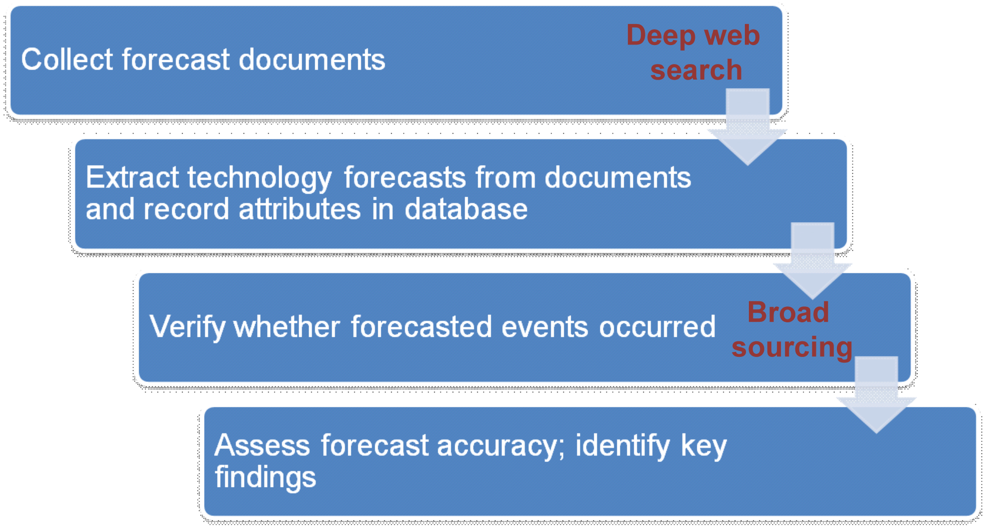

Collect forecast documents
Deep web search
Extract technology forecasts from documents and record attributes in database
Verify whether forecasted events occurred
Broad sourcing
Assess forecast accuracy; identify key findings

FIGURE 1. STUDY METHODOLOGY

2.1 COLLECT FORECAST DOCUMENTS

During the previous report, we collected forecast documents that were easily acquired from academia, industry, popular media, and Internet searches in the U.S. domain. This effort required that we maximize the number of forecast documents collected in order to provide an adequate pool of forecasts with which to meet the increase sample size (at least 1,000 forecasts). Having already collected many open source forecast documents during the original study, our document collection method during this iteration required that we employ alternative methods that would allow us to find as many previously-unidentified forecast documents as possible. These methods include:

- Searching the “deep web” to find forecast documents that are not indexed by standard search engines and therefore cannot be retrieved using standard search engines.

- Searching in foreign web domains for English-language forecast documents that are not available in the U.S. domain
- Purchasing used books and published reports that contain technological forecasts

To search the deep web, we contracted with Bright Planet, which used proprietary methods to search "deep web" sites for forecast-related dynamic content that was not indexed, then cataloged the search results in an indexed repository (a "silo"). Bright Planet retrieved more than 3,000 potential forecast documents using forecast-specific queries and websites of interest provided by the Study Team.

Searches in foreign web domains yielded many forecast documents, most of which were authored or published outside of North America, which was the dominant geographic region in the original study. Used books proved to be a rich source of forecasts, both in quantity and in underrepresented forecast time horizons and methodologies. Vendors like Amazon.com and Abebooks.com provided a wide selection of affordable used books and reports, many of which were published by strategic analysis firms—an underrepresented publication type in the previous study due to their price. All sources yielded a greater number of forecast documents about computer and communications technologies—the dominant technology area tags in the previous study—than other technologies, suggesting that the general pool of available forecast documents favors these two technologies.

The most successful methods for obtaining forecasts were conducting Internet searches and purchasing books and reports, since Internet searches allowed researchers to retrieve documents with forecast-specific language[1] and books and reports typically contained numerous assessable forecasts. Deep web services were not as successful for obtaining forecasts due to their inability to perform contextual data mining of the websites and documents they searched in an automated fashion. Consequently, many of the documents retrieved by the deep web service lacked technological forecasts or assessable forecasts. At present, it is more efficient to have a researcher conduct Internet searches of forecast-specific language because there is not yet an automated system capable of discerning whether forecast-related language in a document is associated with an actual forecast, much less an assessable one.

2.2 EXTRACT FORECASTS FROM DOCUMENTS AND RECORD ATTRIBUTES

After reading, assessing, and recording the forecast documents, analysts extracted forecasts and related information from the documents. The forecast extraction process comprised two steps, which are described in the previous report and are consistent with the original study: 1) identify assessable forecasts and 2) interpret and record forecast data and attributes.

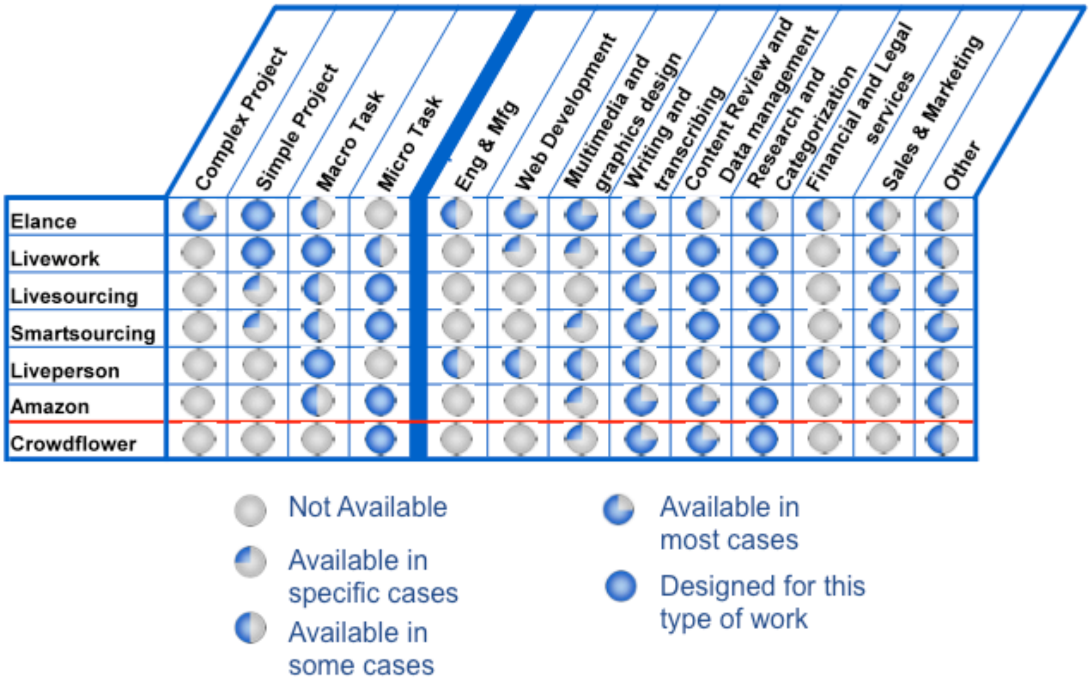
2.3 VERIFY WHETHER FORECASTED EVENTS OCCURRED

After entering forecast information into the database, analysts attempted to verify whether the forecasted event occurred and, if so, when it occurred. In order to verify the more than 1,000 forecasts that were the goal of this study, we employed broad sourcing as an alternative verification technique. This method is described below.

[1] Forecast-specific language includes such phrases as, "in the next X years," "in the short term," and "by X year."

2.3.1 SELECTING BROAD SOURCING SERVICES

Broad sourcing is a method that involves outsourcing a task to a diffuse group of people. The previous work used Tauri personnel and subject matter experts to verify forecasts. Using this method, 38% of assessable forecasts were verified. Since this iteration of the project required double the number of newly verified forecasts in half the time, we needed to increase the rate at which forecasts were identified as well as improve the verification rate. To achieve these improvements, we used broad sourcing to provide us with access to a larger pool of researchers. Several broad sourcing sites exist, most of which are tailored for marketing, web development, graphic design, or simple research projects. Figure 2 summarizes the functionality of seven popular broad sourcing services.



	Complex Project	Simple Project	Macro Task	Micro Task	Eng & Mfg	Web Development	Multimedia and graphics design	Writing and transcribing	Content Review and Data management	Research and Categorization	Financial and Legal services	Sales & Marketing	Other
Elance	◕	●	◑	○	◑	◕	◕	◕	◑	◑	◑	◑	◑
Livework	○	●	●	◑	○	◔	◔	◕	●	●	○	◕	◑
Livesourcing	○	◔	◑	●	○	○	○	◕	●	●	○	◕	◕
Smartsourcing	○	◔	◑	●	○	○	◔	◕	●	●	○	◑	◕
Liveperson	○	○	●	○	◑	◑	◑	◑	◑	◑	◑	◑	◑
Amazon	○	○	◑	●	○	○	◔	◕	◕	●	○	○	◑
Crowdflower	○	○	○	●	○	○	◔	◕	◕	●	○	○	◑

○ Not Available

◔ Available in specific cases

◑ Available in some cases

◕ Available in most cases

● Designed for this type of work

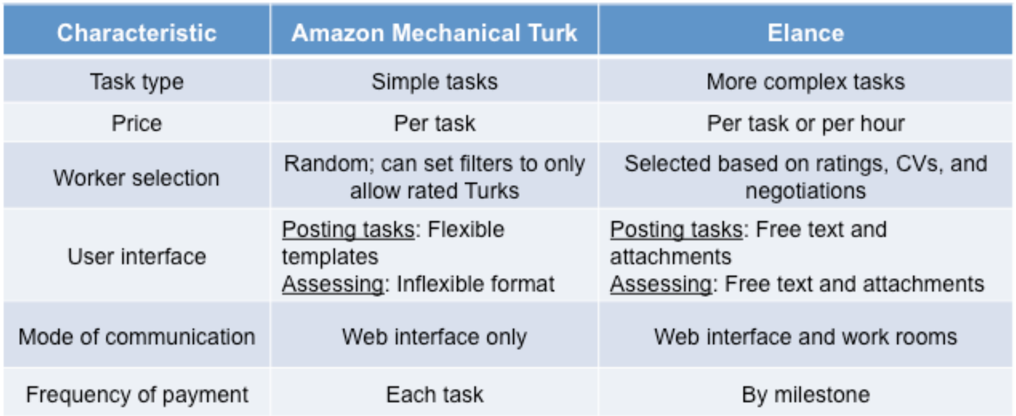
FIGURE 2. SUMMARY OF BROADSOURCE WEBSITE SPECIALTIES[2]

Elance and Amazon Mechanical Turk (AMT) were the broad sourcing sites best suited for forecast verification, which is considered a research-based macrotask. Depending on how forecasts are disseminated to researchers, verification can be either a microtask or a macrotask.[3] Asking a researcher to verify whether a single forecast event has occurred (and, if so, when it occurred) is an example of a microtask, whereas asking a researcher to determine whether multiple forecast events have occurred is a macrotask. Elance is most conducive to macrotasks such as those required for forecast verification, while Amazon Mechanical Turk was more conducive to microtasks.

[2] B. Frei, "Paid Crowdsourcing: Current State and Progress Toward Mainstream Business Use," September 15, 2009, accessed January 14, 2012, www.smartsheet.com

[3] Microtasks are simple tasks that do not typically require much skill or time on the part of the researcher, such as identifying logos or objects in pictures, whereas macrotasks are more complex research tasks. Verifying a single forecast, as was attempted when we used Amazon Mechanical Turk, is a microtask, whereas verifying ten forecasts for a flat rate, as was possible using Elance, is a macrotask.

TABLE 1. COMPARISON BETWEEN AMAZON MECHANICAL TURK AND ELANCE



Characteristic	Amazon Mechanical Turk	Elance
Task type	Simple tasks	More complex tasks
Price	Per task	Per task or per hour
Worker selection	Random; can set filters to only allow rated Turks	Selected based on ratings, CVs, and negotiations
User interface	Posting tasks: Flexible templates Assessing: Inflexible format	Posting tasks: Free text and attachments Assessing: Free text and attachments
Mode of communication	Web interface only	Web interface and work rooms
Frequency of payment	Each task	By milestone

AMT's researchers ("Turks") are paid per task and their identities remain unknown to employers, whereas Elance allows employers to select researchers based on their curriculum vitae and user ratings. Researchers on Elance are typically highly specialized and are generally paid by the hour. The differences between the two services are summarized in table 1.

After identifying the broad sourcing sites most likely to yield successfully verified forecasts, we created a template that contained forecast statements and detailed instructions of acceptable forecast verification research and disseminated forecasts on the AMT site.[4] We initially chose to post forecasts on AMT because of its large population of workers (over two million) and the ability to pay researchers only for successfully verified forecasts. However, we found the site to be ineffective for the type of work we were conducting for the following reasons:

- In spite of using only highly rated Turks (as assessed by AMT's internal rating system), the quality of work from many researchers was well below our expectations and resulted in a success rate of less than 10%.
- Through the AMT interface, work performed by a Turk can be only accepted or rejected. Accepted work is immediately paid for with no chance to communicate with the worker. It is possible to communicate with Turks whose work has been rejected to explain why the work was not adequate, but it is not possible to reassign the same task to the same Turk.
- The AMT interface did not allow us to efficiently evaluate the work performed by Turks.

Elance was better suited for our task, since it allowed us to easily communicate with researchers and batch our tasks into groups of 10 or 20 forecasts instead of one at a time, enabling us to pay researchers by task. We also found that many of the workers on Elance were willing to be paid per accepted forecast, rather than per hour. While the available labor pool on Elance is smaller (500,000) and more expensive than AMT, the quality of work and ability to effectively communicate with researchers resulted in a substantially higher acceptance rate (approximately 50%) for the work performed.

[4] The template is provided in appendix D.

In the end, broad sourcing turned out to be more of a “broad hiring” process in which AMT and Elance were used to identify a potentially qualified workforce to which we would not have had access otherwise. In spite of the difficulties associated with AMT, we were able to find and establish a long-term working relationship with five Turks (out of the 53 who responded to posted tasks). Using Elance, we posted several jobs and invited roughly 40 people to participate. Of the 15 workers with whom we contracted to work, we were able to establish long-term productive relationships with 10 of them. At the peak of the effort, we had 18 concurrent virtual assistants providing verifying information on forecasts. Broad sourcing was highly successful in terms of forecast verification rate, indicating that the method is well-suited for complex research tasks.

The process for verifying forecasts using broad sourcing is summarized in figure 3.

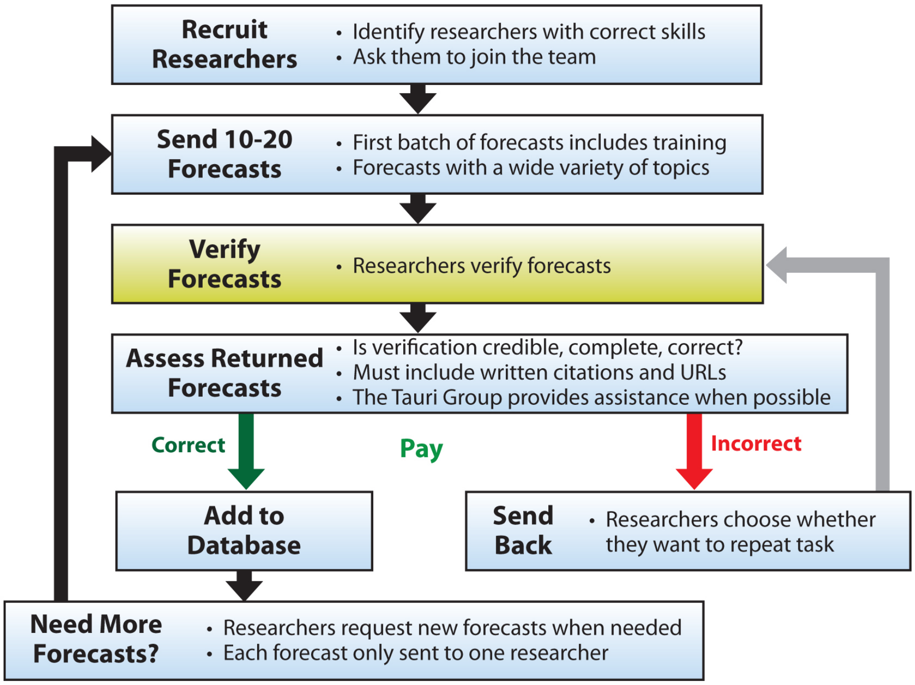

Recruit Researchers
• Identify researchers with correct skills
• Ask them to join the team
Send 10-20 Forecasts
• First batch of forecasts includes training
• Forecasts with a wide variety of topics
Verify Forecasts
• Researchers verify forecasts
Assess Returned Forecasts
• Is verification credible, complete, correct?
• Must include written citations and URLs
• The Tauri Group provides assistance when possible
Correct
Pay
Incorrect
Add to Database
Send Back
• Researchers choose whether they want to repeat task
Need More Forecasts?
• Researchers request new forecasts when needed
• Each forecast only sent to one researcher

FIGURE 3. WORKFLOW FOR FORECAST VERIFICATION USING BROADSOURCING

2.3.2 FORECAST VERIFICATION USING BROAD SOURCING

Researchers verified forecasts using open-source research. Only information from credible sites and sources was accepted from researchers. These sources include conference papers, articles, popular magazines with a reputable editing record, industry sites, electronic books, and other sources with journalistic standards and appropriate subject matter expertise. Researchers were required to find verifying information from a minimum of two independent sources to ensure verification information was correct and unbiased.

Verifying information obtained through broad sourcing, including the citations and page numbers for ground truth sources as well as analyst comments and explanations, was reviewed by Tauri analysts to ensure the research was credible and relevant before broad sourcing researchers were paid and verifying information was entered into the database. Forecasts that were credibly verified were entered into the database by separate researchers in Elance who had been trained to enter verifying information into the database.

2.3.3 VALIDATING FORECASTS

After verified forecasts were entered in the database, a senior Tauri Group analyst validated that a forecast and its attributes had been characterized properly and that verifying information was clear, credible, and reproducible. Where needed, additional research through broad sourcing was conducted on forecasts that required additional sources. This additional research was conducted by a different researcher than the one who provided the original verifying information.

The broad sourcing method of forecast verification yielded 1,058 verified forecasts out of 2,092 total assessable forecasts—an overall verification rate of 51%, which is a significant improvement over the method employed in the original study. The remaining 1,034 assessable forecasts were not included in the statistical analysis because there was not sufficient information to characterize their ground truth with a high degree of confidence.

2.4 ASSESS FORECAST ACCURACY

The purpose of the data analysis is to quantify and characterize the forecasts, identify which forecasting methods are the most accurate, and determine the likelihood that a forecasted event will occur within the specified time frame, given certain conditions.

To facilitate this analysis, we identified six metrics that could be consistently applied to each forecast, a set of criteria by which we could assess each forecast, and an analytical plan that allowed us to answer questions about the accuracy of forecasts.

2.4.1 FORECAST METRICS

We identified six metrics to facilitate the characterization of forecasts:

- ***Success***. A forecast was considered successful if the forecasted event was realized (occurred) within an allowable time frame. The allowable time frame was calculated as +/- 30% of the forecast time frame centered around the forecasted date. If a forecast was made in 1990 and the predicted year of occurrence was 2000 (a ten-year forecast), the forecast would be a success if the actual event occurred sometime between 1997 and 2003 (10 years +/- 3 years). For forecasts that provided an explicit range, we used the provided range as the criteria for success.
- ***Realization.*** This binary metric captures whether the forecasted event has occurred. A forecast that predicts that flying cars will exist by 2010 has been realized because flying cars had been developed by that year. However, a forecast that predicts consumer adoption of flying cars by 2010—for example, by predicting that flying cars will be commonplace by 2010—was not realized on the forecast date and has not yet been realized.
- ***Degree of realization on forecasted year.*** This metric captures the degree to which a complex forecast was realized. In some cases, a forecast may be unrealized but not entirely inaccurate; in these cases, we characterized forecasts as partially accurate. For example, if a forecast predicted solar cell efficiency would increase to 40% by 2005 and the ground truth data revealed that it only increased to 36%, the forecast was characterized as having been partially realized.[5]

[5] Analysts captured the degree of realization on a Likert scale, which included five metrics: not realized at all, somewhat realized, almost realized, mostly realized, and fully realized. A comments box was provided for analysts to explain their selection.

- ***Degree of interpretation.*** This metric captures the amount of interpretation involved in verifying that a forecasted event did or did not occur. Not all ground truth sources provided unambiguous information, and not all forecasts were easily interpreted. This metric allowed analysts to provide insight about potential error introduced during the forecast verification process. [6] For example, when verifying a forecast that stated that by a certain time frame a single product that combined television, VCR, and Internet capabilities would be commonplace, analysts had to loosen their definition of television and VCR to include Internet streaming of television shows and movies and consider the ability of computers to play DVDs.[7]
- ***Signed temporal forecast error (STFE):*** The STFE measures the difference between the date the forecast was realized and the forecast date. As such, it measures the temporal accuracy of a forecast. A forecast that predicts that a technology will emerge in 1990 would have an STFE of -3 years if ground truth revealed that the technology emerged in 1987, three years before the predicted date. For forecasts that provide a range of dates, the STFE is calculated from the midpoint in the range. The STFE provides a consistently comparable metric to evaluate precision and serves as the basis for analytical statements like "medium-term expert sourcing forecasts tend to occur two years sooner, on average, then predicted."
- ***Temporal forecast error (TFE):*** The TFE is the absolute value of the STFE. It measures the magnitude of the error: the larger the average TFE, the larger the error. A forecast that predicts that a technology will emerge in 1990 would have an STFE of -3 years and a TFE of 3 years, if ground truth revealed that the technology emerged in 1987. The TFE complements the STFE, describing the magnitude of errors among forecasts, rather than bias towards over- or underestimating forecast events. The TFE is the basis for analytical statements like "medium-term expert sourcing forecasts on average miss event dates by more than four years."

These metrics were used during and after the forecast verification process to characterize analysts' confidence in verification and the accuracy of forecasts.

2.4.2 ASSESSMENT PROCESS

Figure 4 illustrates the process used to assess the forecast results. We first determined if a forecast could be verified, whether it fell within the time frame limits set using the 30% rule, and whether it met the other three criteria for viability (complete, specificity of language, and relevance). We collected a total of 2,092 forecasts that met this requirement. These forecasts were then verified to determine if the forecasted event was realized (if it occurred). Of the 2,092 forecasts, we were able to verify whether 1,058 of them were realized or not. Using the metrics described in section 2.4.1, we further analyzed forecasts for success and accuracy. Forecasts that were either not realized or realized outside of the allowable range were considered failures. The STFE and TFE metrics described above were then analyzed to determine what types of forecasts provided the best results.

[6] Analysts characterized the degree of interpretation using a Likert scale, which included five metrics: all interpretation, a lot of interpretation, moderate interpretation, little interpretation, and no interpretation. Forecasts with all interpretation lacked the specificity of language necessary to be assessable. A comments box was provided for analysts to explain their selection.

[7] Forecast was derived from David W. Schumann, Andy Artis, and Rachel Rivera, "The Future of Interactive Advertising Viewed Through an IMC Lens," *Journal of Interactive Advertising*, 1:2 (2001).

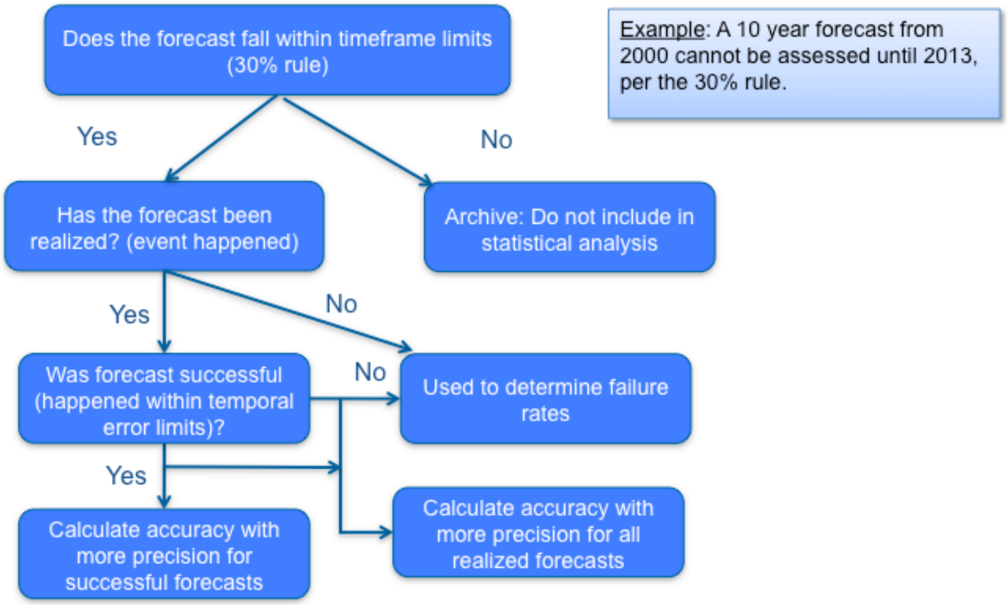

Does the forecast fall within timeframe limits (30% rule)
Example: A 10 year forecast from 2000 cannot be assessed until 2013, per the 30% rule.
Yes
No
Has the forecast been realized? (event happened)
Archive: Do not include in statistical analysis
Yes
No
Was forecast successful (happened within temporal error limits)?
No
Used to determine failure rates
Yes
Calculate accuracy with more precision for successful forecasts
Calculate accuracy with more precision for all realized forecasts

FIGURE 4. FORECAST ASSESSMENT PROCESS

2.4.3 STATISTICAL ANALYSIS PLAN

A statistical analysis was conducted to determine if a given forecast methodology was better than others given various conditions. The intent of the statistical analysis was to inform decision makers as to the most accurate forecasting method to use in the future. Our analytical plan consisted of three parts: testing forecast success rates, testing for the most accurate forecast methodology, and determining the key attributes of successful forecasts. These analyses are described in the following sections.

2.4.3.1 SUCCESS RATE ANALYSIS

We used a binomial test to compare each forecast methodology’s success rate with a hypothesized probability of success based on a random guess. Forecasting methods that failed to outperform the hypothesized test were considered as poor choices for technology forecasts. We continued the analysis by comparing success rates for the methodologies against each other, to determine if there was statistically significant evidence that some methods were more accurate than other methods.

2.4.3.2 TEMPORAL ERROR ANALYSIS

The TFE describes the magnitude of the forecast error and the STFE indicates the tendency to forecast early or late. Therefore, forecasting methodologies with a small average TFE are considered more accurate than methodologies with a large average TFE and those with positive average STFE are considered to generally predict too early while those with a negative average STFE are considered to generally predict too late. For all methodologies, we assessed whether one methodology was better (more accurate), than another by comparing mean accuracy of TFE and STFE.

2.4.3.3 KEY ATTRIBUTES ANALYSIS

The previous tests were designed to provide insights about the most appropriate forecasting methodology given a technology area tag and time frame. The last phase of the analysis was designed to elucidate whether, given a forecast with specific attributes, it is possible to predict with any confidence the accuracy of the forecast. To accomplish this, we conducted a multiple regression analysis as well as several data mining techniques to determine the influence (if any) of each attribute on the accuracy of a forecast. This analysis was conducted for each forecasting method and each technology area tag. We extended this analysis by including some of the other forecast attributes, such as TRL, publication type, and geographic region.

3.0 STUDY RESULTS

This section provides an overview of the results of all analyses, including forecast metrics and statistical analyses.

3.1.1 DOCUMENT COLLECTION METRICS

We extracted 2,278 forecasts from 300 forecast documents. Of these forecasts, 2,092 were assessable; these 2,092 came from 277 forecast documents. Ultimately, we were able to verify 1,058 statistically assessable forecasts from 217 documents (including three forecasts that were later determined to be statistical outliers). Of the 1,058 verified forecasts, approximately 70% (736) of the predicted events had been realized. Only five forecast documents contained more than 30 verified forecasts and the largest number of verified forecasts from a single document was 46. Each forecast document is carefully cited; a full list of the documents included in the analysis is provided in appendix C.[8]

Nearly 60% of the forecast documents in our sample set of 300 originated from North America. Approximately 25% of the documents in our sample originated from Asia and Europe and only 5% originated from Australia and New Zealand. We were unable to extract forecasts from documents originating in Africa or South America. Figure 5 shows the distribution of forecasts by region.

[8] Appendix C is provided in electronic format only as a set of four comma-separated value files.

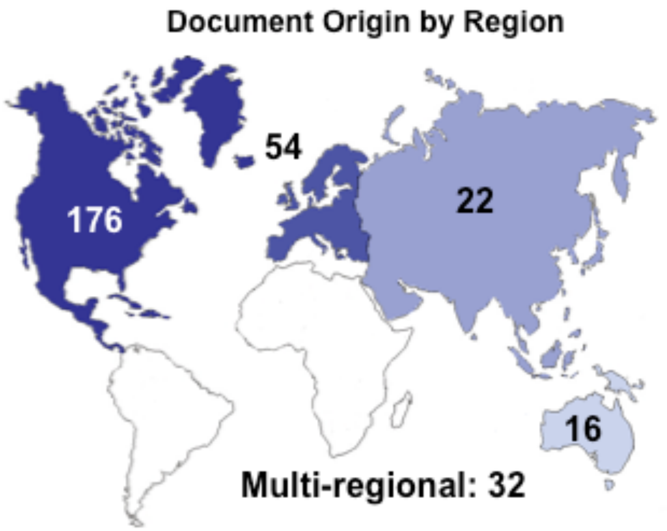


FIGURE 5. NUMBER OF FORECAST DOCUMENTS BY REGION OF ORIGIN

With the exception of market research firms, all publication types are well represented in our sample set. Academic publications represent 30% of the total forecast documents, with most other documents split between government publications, industry, and trade press, with a slightly smaller percentage for strategic analysis firms. Although market research firms produce many forecast documents, these documents are typically available for purchase only and there is sufficient information available from open sources to evaluate these documents. Figure 6 shows a breakout of forecast documents by publication type.

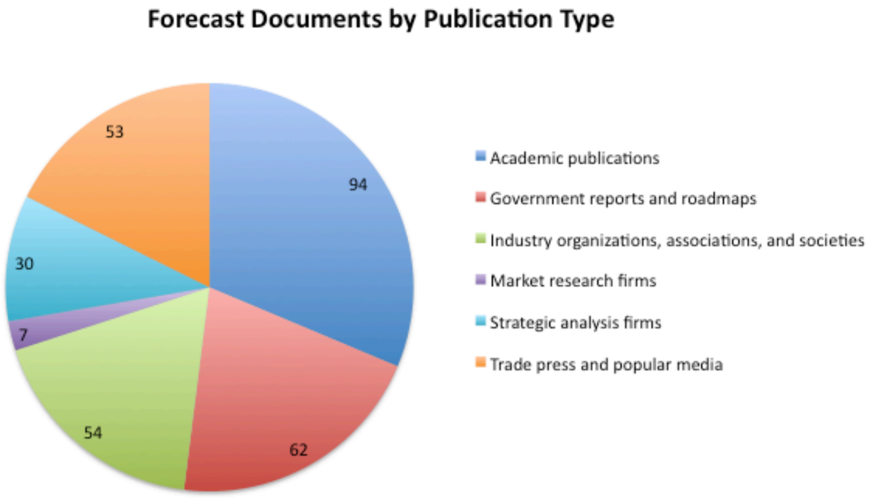


FIGURE 6. NUMBER OF FORECAST DOCUMENTS BY PUBLICATION TYPE

3.1.2 FORECAST COLLECTION METRICS

Each of the assessable forecasts was characterized by nine attributes identified in the initial study.[9] This section highlights some of the more informative findings associated with technological forecasts and their distribution among the nine attributes.

The forecast methodology attribute forms the basis for many of the statistical analyses described in section 4. Through this expansion effort we were able to collect a sufficient number of verified forecasts in each methodology to allow for comparative analysis. Our findings indicate that expert sourcing and expert analysis methods are the dominant forecast methodologies, with about 50% of assessable and verified forecasts falling into these methods. The gaming and scenarios methodology remains the least represented. However, the percentage of verified forecasts using this methodology increased significantly during this expansion. There are approximately 100 verified forecasts associated with all other methodologies, greatly increasing the statistical confidence in our findings as compared to the initial study. Figure 7 shows the distribution of assessable and verified forecasts by methodology.

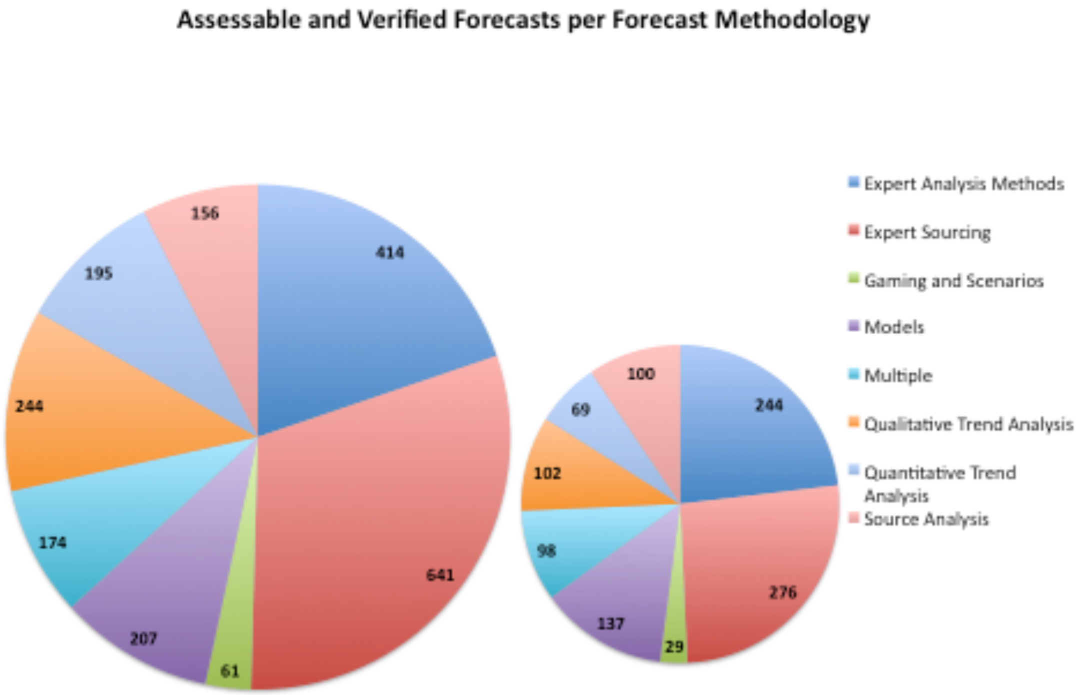

Assessable and Verified Forecasts per Forecast Methodology

FIGURE 7. NUMBER OF ASSESSABLE AND VERIFIED FORECASTS PER FORECAST METHODOLOGY

Compared to the initial study, the number of assessable and verified forecasts within most technology area tags increased significantly. Exceptions include production, photonic and phononic, and maritime transportation technologies, which have a relatively small number of

[9] The attributes are: forecast methodology, technology area tag, time frame, prediction type, geographic origin, geographic region forecasted about, technology readiness level (TRL), and technology complexity (system- versus component-level technologies).

assessable forecasts. The verification rate for forecasts about maritime transportation technologies was unusually low at 23%, compared to the average verification rate of between 40 and 60% for other technology area tags. More new forecasts for computer and communication technologies were identified than any other technology area and these areas now represent about 45% of the sample set. Figure 8 shows the distribution of assessable and verified forecasts by technology area.

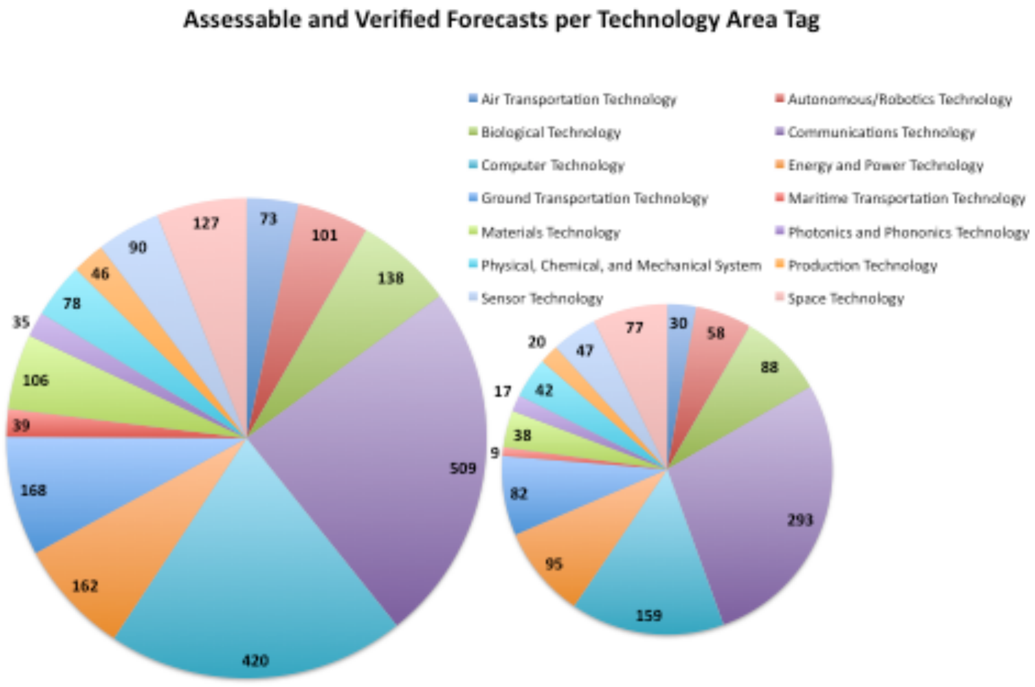


FIGURE 8. NUMBER OF FORECASTS PER TECHNOLOGY AREA TAG

The short term continues to be the dominant time frame used in forecasts, with more 60% of assessable and verified forecasts having predicted timeframes of less then five years. There are approximately 30% fewer verified long-term forecasts than medium-term forecasts in our sample set. Figure 9 shows the distribution of assessable and verified forecasts by timeframe.

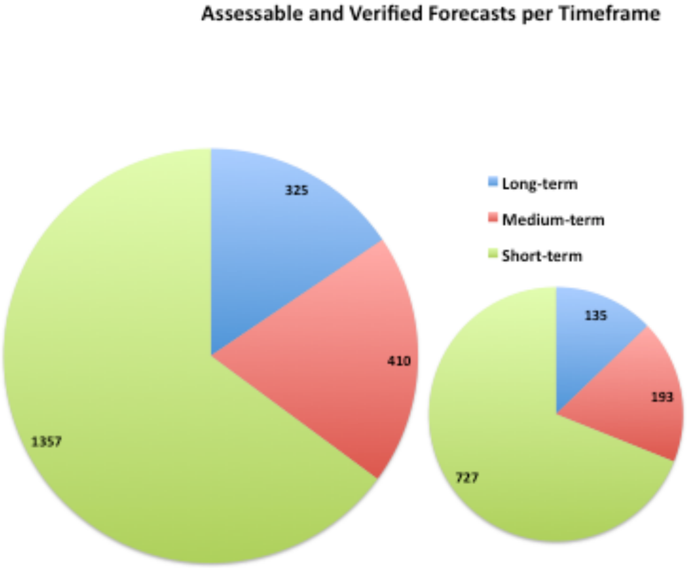


FIGURE 9. NUMBER OF ASSESSABLE AND VERIFIED FORECASTS PER TIME FRAME

Generally, there are more forecasts about mature technologies; our sample contains almost three times as many assessable forecasts associated with high-TRL technologies than low-TRL technologies—a trend that continued through forecast verification. A similar ratio is seen in technology complexity, with nearly three times as many verified forecasts that predict about systems or systems of systems technologies compared to subsystems and components. This ratio is closer to two for assessable forecasts.

3.2 COMPARATIVE ANALYSIS OF FORECAST METRICS

This section provides highlights from a comparative analysis of forecast metrics across the 1055 verified forecasts, excluding three outliers. These observations build on the findings of previous research using an initial dataset of 310 verified forecasts. Section 4 discusses the statistical significance of observations highlighted in this section.

Forecast time frame is positively correlated with success and forecast realization; both short- and medium-term forecasts are realized significantly more often than long-term forecasts, by almost 20 percentage points. Medium-term forecasts continue to perform slightly better than long-term forecasts with respect to success. However, short-term forecasts are significantly more successful than medium- or long-term forecasts by seven percentage points. The most significant change from the original report's findings with respect to time frame is that long-term forecasts are no longer statistically less successful than medium-term forecasts. Table 2 summarizes the success rates by time frame.

TABLE 2. SUCCESS RATE BY TIME FRAME

Time Frame	Not Realized	Realized	Successful
Short term	28%	72%	35%
Medium term	27%	73%	28%
Long term	46%	54%	27%
Average	30%	70%	33%

The realization rates for quantitative trend analysis and models are both well above average, with 81% (56 of 69) and 83% (114 of 137), respectively. Source analysis and gaming and scenarios also were above average, at about 75% realization compared to an average of 70%. The remaining methodologies were roughly consistent at about 62-65%, with expert sourcing methods slightly higher at 68%. These observations are roughly consistent with respect to success rates, where quantitative trend analysis performs the best at a 45% success rate. At 37%, models are also above the average success rate. For the most part, the remaining methodologies fall slightly below the average rate of 33%. Qualitative trend analysis is an exception, with a statistically significant low rate of 24%. The consistency of these findings differ from the initial study of 310 forecast, which indicated qualitative trend analysis and expert sourcing had significantly high realization rates and average success rates, further indicating convergence as additional forecasts were added.

When accounting for time frame, quantitative trend analysis—and models, to a lesser degree—remain top-performing methodologies. Quantitative trend analysis and models have the two highest realization rates in the short and medium terms, although both are below average in the long term. For success rates, quantitative trend analysis (at 56%) outperforms all other methodologies across all time frames except for models (at 60%) in the medium term. Forecasts generated from quantitative trend analysis do have a significantly high percentage of predictions about computer technologies (the technology area with the highest statistically significant success rate), with 46% of quantitative trend analysis forecasts falling into this technology area tag, compared to 21% for expert analysis methods, the methodology with the second highest percentage of forecasts about computer technologies. However, when comparing success rates for methodologies solely within the computer technology area tag, quantitative trend analysis performs slight below average, indicating the predominance of forecasts about computer technologies is not influencing the success rates of associated forecast methodologies. When quantitative trend analysis was compared to all other methodologies while correcting for technology area tag and time frame, it did not demonstrate a statistically better success rate. This is due in part to the small sample size of quantitative forecasts that did not project over the short term or make predictions about computer technologies.

TABLE 3. SUCCESS RATES BY METHODOLOGY

Methodology	Not Realized	Realized	Successful
Quantitative trend analysis	19%	81%	45%
Multiple	38%	62%	33%
Qualitative trend analysis	35%	65%	24%
Expert sourcing	32%	68%	32%
Models	17%	83%	37%
Expert analysis methods	37%	63%	32%
Gaming and scenarios	24%	76%	28%
Source analysis	26%	74%	31%
Average	30%	70%	33%

In general, success rates for the technology area tags are similar to the average success rate of 33%, indicating that technology types have limited impact on forecast success. Maritime transportation technology has the largest percentage of successful forecasts at 56%. However, with only nine forecasts about this technology, the high success rate is likely a product of small sample size. The computer technology tag has the largest percentage of successful forecasts with a statistically significant number of forecasts, with 40% of the 159 forecasts successful (seven points above average). Photonics and phononics technology forecasts had the lowest success rate at 18%. However, this technology area tag also had a low sample size of 17 forecasts. Sensor technology had the second lowest success rate at 9% below average. Realization rates showed less convergence, with computer, energy and power, maritime, and materials technologies above 80%. Interestingly, both energy and power technologies and materials technologies have below average success rates of 28 and 29%, respectively.

The degree of interpretation metric was developed to provide insight into the vagueness of forecasts. However, in practice it proved to be impossible to separate forecast vagueness from the effort required to verify a forecast, which is sometimes due to reasons other than ambiguous language. We found that degree of interpretation does not provide a useful comparison for attributes, especially through the broad sourcing process. In general, our process was able to exclude forecast that were too vague or required significant effort to verify, as indicated by the low number of forecasts that required a lot of interpretation. The remaining values for degree of interpretation are fairly consistent across publication types, as shown in figure 10. As noted above, our sample set was comprised of few forecasts from market research firms; forecasts in this category were extracted from seven documents. There is therefore too little data to draw conclusions about the different distributions.

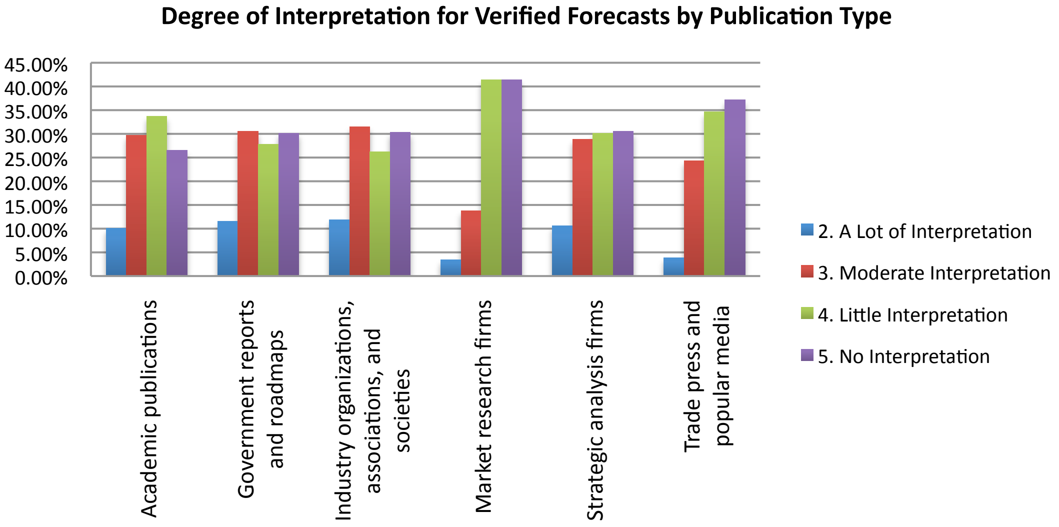


FIGURE 10. DEGREE OF INTERPRETATION BY PUBLICATION TYPE

4.0 KEY FINDINGS

Section 3 described the data in our sample. We applied two series of analysis—success test and temporal error test—to determine if the observations made in section 3 could be extended to the population of forecasts, thereby allowing us to draw conclusions about the population of forecasts. We also conducted classical analysis of potential predictive models that could be used to assess forecast accuracy based on key attributes. Developing a meaningful model would enable the prediction of forecast accuracy with some level of confidence. This section provides the results of these analyses.

4.1 TEST FOR SUCCESS

We defined a successful forecast as one whose prediction was realized within 30% of the forecasted timeframe. Any verified forecast whose prediction had not yet been realized or was realized outside of the 30% range was considered a failed forecast. Our analysis considered multiple factors and attributes to determine whether some types of forecasts demonstrated significantly higher success rates than other types. The success analysis consisted of two parts:

1. A test against a control to determine if forecasts provide more information than an uninformed guess, and
2. Tests against each other to determine if some forecasting attributes have higher success rates than others

4.1.1 TEST AGAINST A CONTROL

Consistent with the previous study, we bounded the ranges of the uninformed random guess by the 99th percentile of forecast horizons and the 95th percentile of the temporal errors (see appendix E for the derivation of the values of the revised uninformed random guess). Table 4 shows the comparison of the sample set for this study and the previous study.

TABLE 4. SAMPLE SET COMPARISON

Measure	Previous Study	Current Study
Upper limit on forecast horizon	20 years	25 years
Upper limit on forecast temporal error	10 years	15.5 years
Forecasts excluded for lengthy horizon	4	4
Forecasts excluded for lengthy TFE	11	36
Adjusted sample size	**295**	**1,018**
ρ	.247	.221

Appendix E provides a list of all forecasts excluded from this test and the reason that each was excluded.

To evaluate success, we conducted a two-tailed binomial test and developed our confidence intervals using Wilson's method with a continuity correction. We allowed ρ to be the success rate for an uninformed random guess, which we compared to the number of successes and observations for each methodology. Thus, our hypothesis test was:

$$H_0 : r_i = \rho$$
$$H_A : r_i \neq \rho$$

Where: r_i is the observed success rate for methodology i, and
ρ is the expected success rate for a random guess

In this test, we rejected the null hypothesis if there was an observed success rate for a given methodology that was so low that it was unlikely to have been generated by a binomial distribution with a probability of success equal to ρ. Failing to reject the null hypothesis indicated that there was sufficient evidence that r_i was approximately ρ. Table 5 provides the results of the binomial test for each of the methodologies compared to ρ=.221.

TABLE 5. BINOMIAL TEST FOR FORECAST METHODOLOGIES

Method	Successes	Failures	Success Rate	p-value
Quantitative trend analysis	31	37	0.456	0.000
Models	51	84	0.378	0.000
Expert sourcing	88	178	0.331	0.000
Expert analysis methods	79	160	0.331	0.000
Multiple	32	66	0.327	0.015
Source analysis	31	64	0.326	0.018
Gaming and scenarios	7	19	0.269	0.635
Qualitative trend analysis	24	67	0.264	0.314
Random guess	**-**	**-**	***0.221***	**-**

Gaming and scenarios and qualitative trend analysis were the two forecasting methods for which we failed to reject the null hypothesis; there is insufficient evidence in the observed sample set to assert that these two methods perform better than an uninformed guess. In the previous study, the source analysis method was the only one for which we failed to reject the null; its success rate is substantially greater in this larger data set. During the previous study, qualitative trend analysis (QLA) was one of the better performing methods during this study. In this data set, however, QLA is the worst performing method. We attribute this significant shift in performance to the fact that QLA had a relatively small sample size (14 out of 295) in the previous study. Finally, in the previous study, quantitative trend analysis (QTA) was the best performing method with a 64% success rate. Its success rate has dropped significantly (nearly 20 percentage points) but it still presents the best success rate of all methods studied.

Table 6 compares the success rates of forecast methodologies during the previous study to the new success rates. We note in yellow the methods with significant changes in success rates between the two studies (QTA, multiple forecasting methodologies, source analysis, and QLA). The success rate associated with QTA methods decreased by nearly 20% but is still significantly higher than all other methods. The multiple and qualitative trend analysis methods also experienced decreased success rates. The changes for the multiple and qualitative trend analysis were expected since they had small sample sizes in the previous study.

TABLE 6. COMPARISON OF FORECAST METHODOLOGY SUCCESS RATE

Method	New Sample Set		Previous Sample Set	
	Population	Success Rate	Population	Success Rate
Quantitative trend analysis	68	0.46	28	0.64
Models	135	0.38	17	0.35
Multiple	266	0.33	6	0.5
Expert analysis methods	239	0.33	71	0.32
Expert sourcing	98	0.33	115	0.38
Source analysis	95	0.33	28	0.14
Qualitative trend analysis	26	0.27	14	0.43
Gaming and scenarios	91	0.27	16	0.31
Overall	1,018	0.34	295	0.37

Our probability of success for a randomly generated uninformed guess assumes a uniform distribution from one to 25 years for the forecast horizon. In other words, to be completely uninformed, we assume that the population of forecast time horizons is evenly distributed among the 25-year span. However, our data set indicates that the distribution of forecasts is weighted more towards the short-term horizon than the medium- and long-term horizons, as seen in figure 11.

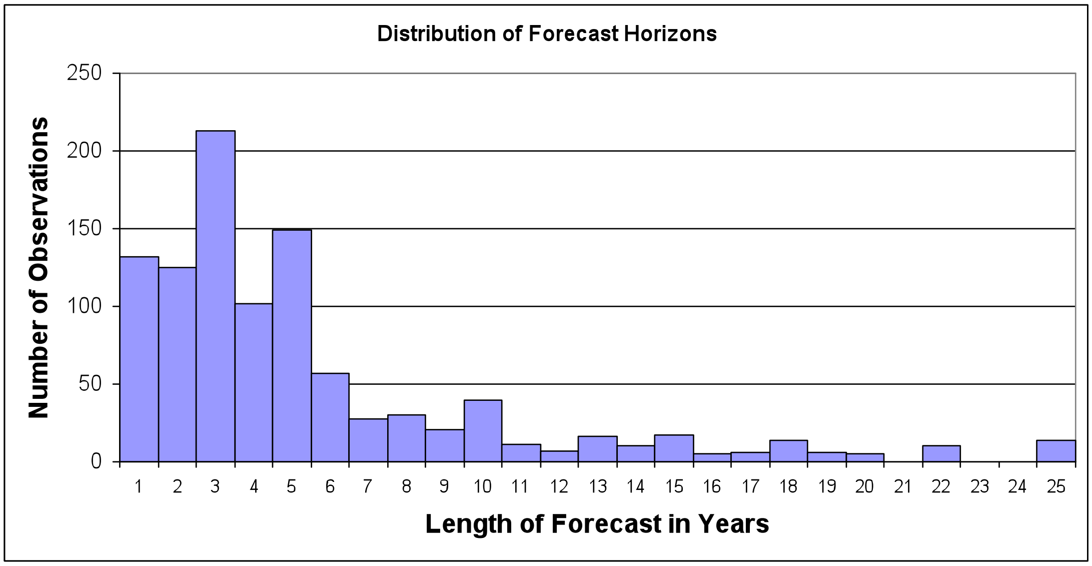


FIGURE 11. DISTRIBUTION OF THE FORCAST HORIZON

Thus, our assumption of uniformity unintentionally biases the test in favor of the random guess since it places more guesses in the long-term than what could be expected given our observations. The bias occurs because of the allowable range. Given our definition of success, forecasts with a longer horizon have a higher probability of success than do short-term forecasts. For example, a forecast with a two-year horizon has a 7.5% chance of being successful, whereas a 20-year forecast has a 32.5% chance of being successful. By placing more forecasts in the long term than what could be expected, we inflate the random guess' probability of success. If we allow our random guess to be informed by the empirical distribution of forecast horizons (weight the probability of each year's outcome by its observed frequency), our informed random guess has a much lower success rate ($\rho = .111$) since it has a higher proportion of short-term guesses. When testing against the success rate of the informed random guess, we reject H_0 for all methods (sufficient evidence that all forecast methods are better than a guess if our data set is representative of the distribution of technology forecast horizons). Figure 12 shows the observed success rate for each method along with a 95% confidence interval for each success rate. The horizontal lines represent the theoretical success rates for both the uninformed and informed random guesses.

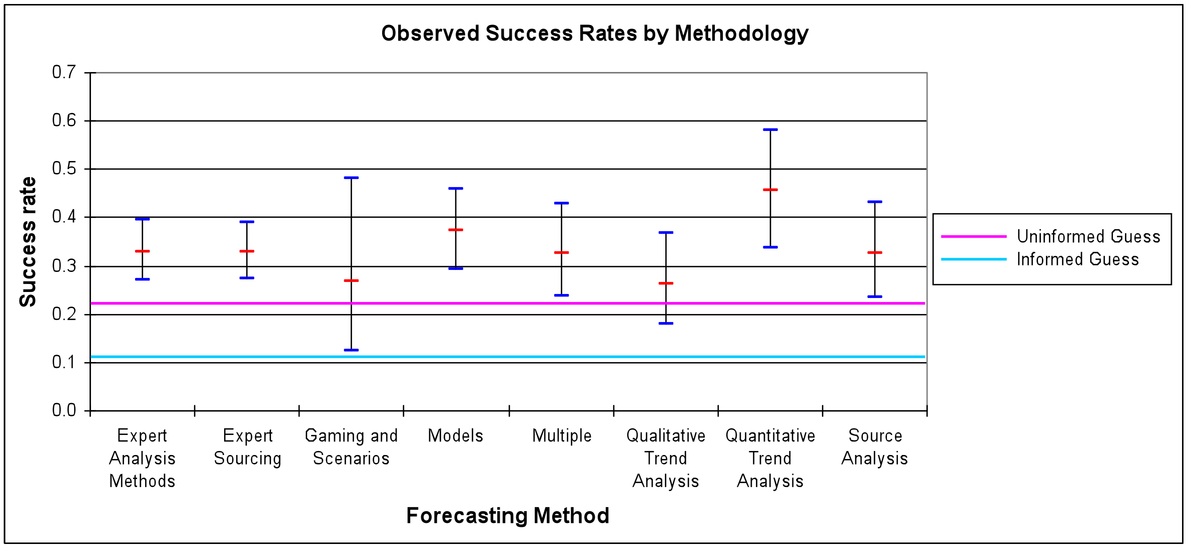


FIGURE 12. OBSERVED SUCCESS RATES BY FORECASTING METHODOLOGY COMPARED TO BOTH THE UNINFORMED AND INFORMED RANDOM GUESS.

The 30% allowable range used to determine the success or failure of a forecast is based upon a subjective assessment by the government that a forecast with an error outside of that range would be of limited value. To determine the sensitivity of the results with respect to the allowable range, we varied the allowable range in 10% increments from 0% allowable range (wherein the forecaster must get the year exactly right) to 100% allowable range (wherein the predicted event must occur anytime within the forecast horizon) and reevaluated the success/failure metric for each methodology as well as the hypothetical probability of success given an uninformed guess and informed guess. Table 7 provides the results of the sensitivity analysis. The values in the table indicate the range over which the findings reported in figure 11 are valid.

TABLE 7. SENSITIVITY ANALYSIS

Method	Uninformed Guess		Informed Guess	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit
Expert analysis methods	0%	40%	0%	100%
Expert sourcing	0%	50%	0%	100%
Gaming and scenarios	0%	100%	10%	100%
Models	0%	100%	0%	100%
Multiple	0%	30%	0%	100%
Qualitative trend analysis	30%	100%	0%	80%
Quantitative trend analysis	0%	100%	0%	100%
Source analysis	0%	30%	0%	100%

Using the uninformed guess, these results mean that if a 30% allowable range is too high, then all findings would hold except for the finding that QLA may not perform better than an uninformed guess. If a 30% allowable range is considered too low, the current findings will change as the allowable range increases. For example, the finding that forecasts that were generated from multiple methodologies are better than a guess does not hold for allowable ranges above 30%. The number of changes close to 30% indicates that these findings are sensitive to the selection of the allowable range value.

When considering the informed guess, our reported findings are robust with respect to the selection of the allowable range values, indicating that if the distribution of the forecasts in the sample set is representative of the population of forecasts, then all forecast methods are better than a guess.

4.1.2 COMPARING METHODS TO EACH OTHER

For the test against the control (success test), we used a trimmed data set of 1,018 forecasts to maintain consistency with the previous study. For this section of the analysis, we use the full data set (1,058 forecasts minus three outliers). Table 8 shows the distribution of all 1,055 forecasts by methodology and each method's success rate. The values are not significantly different than the values reported under the control analysis and all observations that hold for the reduced set of 1,018 forecasts hold for this data set as well. The success and failure observations from the full data set shown in table 8 will be used for subsequent analysis.

TABLE 8. DISTRIBUTION OF THE UNTRIMMED FORECAST DATA SET BY METHODOLOGY

Method	Total	Success	Failure	Success Rate
Quantitative trend analysis	69	31	38	0.45
Models	137	51	86	0.37
Multiple	98	32	66	0.33
Expert analysis methods	244	79	165	0.32
Expert sourcing	276	88	188	0.32
Source analysis	100	31	69	0.31
Gaming and scenarios	29	8	21	0.28
Qualitative trend analysis	102	24	78	0.24

QTA has a higher success rate than the other methods. To determine whether this was a statistically significant observation, we used Fisher's Exact Test to compare the success rate of QTA to those of the other methodologies. We used the following hypothesis test for this comparison:

$H_0 : r_t \le r_s$ Where: r_t is the observed success rate of the QTA method, and
$H_A : r_t > r_s$ r_s is the observed success rate of each other method

We rejected the null hypothesis if there was sufficient evidence to indicate that it would be unlikely to observe a large difference in the number of successes if both methods had similar probabilities of success. This would imply that QTA has a true probability of success greater than the other methods. We conducted these tests with the success rates associated with the 30% rule. Table 9 provides the results of this test.

TABLE 9. FISHER'S EXACT TEST COMPARING QTA TO OTHER FORECASTING METHODS

QTA compared to	*p-value*
Qualitative trend analysis	0.003
Expert sourcing	0.030
Expert analysis	0.038
Source analysis	0.046
Multiple	0.074
Gaming and scenarios	0.083
Models	0.180

There is statistically significant evidence (at a level of significance of .05) that QTA methods have a true probability of success greater than expert sourcing, expert analysis, source analysis, and qualitative trend analysis. The gaming and scenarios method has a success rate that is nearly 20% lower than the QTA method. However, we fail to reject H_0, meaning the observation that QTA is better than the gaming and scenarios method is not statistically relevant. The high p-value, however, is a product of the small sample size (29) for the gaming and scenario method relative to the other sample sizes.

4.2 SUCCESS RATE ANALYSIS BY TECHNOLOGY AREA TAG

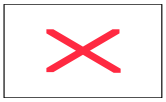
The technology area tag that is predicted about is a key attribute associated with forecasts, and we sought to determine whether some technology area tags have a higher success rate than others. Table 10 describes the success and failures of all 1,055 forecasts distributed by technology area tags.

TABLE 10. DISTRIBUTION OF FORECAST SUCCESS AND FAILURES BY TECHNOLOGY AREA TAG

Technology Area Tag	Total	Success	Failure	Success Rate
Maritime transportation technology	9	5	4	0.555
Computer technology	159	63	96	0.396
Autonomous/robotics technology	58	21	37	0.362
Communications technology	293	104	189	0.354
Production technology	20	7	13	0.350
Air transportation technology	30	10	20	0.333
Physical, chemical, and mechanical system	42	13	29	0.309
Space technology	77	23	54	0.298
Materials technology	38	11	27	0.289
Biological technology	88	25	63	0.284
Energy and power technology	95	27	68	0.284
Ground transportation technology	82	21	61	0.256
Sensor technology	47	11	36	0.234
Photonics and phononics technology	17	3	14	0.176

Maritime transportation technology has a very high success rate compared to all other forecasts but its sample size is very small relative to the other technology areas. We therefore exclude maritime transportation for consideration as having the highest success rate.

In the previous study, computer and autonomous/robotics (CAR) technology area tags had statistically higher success rates than other technology area tags. In the larger data set, CAR again is the most successful. To determine whether this was a statistically significant observation, we used Fisher's Exact Test to compare CAR to the success rate of all other technology area tags combined, and then against each of the other methods. Finally, we compared the computer technology area tag by itself against each of the other technology tags. We used the following hypothesis test for this comparison:



Where: r_{car} is the observed success rate of the CAR technology area tags, and
r_o is the observed success rate of each/all other technology area tags

Table 11 provides the results of the CAR tests while table 12 provides results of the comparison of computer technology forecasts to each of the technology area tags. Given the small sample size for maritime transportation technology forecasts, we excluded it from individual technology area tag comparisons. Instead, it is included in the "all other technology area tags" category.

TABLE 11. FISHER'S EXACT TEST COMPARING CAR TECHNOLOGY AREA TAGS TO OTHER TECHNOLOGY TAGS

CAR compared to	*p-value*
All other technology area tags	0.02
Ground transportation technology	0.02
Sensor technology	0.03
Energy and power technology	0.05
Biological technology	0.06
Photonics and phononics technology	0.07
Space technology	0.11
Materials technology	0.17
Physical, chemical, and mechanical system	0.22
Communications technology	0.26
Air transportation technology	0.36
Production technology	0.47

Table 12 indicates that CAR technology area tags have a higher success rate than all other tags combined, which is consistent with the findings of the previous study. However, when compared against each individual technology area tag, its success rate is not significantly better than the majority of the technology area tags.

TABLE 12. COMPARISON OF COMPUTER TECHNOLOGY AREA TAG FORECASTS TO ALL OTHER TECHNOLOGY AREA TAGS

Computer technology compared to	*p-value*
All other technology areas	0.03
Ground transportation technology	0.02
Sensor technology	0.03
Biological technology	0.05
Energy and power technology	0.05
Photonics and phononics technology	0.06
Space technology	0.09
Materials technology	0.15
Physical, chemical, and mechanical system	0.20
Communications technology	0.22
Air transportation technology	0.33
Autonomous/robotics technology	0.38
Production technology	0.44

Between the two technology areas included in CAR, the computer technology area tag had the highest success rate. However, the results indicate that while computer technology forecasts do appear to come from a different distribution than do all other forecasts consolidated, we cannot assert with statistical significance that forecasts about computer technologies have a higher success rate than each of the other technology area tags.

4.3 SUCCESS RATE ANALYSIS BY TIMEFRAME

Timeframe is the third key attribute of technological forecasts. Table 13 provides the success rates associated with the 1,055 forecasts distributed by timeframe.

TABLE 13. DISTRIBUTION OF FORECAST SUCCESS RATES BASED ON TIMEFRAME

Time Frame	Total	Success	Failure	Success Rate
Short term	727	254	473	0.35
Medium term	193	54	139	0.28
Long term	135	36	99	0.27

Short-term forecasts have a higher success rate. To determine if there was statistical significance associated with this observation, we used Fisher's Exact Test and the following hypothesis test to evaluate the observation:

$H_0 : r_s \le r_o$ Where: r_s is the observed success rate of the short-term forecasts, and

$H_A : r_s > r_o$ r_o is the observed success rate of each timeframe

Table 14 provides the results of the test. There is statistical significance to the observation that short-term forecasts have a higher success rate than do medium-term and long-term forecasts.

TABLE 14. FISHER'S EXACT TEST COMPARING SHORT-TERM FORECAST SUCCESS RATES TO OTHER TIMEFRAME SUCCESS RATES

Short-term compared to	*p-value*
Long term	0.037
Medium term	0.040

4.4 SUCCESS RATE ANALYSIS: AD HOC QUERIES

During the original study, we noticed that QTA had a significantly higher percentage of short-term forecasts than did other forecast methods. Table 15 describes the distribution of forecast methods by time frame for the new data set. QTA no longer stands out as having a disproportionately high number of short-term forecasts. Moreover, two other methods (gaming and scenarios and QLA) have a disproportionately high number of long-term forecasts. Given that the length of a time horizon does indicate a difference in success rate, we reviewed the impact of forecast time frame with respect to methodology. The last three columns of table 15 provide the success rate for each method broken out by time frame.

TABLE 15. DISTRIBUTION OF FORECASTS BY METHODOLOGY AND TIMEFRAME

Method	% by Time Frame			Total	Success Rates			
	Short term	Medium term	Long term		Overall	Short term	Medium term	Long term
Quantitative trend analysis	71.0%	13.0%	15.9%	69	0.449	0.449	0.556	0.364
Models	86.9%	10.2%	2.9%	137	0.372	0.378	0.357	0.250
Multiple	83.7%	5.1%	11.2%	98	0.327	0.317	0.600	0.273
Expert analysis	75.4%	20.5%	4.1%	244	0.324	0.337	0.300	0.200
Expert sourcing	56.5%	30.1%	13.4%	276	0.319	0.340	0.277	0.324
Source analysis	80.0%	11.0%	9.0%	100	0.310	0.375	0.000	0.111

Method	% by Time Frame			Total	Success Rates			
	Short term	Medium term	Long term		Overall	Short term	Medium term	Long term
Gaming and scenarios	69.0%	0.0%	31.0%	29	0.276	0.250	-	0.333
Qualitative trend analysis	36.3%	20.6%	43.1%	102	0.235	0.297	0.143	0.227

Regardless of timeframe, QTA maintains its rank as the method with the highest success rate. However, from a statistical significance perspective, even at a level of significance of 0.1, we cannot state that QTA comes from a different distribution than do all other forecasting methods when short-term forecasts are removed from the sample set. This implies that QTA benefits from its performance in short-term forecasts and is not necessarily better at forecasts with longer horizons. We attribute this finding, however, to the small sample size (20) of QTA when short-term forecasts are removed from consideration.

During this analysis, 59% of the forecasts in which the QTA method was applied were in the CAR technology area tag. Since both CAR and QTA exhibited higher than average success rates, we sought to determine if there was a cause and effect relationship between the CAR technology area tag and the QTA method. We first modified the data set by removing all QTA forecasts and then compared CAR forecasts to all other technology area tags. Table 16 provides the results of this test.

TABLE 16. COMPARISON OF CAR TECHNOLOGY AREA TAG FORECASTS TO OTHER FORECASTING TECHNOLOGY TAGS WHEN QUANTITATIVE TREND ANALYSIS FORECASTS ARE REMOVED FROM SAMPLE SET

CAR compared to:	p-value		Success Rate	
	w/QTA	w/o QTA	w/QTA	w/o QTA
CAR technologies			.387	.402
Communications technology	0.257	0.194	0.355	0.352
Production technology	0.472	0.429	0.350	0.350
Air transportation technology	0.361	0.622	0.333	0.409
Physical, chemical, and mechanical system	0.220	0.193	0.310	0.308
Space technology	0.105	0.091	0.299	0.299
Materials technology	0.167	0.102	0.289	0.270
Biological technology	0.057	0.011	0.284	0.238
Energy and power technology	0.052	0.059	0.284	0.290
Ground transportation technology	0.023	0.028	0.256	0.263
Sensor technology	0.032	0.029	0.234	0.234
Photonics and phononics technology	0.066	0.058	0.176	0.176

With the exception of air transportation technology (22 observations), all other technology area tags demonstrated marginal change (increases and decreases) in success rates and p-values. Success rates among air transportation technology forecasts actually improved with the removal of the QTA forecasts. This implies the success rates for forecasts about CAR technologies does not benefit by the disproportionately large number of QTA forecasts.

We next modified the data set by removing CAR forecasts. We then compared QTA forecasts to all other methods to determine if CAR technology forecasts were responsible for QTA's high success rate. Table 17 provides the results of this analysis.

TABLE 17. COMPARISON OF QTA FORECASTS TO OTHER FORECASTING METHODS WHEN CAR TECHNOLOGY TAG FORECASTS HAVE BEEN REMOVED FROM THE SAMPLE SET

QTA compared to:	p-value		Success Rate	
	w/CAR	w/o CAR	w/CAR	w/o CAR
QTA			0.449	0.429
Models	0.180	0.283	0.372	0.35
Multiple	0.074	0.068	0.327	0.25
Expert analysis methods	0.038	0.155	0.324	0.311
Expert sourcing	0.030	0.178	0.319	0.322
Source analysis	0.046	0.209	0.310	0.323
Gaming and scenarios	0.083	0.150	0.276	0.259
Qualitative trend analysis	0.003	0.042	0.235	0.232

With the exception of forecasts that used multiple methodologies (72 observations), all other success rates demonstrated only marginal changes (increases and decreases) when CAR forecasts were removed from the data set. However, p-values did exhibit substantial changes resulting in no statistically significant differences in observed success rates, implying that QTA performance is a result of the disproportionately large number of CAR forecasts. However, the large p-value changes could not have been influenced by marginal shifts in the observed success rates. By removing CAR technology forecasts from the sample set, QTA's sample size dropped from 69 observations in the full data set to 28 observations in the revised data set. We believe these p-value shifts are a result of the small sample size for the QTA forecasts.

Based on the results of these *ad hoc* queries on the success rate of forecasts, QTA method is a better forecasting method than the other methods analyzed. Its higher success rate is not attributed to the timeframes in which the forecasts were distributed and is not dependant upon the technology area tags in which its forecasts were distributed. Additionally, short-term forecasts have a higher success rate than do longer-term forecasts.

4.5 TESTS FOR TEMPORAL ERROR

The previous test evaluated the performance of forecast types with respect to success rates. These success rates were based off of a binary attribute that only indicates whether a forecast produced results within a given threshold. A related but more informative assessment of forecast performance is one based on the difference in time (years for this study) between the expected year of the forecasted event occurring and the year it actually occurred. Forecasts defined by sets of attributes that generally result in small temporal errors would be better than forecasts defined by attributes that generally result in large temporal errors.

For this series of tests, only those forecasts whose predicted events have been realized can be assessed, since these are the only forecasts for which the temporal error is known with some level of certainty. There are 736 forecasts from our sample of 1,055 that have been realized. Figure 13 shows the distribution of the temporal errors associated with each of these forecasts.

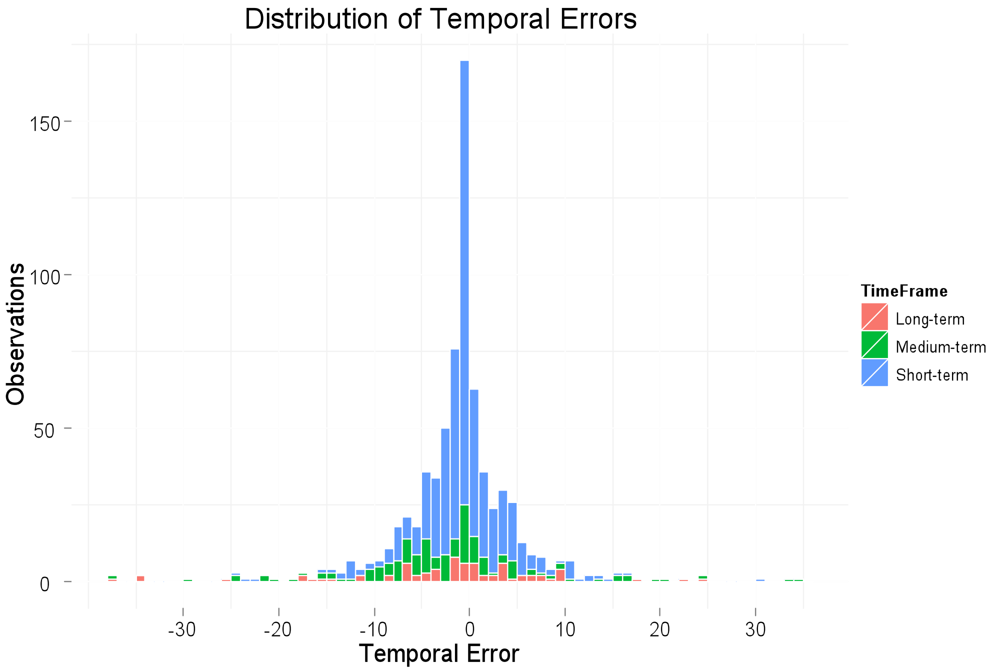


FIGURE 13. DISTRIBUTION OF FORECAST TEMPORAL ERROR BY TIMEFRAME

Visual inspection of the descriptive statistics associated with each method and timeframe in table 18 indicated that some methods and timeframes might have come from different distributions than did other values (candidate values highlighted in yellow). To determine if there was a difference in the means of the underlying distributions of the attributes being investigated, we used the ANOVA test. The ANOVA requires the underlying distributions for each of the attributes being investigated to be normally distributed and have equal variance. Based on the results of both the Wilkes-Shapiro test for normality and the Kolmogorov-Smirnoff goodness of fit test, we have sufficient evidence that the data are not normally distributed. Research suggests that the ANOVA is sufficiently robust to overcome the normality assumption but that it is not robust against violations of equal variance[10]. Table 18 provides the TFE and STFE variances for each method and timeframe in the data set. The underlying variances are far too great for the assumption of equal variance.

TABLE 18. DESCRIPTIVE STATISTICS OF FORECASTS DISTRIBUTED BY FORECAST METHOD

Attribute	Realized Forecasts	Mean		Variance		Range STFE	
		STFE	TFE	STFE	TFE	Min	Max
Methodology							
Expert analysis methods	244	-0.12	3.56	37.17	24.47	-22	35
Expert sourcing	276	-1.21	4.47	48.55	29.92	-37	25
Gaming and scenarios	29	-2.23	6.14	113.04	78.79	-37	8
Models	137	-0.25	1.90	7.20	3.61	-8	9
Multiple	98	0.26	3.90	35.77	20.37	-15	14
Qualitative trend analysis	102	-1.23	7.11	112.99	63.24	-34	31
Quantitative trend analysis	69	-1.41	3.27	28.97	20.13	-23	15
Source analysis	100	-0.59	4.55	66.41	45.80	-29	34

[10] G.E. Box, "Non-Normality and Tests on Variances," *Biometrika*, 40: 3-4 (1953): 318-335.

Attribute	Realized Forecasts	Mean		Variance		Range STFE	
		STFE	TFE	STFE	TFE	Min	Max
Timeframe							
Long term	135	-1.77	7.16	117.27	68.52	-37	25
Medium term	193	-2.06	6.84	98.80	55.91	-37	35
Short term	727	-0.17	2.84	21.97	13.91	-24	31

When data are not normally distributed, the Kruskal-Wallis (KW) test can be used as a non-parametric equivalent of the ANOVA. The KW test requires the underlying distributions have similar shapes and ranges. Visual inspection of the data distributed by attribute reveals unimodal distributions with approximately similar ranges, but different directions and magnitude of skewness, indicating the underlying distributions are not similar.

As the data do not fully support the use of either test, and lacking a credible alternative test for non-normally distributed and non independent and identically distributed variables, we default to the ANOVA test since generally parametric test have a higher power than do non-parametric tests. To compensate for the significant differences in variance, we transformed the data by taking the $\log_{10}$ of each STFE and TFE value and conducted the tests on the transformed data. Results indicated that for the TFE, there were statistically significant differences based on the technology area tag, methodology, and time frame. With respect to the STFE, only technology area tag and time frame resulted in statistically significant differences in the underlying distribution. Table 19 provides the result of all six tests.

TABLE 19. RESULTS OF THE ANOVA TEST WITH RESPECT TO DIFFERENCES IN FORECAST MEANS

Attribute	Variable of Interest	F statistic	p-value
Method	TFE	7.68	0.000
Tech area tag	TFE	3.12	0.001
Time frame	TFE	36.77	0.000
Method	STFE	1.68	0.115
Tech area tag	STFE	2.25	0.012
Time frame	STFE	5.57	0.005

The ANOVA informs there are differences, but does not indicate which methods, technology areas, and timeframes are different. To determine this, we used the Tukey-Kramer Honestly Significant Difference Test (TKHSD) to determine which attribute values were different. Based on the results of this test there was statistical evidence that shows:

1) QLA has the worst (largest) temporal error of all methods assessed, and
2) Short-term forecasts have the best (smallest) temporal error of the three timeframes assessed.

All other comparisons resulted in inconclusive results. While the gaming and scenarios method clearly appears to be similar to the QLA method with respect to performance, its small sample size countered any differences that may have been present. The models method only exhibited superiority when compared to expert sourcing and expert analysis. In all other comparisons between the modeling method and the remaining methods, there was insufficient evidence to draw any conclusions at a significance level of 0.10.

Surprisingly, in spite of its performance on the success/failure tests, the QTA method did not come from a different distribution with respect to TFE.[11] After inspecting the data for a possible cause, we found that six forecasts that used QTA had very large STFE/TFE (in excess of 10 years) that resulted in a large variance that its sample size could not overcome. While the ANOVA test did indicate there were significant differences in the technology area tag attribute and the TKHSD test revealed some pairings of technology area tag means that were statistically different, no information from either test indicated there was a universally better technology area tag with respect to forecast temporal error.

Given the two findings from this test, we inspected the data more closely to see if we could discern any cause for these findings. While inspecting the data, we noticed that only 3% of all forecasts for the QLA method had a horizon of one year. This is about four times smaller than the average for all other methods (12%). Many of these one-year forecasts were unsuccessful, but only had a TFE of one year.[12] A one-year forecast is perhaps predicting an event for which there is substantial evidence, and it therefore may be easier to generate an accurate forecast than it would with a lengthy time frame. In fact, both the gaming and scenarios method (0% of which were one-year forecasts) and the QLA method (3% of which were one-year forecasts) had very large TFE and STFE variances while the modeling method (30% of which were one-year forecasts) had very tight variances for the TFE and STFE.

To determine whether the number of one-year forecasts was influencing the results of the TKHSD test, we conducted the ANOVA and TKHSD on only those forecasts with a horizon in excess of one year. While the ANOVA did indicate there were underlying differences in the methods, the TKHSD did not reveal any differences at the 95% confidence level. This finding indicates that the results of the previous tests (with all forecasts) were influenced by the number of one-year forecasts within the sample set. Results of the ANOVA and TKHSD on the modified data set with respect to timeframe confirmed the previous finding that short-term forecasts were in fact more accurate than were other longer-term forecasts. As a result, the only finding of statistical significance with respect to TFE and STFE is that short-term forecasts exhibit the least temporal error.

4.6 IDENTIFICATION OF KEY ATTRIBUTES

The final component of the analysis consisted of developing a model based on the nine attributes that could be use to predict forecast accuracy. Similar to the previous study, one- and two-dimensional plots of data revealed no identifiable trends. Figure 14 shows the distribution of all forecasts by method and time frame in years. While the data do appear to be symmetric about the abscissa, in contrast to the previous study, there is statistical evidence that forecasts are generally more pessimistic (predict events will occur later than they actually do), with 25% more forecasts having a negative STFE. The negative slope on the regression line is based on the fact that there are nearly twice as many pessimistic forecasts in the medium-term as there are optimistic forecasts and the presence of significant STFE values in the 20- through 25-year time frame.

[11] The tests were not run on STFE since the ANOVA found no statistically significant differences in the method attribute.

[12] A one-year forecast has to get the year exactly right to be successful.

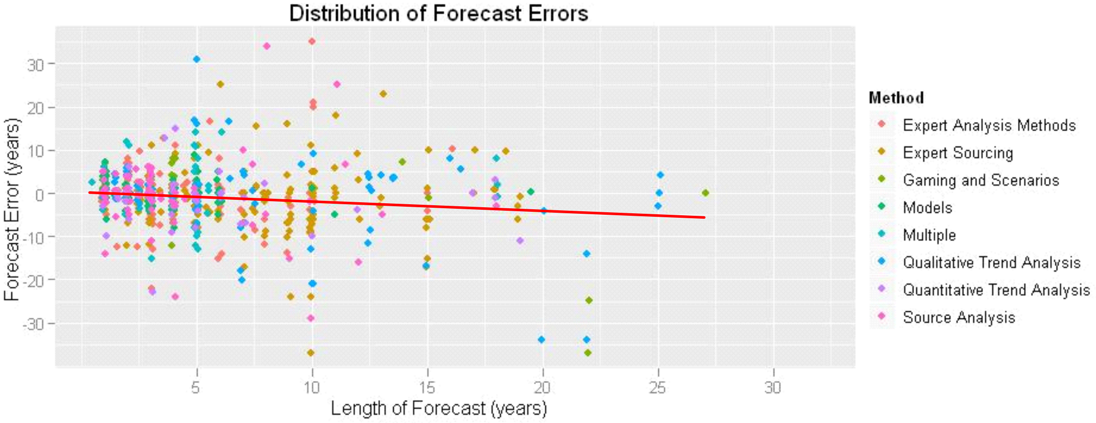


FIGURE 14. DISTRIBUTION OF FORECASTS BY METHOD AND TIME FRAME

REGRESSION ANALYSIS

We conducted an exhaustive regression analysis of the full data set using both linear and non-linear regression models. Analysis included transformations of numerical data in an effort to identify a predictive model with an acceptable R^2 value. The best performing regression on the data set returned an R^2 value of .35, suggesting a low correlation between the attributes and accuracy of a forecast. This value is lower than the best R^2 value returned from the previous data, indicating there was more randomness introduced into the dataset by adding more information. The implication of this finding is that the temporal error associated with a forecast is largely a function of random error.

DATA MINING

Failing to find a pattern in the data using traditional statistical methods, we turned to data mining techniques to determine if there were underlying patterns in the data that could be uncovered using machine learning algorithms. We applied the k-nearest neighbor (KNN) algorithm, partitioning trees (PT), support vector machines (SVM), and neural networks (NN) to discover rules for classifying forecasts. We chose to use classification algorithms as opposed to regression algorithms since we had limited success using exact methods of regression. The exception to this approach was the application of neural networks; we used a classification algorithm for the success test and a regression test for the TFE and STFE.

We used forecast success, TFE, and STFE as the target variables and converted them to classes. Success was divided into two class values—success or failure—based on the 30% rule. TFE and STFE were divided into eight and 16 classes, respectively, of five-year widths. Finally, we divided the data set of 736 realized forecasts into three groups: a training set (60% of the data), a validation set (20% of the data), and a testing set (20% of the data). The training set was used to develop parameters, values, and structures associated with the data mining algorithms. The validation set was used to adjust and modify parameters to refine the model to get the best results, and the test set was used to assess the performance of the refined model given the tuned parameters. Our objective was to develop a classification model using a combination of attributes that could successfully place at least 85% of the test set forecasts into the correct classification. We selected 85% as our threshold, since a naive model of classifying any forecast as a failure would result in a 67% accuracy rate, while classifying any model to have less than a five-year TFE would result in a 58% success rate. Table 20 provides the results of the data mining analysis.

TABLE 20. DATA MINING ANALYSIS RESULTS

Model	Variable	Predictive Power	R^2
KNN	Success	53%	
PT		55%	
SVM		59%	
NN		49%	
PT	TFE	74%	
SVM		71%	
NN			0.3332
PT	STFE	74%	
SVM		71%	
NN			0.3204

With few exceptions, each of the algorithms failed to produce a model that matched the performance of the naive model. Those models that did exceed the performance of the naive model were marginally better (on the order of less than 5%) and not near the minimum threshold of 85%. The rules produced by all of these methods were extensive and indicated an over fitting of data.

Neither classical regression techniques nor data mining techniques were able to discern a meaningful relationship between the attributes included in this study and the temporal errors associated with forecast realization. This implies that the temporal error associated with a forecast is either highly influenced by a random component or some other attribute not collected during the data gathering process of this study.

4.7 SUMMARY OF FINDINGS

Despite substantially larger sample size of 1,055 statistically usable forecasts, the data were still noisy, indicating that there is minimal correlation between forecast accuracy and many of the forecast attributes studied. The two forecast attributes that did correlate to accuracy were time frame and methodology. Short-term forecasts have a higher success rate and smaller temporal error than do other forecasts with longer time frames, and the QTA method has a significantly higher success rate than do other forecasting methods. When correcting for short-term forecasts or well-performing technology area tags, this latter finding is no longer statistically significant, but just as large. We believe this loss of statistical significance is a result of sample size. The implication of this finding is that a forecast derived from quantitative data will generally perform better than a forecast derived from opinion and judgment. The findings about the influence of forecast time frame and methodology on accuracy are consistent with the results of the previous study.

Our study also indicates that there is no combination of attributes we analyzed that could allow forecast consumers to accurately predict how large and in what direction temporal error will be for a given forecast. We believe this is because temporal errors are largely influenced by a random component or some attribute not identified in this study. One significant change from our previous study is that forecasts tend to be pessimistic, underestimating the year of event realization. The previous study found near balance between optimistic and pessimistic forecasts.

As with the previous study, vague language continues to be a problem, with many forecast documents rejected during the document collection phase due to unspecific timeframes or outcomes. Since more forecasts were used during this extension, vague language is likely a common feature among forecasts.

5.0 RECOMMENDATIONS FOR FUTURE FORECASTS

The results of this analysis indicate that consumers of forecasts should seek forecasts that were generated using quantitative methods, as well as those that predict events over the short term (one to five years). Moreover, consumers should keep in mind that most forecasts overestimate the date in which predicted events will be realized.

Producers of technological forecasts should strive to provide enough detail to make the forecasts valuable to users by including a time frame and a clear description of the technology and what precisely will occur in the predicted time frame—technology emergence, evolution, migration, or market penetration.[13] Many forecast documents reviewed during the data collection phase lacked sufficient information to make them useful to the study or assessable by analysts, suggesting that they are also not specific enough to be valuable to users without a significant amount of interpretation.

[13] For example, a forecast document that simply reads, "HDTVs, 2006-2010" may be predicting that the products will emerge on the market, may be adopted, or may be commonplace. In these cases, the analysts made assumptions about what type of prediction was being made.

APPENDIX A. ACRONYMS

AMT	Amazon Mechanical Turk
ANOVA	analysis of variance
ASDR&E	Assistant Secretary of Defense for Research and Engineering
CAR	computer and autonomous/robotics
CV	curriculum vitae
KNN	k-nearest neighbor
KW	Kruskal-Wallis
NN	neural networks
QLA	qualitative trend analysis
QTA	quantitative trend analysis
PT	partitioning trees
STFE	signed temporal forecast error
SVM	support vector machines
TFE	temporal forecast error
TKHSD	Tukey-Kramer Honestly Significant Difference
TRL	technology readiness level
U.S.	United States

APPENDIX B. STANDARD LEXICON AND ANALYSIS RULES

B.1 STANDARD LANGUAGE

Timeframe:

- “A” (as in “a year”) = 1
- “Few” = 3
- “Some” = “few” = 3
- “Couple” = 2
- “Several” = “Multiple” = indicates a range of 4-10
- “Variety/various” = “several” = 4-10
- “Many” = context-specific; can be equivalent to significant (>15%) or several (a range of 4-10)
- “Majority/most” = >50%
- “Short term” = 1 to 5 years
- “Shortly” = 1 to 5 years
- “Near term” = “Short term” = 1 to 5 years
- “Coming years” or “years to come” = “short term” = 1 to 5 years
- “Near Future” = “Short term” = 1 to 5 years
- “Mid term” = 6 to 10 years
- “Not too distant future” = “mid term” = 6 to 10 years
- “Long term” = 11 to 25 years
- “Soon” = near/short term = 1 to 5 years
- “X years away” = evaluated as a negative forecast
- “By the year”= evaluated date is the forecasted year
- “By the 21st century” = 2000
- “In the 21st century” = 2000—2099, can’t be evaluated
- “In the next X years” = “within X years” = a date range from the year of the forecast to forecast + x years, evaluated date is the median
- “In X-Y years” = a date range between x and y with the evaluation data at the median
- “In the future” = not specific enough to include
- “Early next decade” or “early in the 19XXs” = the first 3 years of the decade (i.e., 1970-1973)
- “Middle of the next decade” or “mid 19XXs” = the middle years of the decade (i.e., 1974-1976)
- “End of the next decade” or “late 19XXs” = the last years of the decade (i.e., 1977-1979)
- “Foreseeable” = too vague to be analyzed on its own, requires further support from the document to determine timeframe

Probabilistic Qualifiers:

- “Improbable” = will not happen
- “Maybe,” “may,” “is possible,” “can happen,” “could,” “perhaps” = equivalent to 40-60% and excluded as a not forecast
- “Will,” “should”, “likely”, “probably” = greater than 60% probability and treated as will occur

Technology Maturity:

- “Next Generation” = The version or follow-on model introduced after the technology that dominated its sector at the time of the forecast. For instance, Windows NT is the next generation for Windows 3.1, the 486 is the next generation for the 386. Next generation

may be different for military, civil, and commercial sectors for the same subject area and in the same time frame.
- "Present/current generation" = The version of technology that dominates the sector at the time of the forecast.
- "SOA" = The most advanced technology available at the time of the forecast.
- "Commonplace" = technology is being used and is no longer in development
- "In practical use" = used in consumer products

Growth Rate:
- "Annual rate(s)" = treated as compound annual growth rate
- "Significant" = greater than 15%
- "Substantial" = greater than 30%
- "On the market" = can be purchased commercially
- "Growth until X year" =growth from year of forecast until X year

B.2 ANALYSIS RULES

Exclusion Rules: These rules determine what we will not record in the database.
- Timeframe: This is a temporal filter to avoid recording data that is a waste of time.
    - We will not record recent forecasts.
        - Recent is defined as occurring more recent that 30% of the forecasted timeframe. For example, a forecast in 1950 projecting X event in 2005 will not be recorded. A forecast in 2000 for X event in 2005 will be recorded.
- Specificity: This is a filter to avoid forecasts that are too vague.
    - For instance, at least one piece of vital information (timeframe, event, or technology metric) needs to be explicitly stated in a measurable format in order to be included. If some but not all information is vague or not included, we will track that data down.

Inclusion rules: These rules effect how we populate the database.
- Only include 10 forecasts from a single source.
    - Attempt to select forecasts that are diverse in respect to subject area or timeframe. We can return to these documents at a later date if more forecasts are useful.
- Binary forecasts are treated as two separate forecasts.
    - For example, "In the future there will be more error messages, and the information contained in such messages will be more helpful," will translate to:
        1. There will be error messages.
        2. Error messages in the future will be more helpful.
    - "Rapid growth would follow either the entry of one of more large companies into the field or the emergence of a highly successful nanotechnology company," will translate to:
        1. Rapid growth would follow the entry of one of more large companies.
        2. Rapid growth will follow the emergence of a highly successful nanotechnology company.

Analysis rules: These are rules that affect the final analysis but not data collection.
- An event that occurs 30% out of its time horizon is recorded as a failed forecast.
    - An event forecasted to occur in the near term (1-5 years) that occurs 8 years out (or later) will be a failed forecast.

- An event forecasted to occur in the midterm (6-10 years) that occurs 15 years out (or later) or sooner than 3 years out will be a failed forecast.
- An event forecasted to occur in the long term (11-25 years) that occurs 37 years out (or later) or sooner than 5 years out will be a failed forecast.

Probabilistic Forecasts:

- A probabilistic forecast with less than or equal to a 40% chance of occurring will be treated as a forecasted event that will not occur in the specified timeframe.
- A probabilistic forecast with more than or equal to a 60% chance of occurring with be treated as a forecasted event that will occur.
- A probabilistic forecast between a 40% and 60% chance of occurring will be excluded from analysis as a non-forecast.

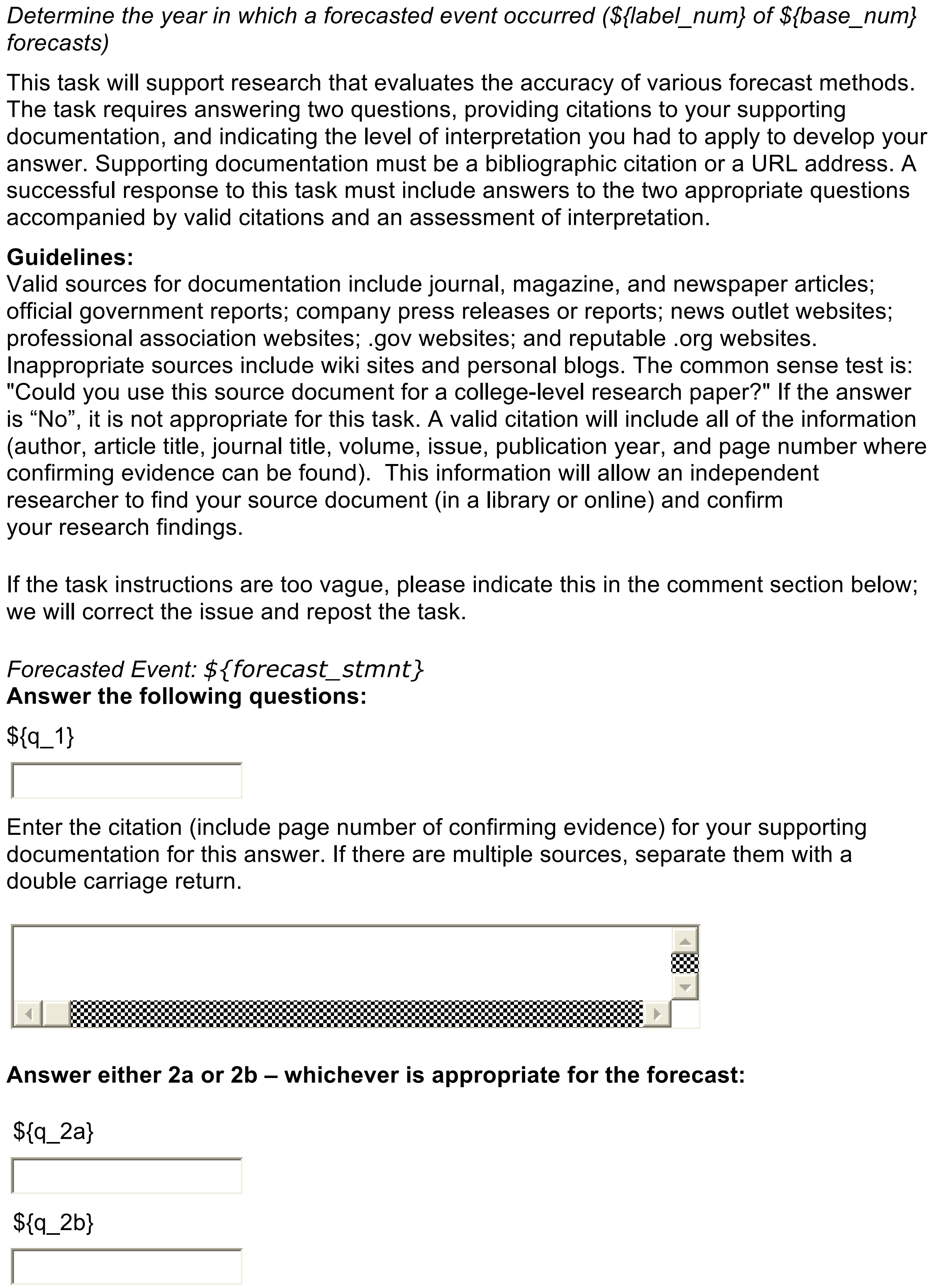
Causal Forecasts:

- We will exclude causal forecasts from the analysis.
    - For instance, forecasts that say, “Because of X, Y will happen.”
    - Due to the limited number of causal forecasts, we will document them but we will not perform analysis on them. They are not verifiable.

APPENDIX C. FORECAST SOURCES

A full list of all forecast sources in a database-readable format is enclosed in the following file: Appendix C_Forecast Source Listings.xls. This file is included in the disc that accompanies this report.

APPENDIX D. TEMPLATE



Determine the year in which a forecasted event occurred (${label_num} of ${base_num} forecasts)

This task will support research that evaluates the accuracy of various forecast methods. The task requires answering two questions, providing citations to your supporting documentation, and indicating the level of interpretation you had to apply to develop your answer. Supporting documentation must be a bibliographic citation or a URL address. A successful response to this task must include answers to the two appropriate questions accompanied by valid citations and an assessment of interpretation.

Guidelines:
Valid sources for documentation include journal, magazine, and newspaper articles; official government reports; company press releases or reports; news outlet websites; professional association websites; .gov websites; and reputable .org websites. Inappropriate sources include wiki sites and personal blogs. The common sense test is: "Could you use this source document for a college-level research paper?" If the answer is “No”, it is not appropriate for this task. A valid citation will include all of the information (author, article title, journal title, volume, issue, publication year, and page number where confirming evidence can be found). This information will allow an independent researcher to find your source document (in a library or online) and confirm your research findings.

If the task instructions are too vague, please indicate this in the comment section below; we will correct the issue and repost the task.

Forecasted Event: ${forecast_stmnt}
Answer the following questions:

${q_1}

Enter the citation (include page number of confirming evidence) for your supporting documentation for this answer. If there are multiple sources, separate them with a double carriage return.

Answer either 2a or 2b – whichever is appropriate for the forecast:

${q_2a}

${q_2b}

Enter the citation (include page number of confirming evidence) for your supporting documentation for your answer to 2a or 2b. If there are multiple sources, separate them with a double carriage return.

4. Approximately how much interpretation did you have to apply to reach your conclusion?

● None ○ A little ○ A moderate amount ○ A lot ○ Significant

Please provide a brief narrative of what you did to interpret the data

5. Please provide feedback on how we can improve this and similar HITs.

${forecast_num}

APPENDIX E. STATISTICAL ANALYSIS

All statistical analysis was conducted using R version 2.13.0, published on April 13, 2011. R is an open source statistical software package available through The R Foundation for Statistical Computing. Binaries for the application can be accessed via the following URL: http://cran.r-project.org/.

We present the elements of analysis for this appendix in the same sequence in which the analysis is referenced in the main report. This appendix is accompanied by two additional files:

1. *Load for R.csv*: This file contains an extract of the RATF database used to conduct all but one of the analytical tests executed during the conduct of this study
2. *Sensitivity Methods.csv*: This file contains the data read in by R used for sensitivity analysis of allowable range.
3. *ISEScript*: This file contains the scripts which load and execute the statistical tests conducted in this study

E.1.1 Outliers in the data set

After research, verification, and validation, there were 1,058 verified forecasts. Using Cook's distance as the basis for identifying outliers, we determined there were three forecasts that qualified for further investigation. The outliers are presented below:

Table E-1. Verified forecasts excluded form analysis due to vagueness in forecast context

Record Number	Technology Area	Method	Time Frame	Year Made	Predicted Year	Year Realized	Time Frame	TFE
973	Energy and Power Technology	Source Analysis	Short-term	1982	1984	1901	2	83
972	Energy and Power Technology	Source Analysis	Short-term	1982	1984	1891	2	93
334	Physical, Chemical, and Mechanical System	Quantitative Trend Analysis	Short-term	2007	2010	1968	3	42

Further research into the cause of the outliers revealed the forecast statements were vague and did not offer sufficient information for us to determine what the author was stating. Forecasts 972 and 973 predicted that certain car batteries would be demonstrated in electric vehicles in the U.S. by 1982. However, those two technologies were demonstrated at least 80 years prior to the forecast being made. It is possible the author was talking about modern electric cars, but the lack of specification would have caused too much inference on our part. Forecast 334 predicted 90K BTU air conditioners would be in use to cool buildings. However, 90K BTU AC's had been in use since 1968. There was perhaps a specific technology associated with the 90K BTU ACs, but the author did not mention it. We chose to exclude these forecasts from the analysis as the error was in forecast vagueness, not forecast accuracy.

E.1.2 Forecasts removed from test against a control

The following forecasts were temporarily removed from the data set for the analysis regarding the test against a control. Forecasts in table E-2 were removed due to excessive TFE while forecasts in table E-3 were removed due to excessive forecast horizon.

Table E-2. Verified forecasts excluded from test against a control due to long horizons

Record Number	Technology Area	Method	Time Frame	Year Made	Predicted Year	Year Realized	Time Frame	TFE
844	Ground Transportation Technology	Gaming and Scenarios	Long-term	1973	2000	2000	27	0
831	Ground Transportation Technology	Expert Sourcing	Long-term	1969	1998		29	
454	Air Transportation Technology	Models	Long-term	1968	2000		32	
453	Air Transportation Technology	Models	Long-term	1968	2000		32	

Table E-3. Verified forecasts excluded form test against a control due to large TFE's

Record Number	Technology Area	Method	Time Frame	Year Made	Predicted Year	Year Realized	Time Frame	TFE
11	Materials Technology	Expert Sourcing	Long-term	1970	1985	1968	15	17
65	Computer Technology	Expert Analysis Methods	Mid-term	1962	1972	2007	10	35
91	Energy and Power Technology	Expert Sourcing	Long-term	1956	1969	1992	13	23
335	Physical, Chemical, and Mechanical System	Quantitative Trend Analysis	Short-term	2007	2010	1987	3	23
414	Production Technology	Expert Sourcing	Mid-term	1985	1994	1970	9	24
449	Air Transportation Technology	Source Analysis	Mid-term	1968	1976	2010	8	34
450	Air Transportation Technology	Source Analysis	Long-term	1968	1979	2004	11	25
601	Autonomous Robotics Technology	Qualitative Trend Analysis	Short-term	1984	1989	2005	5	16
649	Computer Technology	Expert Sourcing	Mid-term	1983	1992	2008	9	16
730	Materials Technology	Source Analysis	Short-term	1986	1990	1966	4	24
745	Materials Technology	Source Analysis	Mid-term	1986	1996	1967	10	29
750	Biological Technology	Source Analysis	Long-term	1986	1998	1982	12	16
752	Energy and Power Technology	Expert Sourcing	Mid-term	1974	1980	2005	6	25
937	Computer Technology	Expert Sourcing	Long-term	1969	1980	1998	11	18

Record Number	Technology Area	Method	Time Frame	Year Made	Predicted Year	Year Realized	Time Frame	TFE
1130	Communications Technology	Expert Analysis Methods	Short-term	2005	2008	1986	3	22
1353	Biological Technology	Expert Sourcing	Mid-term	2001	2008	1991	7	17
1379	Computer Technology	Expert Sourcing	Mid-term	1996	2006	1982	10	24
1635	Computer Technology	Expert Analysis Methods	Mid-term	1989	1994.5	2011	5.5	16.5
1937	Computer Technology	Expert Analysis Methods	Mid-term	1970	1980	2000	10	20
1938	Computer Technology	Expert Analysis Methods	Mid-term	1970	1980	2001	10	21
2002	Space Technology	Qualitative Trend Analysis	Mid-term	1975	1985	1964	10	21
2010	Autonomous Robotics Technology	Qualitative Trend Analysis	Long-term	1975	1995	1961	20	34
2025	Photonics and Phononics Technology	Qualitative Trend Analysis	Mid-term	1975	1985	1964	10	21
2033	Computer Technology	Qualitative Trend Analysis	Short-term	1975	1980	2011	5	31
2040	Space Technology	Qualitative Trend Analysis	Mid-term	1975	1981.5	1998	6.5	16.5
2044	Space Technology	Qualitative Trend Analysis	Long-term	1975	1990	1973	15	17
2047	Space Technology	Qualitative Trend Analysis	Short-term	1975	1980	1997	5	17
2088	Physical, Chemical, and Mechanical System	Expert Sourcing	Mid-term	1997	2007	1970	10	37
2217	Ground Transportation Technology	Gaming and Scenarios	Long-term	1978	2000	1975	22	25
2218	Ground Transportation Technology	Qualitative Trend Analysis	Mid-term	1978	1985	1965	7	20
2219	Ground Transportation Technology	Qualitative Trend Analysis	Long-term	1978	2000	1966	22	34
2222	Ground Transportation Technology	Qualitative Trend Analysis	Mid-term	1978	1985	1967	7	18
2224	Ground Transportation Technology	Gaming and Scenarios	Long-term	1978	2000	1963	22	37

E.1.3 Derivation of 25-year Theoretical Random Guess Probability of Success

Using the same methods and formulas as reported in the previous study, we updated the 20-year random guess to reflect a 25-year random guess, since this was consistent with the data set (99 percentile forecast horizon and 95 percentile temporal error). Table E-4 reflects the number of years in a single direction associated with an allowable range of X% time horizon.

Table E-4. Error in years for a given forecast horizon and allowable range

	Error									
Length of Forecast (years)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
1	0	0	0	0	1	1	1	1	1	1
2	0	0	1	1	1	1	1	2	2	2
3	0	1	1	1	2	2	2	2	3	3
4	0	1	1	2	2	2	3	3	4	4
5	1	1	2	2	3	3	4	4	5	5
6	1	1	2	2	3	4	4	5	5	6
7	1	1	2	3	4	4	5	6	6	7
8	1	2	2	3	4	5	6	6	7	8
9	1	2	3	4	5	5	6	7	8	9
10	1	2	3	4	5	6	7	8	9	10
11	1	2	3	4	6	7	8	9	10	11
12	1	2	4	5	6	7	8	10	11	12
13	1	3	4	5	7	8	9	10	12	13
14	1	3	4	6	7	8	10	11	13	14
15	2	3	5	6	8	9	11	12	14	15
16	2	3	5	6	8	10	11	13	14	16
17	2	3	5	7	9	10	12	14	15	17
18	2	4	5	7	9	11	13	14	16	18
19	2	4	6	8	10	11	13	15	17	19
20	2	4	6	8	10	12	14	16	18	20
21	2	4	6	8	11	13	15	17	19	21
22	2	4	7	9	11	13	15	18	20	22
23	2	5	7	9	12	14	16	18	21	23
24	2	5	7	10	12	14	17	19	22	24
25	3	5	8	10	13	15	18	20	23	25

Table E-5 shows the number of years contained within the allowable range for a given forecast length and allowable range

Table E-5. The allowable range measured in years

	Allowable Error									
Length of Forecast (years)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
1	1	1	1	1	3	3	3	3	3	3
2	1	1	3	3	3	3	3	5	5	5
3	1	3	3	3	5	5	5	5	7	7
4	1	3	3	5	5	5	7	7	9	9
5	3	3	5	5	7	7	9	9	11	11
6	3	3	5	5	7	9	9	11	11	13
7	3	3	5	7	9	9	11	13	13	15
8	3	5	5	7	9	11	13	13	15	17
9	3	5	7	9	11	11	13	15	17	19
10	3	5	7	9	11	13	15	17	19	21
11	3	5	7	9	13	15	17	19	21	23
12	3	5	9	11	13	15	17	21	23	25
13	3	7	9	11	15	17	19	21	25	27
14	3	7	9	13	15	17	21	23	27	29
15	5	7	11	13	17	19	23	25	29	31
16	5	7	11	13	17	21	23	27	29	32
17	5	7	11	15	19	21	25	29	31	33
18	5	9	11	15	19	23	27	29	32	34
19	5	9	13	17	21	23	27	31	33	35
20	5	9	13	17	21	25	29	32	34	36
21	5	9	13	17	23	27	31	33	35	37
22	5	9	15	19	23	27	31	34	36	38
23	5	11	15	19	25	29	32	34	37	39
24	5	11	15	21	25	29	33	35	38	40
25	7	11	17	21	27	31	34	36	39	40

Table E-6 shows the probability of success associated with a given forecast length if event realization is uniformly distributed over 40 years

Table E-6. The probability of getting a forecast correct given a random guess

		Allowable Error									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Length of Forecast (years)	1	0.025	0.025	0.025	0.025	0.075	0.075	0.075	0.075	0.075	0.075
	2	0.025	0.025	0.075	0.075	0.075	0.075	0.075	0.125	0.125	0.125
	3	0.025	0.075	0.075	0.075	0.125	0.125	0.125	0.125	0.175	0.175
	4	0.025	0.075	0.075	0.125	0.125	0.125	0.175	0.175	0.225	0.225
	5	0.075	0.075	0.125	0.125	0.175	0.175	0.225	0.225	0.275	0.275
	6	0.075	0.075	0.125	0.125	0.175	0.225	0.225	0.275	0.275	0.325
	7	0.075	0.075	0.125	0.175	0.225	0.225	0.275	0.325	0.325	0.375
	8	0.075	0.125	0.125	0.175	0.225	0.275	0.325	0.325	0.375	0.425
	9	0.075	0.125	0.175	0.225	0.275	0.275	0.325	0.375	0.425	0.475
	10	0.075	0.125	0.175	0.225	0.275	0.325	0.375	0.425	0.475	0.525
	11	0.075	0.125	0.175	0.225	0.325	0.375	0.425	0.475	0.525	0.575
	12	0.075	0.125	0.225	0.275	0.325	0.375	0.425	0.525	0.575	0.625
	13	0.075	0.175	0.225	0.275	0.375	0.425	0.475	0.525	0.625	0.675
	14	0.075	0.175	0.225	0.325	0.375	0.425	0.525	0.575	0.675	0.725
	15	0.125	0.175	0.275	0.325	0.425	0.475	0.575	0.625	0.725	0.775
	16	0.125	0.175	0.275	0.325	0.425	0.525	0.575	0.675	0.725	0.8
	17	0.125	0.175	0.275	0.375	0.475	0.525	0.625	0.725	0.775	0.825
	18	0.125	0.225	0.275	0.375	0.475	0.575	0.675	0.725	0.8	0.85
	19	0.125	0.225	0.325	0.425	0.525	0.575	0.675	0.775	0.825	0.875
	20	0.125	0.225	0.325	0.425	0.525	0.625	0.725	0.8	0.85	0.9
	21	0.125	0.225	0.325	0.425	0.575	0.675	0.775	0.825	0.875	0.925
	22	0.125	0.225	0.375	0.475	0.575	0.675	0.775	0.85	0.9	0.95
	23	0.125	0.275	0.375	0.475	0.625	0.725	0.8	0.85	0.925	0.975
	24	0.125	0.275	0.375	0.525	0.625	0.725	0.825	0.875	0.95	1
	25	0.156	0.244	0.378	0.467	0.6	0.689	0.756	0.8	0.867	1

Table E-7 shows the overall probability of success for a given allowable range

Table E-7. Probability of success for an uninformed guess

	Sensitivity Analysis of Allowable Range										
Allowable Range	**0%**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
ρ	3.3%	9.0%	15.4%	22.1%	28.3%	36.0%	41.2%	47.3%	52.3%	57.5%	61.9%

There was a discussion as to whether a completely uninformed guess was the correct model to use and whether an empirically informed distribution based on how forecasts were actually generated would be better. We chose to model all forecasts in one group as opposed to create a separate distribution (and thus r) for each method. Table E-8 shows the observed distribution of forecast horizons from one through 25 years for each level of the allowable ranges.

Table E-8. Distribution of allowable forecasts given horizon and allowable range

		Allowable Error										
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Length of Forecast (years)	1	132	132	132	132	132	132	132	132	132	132	132
	2	125	125	125	125	125	125	125	125	125	125	125
	3	213	213	213	213	213	213	213	213	213	208	208
	4	102	102	102	102	102	102	102	101	101	100	100
	5	149	149	149	149	149	122	122	119	119	103	103
	6	57	57	57	57	57	47	47	47	43	43	41
	7	28	28	28	28	27	27	27	27	27	27	27
	8	30	30	30	30	26	26	26	26	26	26	26
	9	21	21	21	21	21	21	21	21	20	13	13
	10	40	40	40	40	40	40	37	26	16	16	12
	11	11	11	11	11	11	10	9	8	8	5	5
	12	7	7	7	7	7	7	7	7	6	6	6
	13	16	16	16	16	16	15	13	13	13	13	13
	14	10	10	10	10	10	7	7	7	7	6	6
	15	17	17	17	17	17	13	13	13	13	12	12
	16	5	5	5	5	5	5	5	5	4	4	4
	17	6	6	6	6	6	6	6	6	3	3	3
	18	14	14	14	14	12	12	12	10	10	10	10
	19	6	6	6	6	6	6	6	5	5	5	5
	20	5	5	5	5	5	5	5	4	4	3	3
	21	0	0	0	0	0	0	0	0	0	0	0
	22	10	10	10	10	10	10	10	10	10	10	10
	23	0	0	0	0	0	0	0	0	0	0	0
	24	0	0	0	0	0	0	0	0	0	0	0
	25	14	14	14	14	14	14	14	14	14	14	14

Table E-9 shows the result of taking the dot product of each column vectors in table E-9 with the appropriate column vector in table E-8, and then normalizing the resulting dot product by the number of observed allowable forecasts associated with each allowable range.

Table E-9. Probability of success for an informed guess

Allowable Range	Sensitivity Analysis of Allowable Range										
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
ρ	2.5%	5.1%	8.1%	11.1%	13.1%	17.3%	18.6%	20.8%	22.3%	25.4%	26.6%

E.1.4 Sensitivity analysis of observations given changes in allowable range

We wanted to know how sensitive the results of our observations were relative to the allowable range. We repeated our analysis against the control using different values for the allowable range that varied from 0% to 100% in increments of 10%.

We used a two-tailed binomial test with the Wilson continuity correction factors applied. The hypothesis test follows:

$$H_0 : r_i = \rho$$
$$H_A : r_i \neq \rho$$

Where: r_i is the observed success rate for methodology i, and
r is the expected success rate for a random guess

Rejecting the null hypothesis indicates there is sufficient evidence the tests are different, but one must look at the success rate to determine if the success rate is better or worse than a random guess.

Tables E-10 through E-20 show the results of the sensitivity analysis using the **uninformed** random guess. Blocks highlighted in red indicate those tests in which we would have failed to reject the null hypothesis at a level of significance of a=.10. Blocks highlighted in yellow indicate those tests we would have failed to reject the null hypothesis at a level of significance of a = .05.

Table E-10. Binomial test results with 0% allowable range vs. an uninformed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
0%	0.033	Expert Analysis Methods	65	239	0.000	0.272	0.218	0.334
0%	0.033	Expert Sourcing	60	266	0.000	0.226	0.178	0.281
0%	0.033	Gaming and Scenarios	2	26	0.211	0.077	0.013	0.266
0%	0.033	Models	34	135	0.000	0.252	0.183	0.335
0%	0.033	Multiple	31	98	0.000	0.316	0.228	0.419
0%	0.033	Qualitative Trend Analysis	20	91	0.000	0.220	0.142	0.321
0%	0.033	Quantitative Trend Analysis	18	68	0.000	0.265	0.168	0.388
0%	0.033	Source Analysis	27	95	0.000	0.284	0.199	0.387

Table E-11. Binomial test results with 10% allowable range vs. an uninformed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
10%	0.09	Expert Analysis Methods	68	239	0.000	0.285	0.229	0.347
10%	0.09	Expert Sourcing	74	266	0.000	0.278	0.226	0.337
10%	0.09	Gaming and Scenarios	4	26	0.288	0.154	0.050	0.357
10%	0.09	Models	37	135	0.000	0.274	0.203	0.359
10%	0.09	Multiple	31	98	0.000	0.316	0.228	0.419
10%	0.09	Qualitative Trend Analysis	20	91	0.000	0.220	0.142	0.321
10%	0.09	Quantitative Trend Analysis	20	68	0.000	0.294	0.193	0.419
10%	0.09	Source Analysis	28	95	0.000	0.295	0.208	0.398

Table E-12. Binomial test results with 20% allowable range vs. an uninformed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
20%	0.154	Expert Analysis Methods	74	239	0.000	0.310	0.252	0.373
20%	0.154	Expert Sourcing	82	266	0.000	0.308	0.254	0.368
20%	0.154	Gaming and Scenarios	7	26	0.106	0.269	0.124	0.481
20%	0.154	Models	45	135	0.000	0.333	0.256	0.420
20%	0.154	Multiple	32	98	0.000	0.327	0.237	0.430
20%	0.154	Qualitative Trend Analysis	22	91	0.028	0.242	0.161	0.345
20%	0.154	Quantitative Trend Analysis	29	68	0.000	0.426	0.309	0.552
20%	0.154	Source Analysis	30	95	0.000	0.316	0.226	0.420

Table E-13. Binomial test results with 30% allowable range vs. an uninformed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
30%	0.221	Expert Analysis Methods	79	239	0.000	0.331	0.272	0.395
30%	0.221	Expert Sourcing	88	266	0.000	0.331	0.275	0.391
30%	0.221	Gaming and Scenarios	7	26	0.635	0.269	0.124	0.481
30%	0.221	Models	51	135	0.000	0.378	0.297	0.466
30%	0.221	Multiple	32	98	0.015	0.327	0.237	0.430
30%	0.221	Qualitative Trend Analysis	24	91	0.314	0.264	0.179	0.368
30%	0.221	Quantitative Trend Analysis	31	68	0.000	0.456	0.336	0.581
30%	0.221	Source Analysis	31	95	0.018	0.326	0.236	0.431

Table E-14. Binomial test results with 40% allowable range vs. an uninformed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
40%	0.283	Expert Analysis Methods	81	234	0.035	0.346	0.286	0.411
40%	0.283	Expert Sourcing	93	264	0.014	0.352	0.295	0.414
40%	0.283	Gaming and Scenarios	7	26	1.000	0.269	0.124	0.481
40%	0.283	Models	53	135	0.007	0.393	0.311	0.481
40%	0.283	Multiple	34	98	0.178	0.347	0.255	0.451
40%	0.283	Qualitative Trend Analysis	24	91	0.728	0.264	0.179	0.368
40%	0.283	Quantitative Trend Analysis	36	68	0.000	0.529	0.405	0.650
40%	0.283	Source Analysis	32	95	0.255	0.337	0.245	0.442

Table E-15. Binomial test results with 50% allowable range vs. an uninformed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
50%	0.36	Expert Analysis Methods	89	233	0.495	0.382	0.320	0.448
50%	0.36	Expert Sourcing	102	245	0.072	0.416	0.354	0.481
50%	0.36	Gaming and Scenarios	8	22	1.000	0.364	0.180	0.592
50%	0.36	Models	60	124	0.005	0.484	0.394	0.575
50%	0.36	Multiple	35	87	0.435	0.402	0.300	0.513
50%	0.36	Qualitative Trend Analysis	26	91	0.156	0.286	0.198	0.391
50%	0.36	Quantitative Trend Analysis	36	68	0.005	0.529	0.405	0.650
50%	0.36	Source Analysis	34	95	1.000	0.358	0.264	0.463

Table E-16. Binomial test results with 60% allowable range vs. an uninformed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
60%	0.412	Expert Analysis Methods	97	233	0.894	0.416	0.353	0.483
60%	0.412	Expert Sourcing	113	239	0.057	0.473	0.408	0.538
60%	0.412	Gaming and Scenarios	9	22	1.000	0.409	0.215	0.633
60%	0.412	Models	70	124	0.001	0.565	0.473	0.652
60%	0.412	Multiple	38	87	0.664	0.437	0.332	0.547
60%	0.412	Qualitative Trend Analysis	29	91	0.088	0.319	0.227	0.426
60%	0.412	Quantitative Trend Analysis	37	68	0.035	0.544	0.419	0.664
60%	0.412	Source Analysis	36	95	0.534	0.379	0.283	0.485

Table E-17. Binomial test results with 70% allowable range vs. an uninformed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
70%	0.473	Expert Analysis Methods	99	233	0.149	0.425	0.361	0.491
70%	0.473	Expert Sourcing	118	224	0.109	0.527	0.459	0.593
70%	0.473	Gaming and Scenarios	10	22	1.000	0.455	0.251	0.673
70%	0.473	Models	73	123	0.009	0.593	0.501	0.680
70%	0.473	Multiple	38	86	0.591	0.442	0.336	0.553
70%	0.473	Qualitative Trend Analysis	30	91	0.006	0.330	0.237	0.437
70%	0.473	Quantitative Trend Analysis	38	67	0.142	0.567	0.441	0.686
70%	0.473	Source Analysis	37	93	0.177	0.398	0.299	0.505

Table E-18. Binomial test results with 80% allowable range vs. an uninformed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
80%	0.523	Expert Analysis Methods	101	230	0.012	0.439	0.374	0.506
80%	0.523	Expert Sourcing	124	212	0.074	0.585	0.515	0.651
80%	0.523	Gaming and Scenarios	11	22	0.835	0.500	0.288	0.712
80%	0.523	Models	81	122	0.002	0.664	0.572	0.745
80%	0.523	Multiple	38	86	0.160	0.442	0.336	0.553
80%	0.523	Qualitative Trend Analysis	32	90	0.001	0.356	0.259	0.464
80%	0.523	Quantitative Trend Analysis	42	66	0.084	0.636	0.508	0.749
80%	0.523	Source Analysis	38	91	0.046	0.418	0.317	0.526

Table E-19. Binomial test results with 90% allowable range vs. an uninformed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
90%	0.575	Expert Analysis Methods	103	222	0.001	0.464	0.397	0.532
90%	0.575	Expert Sourcing	129	201	0.063	0.642	0.571	0.707
90%	0.575	Gaming and Scenarios	14	19	0.171	0.737	0.486	0.899
90%	0.575	Models	92	121	0.000	0.760	0.673	0.831
90%	0.575	Multiple	38	86	0.016	0.442	0.336	0.553
90%	0.575	Qualitative Trend Analysis	32	89	0.000	0.360	0.263	0.469
90%	0.575	Quantitative Trend Analysis	42	56	0.010	0.750	0.614	0.852
90%	0.575	Source Analysis	38	90	0.004	0.422	0.320	0.531

Table E-20. Binomial test results with 100% allowable range vs. an uninformed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
100%	0.619	Expert Analysis Methods	103	219	0.000	0.470	0.403	0.539
100%	0.619	Expert Sourcing	133	199	0.166	0.668	0.598	0.732
100%	0.619	Gaming and Scenarios	14	19	0.351	0.737	0.486	0.899
100%	0.619	Models	92	121	0.001	0.760	0.673	0.831
100%	0.619	Multiple	38	86	0.001	0.442	0.336	0.553
100%	0.619	Qualitative Trend Analysis	32	89	0.000	0.360	0.263	0.469
100%	0.619	Quantitative Trend Analysis	45	55	0.002	0.818	0.686	0.905
100%	0.619	Source Analysis	38	90	0.000	0.422	0.320	0.531

Tables E-21 through E-31 contain the results of the sensitivity analysis conducted against the **informed** random guess. The hypothesis test and color code for these tables is consistent with the test and color codes for tables E-10 through E-20.

Table E-21. Binomial test results with 0% allowable range vs. an informed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
0%	0.025	Expert Analysis Methods	65	239	0.000	0.272	0.218	0.334
0%	0.025	Expert Sourcing	60	266	0.000	0.226	0.178	0.281
0%	0.025	Gaming and Scenarios	2	26	0.137	0.077	0.013	0.266
0%	0.025	Models	34	135	0.000	0.252	0.183	0.335
0%	0.025	Multiple	31	98	0.000	0.316	0.228	0.419
0%	0.025	Qualitative Trend Analysis	20	91	0.000	0.220	0.142	0.321
0%	0.025	Quantitative Trend Analysis	18	68	0.000	0.265	0.168	0.388
0%	0.025	Source Analysis	27	95	0.000	0.284	0.199	0.387

Table E-22. Binomial test results with 10% allowable range vs. an informed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
10%	0.051	Expert Analysis Methods	68	239	0.000	0.285	0.229	0.347
10%	0.051	Expert Sourcing	74	266	0.000	0.278	0.226	0.337
10%	0.051	Gaming and Scenarios	4	26	0.041	0.154	0.050	0.357
10%	0.051	Models	37	135	0.000	0.274	0.203	0.359
10%	0.051	Multiple	31	98	0.000	0.316	0.228	0.419
10%	0.051	Qualitative Trend Analysis	20	91	0.000	0.220	0.142	0.321
10%	0.051	Quantitative Trend Analysis	20	68	0.000	0.294	0.193	0.419
10%	0.051	Source Analysis	28	95	0.000	0.295	0.208	0.398

Table E-23. Binomial test results with 20% allowable range vs. an informed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
20%	0.081	Expert Analysis Methods	74	239	0.000	0.310	0.252	0.373
20%	0.081	Expert Sourcing	82	266	0.000	0.308	0.254	0.368
20%	0.081	Gaming and Scenarios	7	26	0.004	0.269	0.124	0.481
20%	0.081	Models	45	135	0.000	0.333	0.256	0.420
20%	0.081	Multiple	32	98	0.000	0.327	0.237	0.430
20%	0.081	Qualitative Trend Analysis	22	91	0.000	0.242	0.161	0.345
20%	0.081	Quantitative Trend Analysis	29	68	0.000	0.426	0.309	0.552
20%	0.081	Source Analysis	30	95	0.000	0.316	0.226	0.420

Table E-24. Binomial test results with 30% allowable range vs. an informed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
30%	0.111	Expert Analysis Methods	79	239	0.000	0.331	0.272	0.395
30%	0.111	Expert Sourcing	88	266	0.000	0.331	0.275	0.391
30%	0.111	Gaming and Scenarios	7	26	0.020	0.269	0.124	0.481
30%	0.111	Models	51	135	0.000	0.378	0.297	0.466
30%	0.111	Multiple	32	98	0.000	0.327	0.237	0.430
30%	0.111	Qualitative Trend Analysis	24	91	0.000	0.264	0.179	0.368
30%	0.111	Quantitative Trend Analysis	31	68	0.000	0.456	0.336	0.581
30%	0.111	Source Analysis	31	95	0.000	0.326	0.236	0.431

Table E-25. Binomial test results with 40% allowable range vs. an informed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
40%	0.131	Expert Analysis Methods	81	234	0.000	0.346	0.286	0.411
40%	0.131	Expert Sourcing	93	264	0.000	0.352	0.295	0.414
40%	0.131	Gaming and Scenarios	7	26	0.071	0.269	0.124	0.481
40%	0.131	Models	53	135	0.000	0.393	0.311	0.481
40%	0.131	Multiple	34	98	0.000	0.347	0.255	0.451
40%	0.131	Qualitative Trend Analysis	24	91	0.001	0.264	0.179	0.368
40%	0.131	Quantitative Trend Analysis	36	68	0.000	0.529	0.405	0.650
40%	0.131	Source Analysis	32	95	0.000	0.337	0.245	0.442

Table E-26. Binomial test results with 50% allowable range vs. an informed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
50%	0.176	Expert Analysis Methods	89	233	0.000	0.382	0.320	0.448
50%	0.176	Expert Sourcing	102	245	0.000	0.416	0.354	0.481
50%	0.176	Gaming and Scenarios	8	22	0.043	0.364	0.180	0.592
50%	0.176	Models	60	124	0.000	0.484	0.394	0.575
50%	0.176	Multiple	35	87	0.000	0.402	0.300	0.513
50%	0.176	Qualitative Trend Analysis	26	91	0.009	0.286	0.198	0.391
50%	0.176	Quantitative Trend Analysis	36	68	0.000	0.529	0.405	0.650
50%	0.176	Source Analysis	34	95	0.000	0.358	0.264	0.463

Table E-27. Binomial test results with 60% allowable range vs. an informed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
60%	0.19	Expert Analysis Methods	97	233	0.000	0.416	0.353	0.483
60%	0.19	Expert Sourcing	113	239	0.000	0.473	0.408	0.538
60%	0.19	Gaming and Scenarios	9	22	0.024	0.409	0.215	0.633
60%	0.19	Models	70	124	0.000	0.565	0.473	0.652
60%	0.19	Multiple	38	87	0.000	0.437	0.332	0.547
60%	0.19	Qualitative Trend Analysis	29	91	0.003	0.319	0.227	0.426
60%	0.19	Quantitative Trend Analysis	37	68	0.000	0.544	0.419	0.664
60%	0.19	Source Analysis	36	95	0.000	0.379	0.283	0.485

Table E-28. Binomial test results with 70% allowable range vs. an informed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
70%	0.218	Expert Analysis Methods	99	233	0.000	0.425	0.361	0.491
70%	0.218	Expert Sourcing	118	224	0.000	0.527	0.459	0.593
70%	0.218	Gaming and Scenarios	10	22	0.016	0.455	0.251	0.673
70%	0.218	Models	73	123	0.000	0.593	0.501	0.680
70%	0.218	Multiple	38	86	0.000	0.442	0.336	0.553
70%	0.218	Qualitative Trend Analysis	30	91	0.015	0.330	0.237	0.437
70%	0.218	Quantitative Trend Analysis	38	67	0.000	0.567	0.441	0.686
70%	0.218	Source Analysis	37	93	0.000	0.398	0.299	0.505

Table E-29. Binomial test results with 80% allowable range vs. an informed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
80%	0.239	Expert Analysis Methods	101	230	0.000	0.439	0.374	0.506
80%	0.239	Expert Sourcing	124	212	0.000	0.585	0.515	0.651
80%	0.239	Gaming and Scenarios	11	22	0.009	0.500	0.288	0.712
80%	0.239	Models	81	122	0.000	0.664	0.572	0.745
80%	0.239	Multiple	38	86	0.000	0.442	0.336	0.553
80%	0.239	Qualitative Trend Analysis	32	90	0.013	0.356	0.259	0.464
80%	0.239	Quantitative Trend Analysis	42	66	0.000	0.636	0.508	0.749
80%	0.239	Source Analysis	38	91	0.000	0.418	0.317	0.526

Table E-30. Binomial test results with 90% allowable range vs. an informed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
90%	0.274	Expert Analysis Methods	103	222	0.000	0.464	0.397	0.532
90%	0.274	Expert Sourcing	129	201	0.000	0.642	0.571	0.707
90%	0.274	Gaming and Scenarios	14	19	0.000	0.737	0.486	0.899
90%	0.274	Models	92	121	0.000	0.760	0.673	0.831
90%	0.274	Multiple	38	86	0.001	0.442	0.336	0.553
90%	0.274	Qualitative Trend Analysis	32	89	0.075	0.360	0.263	0.469
90%	0.274	Quantitative Trend Analysis	42	56	0.000	0.750	0.614	0.852
90%	0.274	Source Analysis	38	90	0.003	0.422	0.320	0.531

Table E-31. Binomial test results with 100% allowable range vs. an informed guess

Allowable	Rho	Method	Success	Observations	p-value	Rate	95 CI lower	95 CI upper
100%	0.29	Expert Analysis Methods	103	219	0.000	0.470	0.403	0.539
100%	0.29	Expert Sourcing	133	199	0.000	0.668	0.598	0.732
100%	0.29	Gaming and Scenarios	14	19	0.000	0.737	0.486	0.899
100%	0.29	Models	92	121	0.000	0.760	0.673	0.831
100%	0.29	Multiple	38	86	0.003	0.442	0.336	0.553
100%	0.29	Qualitative Trend Analysis	32	89	0.161	0.360	0.263	0.469
100%	0.29	Quantitative Trend Analysis	45	55	0.000	0.818	0.686	0.905
100%	0.29	Source Analysis	38	90	0.007	0.422	0.320	0.531

E.1.5 Results of comparative assessments

To compare method success rates against each other, we conducted Fisher's Exact Test comparing success rates of different methods to each other, different technology area tags, and timeframes. The following is the hypothesis test used for this assessment.

$H_0 : r_t \leq r_s$ Where: r_t is the observed success rate of the QTA method, and
$H_A : r_t > r_s$ r_s is the observed success rate of each other method

Rejecting the null hypothesis means there is sufficient evidence the success rate of the attribute value on the left hand side of the test is statistically higher than the success rate of the attribute on the RHS. For tables E-32 through E-34, blocks with highlighted p-values indicate comparisons for which we rejected the null hypothesis

Table E-32: QTA compared to all other methods

Method	Successes	Failures	p-value
Quantitative Trend Analysis compared to	31	38	
Qualitative Trend Analysis	24	78	0.003
Expert Sourcing	88	188	0.030
Expert Analysis Methods	79	165	0.038
Source Analysis	31	69	0.046
Multiple	32	66	0.074
Gaming and Scenarios	8	21	0.083
Models	51	86	0.180

Results indicate QTA is better than at least for methods. Failure to reject gaming and scenarios may be a result of small sample size

Table E-33: Short-term compared to all other timeframes

Time Frame	Successes	Failures	p-value
Short term compared to	254	473	
Medium term	54	139	0.040
Long term	36	99	0.037

Results indicate the success rate for short term forecasts are better than other time frames

Table E-34: CAR compared to all other technology area tags

Technology Area Tags	Successes	Failures	p-value
Computer/Autonomous/Robotics compared to	84	133	
Others	260	578	0.020
Ground Transportation Technology	21	61	0.023
Sensor Technology	11	36	0.032
Energy and Power Technology	27	68	0.052
Biological Technology	25	63	0.057
Photonics and Phononics Technology	3	14	0.066
Space Technology	23	54	0.105
Materials Technology	11	27	0.167
Physical, Chemical, and Mechanical System	13	29	0.220
Communications Technology	104	189	0.257
Air Transportation Technology	10	20	0.361
Production Technology	7	13	0.472

Results indicate that while CAR is better than a few technology area tags, it is not universally better than all other technology area tags. This is different than the results of the previous study, which indicated CAR was substantially better than all other technology area tags with large sample sizes.

E.1.6 Results of comparative assessments between CAR and QTA

There were many QTA-derived forecasts in the CAR technology sample set. We conducted the Fisher's Exact Test again on methods but this time removed the CAR forecasts and the technology are tags but with QTA forecasts removed. Tables E-35 and E-36 provide the results of these tests.

Table E-35: CAR compared to all other technology area tags without QTA

Technology Area Tags	Successes	Failures	p-value
Computer/Autonomous/Robotics compared to	51	76	
Biological Technology	19	61	0.011
Ground Transportation Technology	21	59	0.028
Sensor Technology	11	36	0.029
Photonics and Phononics Technology	3	14	0.058
Energy and Power Technology	27	66	0.059
Space Technology	23	54	0.091
Materials Technology	10	27	0.102
Physical, Chemical, and Mechanical System	12	27	0.193
Communications Technology	102	188	0.194
Production Technology	7	13	0.429
Air Transportation Technology	9	13	0.622

Results of this test are not significantly different than the results of the test with QTA forecasts. This indicates CAR's high success rate is not due QTA derieved forecasts.

Table E-36: QTA compared to all other methods without CAR

Technology Area Tags	Successes	Failures	p-value
Quantitative Trend Analysis compared to	12	16	
Qualitative Trend Analysis	19	63	0.042
Multiple	18	54	0.068
Gaming and Scenarios	7	20	0.150
Expert Analysis Methods	56	124	0.155
Expert Sourcing	75	158	0.178
Source Analysis	30	63	0.209
Models	43	80	0.283

Results are significantly different than test results including CAR. Initially, we interpreted this to mean that QTA success rate was influenced by CAR technology area forecasts. Upon closer inspection, however, the resulting changes in success rates for all but one method were marginal. This indicated something else was contributing to test results. We believe these observations are a result of the small sample size for QTA methods.

E.1.7 Results of comparative assessments between Timeframe and QTA

We noticed different distributions of short-term, medium-term, and long-term forecasts when we looked at forecasts by method and timeframe. To determine if forecasts horizon was influencing results, we again conducted the Fisher's Exact Test with various timeframes missing or with only one timeframe available. Tables E-37 through E-42 provide results of these tests.

Table E-37. QTA against all other methods without short-term forecasts

Technology Area Tags	Successes	Failures	p-value
Quantitative Trend Analysis compared to	9	11	
Source Analysis	1	19	0.004
Qualitative Trend Analysis	13	52	0.029
Expert Sourcing	35	85	0.126
Expert Analysis Methods	17	43	0.136
Models	6	12	0.345
Gaming and Scenarios	3	6	0.432
Multiple	6	10	0.456

Results indicate short-term forecasts were responsible for QTA's high success rates. Again, however, success rate changes were too marginal to account for the drastic change in p-values. We believe the results of this test are influenced by the sample size for all forecasting methods, as opposed to just the impact of short-term forecasts.

Table E-38. QTA against all other methods without medium-term forecasts

Technology Area Tags	Successes	Failures	p-value
Quantitative Trend Analysis compared to	26	34	
Qualitative Trend Analysis	21	60	0.024
Multiple	29	64	0.088
Expert Analysis Methods	64	130	0.096
Expert Sourcing	65	128	0.114
Gaming and Scenarios	8	21	0.114
Source Analysis	31	58	0.191
Models	46	77	0.270

Even though medium-term forecasts constitute a small portion of the sample set, QTA's overall success rate and performance against other methods is dependent upon its performance in medium-term forecasts.

Table E-39. QTA against all other methods without long-term forecasts

Technology Area Tags	Successes	Failures	p-value
Quantitative Trend Analysis compared to	27	31	
Qualitative Trend Analysis	14	44	0.010
Expert Sourcing	76	163	0.026
Expert Analysis Methods	77	157	0.038
Source Analysis	30	61	0.068
Gaming and Scenarios	5	15	0.075
Multiple	29	58	0.077
Models	50	83	0.159

Long-term forecasts have some impact on QTA's performance but only marginally so. In general, trends found with all forecasts are consistent with trends found in short -and medium-term forecasts.

Table E-40. QTA against all other methods short-term forecasts only

Technology Area Tags	Successes	Failures	p-value
Quantitative Trend Analysis compared to	22	27	
Multiple	26	56	0.092
Expert Analysis Methods	62	122	0.100
Gaming and Scenarios	5	15	0.102
Expert Sourcing	53	103	0.113
Qualitative Trend Analysis	11	26	0.113
Models	45	74	0.248
Source Analysis	30	50	0.259

Results indicate that short-term forecast success rates are not the factor driving QTA's high performance in success rate.

Table E-41. QTA against all other methods medium-term forecasts only

Technology Area Tags	Successes	Failures	p-value
Quantitative Trend Analysis compared to	5	4	
Source Analysis	0	11	0.008
Qualitative Trend Analysis	3	18	0.032
Expert Sourcing	23	60	0.093
Expert Analysis Methods	15	35	0.135
Models	5	9	0.306
Multiple	3	2	0.762
Gaming and Scenarios	0	0	1.000

Results indicate that medium-term forecasts could be a driver in QTA's high success rate, but small sample sizes could be responsible for the findings.

Table E-42. QTA against all other methods long-term forecasts only

Technology Area Tags	Successes	Failures	p-value
Quantitative Trend Analysis compared to	4	7	
Source Analysis	1	8	0.221
Qualitative Trend Analysis	10	34	0.285
Expert Analysis Methods	2	8	0.367
Multiple	3	8	0.500
Expert Sourcing	12	25	0.539
Models	1	3	0.593
Gaming and Scenarios	3	6	0.630

Results indicate that long-term forecasts alone are not responsible for QTA's high success rate. While each test regarding QTA and timeframes is inconclusive in its own right, the consolidation of all results indicates there is no causal relationship with timeframes that cause QTA to perform better than other methods. We therefore conclude that QTA is simply a better method overall.

E.1.8 Results of tests for normality

We planned to use the ANOVA as an omnibus test to compare temporal errors between attribute values in an effort to determine if some attributes indicated a higher precision. A criterion for using the ANOVA is an assumption of normality. We conducted two tests—the Shapiro-Wilk and the Kolmogorov-Smirnoff tests—to determine if the sub sample datasets were normally distributed.

The hypothesis tests for this analyses is:

$$H_0 : Data\,are\,N(\mu,\sigma^2)$$
$$H_A : Data\,are\,not\,N(\mu,\sigma^2)$$

Results of each test are presented in tables E-43 and E-44, respectively.

Table E-43. Results of the Shapiro-Wilk test for normality

Data set	N	p-value	Mean	s	Kurtosis	Skewness
All Data	736	0.000	-0.694	6.823	10.519	-0.403
Expert Analysis Methods	154	0.000	-0.120	6.097	11.824	1.210
Expert Sourcing	189	0.000	-1.206	6.968	8.528	-0.362
Gaming and Scenarios	22	0.000	-2.227	10.632	6.956	-2.043
Models	114	0.001	-0.246	2.683	4.102	0.435
Multiple	61	0.001	0.262	5.981	3.676	0.090
Qualitative Trend Analysis	66	0.000	-1.227	10.629	5.339	-0.665
Quantitative Trend Analysis	56	0.000	-1.411	5.383	7.943	-0.632
Source Analysis	74	0.000	-0.588	8.149	9.450	0.350

Results indicate that with a high level of confidence we reject H_0 for the entire data set, as well as the underlying sample sets

Table E-44. Results of the Kolmogorov-Smirnoff Goodness of Fit test

Data set	N	p-value	Mean	s	Kurtosis	Skewness
All Data	736	0.000	-0.694	6.823	10.519	-0.403
Expert Analysis Methods	154	0.001	-0.120	6.097	11.824	1.210
Expert Sourcing	189	0.000	-1.206	6.968	8.528	-0.362
Gaming and Scenarios	22	0.050	-2.227	10.632	6.956	-2.043
Models	114	0.001	-0.246	2.683	4.102	0.435
Multiple	61	0.026	0.262	5.981	3.676	0.090
Qualitative Trend Analysis	66	0.044	-1.227	10.629	5.339	-0.665
Quantitative Trend Analysis	56	0.017	-1.411	5.383	7.943	-0.632
Source Analysis	74	0.005	-0.588	8.149	9.450	0.350

Results of this test are consistent with the Shapiro-Wilk test. There is sufficient evidence the data are not normally distributed. There is evidence the ANOVA is robust against the assumption of normality but not against the assumption of equal variance. We conducted the F Test for equal variances between the data sets. The hypothesis test is:

$$H_0 : \sigma_l^2 = \sigma_s^2$$
$$H_A : \sigma_l^2 \neq \sigma_s^2$$

where, *s* is the standard deviation,
l is the method with the largest standard deviation, and
s is the method with the smallest standard deviation.

We reject H_0 if the *F-statistic* is larger than the critical value. Table E-45 has provides the variances for each method.

Table E-45. Variance and *log*$_{10}$ of the variance for each method

Method	*df*	σ^2	$log_{10}(\sigma^2)$
Expert Analysis Methods	153	37.17	1.57
Expert Sourcing	188	48.55	1.69
Gaming and Scenarios	21	113.04	2.05
Models	113	7.20	0.86
Multiple	60	35.77	1.55
Qualitative Trend Analysis	65	112.99	2.05
Quantitative Trend Analysis	55	28.97	1.46
Source Analysis	73	66.41	1.82

The yellow highlighted rows identify the largest and smallest variances in the sample set. Dividing the largest variance by the smallest variance results in an *F statistic* of 15.7. Comparing this value to *F Distribution* returns a p-value of 1.8 x 10^{-24} indicating with high confidence we can reject H_0. Data can sometimes be transformed to approach normality. A standard transformation is to take the log of the data. Rerunning the test with the transformed data results in an *F statistic* of 2.39. Comparing this value to the F-Distribution returns a p-value of .002 – indicating the transformed data do not exhibit equal variances.

E.1.9 Results of tests for similar distributions (i.i.d)

A non-parametric analog to the ANOVA is the Kruskal-Wallis (KW) test. The requirements for the KW test are similar shapes and ranges of the underlying distributions. Figure E-1 shows the histograms of the underlying data broken out by method.

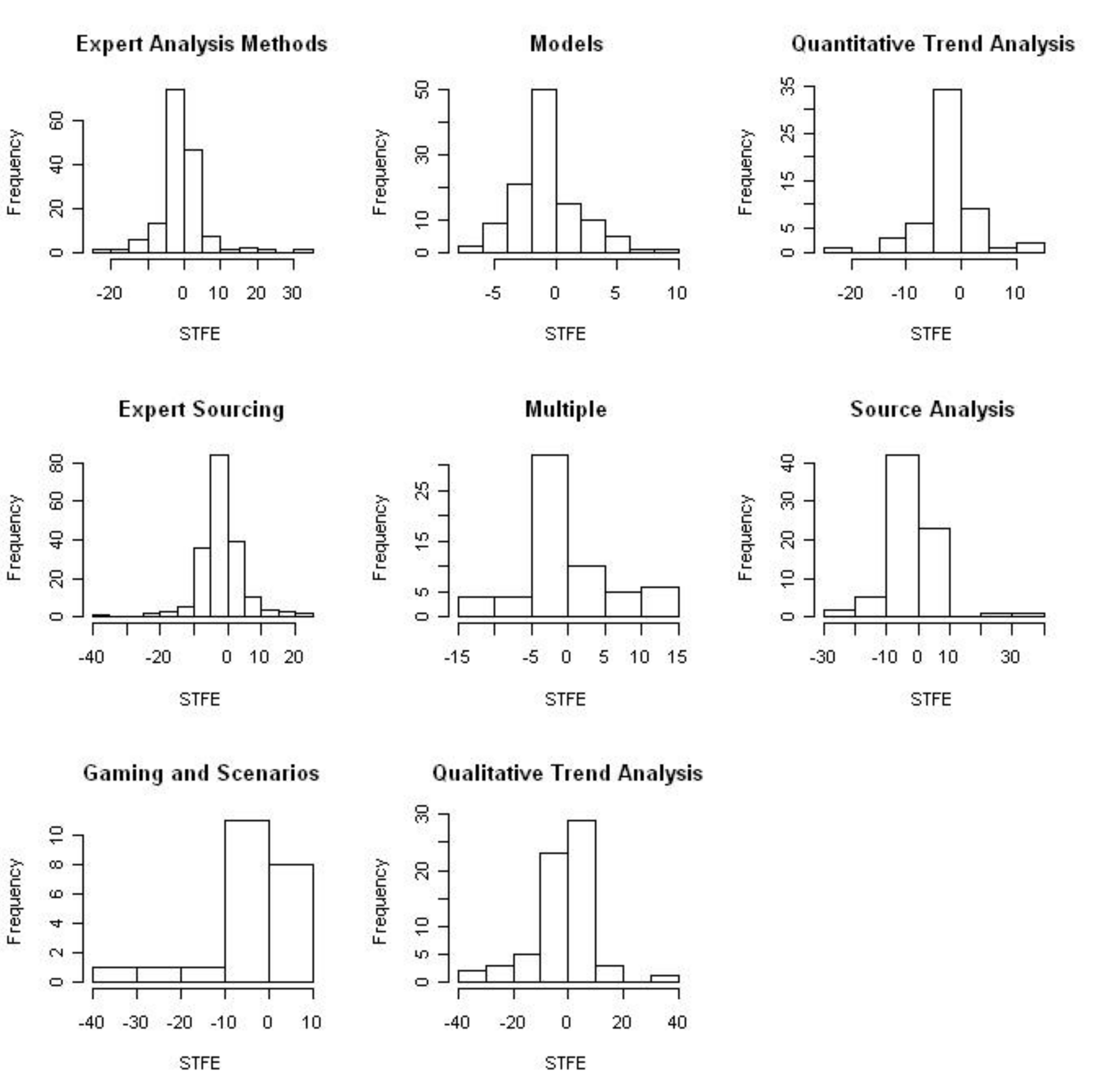


Figure E-1. Histograms of the data broken out by method.

While the data exhibit unimodal distributions and the ranges are within an order of magnitude of each other (which may not be sufficient to determine similar ranges), they certainly exhibit skewness in different directions, which does indicate the data may not have the same shape.

E.1.10 Results of ANOVA tests

Since there are no standard parametric or non-parametric tests to use, we caveat our findings with the results above and apply the ANOVA test as an omnibus to determine if there are differences in the underlying distributions. If so, this could be used to determine whether some methods (or technology areas or timeframes) are better than others. The hypothesis test for this analysis follows:

$H_0; \mu_a = \mu_b = \ldots = \mu_g$ where, μ_i represents the mean of the i values each attribute can take

$H_A; \mu_a \neq \mu_b \neq \ldots \neq \mu_g$

Rejecting H0 indicates there is sufficient evidence that there are differences in the means. However, we caution that results from this test may not be accurate, since we failed to satisfy either criteria for using the ANOVA.

Table E-46. Results of the ANOVA on each of the three primary attributes

Attribute	Variable of interest	F-Statistic	Numerator df	Denominator df	p-Value
TechArea	TFE	3.399	13	111.20	0.000
TechArea	STFE	2.744	13	110.60	0.002
Method	TFE	10.939	7	180.15	0.000
Method	STFE	0.975	7	181.31	0.451
TimeFrame	TFE	27.358	2	136.85	0.000
TimeFrame	STFE	3.068	2	135.96	0.050

Results indicate that there is sufficient evidence that differences in the underlying means for all three primary attributes exist when looking at the TFE (magnitude of error).

E.1.11 Results of TKHSD

The ANOVA only indicates that there are underlying differences in the means; it does not indicate which attributes may have different means. We used the TKHSD to determine which attribute pairings resulted in statistically significant means. We note the TKHSD assumes normality, which we have shown is not a valid assumption given our data.

Table E-47. Results of TKHSD for methods

Conf Level	Treatment	Difference	Lower Bound	Upper Bound	Adjusted o-value
0.95	Qualitative Trend Analysis-Expert Analysis Methods	3.551	1.137	5.965	0.000
0.95	Qualitative Trend Analysis-Models	5.203	2.665	7.740	0.000
0.95	Models-Expert Sourcing	-2.567	-4.513	-0.622	0.002
0.95	Quantitative Trend Analysis-Qualitative Trend Analysis	-3.838	-6.819	-0.857	0.003
0.95	Qualitative Trend Analysis-Expert Sourcing	2.635	0.289	4.981	0.015
0.95	Models-Gaming and Scenarios	-4.233	-8.054	-0.412	0.018
0.95	Qualitative Trend Analysis-Multiple	3.204	0.290	6.119	0.020
0.95	Source Analysis-Models	2.644	0.194	5.093	0.024
0.95	Source Analysis-Qualitative Trend Analysis	-2.559	-5.337	0.219	0.097
0.95	Models-Expert Analysis Methods	-1.652	-3.679	0.376	0.207
0.95	Multiple-Models	1.998	-0.605	4.601	0.277
0.95	Quantitative Trend Analysis-Gaming and Scenarios	-2.869	-6.997	1.260	0.408
0.95	Gaming and Scenarios-Expert Analysis Methods	2.581	-1.159	6.321	0.417
0.95	Multiple-Gaming and Scenarios	-2.235	-6.315	1.846	0.710
0.95	Expert Sourcing-Expert Analysis Methods	0.916	-0.866	2.697	0.772
0.95	Quantitative Trend Analysis-Models	1.364	-1.313	4.042	0.780
0.95	Quantitative Trend Analysis-Expert Sourcing	-1.203	-3.700	1.293	0.826
0.95	Gaming and Scenarios-Expert Sourcing	1.665	-2.031	5.362	0.871
0.95	Source Analysis-Quantitative Trend Analysis	1.279	-1.627	4.186	0.884
0.95	Source Analysis-Expert Analysis Methods	0.992	-1.329	3.313	0.899
0.95	Source Analysis-Gaming and Scenarios	-1.589	-5.574	2.395	0.928
0.95	Qualitative Trend Analysis-Gaming and Scenarios	0.970	-3.070	5.009	0.996
0.95	Multiple-Expert Sourcing	-0.569	-2.986	1.847	0.997
0.95	Source Analysis-Multiple	0.646	-2.192	3.483	0.997
0.95	Quantitative Trend Analysis-Multiple	-0.634	-3.670	2.403	0.998
0.95	Multiple-Expert Analysis Methods	0.346	-2.136	2.829	1.000
0.95	Quantitative Trend Analysis-Expert Analysis Methods	-0.287	-2.848	2.273	1.000
0.95	Source Analysis-Expert Sourcing	0.076	-2.174	2.326	1.000

The values highlighted in red on the table are those for which there is sufficient evidence that the two underlying distributions are different. Only the QLA method indicates a universal rejection of all pairings, suggesting that it is perhaps worse than all other methods. The models method does exhibit superior performance over three other methods.

Table E-48. Results of TKHSD test for technology area tags

CI	Treatment	Difference	Lower Bound	Upper Bound	Adjusted p-value
0.95	Materials Technology-Computer Technology	3.340	-0.259	6.939	0.102
0.95	Ground Transportation Technology-Energy and Power Technology	2.797	-0.377	5.970	0.153
0.95	Ground Transportation Technology-Biological Technology	3.059	-0.544	6.663	0.199
0.95	Ground Transportation Technology-Computer Technology	2.436	-0.443	5.315	0.204
0.95	Ground Transportation Technology-Communications Technology	3.793	1.083	6.503	0.255
0.95	Space Technology-Communications Technology	2.420	-0.581	5.421	0.273
0.95	Materials Technology-Maritime Transportation Technology	5.594	-1.621	2.809	0.336
0.95	Sensor Technology-Materials Technology	-3.576	-8.299	1.147	0.377
0.95	Materials Technology-Autonomous/Robotics Technology	3.149	-1.285	7.584	0.489
0.95	Materials Technology-Communications Technology	4.697	1.231	8.163	0.506
0.95	Communications Technology-Air Transportation Technology	-3.496	-8.536	1.543	0.530
0.95	Maritime Transportation Technology-Ground Transportation Technology	-4.690	-11.573	2.194	0.561
0.95	Computer Technology-Communications Technology	1.357	-0.679	3.393	0.598
0.95	Sensor Technology-Ground Transportation Technology	-2.672	-6.872	1.528	0.673
0.95	Materials Technology-Energy and Power Technology	3.701	-0.138	7.540	0.720
0.95	Ground Transportation Technology-Autonomous/Robotics Technology	2.245	-1.628	6.118	0.797
0.95	Physical, Chemical, and Mechanical System-Materials Technology	-2.654	-7.334	2.026	0.822
0.95	Physical, Chemical, and Mechanical System-Communications Technology	2.043	-1.575	5.661	0.826
0.95	Maritime Transportation Technology-Air Transportation Technology	-4.393	-12.482	3.697	0.863
0.95	Space Technology-Materials Technology	-2.277	-6.498	1.944	0.869
0.95	Photonics and Phononics Technology-Communications Technology	3.035	-2.612	8.682	0.872
0.95	Materials Technology-Biological Technology	3.963	-0.238	8.165	0.879
0.95	Production Technology-Communications Technology	2.623	-2.595	7.841	0.919
0.95	Biological Technology-Air Transportation Technology	-2.762	-8.334	2.809	0.926
0.95	Space Technology-Maritime Transportation Technology	3.317	-3.686	0.320	0.948
0.95	Energy and Power Technology-Air Transportation Technology	-2.500	-7.803	2.803	0.949
0.95	Communications Technology-Autonomous/Robotics Technology	-1.548	-4.843	1.747	0.951
0.95	Photonics and Phonoics Technology-Maritime Transportation Technology	3.932	-4.549	2.413	0.956
0.95	Space Technology-Biological Technology	1.686	-2.141	5.513	0.970
0.95	Production Technology-Maritime Transportation Technology	3.519	-4.682	1.721	0.976

CI	Treatment	Difference	Lower Bound	Upper Bound	Adjusted p-value
0.95	Physical, Chemical, and Mechanical System-Ground Transportation Technology	-1.750	-5.901	2.401	0.980
0.95	Computer Technology-Air Transportation Technology	-2.139	-7.271	2.993	0.982
0.95	Space Technology-Energy and Power Technology	1.424	-2.001	4.849	0.982
0.95	Energy and Power Technology-Communications Technology	0.996	-1.439	3.431	0.984
0.95	Physical, Chemical, and Mechanical System-Maritime Transportation Technology	2.940	-4.349	0.229	0.986
0.95	Sensor Technology-Air Transportation Technology	-2.375	-8.349	3.599	0.988
0.95	Space Technology-Ground Transportation Technology	-1.373	-4.999	2.253	0.992
0.95	Photonics and Phononics Technology-Biological Technology	2.301	-3.825	8.427	0.993
0.95	Photonics and Phononics Technology-Energy and Power Technology	2.039	-3.844	7.922	0.997
0.95	Production Technology-Materials Technology	-2.075	-8.078	3.929	0.997
0.95	Maritime Transportation Technology-Autonomous/Robotics Technology	-2.444	-9.579	4.690	0.997
0.95	Maritime Transportation Technology-Computer Technology	-2.254	-8.901	4.393	0.997
0.95	Autonomous/Robotics Technology-Air Transportation Technology	-1.948	-7.697	3.801	0.997
0.95	Space Technology-Computer Technology	1.063	-2.091	4.217	0.997
0.95	Production Technology-Biological Technology	1.889	-3.844	7.622	0.998
0.95	Sensor Technology-Communications Technology	1.121	-2.553	4.795	0.999
0.95	Physical, Chemical, and Mechanical System-Biological Technology	1.309	-3.019	5.637	0.999
0.95	Production Technology-Energy and Power Technology	1.626	-3.847	7.099	0.999
0.95	Space Technology-Sensor Technology	1.299	-3.095	5.692	0.999
0.95	Sensor Technology-Photonics and Phononics Technology	-1.914	-8.409	4.581	0.999
0.95	Photonics and Phononics Technology-Computer Technology	1.678	-4.052	7.408	0.999
0.95	Maritime Transportation Technology-Energy and Power Technology	-1.893	-8.673	4.887	1.000
0.95	Sensor Technology-Maritime Transportation Technology	2.018	-5.299	9.335	1.000
0.95	Physical, Chemical, and Mechanical System-Energy and Power Technology	1.047	-2.930	5.024	1.000
0.95	Photonics and Phononics Technology-Materials Technology	-1.662	-8.041	4.717	1.000
0.95	Communications Technology-Biological Technology	-0.734	-3.708	2.240	1.000
0.95	Sensor Technology-Production Technology	-1.501	-7.627	4.624	1.000
0.95	Physical, Chemical, and Mechanical System-Air Transportation Technology	-1.453	-7.393	4.487	1.000
0.95	Production Technology-Computer Technology	1.265	-4.042	6.573	1.000
0.95	Photonics and Phononics Technology-Autonomous/Robotics Technology	1.487	-4.801	7.775	1.000
0.95	Maritime Transportation Technology-Biological Technology	-1.630	-8.622	5.361	1.000
0.95	Materials Technology-Ground Transportation	0.904	-3.115	4.923	1.000

CI	Treatment	Difference	Lower Bound	Upper Bound	Adjusted p-value
	Technology				
0.95	Space Technology-Autonomous/Robotics Technology	0.872	-3.209	4.954	1.000
0.95	Production Technology-Ground Transportation Technology	-1.170	-6.771	4.431	1.000
0.95	Materials Technology-Air Transportation Technology	1.201	-4.648	7.050	1.000
0.95	Biological Technology-Autonomous/Robotics Technology	-0.814	-4.876	3.248	1.000
0.95	Computer Technology-Biological Technology	0.623	-2.505	3.752	1.000
0.95	Space Technology-Air Transportation Technology	-1.076	-6.662	4.509	1.000
0.95	Sensor Technology-Physical, Chemical, and Mechanical System	-0.922	-5.758	3.914	1.000
0.95	Physical, Chemical, and Mechanical System-Computer Technology	0.686	-3.060	4.432	1.000
0.95	Production Technology-Autonomous/Robotics Technology	1.075	-4.831	6.981	1.000
0.95	Physical, Chemical, and Mechanical System-Photonics and Phononics Technology	-0.992	-7.455	5.471	1.000
0.95	Energy and Power Technology-Autonomous/Robotics Technology	-0.552	-4.237	3.134	1.000
0.95	Maritime Transportation Technology-Communications Technology	-0.897	-7.473	5.679	1.000
0.95	Energy and Power Technology-Computer Technology	-0.361	-2.982	2.260	1.000
0.95	Ground Transportation Technology-Air Transportation Technology	0.297	-5.138	5.732	1.000
0.95	Photonics and Phononics Technology-Air Transportation Technology	-0.461	-7.815	6.893	1.000
0.95	Production Technology-Air Transportation Technology	-0.874	-7.904	6.157	1.000
0.95	Computer Technology-Autonomous/Robotics Technology	-0.191	-3.625	3.244	1.000
0.95	Physical, Chemical, and Mechanical System-Autonomous/Robotics Technology	0.495	-4.059	5.050	1.000
0.95	Sensor Technology-Autonomous/Robotics Technology	-0.427	-5.026	4.173	1.000
0.95	Energy and Power Technology-Biological Technology	0.262	-3.139	3.664	1.000
0.95	Sensor Technology-Biological Technology	0.387	-3.988	4.762	1.000
0.95	Sensor Technology-Computer Technology	-0.236	-4.036	3.564	1.000
0.95	Sensor Technology-Energy and Power Technology	0.125	-3.903	4.153	1.000
0.95	Photonics and Phononics Technology-Ground Transportation Technology	-0.758	-6.760	5.245	1.000
0.95	Production Technology-Photonics and Phononics Technology	-0.413	-7.890	7.065	1.000
0.95	Space Technology-Photonics and Phononics Technology	-0.615	-6.754	5.524	1.000
0.95	Production Technology-Physical, Chemical, and Mechanical System	0.580	-5.513	6.672	1.000
0.95	Space Technology-Physical, Chemical, and Mechanical System	0.377	-3.969	4.723	1.000
0.95	Space Technology-Production Technology	-0.203	-5.950	5.545	1.000

In spite of the results of the ANOVA, which indicated there were differences in the underlying distributions at the 95% CI, the TKHSD test reveals no statistically significant different pairings.

Table E-49. Results of TKHSD for timeframes

CI	Treatment	Difference	Lower Bound	Upper Bound	Adjusted p-value
0.95	Short-term-Long-term	-4.316	-5.849	-2.782	0.000
0.95	Short-term-Medium-term	-4.002	-5.167	-2.838	0.000
0.95	Medium-term-Long-term	-0.314	-2.083	1.456	0.909

Results indicate short-term forecasts have statistically different temporal error rates than do longer-term forecasts.

E.1.11 Results of TKHSD when one-year forecasts are removed from the data set

We observed that 30% of all model forecasts had a one-year forecast horizon, which was substantially larger than the number of one-year forecasts for all other methods. We wanted to determine if the differences observed in the model method compared to the other methods could be attributed to the number of one-year forecasts. For this test we removed all forecasts with a horizon of one year and reran the ANOVA and TKHSD.

Table E-50. Results of ANOVA with no one-year forecasts

Attribute	Variable of interest	F-Statistic	Numerator df	Denominator df	p-Value
Method	TFE	11.829	7	107.13	0.000
Method	STFE	0.899	7	108.91	0.510
TimeFrame	TFE	19.300	2	104.11	0.000
TimeFrame	STFE	2.769	2	105.38	0.067

Results indicate there are differences within the method and timeframe even with the removal of one-year forecasts.

Table E-51. Results of TKHSD by methods with no one year forecasts

CI	Treatment	Difference	Lower Bound	Upper Bound	Adjusted p-value
0.95	Multiple-Models	3.575	-0.443	7.594	0.122
0.95	Qualitative Trend Analysis-Expert Analysis Methods	2.667	-0.380	5.714	0.136
0.95	Models-Gaming and Scenarios	-5.767	-11.170	-0.365	0.270
0.95	Source Analysis-Models	3.879	0.221	7.538	0.289
0.95	Gaming and Scenarios-Expert Analysis Methods	3.667	-1.497	8.832	0.376
0.95	Models-Expert Analysis Methods	-2.100	-5.182	0.982	0.433
0.95	Quantitative Trend Analysis-Qualitative Trend Analysis	-2.389	-6.510	1.732	0.644
0.95	Quantitative Trend Analysis-Gaming and Scenarios	-3.389	-9.252	2.473	0.647
0.95	Quantitative Trend Analysis-Models	2.378	-1.769	6.525	0.657
0.95	Expert Sourcing-Expert Analysis Methods	1.408	-1.100	3.915	0.681
0.95	Source Analysis-Expert Analysis Methods	1.779	-1.517	5.076	0.723
0.95	Qualitative Trend Analysis-Models	4.767	1.331	8.203	0.758
0.95	Models-Expert Sourcing	-3.507	-6.475	-0.540	0.845
0.95	Gaming and Scenarios-Expert Sourcing	2.260	-2.837	7.357	0.879
0.95	Qualitative Trend Analysis-Expert Sourcing	1.260	-1.671	4.191	0.895
0.95	Multiple-Expert Analysis Methods	1.475	-2.216	5.167	0.927
0.95	Multiple-Gaming and Scenarios	-2.192	-7.964	3.580	0.943
0.95	Source Analysis-Quantitative Trend Analysis	1.501	-2.808	5.810	0.964
0.95	Source Analysis-Gaming and Scenarios	-1.888	-7.416	3.640	0.968
0.95	Quantitative Trend Analysis-Expert Sourcing	-1.129	-4.869	2.611	0.984
0.95	Qualitative Trend Analysis-Multiple	1.192	-2.800	5.184	0.985
0.95	Quantitative Trend Analysis-Multiple	-1.197	-5.816	3.421	0.994
0.95	Source Analysis-Qualitative Trend Analysis	-0.888	-4.517	2.741	0.996
0.95	Qualitative Trend Analysis-Gaming and Scenarios	-1.000	-6.383	4.383	0.999
0.95	Source Analysis-Expert Sourcing	0.372	-2.818	3.561	1.000
0.95	Quantitative Trend Analysis-Expert Analysis Methods	0.278	-3.553	4.110	1.000
0.95	Source Analysis-Multiple	0.304	-3.881	4.489	1.000
0.95	Multiple-Expert Sourcing	0.068	-3.529	3.665	1.000

Even though the ANOVA indicated differences, the TKHSD did not identify any at a 95% confidence level. This means the results of the previous TKHSD was in part a result of the one-year forecasts.

Table E-52. Results of TKHSD by timeframe with no one year forecasts

CI	Treatment	Difference	Lower Bound	Upper Bound	Adjusted p-value
0.95	Short-term-Long-term	-4.633	-6.647	-2.619	0.000
0.95	Short-term-Medium-term	-3.908	-5.436	-2.379	0.000
0.95	Medium-term-Long-term	-0.725	-2.983	1.532	0.730

Results indicate no difference between the two TKHSD tests. This indicates the results with respect to timeframe do not depend on the one-year forecasts.

E.1.13 Results of regression analysis

R has the ability to create the best linear model from the full set of variables presented. It has two modes of doing this: 1) reverse (indicated by “R” in table E-53 1st column), where is takes a full model with all available attributes and then removes variables one at a time until it finds the best performing model with respect to the adjusted R^2 value; and two) forward (indicated by “F” in table E-53, first column), where the model starts with no variables and then build to the best model by introducing variables as necessary until it gets the best adjusted R2.

In addition to building the model looking for the best fit, we also used several transformations on the continuous variables in an effort to see if model performance could benefit. We applied a combination of exponential and natural log transformations to the STFE (dependant variable) and timeframe in years (independent variable). The first column of table E-53 indicates which transformation was applied to the variable(s) and also the direction of the model build (forward or reverse). The second column indicates which attributes were included in the best performing model. The third column indicates how much of the change in the dependent variable is accounted for by the independent variables.

Table E-53. Results of the regression analysis with variable transformation

Variable Transformation	Included attributes	Adjusted R^2
Log Time R	Region of Origin, Tech Area, Method, Timeframe, Publication type, Prediction type, TRL, Degree of Realization	0.3523
Exp Time R	Tech Area, Method, Timeframe, Publication type, Prediction type, TRL, Degree of Realization, Timeframe in years	0.3483
Log STFE F	Timeframe in years, Tech area, Timeframe, TRL	0.1039
Log STFE R	Tech area, Timeframe, TRL, Timeframe in years	0.1039
No mods F	Tech Area, Timeframe in years, Timeframe, TRL, Region of Origin	0.0995
No mods R	Tech Area, Timeframe, TRL, Region of Origin, Timeframe in years	0.0995
Log Time F	Tech Area, TRL, Region of Origin, Timeframe in years	0.0816
Exp Time F	Tech Area, Timeframe, TRL, Region of Origin	0.0780
Log STFE Log Time F	Timeframe in years, Tech area, Region of Origin	0.0712
Log STFE Log Time R	Tech Area, TRL, Timeframe in years	0.0712
Log STFE Exp Time F	Tech Area, Timeframe, TRL	0.0662
Log STFE Exp Time R	Tech Area, Timeframe, TRL	0.0662
Exp STFE F	Timeframe, Publication Type	0.0155
EXP STFE R	Timeframe, Publication Type	0.0155
Exp STFE Log Time F	Timeframe, Publication Type	0.0155
Exp STFE Log Time R	Timeframe, Publication Type	0.0155
Exp STFE Exp Time F	Timeframe, Publication Type	0.0155
Exp STFE Exp Time R	Timeframe, Publication Type	0.0155

The low R^2 value associated with all of the model runs indicates a poor fit. In addition to linear regression models, we also developed polynomial models in the second and third degree. No model structure and attribute selection resulted in an adjusted R^2 value that met our threshold of .80.

E.1.14 Results of data mining algorithms

In conjunction with the classical regression techniques, we also applied four data mining algorithms to the data set to determine if we could develop classifying algorithms that could provide better results than the regression models. These four algorithms were the K-Nearest Neighbor (KNN), Partitioning Trees, Support Vector Machines (SVM), and Neural Networks. For all algorithms, we divided the data set into three randomly selected data sets (training set with 60% of the records, validation set with 20% of the records, and the testing set with 20% of the records). Models were developed using the training set and parameters refined using the validation set. Once we found a combination of parameters that provided the best predictive value, we ran the model against the test set and reported the value obtained from that single run. This is standard practice in data mining and precludes the biasing of models by exposing them to the test data set. We used the same data set for all analysis.

Because we were going to analyze both the temporal error and success rates, we used a single data source of 736 records. This data set consisted only of the forecasts for which the event had been realized. We did this so we could compare results between model performance for both success classification and error magnitude classification. We used the seed value of 42 for our sampling algorithm to ensure a consistent allocation of records across multiple days of analysis.

The classifying algorithms work by looking at attributes and then predicting which class the record should belong to. TFE and STFE had to be classified into different ranges to enable the classifying algorithms to work. Table E-54 shows the rules used to convert STFE and TFE into classes.

Table E-54. Conversion table for TFE and STFE into Classes

TFE	Category	STFE Years	Class Is
0	A	-37.5	A
5	B	-32.5	B
10	D	-27.5	C
15	E	-22.5	D
20	F	-17.5	E
25	G	-12.5	F
30	H	-7.5	G
35	I	-2.5	H
		2.5	I
		7.5	J
		12.5	K
		17.5	L
		22.5	M
		27.5	N
		32.5	O

KNN: R does not have a KNN package that accommodates nominal variables. Given the simple nature of a KNN algorithm, we developed our own code in VBA that allowed us to use Hamming distances and value metric differences (VDM)[14] as substitutes for nominal values. We compared and contrasted results of both the Hamming distance and VDM along with different values for K (the number of records in the training set used to determine what the classification should be). We let K = (1, 3, 5, 7). Our selection of odd numbers for K was deliberate to avoid ties in voting. We did not use the KNN algorithm on the STFE and TFE analysis since the code was written specifically for Success/Fail analysis. Tables E-55 through E-58 provide the results of the KNN test using various values of K for different numbers of attributes concurrently considered in the distance formula. 'Attribute applied' indicates which attribute set provided the best error rate for a given set of nearest neighbors. The next four columns present the confusion matrix. 'P success' indicates how many were predicted to be successful while 'P fail' indicates how many were predicted to have been failed forecasts. These predicted fails and successes are compared to the actual classification. 'Error rate' indicates the percentage of forecasts that were misclassified.

[14] Wang, Hui, Nearest Neighbors by Neighborhood Counting, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:6 (2006).

Table E-55. Results of KNN test using only 1 attribute

K	Attribute Applied	Actually Successful		Actually Failed		Error rate
		P Success	P Fail	P Success	P Fail	
1	Sys V Comp	36	25	48	38	0.497
3	Sys V Comp	40	21	45	41	0.449
5	TRL	40	21	46	40	0.456
7	Pub Type	37	24	45	41	0.469

Using three nearest neighbors to vote on the classification of the forecasts results in marginally better result when using only one attribute.

Table E-56. Results of KNN test using two attributes

K	Attributes Applied	Actually Successful		Actually Failed		Error rate
		P Success	P Fail	P Success	P Fail	
1	Timeframe Sys V Comp	35	26	40	46	0.449
3	Tech Area TRL	37	24	37	49	0.415
5	Tech Area Timeframe	36	25	42	44	0.456
7	Pred Type Sys V Comp	40	21	45	41	0.449

Using three nearest neighbors to vote on the classification of a forecast results in marginally better predictions when using two attributes.

Table E-57. Results of KNN test using three attributes

K	Attributes Applied	Actually Successful		Actually Failed		Error rate
		P Success	P Fail	P Success	P Fail	
1	Method Timeframe Pred Type	40	21	41	45	0.422
3	Timeframe Pub Type Sys V Comp	38	23	44	42	0.456
5	Tech Area Method Pred Type	38	23	38	48	0.415
7	Tech Area Method Sys V Comp	37	24	41	45	0.442

Using five nearest neighbors to vote on the classification of a forecast results in marginally better predictions when using three attributes.

Table E-58. Results of KNN test using 4 attributes

K	Attributes Applied	Actually Successful		Actually Failed		Error rate
		P Success	P Fail	P Success	P Fail	
1	Method Timeframe Pub Type Sys V Comp	41	20	34	52	0.367
3	Method Timeframe TRL Sys V Comp	39	22	43	43	0.442
5	Tech Area Method Timeframe Sys V Comp	40	21	34	52	0.374
7	Tech Area Method Timeframe TRL	37	24	47	39	0.483

Using a single nearest neighbors to vote on the classification of a forecast resulted in marginally better predictions when using two attributes. The yellow highlighted row in table 58 resulted in the best KNN solution for Hamming distances.

Applying the VDM method for converting categorical values into numerical distances resulted in a consistently (but marginally) worse error rate. We do not provide the results of these training runs in this appendix.

Partitioning Trees: We used the RPart package for R to conduct the tree analysis. There are four variables to adjust when refining the model:

1. Min Split- the minimum number of records that must exist in a node before it can be a candidate for partitioning,
2. Min Bucket – the minimum number of records that must exist in a leaf node
3. Max Depth – the maximum depth associated with a tree solution
4. Complexity – controls the sensitivity of the model to the number of partitions it creates. Small values of complexity result in very large complex trees while large values of complexity result in small simple trees. Generally, more complexity provides better coverage of the underlying data, but can result in over fitting. Less complex trees result in less coverage of the data but rarely result in over fitting.

We varied all four variables over some specified ranges until we found the model with the best predictive results. Table E-59 provides the results of our investigation of the best parameter and variable selections for the tree algorithm.

Table E-59. Results of the Tree Partitioning algorithm

Conditions	Target Variable	Error Rate	Parameter values (Min Split, Min Bucket, Max Depth, Complexity)
All Categorical	Success	47%	(20,10, 10, .01)
All Cat (-TF) + TFYears	Success	45%	(20,10, 15, .01)
All categorical	TFE	26%	(20,10, 10, .01)
All Cat (-TF) + TFYears	TFE	26%	(20,10, 10, .01)
All categorical	STFE	51%	(10, 10, 20, .01)
All Cat (-TF) + TFYears	STFE	51%	(5, 5, 20, .01)

The ‘conditions’ column indicates that if we used all categorical (nominal) variables in our model or if we substituted timeframe in years for the categorical variable timeframe (short, medium, long term). The second column indicates which was the target variable we were trying to classify. Error rate is the percentage of classifications that were wrong; lower numbers are better for this category. ‘Parameter values’ shows the values associated with the four parameters discussed above.

SVM: We used the Kernlab package for R and the *ksvm* routine to create our SVM models. Under this package, there is only one parameter to set which is the kernel used to process the data. There are eight different kernels available. We used all eight kernels and evaluated which ones provided the best results given the conditions and target variable. We evaluated model performance based on the lowest error rate. Table E-60 provides results for the SVM analysis.

Table E-60. Results of the support vector machine analysis

Conditions	Target Variable	Error Rate	Parameter values (kernel)
All categorical variables only	Success	41%	Poly -2
All Cat (-TF) + TFYears	Success	41%	Laplacian
All categorical variables only	TFE	29%	anovadot
All Cat (-TF) + TFYears	TFE	100%	Poly-2
All categorical variables only	STFE	77%	vanilladot
All Cat (-TF) + TFYears	STFE	68%	anovadot

The number following the “Poly” term indicates the degree associated with the polynomial used to process the data.

Neural Networks: We used the *nnet* package in R to develop our neural networks. There is only one variable in the neural network model that we adjusted. This variable is the number of hidden layers between the input nodes and the output node. We varied the number of hidden layers from one to 10 to find the best performing model. Unlike the other models considered in this analysis, the neural networks used the actual numerical values of the TFE and STFE to train the model. Therefore, instead of generating an error rate, models developed for these two target variables resulted in correlation coefficients (R^2). Table E-61 contains the results of the success variable classification while table E-62 contains the results of the TFE and STFE analysis

Table E-61. Results of applying neural networks on success as the target variable

Conditions	Variable	Error Rate	Parameter values Hidden Layers
All categorical variables only	Success	39%	HL = 1
All Cat (-TF) + TFYears	Success	41%	HL = 5

When attempting to classify the success variable with neural networks, a single hidden layer on all categorical variables resulted in the best solution.

Table E-62. Results of applying neural networks on TFE and STFE as the target variable

Conditions	Variable	R^2	Parameter values Hidden Layers
All categorical variables only	TFE	0.3332	HL = 3
All Cat (-TF) + TFYears	TFE	0.2990	HL = 3
All categorical variables only	STFE	0.1685	HL = 4
All Cat (-TF) + TFYears	STFE	0.3204	HL = 4

Neural networks performed on par with classical regression techniques in the development of predictive models. The low R2 value indicates all four of these are poor models for accurately predicting forecast errors.

Based on the poor performance of the classical regression techniques as well as the data mining techniques, we conclude the error rate association with these forecasts is either driven by a random component not considered in the study.