BRYCE

space and technology

Retrospective Analysis of Long-Term Forecasts

20 July 2018

Submitted To:

Luke Muehlhauser
Open Philanthropy Project

Submitted By:

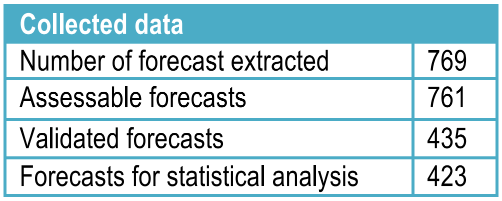
Carie A. Mullins
Senior Engineer
Bryce Space and Technology
724-766-2796
brycetech.com

Executive Summary

The purpose of this study was to provide a rigorous statistical evaluation of the accuracy of long-term forecasting by (1) verifying and validating at least 400 retrospective, long-term forecasts, and (2) conducting statistical analysis to determine if any forecasting methods, time frames, domain areas, and other key attributes yield a higher (or lower) probability of success. This work builds upon a previous retrospective analysis of forecasts conducted by Bryce Space and Technology and The Tauri Group in 2012, although the current effort is broader, focusing on predictions about economics, demographics, and other domains—not just technologies—and on predictions made over the long term (10 or more years).

Methodology and Results

Bryce Space and Technology extracted and analyzed long-term forecasts from 17 documents chosen by the Open Philanthropy Project. We extracted 769 forecasts from those documents; of the 769 forecasts, 761 were found to be timely, specific, complete, and relevant enough to be further verified and assessed for accuracy. Figure ES-1 summarizes data collection metrics.



Collected data	
Number of forecast extracted	769
Assessable forecasts	761
Validated forecasts	435
Forecasts for statistical analysis	423

Figure ES-1. Collected Documents and Forecasts

We categorized forecasts based on six objective attributes: forecast methodology, time frame, geographic origin of the forecast, geographic region forecasted about, publication type, and domain area. We were able to verify 435 forecasts by determining if the predicted event occurred and, if so, when it occurred. The 435 validated forecasts were used to provide descriptive statistics of our data. Of these 435 forecasts, 12 were outliers that were excluded during the statistical analysis. Consequently, 423 forecasts were used for the statistical analysis to determine if a given forecast methodology was better than others given various conditions.

Key Findings

- ✓ All forecast methodologies provide more accurate predictions than uninformed guesses
- ✓ Forecasts based on quantitative and qualitative trend analyses are more accurate than forecasts based on expert opinion
- ✓ Forecasts of the economy are the least accurate in comparison to all other domains
- ✓ All forecasts time frames (10 years, 11-20 years, and >20 years or more) are more accurate than an uninformed guess, but no time frame was more accurate than another
- ✓ As a whole, long-term forecasts are equally likely to over or underestimate the event date
- ✓ A predictive model of forecast accuracy could not be developed

1. Introduction

1.1. Background

This report describes the retrospective analysis of long-term forecasts that Bryce Space and Technology conducted for the Open Philanthropy Project. The Open Philanthropy Project is a organization that provides grants for, among other things, research addressing long-term issues with widespread effects but minimal support, such as pandemic preparedness. One part of Open Philanthropy Project's undertaking is to identify funding opportunities for long-term causes. As part of identifying long-term funding opportunities, Open Philanthropy reviews documents that forecast the future state of potential research areas. However, Open Philanthropy Project seeks to better understand the accuracy of these forecasts in order to better identify funding opportunities.

To determine which forecast documents merit the most consideration, the Open Philanthropy Project asked Bryce Space and Technology to conduct a retrospective analysis of forecasts in multiple areas, including economics, environment, and agriculture. Bryce Space and Technology (then under the name of "The Tauri Group") had previously developed a methodology to retrospectively analyze technology forecasts for the Assistant Secretary of Defense for Research & Engineering (ASDR&E).[1] The analyses showed that long-term technology forecasts (more than 10 years) were less accurate than short- or medium-term forecasts, and that technology forecasts generated using quantitative information were more accurate than forecasts generated using other methods. The Open Philanthropy Project wished to determine whether these findings held true for long-term forecasts in domains outside of the technology domain.

1.2. Objectives

The purpose of this study was to provide a rigorous statistical evaluation of the accuracy of long-term forecasting by (1) verifying and validating at least 400 retrospective, long-term forecasts extracted from client-selected documents, and (2) conducting a statistical analysis to determine if any forecasting methods, time frames, domain areas, and other key attributes yield a higher (or lower) probability of success.

This work builds upon a previous retrospective analysis of forecasts conducted by Bryce Space and Technology and The Tauri Group in 2012, although the current effort is broader, focusing on predictions about economics, demographics, and other domains—not just technologies—and on predictions made over the long term (10 years or more). The results of the study will be used by the Open Philanthropy Project to more reliably identify long-term funding opportunities.

[1] In 2011, Bryce Space and Technology analyzed 300 technology forecasts for the ASDR&E. That project culminated in the following report: https://ideas.repec.org/a/eee/tefoso/v80y2013i6p1222-1231.html. In

2. Methods

The Open Philanthropy Project reviewed and selected 17 documents from government, industry organizations, and strategic analysis firms that made long-term forecasts, defined as those with a time frame of 10 or more years. Bryce Space and Technology was asked to analyze at least 400 forecasts from these documents. The study was conducted in three phases:

1. ***Extract technology forecasts from documents and record attributes in database.*** During this phase, we extracted forecasts from the 17 documents, characterized each forecast based on six common forecasting attributes (methodology, time frame, geographic origin of the forecast, geographic region forecasted about, publication type, and domain area) and added the data to a Quickbase database that served as a repository for analysis data.

2. ***Verify whether forecasted events occurred.*** During this phase, we conducted research to verify whether (1) forecasted events had occurred and, if so, (2) when they occurred.

3. ***Assess forecast accuracy and identify key findings.*** During this phase, we conducted statistical tests—developed during the 2011 and 2012 studies—to ascertain the accuracy of forecasts and determine which of the six attributes are most associated with forecast accuracy. We also refined our previously-developed dictionary of ambiguous language to develop clear and concise interpretations of forecasts. The intent of this dictionary is to help users read and understand future forecasts.

Figure 1 summarizes the study methodology.

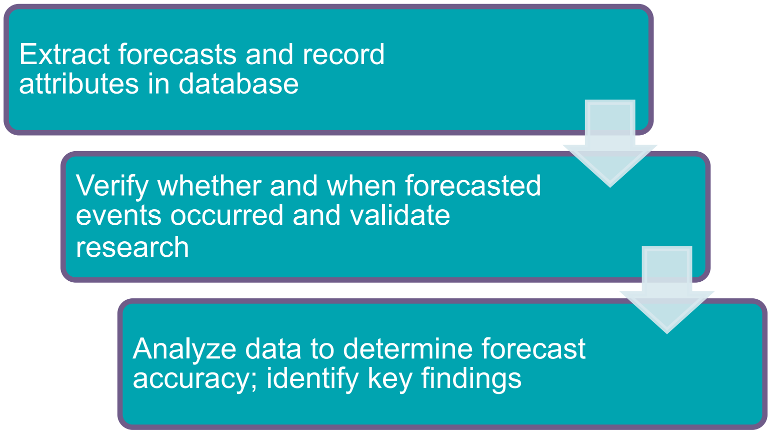

Extract forecasts and record attributes in database
Verify whether and when forecasted events occurred and validate research
Analyze data to determine forecast accuracy; identify key findings

Figure 1. Study Methodology

The following sections provide more detail about the three phases of the study.

2.1. Phase I: Forecast Extraction

2.1.1. Document selection

As mentioned above, the Open Philanthropy Project selected 17 documents that made long-term forecasts in six domains: technology, economy, demographics, energy, environment, and natural resources. The documents were selected based on internal criteria for analytical rigor and relevance to Open Philanthropy Project's mission. Table 1 shows the initial 17 documents that Bryce analysts reviewed for use in this study.

Author	Organization	Title	Year	Publication Type
Barney	Council on Environment Quality	The Global 2000 Report to the President	1980	Government reports & roadmaps
Battelle Memorial Institute	Battelle Memorial Institute	Agriculture 2000	1983	Industry organizations, associations, & societies
Brundtland Commission	United Nations (UN)	Report of the World Commission on Environment and Development: Our Common Future	1987	Government reports & roadmaps
Food and Agriculture Organization (FAO)	UN	Agriculture: Toward 2000	1980	Government reports & roadmaps
Gordon & Helmer	RAND	Report on a Long-Range Forecasting Study	1964	Industry organizations, associations, & societies
Intergovernmental Panel on Climate Change (IPCC)	IPCC	First Assessment Report	1990	Government reports & roadmaps
Kahn & Weiner	Hudson Institute	The Year 2000	1967	Strategic analysis firms
Leontief et al.	United Nations	The Future of the World Economy	1978	Government reports & roadmaps
Lesourne et al.	Organisation for Economic Cooperation and Development (OECD)	Facing the Future	1979	Government reports & roadmaps
Meadows et al.	Club of Rome	Limits to Growth	1972	Industry organizations, associations, & societies
Mesarovic & Pestel	Club of Rome	Mankind at the Turning Point	1974	Industry organizations, associations, & societies
National Aeronautics and Space Administration (NASA)	NASA	Outlook for Space	1972	Government reports & roadmaps
NASA	NASA	A Forecast of Space Technology 1980-2000[2]	1976	Government reports & roadmaps
Tenet	National Intelligence Council	Global Trends 2015: A Dialogue About the Future with Nongovernment Experts	2000	Government reports & roadmaps

[2] This document was initially assessed by Bryce Space and Technology and The Tauri Group in the 2012 study for ASDR&E. The Open Philanthropy Project asked that we analyze the document's long-term forecasts for this effort.

Author	Organization	Title	Year	Publication Type
UN	UN	Growth of the World's Urban and Rural Population, 1920-2000	1969	Government reports & roadmaps
UN	UN	Overall Socio-Economic Perspective of the World Economy to the Year 2000	1990	Government reports & roadmaps
Wilson	Workshop on Alternative Energy Strategies	Energy: Global Prospects 1985-2000	1977	Industry organizations, associations, & societies

Table 1. Long-term forecast documents identified for analysis

One of the initial 17 documents, the Brundtland Commission's "Report of the World Commission on Environment and Development: Our Common Future" was not included in the study because it cited other forecast documents rather than providing original forecasts. As a result, we extracted forecasts from 16 documents.

2.1.2. Extraction and Attribute Recording

Before extracting forecasts from the 16 documents, we first needed to determine their assessability. To be assessable, forecasts needed to meet four inclusion criteria. First, they needed to be accompanied by a clear, 10-year-or-greater time frame. Second, they needed to be old enough to allow their predictions to be evaluated. Third, the methodology by which they were produced needed to be identifiable. Fourth, their language needed to be specific enough to allow for subsequent analysis. We extracted 769 forecasts from the 16 documents; 761 of these were assessable. That is, they were timely, specific, complete, and relevant enough to proceed through the verification process.[3]

Forecast language is often vague or ambiguous, requiring interpretation. In many cases, extracted forecasts were specific enough for analysis but contained words or phrases that required interpretation before analysis could begin. During the 2012 study for ASDR&E, we developed a standard lexicon to enable consistent interpretation of vague terminology commonly used in forecasts. Analysis rules were also included in the standard lexicon to ensure different types of forecasts, such as multi-variable and contingent forecasts, were evaluated consistently.

During this study, we refined and updated the standard lexicon and analysis rules as needed. For example, we added a rule that for forecasts predicting growth each year within a specific time frame, the last year in that time frame should be the evaluation year.[4] We also added a rule that we would not attempt to find year of realization for steady-state/as-is forecasts if the event did not occur on the forecasted year.[5] Appendix A provides the complete standard lexicon and

[3] Eight forecasts that were not assessable were causal forecasts or used vague terms like "perhaps." Per our Standard Lexicon and Analysis Rules (Appendix A), these forecasts are to be excluded from the analysis.

[4] For example, forecast 2863 predicts the world's GDP will grow on average by 3.5% per year from 1990-2000. Rather than using the median year as the evaluation year, as is done for events forecasted within a range of time, we used 2000 as the year to evaluate the success of the forecast, because the forecast predicts that each year in the range (i.e., 1990 through 2000) would experience 3.5% growth).

[5] For example, forecast 2470 predicts that the bulk of the world population would continue to be Asian in the year 2000. If the bulk of the population in 2000 was not Asian, then the verifier would not seek to find the first year of realization (i.e., the first year the bulk of the world population was Asian) because the

analysis rules.

After extracting forecasts and identifying those that were assessable, we characterized assessable forecasts based on the six attributes we hypothesized could influence accuracy: forecast methodology, time frame, geographic origin of the forecast, geographic region forecasted about, publication type, and domain area. Domain area refers to the subject of the forecast, including technology, environment, economy, demographics, energy, and natural resources. We entered assessable forecasts and their associated attributes into a Quickbase database that served as a repository for all analysis information.

2.2. Phase II: Forecast Verification and Validation

We verified assessable forecasts by determining (1) whether the forecasted event occurred and, if so, (2) when it occurred. Researchers verified forecasts using open-source research. Only information from credible websites and sources was accepted from researchers. These sources include conference papers, government and international publications, newspaper articles, popular magazines with a reputable editing record, industry websites, electronic books, interviews with experts, and other sources with journalistic standards and appropriate subject matter expertise.

In order to ensure verification information was correct and unbiased, we attempted to find verifying information from at least two independent sources. If sources conflicted—or if we were unable to find verifying information using open-source research—we relied on expert opinion for clarification or verification. However, in the present study, we encountered a number of forecasts that could only be verified using a single source because that source appeared to be the authority on a particular data set. For example, the FAO, an agency of the UN, maintains and publicly shares eight databases on subjects including forestland, caloric consumption, water availability, and more. The World Bank and other reputable publications cite FAO as their sole source for statistics on food and agriculture, further demonstrating that FAO is the authority on these topics. As a result, forecasts related to food and agriculture were typically verified using only FAO data. These types of “soft verifications” comprise 15% of our 435 validated forecasts and were mostly forecasts relating to demographics.

After adding verifying information to the database, we characterized the degree of interpretation involved in verifying the forecast (a measure of how specific the forecast language was, and how closely it aligned with ground truth sources) using a five-point Likert scale.

After a forecast was verified, a senior analyst validated that the forecast and its attributes had been characterized properly and that verifying information was clear, credible, and reproducible. Where needed, further research was conducted on forecasts that required additional sources. A different researcher than the one who originally attempted verification typically conducted this additional research.

Degrees of Interpretation:

1. All interpretation
2. A lot of interpretation
3. Moderate interpretation
4. Little interpretation
5. No interpretation

Using these methods, we verified and validated 57% of the assessable forecasts (435 of the

forecast states that the Asian population would still be the bulk, rather than that it would reach the bulk at some point in the future.

761). Twelve forecasts were outliers that were excluded during the statistical analysis, resulting in a final sample size of 423 forecasts. All verification information, including the degree of interpretation rating, analyst justifications, source citations, and page numbers, is included in the database in sufficient detail to allow each forecast's verification to be reproduced by other analysts.

The remaining 325 unverified forecasts were not included in the statistical analysis due to time constraints and an absence of sufficient information to characterize their ground truth with a high degree of confidence.

2.3. Phase III: Data Analysis

The purpose of the data analysis phase is to quantify and characterize the forecasts in our sample, identify which of the six attributes yield the most accurate forecasts, and determine the likelihood that a forecasted event will occur within the specified time frame, given certain conditions.

2.3.1. Characterization Metrics

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.

- **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.
- **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. For example, when verifying a forecast that stated that the share of the automobile's energy would drop by a certain time frame, analysts had to loosen their definition of automobiles to include not only light-duty, privately owned passenger vehicles but also commercially owned vehicles carrying property, such as buses and trucks.
- **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 "long-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.

2.3.2. Assessment Process

To assess forecasts, 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 two criteria for viability (completeness and language specificity). We collected a total of 761 forecasts that met this requirement. These forecasts were then verified to determine if the forecasted event was realized. Of the 761 forecasts, we were able to verify whether 435 of them were realized or not, and when they were realized. Using the metrics described above, 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.

Figure 2 illustrates the process used to assess the forecast results.

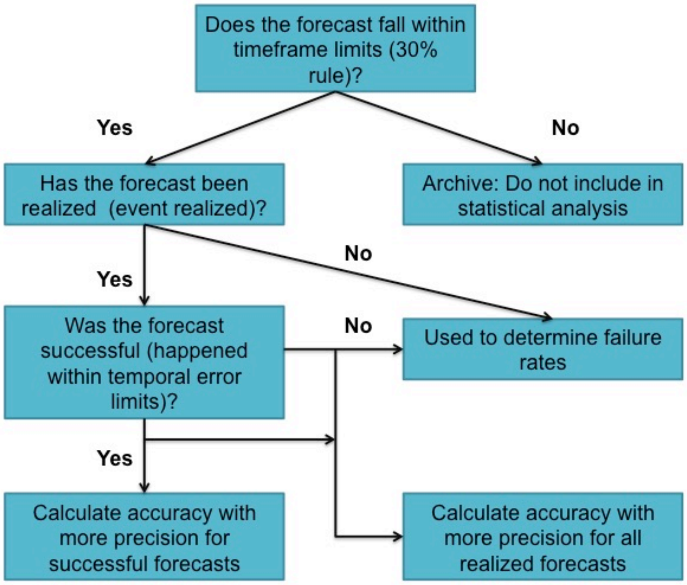

Does the forecast fall within timeframe limits (30% rule)?
Yes
No
Has the forecast been realized (event realized)?
Archive: Do not include in statistical analysis
No
Yes
Was the forecast successful (happened within temporal error limits)?
No
Used to determine failure rates
Yes
Calculate accuracy with more precision for successful forecasts
Calculate accuracy with more precision for all realized forecasts

Figure 2. Forecast Assessment Process

2.3.3. Statistical Analysis

We conducted a statistical analysis 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.

2.3.3.1. 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 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.

2.3.3.2. 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.

2.3.3.2. Key Attributes Analysis

The previous tests were designed to provide insights about the most appropriate forecasting methodology given a domain area 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 to determine the relationship (if any) of each of the six attributes on the accuracy of a forecast.

3. Results

This section describes the data in our sample. We describe metrics for the forecast documents, including geographic origin and publication type. We then describe metrics for the forecasts that we extracted, verified, and validated. These metrics include the geographic origin, methodology, and domain of the forecasts, as well as degree of interpretation used for verification. This section also provides a comparative analysis of metrics.

3.1. Document Collection Metrics

As mentioned in Section 2.1, we extracted forecasts from 16 pre-selected forecast documents. The geographic origin of these 16 documents was primarily worldwide organizations like the United Nations (50%), followed by the Americas (44%) and multiple regions (6%). Figure 3 shows the distribution of forecasts by region.

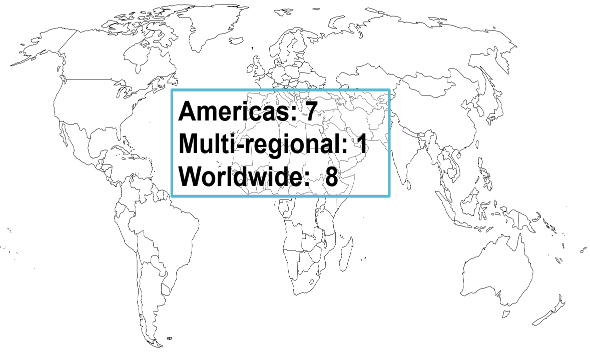


Figure 3. Number of forecast documents by

Most of the 16 documents (63%) were produced by government entities, followed by industry organizations (31%) and strategic analysis firms (6%). Figure 4 shows a breakout of forecast documents by publication type.

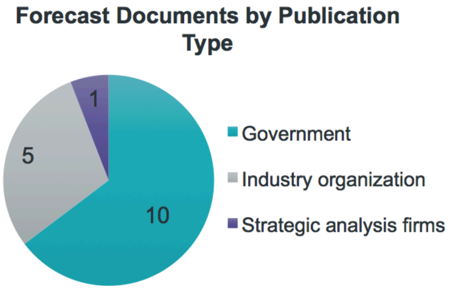


Figure 4. Number of forecast documents by publication type

3.2. Forecast Verification Metrics

As mentioned in Section 2.1, we extracted 769 forecasts from the 16 documents. Eight of these forecasts were deemed unassessable because they used language that was too vague to be assessed.[6] Therefore, the final number of assessable forecasts extracted from the 16 documents was 761. We were able to verify and validate 435 (57%) of these forecasts.

[6] Those forecasts are 2677, 2934, 2935, 2936, 3054, 2465, 2312, and 2428.

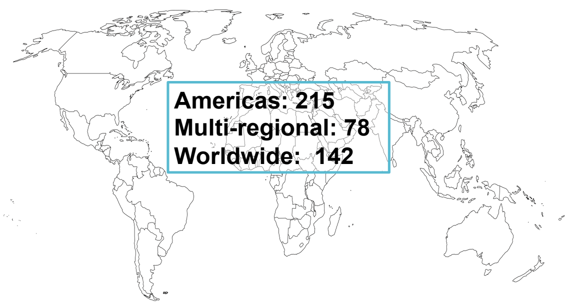
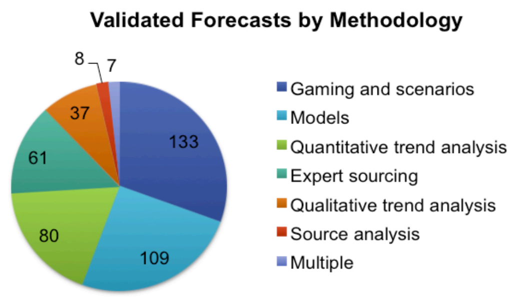


Figure 5. Number of validated forecasts by origin of region.

We characterized each of the 435 validated forecasts by six attributes. This section highlights some of the informative findings associated with the long-term forecasts in our sample and their distribution among the six attributes.

In terms of geographic origin, 215 of the 435 (49%) originated in the Americas, while 142 (33%) originated from worldwide organizations, such as the United Nations. Only 78 forecasts (18%) originated from a multi-regional organization, such as the Organisation for European Economic Co-operation (OECD). Figure 5 shows the geographic origin of forecasts in our sample.

In terms of forecast methodology, 133 (31%) of the 435 forecasts used gaming and scenarios, 109 (25%) used models, 80 (18%) used quantitative trend analysis, 61 (14%) used expert sourcing, 37 (9%) used qualitative trend analysis, eight used source analysis (2%), and seven (2%) used multiple forecasting methodologies.



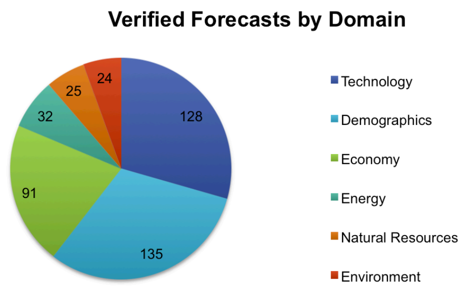


Figure 6. Number of validated forecasts by methodology

Figure 7. Number of validated forecasts by domain

Most of the 435 validated forecasts related to the demographics domain (135, or 31%), followed by technology domain (128, or 29%), economy (91, or 21%), energy (32, or 7%), natural resources (25, or 6%), and environment (24, or 6%). Figures 6 and 7 show the breakout of forecasts by methodology and domain.

Two hundred and thirty-seven (54%) of the 435 validated forecasts were verified using two sources, which is the minimum required for a forecast to be verified. When more data was available or necessary to provide additional context or support, verifiers used additional sources: 75 forecasts (17%) were verified with three sources, 42 (9%) with four sources, 21 (5%) with five sources, and three (0.6%) with six sources. Meanwhile, 67 (15.4%) were “soft verified” using only one source. As mentioned in Section 2.2, in

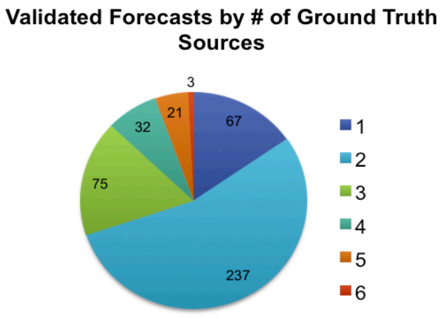


Figure 8. Validated forecasts by number of ground truth sources used for verification

these cases, the verifier relied on a single source because that source appeared to be the authority on a particular data set. Figure 8 shows the breakdown of validated forecasts by number of ground truth sources.

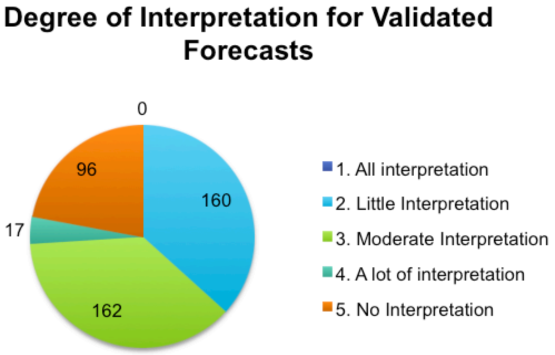


Figure 9. Validated forecasts by degree of interpretation used for verification

The majority of validated forecasts (160, or 37%) were verified using little interpretation. A little interpretation resulted, for example, when a verifier was required to convert metrics or when minor detail was absent from a ground truth source. A total of 41% of forecasts required moderate (37%) to a lot (4%) of interpretation. For example, a forecast required more interpretation when the verifier needed to assume the definition for a key term, such as “automobile” in a forecast about the energy use of automobiles. The remaining 22% of validated forecasts required no interpretation. No forecasts in our sample required “all interpretation” to verify. Figure 9 shows the degree of interpretation involved in verifying the forecasts in our sample.

3.3. Comparative Analysis of Forecast Metrics

This section provides highlights from a comparative analysis of attributes and other metrics from the 435 validated forecasts.

3.3.1. Success Among Forecast Documents

To limit biases based on document selection, we attempted to extract as many forecasts from each document as possible. However, due to assessability criteria for forecasts (timely, specific, complete, and relevant), a variable number of forecasts were extracted from among the documents. To determine whether any of the documents produced forecasts that were more accurate—or less accurate—than others, and whether accuracy or inaccuracy related to the degree of interpretation involved in verifying forecasts, we compared the success of the 16 documents.

Only 15 of the 16 documents produced successful forecasts. Battelle Memorial Institute's “Agriculture 2000” had the highest success rate. However, there were only seven validated forecasts from that document. Mesarovic & Pestel’s “Mankind at the Turning Point” and NASA’s “Outlook for Space” yielded forecasts that were 67% and 68% successful, respectively. However, the Mesarovic & Pestel document had a small sample size (three validated forecasts) while the NASA document yielded 28 validated forecasts. The NASA document used quantitative trend analysis as its forecast methodology, which could have impacted accuracy.

Table 2 compares success for the 16 documents, showing them in order of most successful to least successful.

Source	# Successes	Total # Validated	% Success	Method
Battelle Memorial Institute, “Agriculture 2000”	7	7	100.0%	Multiple
NASA, “Outlook for Space”	19	28	67.9%	Quantitative
Meadows et al., “Limits to Growth”	2	3	66.7%	Models
Mesarovic & Pestel, “Mankind at the Turning Point”	2	3	66.7%	Models
UN, “Growth of the World’s Urban and Rural Population, 1920-2000”	17	40	42.5%	Quantitative
NASA, “A Forecast of Space Technology 1980-2000”	15	37	40.5%	Qualitative
UN, “Overall Socio-Economic Perspective of the World Economy to the Year 2000”	15	38	39.5%	Models
Kahn & Weiner, “The Year 2000”	12	32	37.5%	Gaming & scenarios
Barney, “The Global 2000 Report to the President”	18	49	36.7%	Models
Wilson, “Energy: Global Prospects 1985-2000”	5	15	33.3%	Gaming & scenarios
Leontief et al., “The Future of the World Economy”	4	16	25.0%	Models
Lesourne et al., “Facing the Future”	19	76	25.0%	Gaming & scenarios
FAO, “Agriculture: Toward 2000”	2	10	20.0%	Quantitative
Gordon & Helmer, “Report on a Long-Range Forecasting Study”	11	61	16.4%	Expert source; Gaming & scenarios
Intergovernmental Panel on climate Change (IPCC), “First Assessment Report”	4	15	6.7%	Source analysis; Gaming & scenarios
Tenet, “Global Trends 2015: A Dialogue About the Future with Nongovernment Experts”	0	1	0.0%	Expert source

Table 2. Success compared across forecast documents

3.3.2. Success Among Forecasting Methodologies

Decision makers continue to seek forecasts that will provide strong predictive results, relying on forecasts generated from many different forecasting methods. Methods range from more data-driven quantitative trends to highly qualitative methods like source analyses. A number of studies examining forecast accuracy have found that quantitative methods produce more accurate forecasts than qualitative methods do.[7] Although forecasts derived from expert methods are less accurate than those derived from quantitative methods, expert methods are widely used. Armstrong found that among expert methods, those that rely on the collaborative

[7] J.S. Armstrong, M.C. Grohman, A comparative study for long-range market forecasting, Manag. Sci. 19 (2) (1972) 211–227; K.S. Lorek, C.L. McDonald, D.H. Patz, A comparative examination of management forecasts and Box–Jenkins forecasts of earnings, Account. Rev. 51 (2) (1976) 321–330;
J.S. Armstrong, Long-Range Forecasting: From Crystal Ball to Computer, Wiley, New York, NY, 1978.; W.M. Grove, D.H. Zald, B.S. Lebow, B.E. Snitz, C. Nelson, Clinical versus mechanical prediction: a meta-analysis, Psychol. Assess. 12 (1) (2000) 19–30.

judgment of participants produce markedly less accurate forecasts than those that average expert input to generate forecasts.[8] Furthermore, Carbone and colleagues observed that when experts reviewed and modified quantitative forecasts, accuracy decreased.[9] In our 2012 study for ASDR&E, we showed that forecasts produced using quantitative data are the most accurate, and are more accurate than those produced from expert input.[10]

Our findings corroborate these earlier observations. Specifically, our results indicate that long-term forecasts produced using quantitative trend analysis were the most successful, followed by qualitative trend analysis, models, and gaming and scenarios. Because there were only seven validated forecasts produced using multiple methods (and only one document that used these methods),[11] we cannot draw conclusions about this method's accuracy. Expert sourcing and source analysis had the least success. However, our sample size for the source analysis method was small (eight forecasts), which might have affected our findings on its success.

Table 3 shows the most and least successful forecasting methodologies, as well as the number of forecasts in our study that used these methods. Section 4.1 below, provides detail about whether any of these forecast methodologies are statistically more accurate than a random guess.

Method	# Successes	Total # Validated	% Success	# Documents
Multiple	7	7	100%	1
Quantitative trend analysis	38	80	48%	3
Qualitative trend analysis	15	37	41%	1
Models	41	109	38%	5
Gaming & scenarios	40	133	30%	4
Expert sourcing	10	61	16%	1
Source analysis	1	8	13%	1

Table 3. Success among forecast methodologies

3.3.3. Success Among Time Frames

This study evaluates long-term (10 years or more) forecasts. To determine whether any span of time within this long-term time frame (i.e., 10 years, 11-20 years, or more than 20 years) is more successful than another, we compared the success rates of these time frames. Our results show that the success of these time frames fell within similar ranges of 30% to 36%. However, our sample size for forecasts in the 10-year time frame was much smaller (47) than the other two time frames (260 and 128), which might have affected the results.

[8] J.S. Armstrong, How to make better forecasts and decisions: avoid face-to-face meetings, Foresight 5 (2006) 3–15.

[9] R.Carbone, A.Anderson, Y.Corriveau, P.P.Corson, Comparing different time series methods the value of technical expertise, individualized analysis, and judgmental adjustment, Manag. Sci. 29 (5) (1983) 559–566.

[10] C. Mullins, Retrospective Analysis of Technology Forecasting: In-scope Expansion, August 2012, http://www.dtic.mil/dtic/tr/fulltext/u2/a568107.pdf.

[11] Battelle Memorial Institute's "Agriculture 2000" used multiple methodologies to generate forecasts.

Time Frame	# Successes	Total # Validated	% Success
>20 years	94	260	36%
11-20 years	42	128	33%
10 years	16	47	34%

Table 4. Success among time frames

3.3.4. Success Among Domains

Previous research conducted by Bryce Space and Technology and The Tauri Group studied only forecasts of technology. For this analysis, we included forecasts from other domains, including demographics, the economy, energy, the environment, and natural resources. Table 5 shows the success of forecasts within each of these domains.

Success ranges between 18% (economy) to 41% (demographics). Based on the forecasts in our sample, forecasts about the economy domain were less successful than those from other domains. Section 4.2, below, analyzes whether this is a statistically significant finding. That is, whether economy forecasts are statistically less successful than forecasts in other domain areas.

Domain	# Successes	Total # Validated	% Success	# Documents
Demographics	56	135	41%	11
Economy	17	91	18%	8
Energy	10	32	31%	3
Environment	7	24	30%	3
Natural resources	10	25	40%	3
Technology	47	128	37%	12

Table 5. Success among domains

The majority of the economy forecasts (53%) used the gaming and scenarios methodology, while 37% used models. Only 10% used quantitative methods. The use of gaming and scenarios methods, which are less successful than other methods, may play a role in this result, although we would need to analyze a larger sample of economy forecasts to fully elucidate whether the economy is simply a more difficult domain to predict in the long term.

3.3.5. Degree of Interpretation by Document and Publication Type

Degree of interpretation measures a researcher's confidence in his or her verification. This confidence is a product of (1) the forecast's specificity, (2) ground truth source's specificity, and (3) the similarity of forecast language and ground truth language.

Forecasts that require considerable interpretation are generally more difficult to verify. For example, several forecasts made predictions about population but did not specify whether they were referring to a city population only or to the wider metropolitan area.[12] Several other

[12] Examples include 2476, 2477, and 2478.

forecasts made predictions about geographic areas such as “less developed/developing countries”[13] without specifying which countries those are comprised of, or about the “lower middle class”[14] without specifying what income level constitutes lower middle class. These forecasts required more interpretation to verify than did forecasts with clearer language.

In other cases, degree of interpretation increased because the predicted event appeared to have been realized before the forecast was made. Gordon & Helmer’s “Report on a Long-Range Forecasting Study,” published in 1964, produced a number of these forecasts. For example, one forecast predicted that by 1984, the military would be using (among other things) a “lightweight, rocket-type personnel armament.” However, the bazooka, a lightweight, rocket-type personnel armament, was introduced as early as 1942. This same source predicted that “general immunizations”[15] against both viral and bacterial diseases would be available in 2000. However, vaccines protecting against individual viruses and bacteria have been generally available since the early 1900s. In these examples, degree of interpretation increased because verifying researchers assumed that, because the predicted technologies appeared to exist before 1964, it was likely they were interpreting the forecast incorrectly.

In some cases, the long-term nature of the forecasts resulted increased interpretation, For example, several forecasts in our sample made predictions about Gross National Product in 1980, a term that became obsolete in the 1990s.[16] Similarly, a number of forecasts made predictions about geographic regions that no longer exist, such as Yugoslavia and the Soviet Union.[17] The forecasters could not have anticipated that these terms and regions would become obsolete.

To determine whether any documents in our sample required considerably more interpretation than others, we compared the degree of interpretation for each of the 16 document. Table 6 provides this data. No forecasts from any of the documents required “all interpretation” (score of 1). Only 17 validated forecasts—from five documents—required “a lot of interpretation” (score of 2). For one document—NASA’s “Forecast of Space Technology 1980-2000”—14% of its verified forecasts required “a lot of interpretation.” For two other documents—Gordon & Helmer’s “Report on a Long-Range Forecasting Study” and Kahn & Weiner’s “The Year 2000”—less than 10% of their verified forecasts required a lot of interpretation, while “Facing the Future” produced 4% requiring a lot of interpretation. The vast majority of verified forecasts required moderate interpretation (score of 3) and little interpretation (score of 2).

Documents that yielded forecasts with the least success generally had no forecasts that required a lot of interpretation. However, Gordon & Helmer’s “Report on a Long-Range Forecasting Study” had a success rate of only 16% and produced one of the highest numbers of forecasts requiring a lot of interpretation. This document used expert sourcing to generate the vast majority of its forecasts.[18]

[13] Examples from three different documents include 2353, 2894, and 2493.
[14] Examples include 2719, 2720, and 2721.
[15] This example is from forecast 2450.
[16] Includes forecasts 2711, 2710, 2709, 2512, 2511, 2510, 2497, 2496, 2495, 2494, 2469, 2377, 2375, 2373, 2372, 2371, and 2358.
[17] Includes forecasts 2969, 2574, 2597, 2598, 2612, 2613, 2704, and 2622.
[18] The forecasters used gaming and scenarios to generate one validated forecast in this document; all other forecasts were generated using expert sourcing.

Source	Method	Degree of Interpretation				
		5	4	3	2	1
Barney, “The Global 2000 Report to the President”	Models	27%	33%	39%	2%	0%
Battelle Memorial Institute, “Agriculture 2000”	Multiple	14%	43%	43%	0%	0%
FAO, “Agriculture: Toward 2000”	Quantitative	10%	60%	30%	0%	0%
Gordon & Helmer, “Report on a Long-Range Forecasting Study”	Expert source; Gaming & scenarios	25%	30%	38%	8%	0%
Intergovernmental Panel on climate Change (IPCC), “First Assessment Report”	Source analysis; Gaming & scenarios	27%	40%	33%	0%	0%
Kahn & Weiner, “The Year 2000”	Gaming & scenarios	38%	22%	31%	9%	0%
Leontief et al., “The Future of the World Economy”	Models	19%	31%	50%	0%	0%
Lesourne et al., “Facing the Future”	Gaming & scenarios	21%	41%	34%	4%	0%
Meadows et al., “Limits to Growth”	Models	67%	33%	0%	0%	0%
Mesarovic & Pestel, “Mankind at the Turning Point”	Models	33%	33%	33%	0%	0%
NASA, “A Forecast of Space Technology 1980-2000”	Qualitative	30%	27%	30%	14%	0%
NASA, “Outlook for Space”	Quantitative	0%	21%	79%	0%	0%
Tenet, “Global Trends 2015: A Dialogue About the Future with Nongovernment Experts”	Expert source	0%	100%	0%	0%	0%
UN, “Growth of the World’s Urban and Rural Population, 1920-2000”	Quantitative	0%	58%	43%	0%	0%
UN, “Overall Socio-Economic Perspective of the World Economy to the Year 2000”	Models	42%	34%	24%	0%	0%
Wilson, “Energy: Global Prospects 1985-2000”	Gaming & scenarios	7%	60%	33%	0%	0%

Key:
5 – No interpretation
4 – Little interpretation
3 – Moderate interpretation
2 – A lot of interpretation
1 – All interpretation

Table 6. Degree of interpretation compared across forecast documents

4. Key Findings

While Section 3 described our data sample, this section describes the results of our statistical analyses to determine forecast accuracy. We conducted three series of analyses - success tests, temporal error tests, and regression modeling The three attributes analyzed in this phase of the analysis were forecast methodology, domain area, and time frame. The metrics of interest in this analysis were forecast success rate, TFE, and STFE. During this phase of the analysis, we also wanted to determine whether there was validity behind two common assumptions associated with forecasts: 1) a forecast is better than a guess, and 2) the longer the forecast time frame, the greater the variance in forecast errors will be.

4.1 Test for Success

The outcome of a forecast is a binary event; it is either a success or a failure. Therefore, a methodology's outcome can be represented by a binomial distribution. The binomial test requires three parameters: the number of successes, the number of observations, and the hypothesized probability of success, ρ. For this test, we allowed ρ to be the success rate for a random guess, and we compared the number of successes and observations for each methodology against ρ. Thus, our hypothesis test was:

$$H_0 : r_i \geq \rho$$
$$H_A : r_i < \rho$$

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 probability of success equal to ρ. Failing to reject the null hypothesis indicated there was sufficient evidence that r_i was at least as large as ρ; rejecting the null hypothesis indicated that r_i may be less than ρ.

We labeled a forecast as successful if the forecasted event was realized within ±30% of the forecasted time frame around the forecasted year for a point estimate, or within the predicted range for a range estimate. If a forecasted event occurred outside of those limits, it was not a success.

To prevent outliers in the data from influencing this test, we based our theoretical probability of success on the statistically useful records that fell within the 99th percentile of data for time frame—thereby excluding the extreme outliers in forecast length—and the 95th percentile for temporal error. The trimmed data set included all records with a forecast length no greater than 36 years and no forecasts with a temporal error greater than 23 years.

To compare with the theoretical probability of success, no records had forecast lengths in excess of 36 years and twelve additional forecasts that had temporal errors in excess of 23 years from the original set of 435 records, for a trimmed sample size of 423 forecasts. All of the excluded forecasts were failed forecasts, resulting in a conservative test in favor of the forecasting methods. Details on the excluded forecasts can be found in Appendix C.

We used a uniform distribution from 1 to 36 to represent the predicted event realization. We assumed that the range in which a forecasted event could occur was within a 59-year time frame (longest forecast plus worst temporal error). Given these bounds on our parameters, the expected probability of success, ρ, for a randomly generated guess was 21%. Appendix C provides detailed information on how ρ was derived. Table 7 provides the results of the binomial test for each of the methodologies compared to ρ.

Method	Successes	Failures	Success Rate	p-value
Random guess			.21	
Gaming and scenarios	40	88	0.31	0.998
Models	41	66	0.38	1.000
Qualitative trend analysis	15	21	0.42	0.999
Quantitative trend analysis	38	40	0.49	1.000
Expert sourcing	10	48	0.17	0.302
Multiple	7	0	1.00	1.000
Source analysis	1	7	0.13	0.474

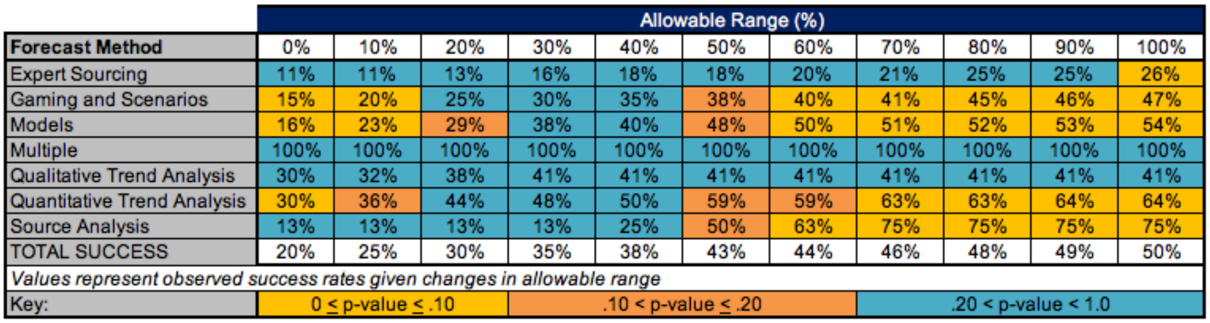
Table 7. Results of binomial test comparing methodologies to a random guess

We failed to reject the null hypothesis for all seven methodologies. We interpret this to mean that all seven methodologies generally perform better than a random guess. Even though expert sourcing and source analysis had lower success rates than a random guess, this result was not statistically significant. It should also be noted that source analysis and multiple methodologies do not have enough data to compare with a large degree of statistical certainty.

The definition of success is based on a subjectively-derived time frame value of 30%. To determine how robust our findings were relative to the definition of success, we conducted a sensitivity analysis where the allowable time frame ranged from zero (forecast must be accurate to within the exact year) to 100% (forecast can occur anytime within twice the length of the forecast). These results are provided in Table 8.

Key Finding: Forecast Methodology

In general, forecasts provide more accurate predictions than uninformed guesses. All seven methodologies are statistically more accurate than a theoretical probability of success. Source analysis and forecasts using multiple methods do not have enough data to compare with a large degree of statistical certainty.



	Allowable Range (%)										
Forecast Method	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Expert Sourcing	11%	11%	13%	16%	18%	18%	20%	21%	25%	25%	26%
Gaming and Scenarios	15%	20%	25%	30%	35%	38%	40%	41%	45%	46%	47%
Models	16%	23%	29%	38%	40%	48%	50%	51%	52%	53%	54%
Multiple	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Qualitative Trend Analysis	30%	32%	38%	41%	41%	41%	41%	41%	41%	41%	41%
Quantitative Trend Analysis	30%	36%	44%	48%	50%	59%	59%	63%	63%	64%	64%
Source Analysis	13%	13%	13%	13%	25%	50%	63%	75%	75%	75%	75%
TOTAL SUCCESS	20%	25%	30%	35%	38%	43%	44%	46%	48%	49%	50%
Values represent observed success rates given changes in allowable range											
Key:	0 ≤ p-value ≤ .10			.10 < p-value ≤ .20				.20 < p-value < 1.0			

Table 8. Results of sensitivity analysis on allowable range

All methods responded to the sensitivity analysis in an intuitive fashion. However, forecast success was fairly robust with high p-values until the allowable range was increased or decreased about +/- 20%. The aberrant results associated with the qualitative trend analysis method are due to the distribution of temporal errors. For this method, temporal errors were extremely small (less than one year) or extremely large (twice as long as the forecast period itself). This means the method's success rate will not benefit from the increased allowable range until the range is large enough to include the realized forecasts with large errors.

Quantitative trend analysis and qualitative trend analysis both had relatively high success rates compared to the other forecast methodologies. To determine whether this was a statistically significant observation, we developed a test to compare the success rates of these methodologies to the other forecasting methods. Our primary interest was in comparing both methods against the two other methods with relatively large sample sizes. While forecasts using multiple methods also had a high success rate, the sample size is too low to determine statistical significance. We excluded both source analysis and multiple from this test due to the low sample size.

We tested both trend methods individually against the other methods using Fisher's Exact Test. The resulting p-value represents the likelihood that the two data sets came from the same binomial distribution. Our hypothesis test was:

Key Finding: Trend Analyses

Forecasts based on quantitative and qualitative trend analyses are more accurate than forecasts based on expert opinion.

Quantitative trends have a higher success rate than other methods but are only statistically more successful than gaming and scenarios and expert sourcing.

$H_0 : r_t \leq r_s$	Where: r_t is the observed success rate of a trend 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 observed successes if both methods had similar probabilities of success. The implication is that the trend method 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 and Table 10 provide the results of this test.

Test	p-value
Quantitative trend analysis tests	
Quantitative trend analysis vs Gaming & scenarios	0.0093
Quantitative trend analysis vs Qualitative trend analysis	0.3093
Quantitative trend analysis vs Expert sourcing	0.0001
Quantitative trend analysis vs Models	0.1035
Qualitative trend analysis tests	
Qualitative trend analysis vs Gaming & scenarios	0.1658
Qualitative trend analysis vs Quantitative trend analysis	0.8167
Qualitative trend analysis vs Expert sourcing	0.0094
Qualitative trend analysis vs Models	0.4342

Table 9. Results of Fisher's Exact Test for Quantitative and Qualitative Trend Analysis

We rejected the null hypothesis for quantitative trend analysis, compared to gaming and scenarios, and expert sourcing, showing that quantitative trends perform better than these methods. However we cannot say quantitative trend analysis is better than qualitative trend analysis or modeling. Qualitative trend analysis also performed better than expert sourcing. However, it did not perform better than other methods. Based on this analysis, expert sourcing did not perform better than either trend method.

4.2 Success Rate Analysis by Domain Area

Domain area is a key attribute associated with forecasts, and we sought to determine whether some domain areas have a higher success rate than others. Table 10 describes the success and failures of all 423 forecasts distributed by domain areas.

Domain Area	Total	Success	Failure	Success Rate
Demographics	135	56	79	0.41
Economy	91	17	74	0.18
Energy	32	10	22	0.31
Environment	24	7	17	0.30
Natural resources	25	10	15	0.40
Technology	128	47	81	0.37

Table 10. Distribution of forecast success and failures by domain

Natural resources and demographics have high success rates compared to all other domains. To determine whether this was a statistically significant observation, we used Fisher's Exact Test to compare the success rate of these two domains to the success rate of all other domains. We used the following hypothesis test for this comparison:

$H_0 : r_{CAR} \leq r_o$ $H_A : r_{CAR} > r_o$	Where: r_{car} is the observed success rate of the natural resources and demographics domains, and r_o is the observed success rate of each/all other domains

Table 11 provides the results of the natural resources tests.

Natural resources compared to…	*p-value*
Demographics	0.58
Economy	0.02
Energy	0.39
Environment	0.60
Technology	0.40

Table 11 Fisher's Exact Test comparing natural resource forecasts to other domain areas

Table 12 provide results of the demographics tests.

Demographics compared to…	*p-value*
Natural resources	0.60
Economy	0.0001
Energy	0.33
Environment	0.59
Technology	0.23

Table 12. Comparison of Demographics forecasts to all other domain areas

There is statistical significance to the observation that forecasts made in both the natural resources and demographics domains have higher success rates than those made in the economy domain. These p-values also suggest that forecasts in the economy domain are less successful than all other domains. To confirm this significance, we compared economy domain forecasts to those in other domain areas. Table 13 provides these results.

Economy compared to…	*p-value*
Demographics	.99
Natural resources	.99
Energy	.94
Environment	.99
Technology	.99

Table 13. Comparison of Economy forecasts to all other domain areas

The results of this test indicate that forecasts made in the economy domain are statistically the least accurate. However, we cannot assert with statistical significance that forecasts about natural resources or demographics have a higher success rate than any of the other domain areas.

4.3 Success Rate Analysis by Timeframe

Time frame is the third key attribute of forecasts. Table 14 provides the success rates associated with the 423 forecasts distributed by time frame.

Time Frame	Total	Success	Failure	Success Rate
10 Years	47	16	31	0.34
11-20 Years	124	42	82	0.33
>20 Years	252	94	158	0.37

Table 14. Distribution of forecast success rates based on time frame

The highest success rates are found in 10-year forecasts and >20-year forecasts. However, the success rate of 11-20 year forecasts is extremely close in proximity. To determine if there was statistical significance among time frame successes, we used Fisher's Exact Test and the following hypothesis test to evaluate the observation:

$H_0 : r_s \le r_o$ $H_A : r_s > r_o$	Where:r_s is the observed success rate of the 10-yr., 11-20 yr, and >20 yr. forecasts, and r_o is the observed success rate of each time frame

Tables 15, 16, and 17 provide the results of these tests. There is no statistical significance to the success rates at different time frames.

10 year compared to	*p-value*
11-20 Years	.56
>20 Years	.72

Table 15. Fisher's Exact Test comparing 10 year forecast success rates to other time frame success rates

11-20 year compared to	*p-value*
10 Years	.58
>20 Years	.78

Table 16. Fisher's Exact Test comparing 11-20 year forecast success rates to other time frame success rates

>20-year compared to	*p-value*
10 Years	.40
11-20 Years	.30

Table 17. Fisher's Exact Test comparing >20 year forecast success rates to other time frame success rates

The analyses above require multiple pair-wise comparisons. In these cases, it may be useful to compensate for the number of tests conducted. The Bonferroni correction is one method to correct for erroneous inferences from a set of statistical observations. In order to ensure that a statistically significant observation has not arisen by chance because of the size of the set of comparisons, this correction is used to deflate the significance level as number of comparisons increases in size. When applied to the p-value outputs of the binomial sensitivity and FET tests, the Bonferroni correction of alpha ranges from .01 (five tests at .05 alpha) to .025 (two tests at .05 alpha). Using a Bonferroni correction on the comparisons above would reduced the limit of statistical sensitivity from .05 to .01 for the binomial sensitivity test (five tests) and from .05 to .025 for the Fisher's Exact tests (two tests). This would not result in any major shifts in an outcome's statistical significance. In these analyses, the p-values are either very significant (far below .05) or very insignificant (far above .05). This test is more effective for comparisons where p-values are closer to (± .01) our value of alpha of .05.

4.4 Test for Temporal Error

The temporal error is the number of years between the actual realization and the predicted realization dates. Forecasts with a small temporal error are more accurate than those with a large temporal error. Forecasting methodologies that produce forecasts with a small average temporal error are more accurate than methodologies associated with large average temporal errors.

Two statistics are important when considering average temporal errors for forecast methods: the absolute temporal error (the TFE) and the STFE. The TFE provides insight about the average magnitude of errors associated with each method. A large average TFE, for example, indicates that a forecast methodology was less accurate. The STFE provides insight about the distribution of optimistic and pessimistic forecasts. A method with a negative average STFE indicates that it tends to overestimate the time frame necessary for the event; in other words, the method tends toward pessimistic forecasts. A positive average STFE indicates a tendency to underestimate timing of events; that is, the method generates optimistic forecasts.

Table 18 provides calculated statistics of the forecasting methods. This analysis is based on the forecasts in the sample set that had realized events, since temporal error can only be calculated if an event occurred.

Forecast Method	# Realized	Mean		Variance		Median		Range STFE	
		TFE	STFE	TFE	STFE	TFE	STFE	Min	Max
Expert sourcing	19	10.05	-6.26	46.05	111.32	10.00	-6.00	-23	10
Gaming and Scenarios	64	6.74	-0.79	43.71	89.26	5.00	0.00	-22	19
Models	63	5.02	-1.37	24.77	48.44	4.00	0.00	-20	14
Qualitative trend analysis	25	7.30	-0.42	48.90	104.22	5.00	0.00	-21	22.5
Quantitative trend analysis	52	5.41	0.61	17.43	46.93	4.50	1.75	-14	16
Multiple	7	2.36	0.93	4.48	9.95	1.50	0.50	-3.5	5.5
Source analysis	6	4.50	-0.50	5.90	29.90	5.00	-2.00	-6	7

Table 18. Description of Temporal Error

Excluding multiple and source analysis due to the relatively low sample sizes, the descriptive statistics in Table 19 show that quantitative trend analysis and models have smaller mean errors, variance, and error range. Previous research by Bryce Space and Technology and The Tauri Group showed that quantitative trend analysis was statistically better than all other forecasting methods. Binomial tests in this analysis comparing quantitative trends shown above did not support this finding since they were only statistically more successful than gaming and scenarios and expert sourcing. To determine whether this is also true based on errors, we compared quantitative trend analysis to all other methods. Lacking a classical test by which to compare these methods, we used an approximation of the Tukey-Kramer Honestly Significant Difference Test to determine if the quantitative trend analysis method is derived from a different distribution than the other methods. Our test was:

$H_0 : \mu_{QT} = \mu_i$ Where μ_{QT} is the mean of the TFE for quantitative trend analysis, and

$H_A : \mu_{QT} \neq \mu_i$ μ_i is the mean of the TFE for forecasting method *i*

We generated a 95% confidence interval for each methodology and then subtracted the confidence interval for the quantitative trends analysis method from each of the other methods. If zero was included in the resulting confidence interval, then we could not reject the null hypothesis. If zero was excluded from the resulting confidence interval, then we could reject the null hypothesis. If the lower bound of the resulting confidence interval was positive, the test implied μ_{QT} was less than μ_i, and if the upper bound of the difference confidence interval was negative, the test implied μ_{QT} was greater than μ_i. Appendix C provides details on the derivation of the confidence intervals and significance levels. The results of the test are shown in Table 19.

Forecast Method	Realized Forecasts	TFE		Difference CI	
		Mean	Variance	Lower	Upper
Expert sourcing	19	10.05	46.05	2.53	6.74
Gaming and Scenarios	64	6.74	43.71	0.84	1.81
Models	63	5.02	24.77	-0.49	-0.31
Qualitative trend analysis	25	7.30	48.90	0.16	3.61
Multiple	7	2.36	4.48	-3.85	-2.27
Source analysis	6	4.50	5.90	-2.3	0.47

Table 19. Results of temporal error test

We reject the null hypothesis for all methods except source analysis. However, source analysis and multiple have sample sizes that are too low to provide statistical accuracy. Because the upper limit for models was negative, there is evidence the distribution of the TFE for the quantitative trend analysis method is larger than that of models. Additionally, because the lower limit for the three remaining methods was positive – expert sourcing, gaming and scenarios, and qualitative trend analysis – there is evidence the distribution of the TFE for the quantitative trend analysis method is smaller that these three methods. This observation is consistent with the descriptive data presented in Table 19, above.

4.5 Identification of Key Attributes

After evaluating temporal error, we conducted a regression analysis to determine if there were combinations of attributes that contributed to forecast accuracy and could be used to produce a predictive model of accuracy. We used forecast attributes—methodology, domain area, time frame, geographic origin, and sub-methodology—as variables for the regression model.

We initially generated one- and two-variable plots of the data to identify potential trends that would facilitate the development of a regression model. However, we could not identify any trends in the data with any combination of attributes; the data generally exhibited a random distribution. Figure 10 presents the results of STFE against time frame (in years).

Key Finding: Over and Underestimating Forecast Dates

Forecasts are equally likely to over- or underestimate the event date. A statistical analysis of the temporal error, defined as the difference between the predicted date and the realized date, shows that long-term forecasts do not tend to error sooner or later than reality.

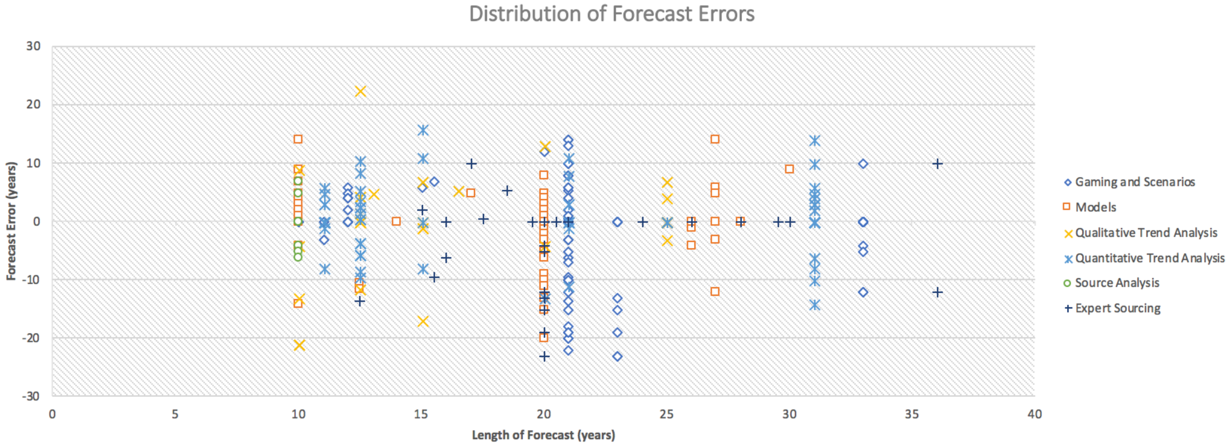


Figure 10. STFE by time frame

As indicated in Figure 10, the STFE is generally symmetric about the abscissa. This implies that forecasts are neither optimistic nor pessimistic. Lastly, the data are random in their placement, do not indicate a positive or negative trend, and do not indicate any type of underlying non-linear trend. There is a slight positive slope of the trendline as a result of the twelve outliers with STFE values in excess of 23 years. We eliminated these twelve records from the regression analysis, as their influence in the model was disproportional to their weight. Analysis of the TFE was consistent with the observations of the STFE.

We explored all combinations up to four variables. To filter out the ill-fitting models, we established the criteria that a model had to produce a *p*–value of .05 or smaller and an adjusted R^2 value of .8 or better. No model produced any results that satisfied the criteria. The details of the regression analysis are provided in Appendix C.

4.6 Summary of Findings

In terms of the accuracy of long-term forecasts:

1. All forecast methodologies provide more accurate predictions than uninformed guesses

2. Forecasts based on quantitative and qualitative trend analyses are more accurate than forecasts based on expert opinion

3. Forecasts of the economy are the least accurate in comparison to all other domains

4. All forecasts time frames (10 years, 11-20 years, and >20 years or more) are more accurate than an uninformed guess, but no time frame was more accurate than another

5. As a whole, long-term forecasts are equally likely to over or underestimate the event date

6. A predictive model of forecast accuracy could not be developed

Appendix A. Standard Language Lexicon

A.1 Standard Language

Time frame:

- "A" (as in "a year") = 1
- "About" or "approximately" or "around" or "almost" = the number specified ("about 15 years" = 15 years)
- "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 time frame

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

Growth Rate:

- "Annual rate(s)" = treated as compound annual growth rate
- "Rate of growth per year" = final year in range treated as evaluation year
- "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

A.2 Analysis Rules

Exclusion Rules: These rules determine what we will not record in the database.

- Time frame: 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 than 30% of the forecasted time frame. 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 (time frame, 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.

- Attempt to select forecasts that are diverse in respect to subject area or time frame from each source.
- 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 frame is recorded as a failed forecast. For example, 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 time frame.
    - A probabilistic forecast with more than or equal to a 60% chance of occurring will 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.

- Steady-state/as-is forecasts:
  - If a steady-state forecast does not occur on the forecasted year, then we will not attempt to find year of realization because the forecast is predicting a condition or event will still be occurring on the predicted year. For instance, forecasts that say, “X will continue to be X in 2000.”

- Forecasts predicting year-by-year growth within a range:
  - The year of realization for forecasts predicting growth each year within a range of time will be the last year in that range, because said growth must occur each year in the range in order for the forecast to be successful. For instance, forecasts that say, “X will have a growth rate of X% per year from 1980-2000.”

- Close but Conflicting Ground Truth Sources
  - If two ground truth sources are close in narrowing down a year of realization, then the average of those two sources is used as the year of realization. For instance, if a one source says that Gross Domestic Product for a particular region reached 2% in 2001 but another source gives the year as 2003, then the average year (2002) will be used as the year of realization.

Appendix B. Forecast Sources

Author	Organization	Title	Year	Publication Type
Barney	Council on Environment Quality	The Global 2000 Report to the President	1980	Government reports & roadmaps
Battelle Memorial Institute	Battelle Memorial Institute	Agriculture 2000	1983	Industry organizations, associations, & societies
Brundtland Commission	United Nations (UN)	Report of the World Commission on Environment and Development: Our Common Future	1987	Government reports & roadmaps
Food and Agriculture Organization (FAO)	UN	Agriculture: Toward 2000	1980	Government reports & roadmaps
Gordon & Helmer	RAND	Report on a Long-Range Forecasting Study	1964	Industry organizations, associations, & societies
Intergovernmental Panel on climate Change (IPCC)	IPCC	First Assessment Report	1990	Government reports & roadmaps
Kahn & Weiner	Hudson Institute	The Year 2000	1967	Strategic analysis firms
Leontief et al.	United Nations	The Future of the World Economy	1978	Government reports & roadmaps
Lesourne et al.	Organisation for Economic Cooperation and Development (OECD)	Facing the Future	1979	Government reports & roadmaps
Meadows et al.	Club of Rome	Limits to Growth	1972	Industry organizations, associations, & societies
Mesarovic & Pestel	Club of Rome	Mankind at the Turning Point	1974	Industry organizations, associations, & societies
National Aeronautics and Space Administration (NASA)	NASA	Outlook for Space	1972	Government reports & roadmaps
NASA	NASA	A Forecast of Space Technology 1980-2000	1976	Government reports & roadmaps
Tenet	National Intelligence Council	Global Trends 2015: A Dialogue About the Future with Nongovernment Experts	2000	Government reports & roadmaps
UN	UN	Growth of the World's Urban and Rural Population, 1920-2000	1969	Government reports & roadmaps
UN	UN	Overall Socio-Economic Perspective of the World Economy to the Year 2000	1990	Government reports & roadmaps
Wilson	Workshop on Alternative Energy Strategies	Energy: Global Prospects 1985-2000	1977	Industry organizations, associations, & societies

Appendix C. Statistical Analysis

All statistical analysis was conducted using R version 3.5.0, published on April, 23rd 2018. 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. There are five data files that support this appendix:

RATF_All_Data.csv: This file contains used by R to tabulate the full set of records in the database. It serves as the base data set for all other data files.

Load For R.csv: This file contains the data read in by R used for the analysis that compared success rates among forecasts, TFE rates, as well as regression analyses.

RATF_Results.xlsx: This Excel® file contains all data tables presented in the appendix.

C.1.1 Outliers in the Data Set

There were 436 records available for analysis in the data set. The following twelve records were extreme outliers with respect to forecast error and were removed from the data set for all statistical analysis conducted in support of this study:

Record Number	Technology Area	Methodology	Year of Forecast	Predicted Year	Year Occurred	TFE	Success
2679	Air Transportation Technology	Expert Sourcing	1964	1984	1942	42	FALSE
2506	Population	Gaming and Scenarios	1967	2000	1965	35	FALSE
2010	Autonomous/Robotics Technology	Qualitative Trend Analysis	1975	1995	1961	34	FALSE
2502	Labor Force	Gaming and Scenarios	1967	2000	1968	32	FALSE
2501	Labor Force	Gaming and Scenarios	1967	2000	1969	31	FALSE
2811	Arable land	Models	1980	2000	1969	31	FALSE
2449	Temperature	Expert Sourcing	1964	2000	1971	29	FALSE
3031	Sensor Technology	Quantitative Trend Analysis	1975	1987.5	2016	28.5	FALSE
3008	Population	Quantitative Trend Analysis	1969	2000	1972	28	FALSE
2652	Energy and Power Technology	Gaming and Scenarios	1979	2000	1973	27	FALSE
2756	Gender	Gaming and Scenarios	1978	1990	2016	26	FALSE
2462	Communications Technology	Expert Sourcing	1964	1984	2009	25	FALSE

Table C-1. Forecast outliers excluded from analysis

C.1.2 Deriving ρ, the Probability for Success of an Uninformed Guess

For this study, we developed a test that compared each of the eight forecast methodologies against a control. The control for this test was the probability of success associated with an uninformed guess. Because we classified a forecast as either a success or a failure, the binomial distribution was an appropriate distribution for this test and we used the probability of success of an uninformed guess as the hypothesized success rate. The statistical test then was:

$$H_0 : r_i \geq \rho$$
$$H_A : r_i < \rho$$

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

Conducting the test required that we develop ρ. We initially considered using the 95th percentile of the data set for both the length of forecasts and the temporal error. Upon inspection, however, we discovered the 95th percentile in forecast length (forecasts in excess of 35 years) would have unfairly biased the analysis in favor of expert sources by removing over 24% of its forecasts (all of which were failed forecasts). We therefore used the 99th percentile of data for timeframe and the 95th percentile of data for temporal forecast error. This selection of percentiles resulted in a forecast range of one to 36 years and a temporal error between zero and 23 years.

No records were removed for excessive forecast lengths.

The twelve outlier forecasts were removed from the data set due to excessive temporal forecast error. All of these were failed forecasts.

The remaining data set consisted of 440 records for this particular phase of statistical analysis.

We provide the following symbolic definition for this section of analysis:

$\langle a \rangle$ represents the function to round the term *a* to the nearest integer. We use the convention that $\langle a \rangle = \lfloor a + 0.5 \rfloor$.

We developed ρ assuming *x*% allowable error in a forecast. This means a forecasted event can occur within *x*% of the predicted time and still be considered a success. The first step was to calculate the allowable range factor using the following equation:

$$\alpha = 1 + 2\langle t \times x \rangle$$

Where α = allowable range measured in years
t = number of years between the prediction and the predicted year of occurrence
1 accounts for the year the event was predicted to occur
2 accounts for the allowable time on either side of the forecast.

For example, a 10-year forecast with a 30% allowable error would be calculated as:

$$\alpha = 1 + 2\langle 10 \times .30 \rangle$$
$$\alpha = 1 + 2\langle 3 \rangle$$
$$\alpha = 7$$

When values of *x* get sufficiently large, it is possible that the allowable range could exceed the 59-year period of the hypothetical guess (36-year forecast + 23-year TFE). Therefore, we bound the upper half of the allowable range by 23 years.

A calculation with the upper bound in place for a 36-year forecast with a 60% allowable is:

$a = 1 + (t \times x) + min([t \times x], 23)$

$a = 1 + (36 \times .6) + min([36 \times .6], 23)$

$a = 1 + (21.6) + min(21.6, 23)$

$a = 1 + 21.6 + 23$

$a = 45.6$

The following Table C-2 provides α for all combinations of years (t) and allowable error (x) in increments of 10%. These values served as the basis of our sensitivity analysis of the allowable range.

Forecast Length (years)	Allowable Error									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
1	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3
2	1.4	1.8	2.2	2.6	3	3.4	3.8	4.2	4.6	5
3	1.6	2.2	2.8	3.4	4	4.6	5.2	5.8	6.4	7
4	1.8	2.6	3.4	4.2	5	5.8	6.6	7.4	8.2	9
5	2	3	4	5	6	7	8	9	10	11
6	2.2	3.4	4.6	5.8	7	8.2	9.4	10.6	11.8	13
7	2.4	3.8	5.2	6.6	8	9.4	10.8	12.2	13.6	15
8	2.6	4.2	5.8	7.4	9	10.6	12.2	13.8	15.4	17
9	2.8	4.6	6.4	8.2	10	11.8	13.6	15.4	17.2	19
10	3	5	7	9	11	13	15	17	19	21
11	3.2	5.4	7.6	9.8	12	14.2	16.4	18.6	20.8	23
12	3.4	5.8	8.2	10.6	13	15.4	17.8	20.2	22.6	25
13	3.6	6.2	8.8	11.4	14	16.6	19.2	21.8	24.4	27
14	3.8	6.6	9.4	12.2	15	17.8	20.6	23.4	26.2	29
15	4	7	10	13	16	19	22	25	28	31
16	4.2	7.4	10.6	13.8	17	20.2	23.4	26.6	29.8	33
17	4.4	7.8	11.2	14.6	18	21.4	24.8	28.2	31.6	35
18	4.6	8.2	11.8	15.4	19	22.6	26.2	29.8	33.4	37
19	4.8	8.6	12.4	16.2	20	23.8	27.6	31.4	35.2	39
20	5	9	13	17	21	25	29	33	37	41
21	5.2	9.4	13.6	17.8	22	26.2	30.4	34.6	38.8	43
22	5.4	9.8	14.2	18.6	23	27.4	31.8	36.2	40.6	45
23	5.6	10.2	14.8	19.4	24	28.6	33.2	37.8	42.4	47
24	5.8	10.6	15.4	20.2	25	29.8	34.6	39.4	44.2	48
25	6	11	16	21	26	31	36	41	46	49
26	6.2	11.4	16.6	21.8	27	32.2	37.4	42.6	47.4	50
27	6.4	11.8	17.2	22.6	28	33.4	38.8	44.2	48.3	51

28	6.6	12.2	17.8	23.4	29	34.6	40.2	45.8	49.2	52
29	6.8	12.6	18.4	24.2	30	35.8	41.6	47.2	50.1	53
30	7	13	19	25	31	37	43	48	51	54
31	7.2	13.4	19.6	25.8	32	38.2	44.4	48.8	51.9	55
32	7.4	13.8	20.2	26.6	33	39.4	45.8	49.6	52.8	56
33	7.6	14.2	20.8	27.4	34	40.6	47.1	50.4	53.7	57
34	7.8	14.6	21.4	28.2	35	41.8	47.8	51.2	54.6	58
35	8	15	22	29	36	43	48.5	52	55.5	59
36	8.2	15.4	22.6	29.8	37	44.2	49.2	52.8	56.4	60

Table C-2. Width of allowable error in years

Given the 59-year boundary of an event occurring and a uniform chance of the event occurring any time within the 59 years, there was a 1/59 = .016 probability of an event occurring in any given year. We used the following formula to determine the probability that a forecast would be a success given a specified allowable error:

$\rho = a / 59$

For example, the probability of success for a 15-year forecast with an allowable range of 40% would be:

$\rho = a / 59$

$\rho = 13 / 59 = .22$

Table C-3 provides *p* for all combinations of years (*t*) and allowable error (*x*) in increments of 10%.

Forecast Length (years)					**Allowable Error**					
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
1	0.020	0.024	0.027	0.031	0.034	0.037	0.041	0.044	0.047	0.051
2	0.024	0.031	0.037	0.044	0.051	0.058	0.064	0.071	0.078	0.085
3	0.027	0.037	0.047	0.058	0.068	0.078	0.088	0.098	0.108	0.119
4	0.031	0.044	0.058	0.071	0.085	0.098	0.112	0.125	0.139	0.153
5	0.034	0.051	0.068	0.085	0.102	0.119	0.136	0.153	0.169	0.186
6	0.037	0.058	0.078	0.098	0.119	0.139	0.159	0.180	0.200	0.220
7	0.041	0.064	0.088	0.112	0.136	0.159	0.183	0.207	0.231	0.254
8	0.044	0.071	0.098	0.125	0.153	0.180	0.207	0.234	0.261	0.288
9	0.047	0.078	0.108	0.139	0.169	0.200	0.231	0.261	0.292	0.322
10	0.051	0.085	0.119	0.153	0.186	0.220	0.254	0.288	0.322	0.356
11	0.054	0.092	0.129	0.166	0.203	0.241	0.278	0.315	0.353	0.390
12	0.058	0.098	0.139	0.180	0.220	0.261	0.302	0.342	0.383	0.424
13	0.061	0.105	0.149	0.193	0.237	0.281	0.325	0.369	0.414	0.458
14	0.064	0.112	0.159	0.207	0.254	0.302	0.349	0.397	0.444	0.492

15	0.068	0.119	0.169	0.220	0.271	0.322	0.373	0.424	0.475	0.525
16	0.071	0.125	0.180	0.234	0.288	0.342	0.397	0.451	0.505	0.559
17	0.075	0.132	0.190	0.247	0.305	0.363	0.420	0.478	0.536	0.593
18	0.078	0.139	0.200	0.261	0.322	0.383	0.444	0.505	0.566	0.627
19	0.081	0.146	0.210	0.275	0.339	0.403	0.468	0.532	0.597	0.661
20	0.085	0.153	0.220	0.288	0.356	0.424	0.492	0.559	0.627	0.695
21	0.088	0.159	0.231	0.302	0.373	0.444	0.515	0.586	0.658	0.729
22	0.092	0.166	0.241	0.315	0.390	0.464	0.539	0.614	0.688	0.763
23	0.095	0.173	0.251	0.329	0.407	0.485	0.563	0.641	0.719	0.797
24	0.098	0.180	0.261	0.342	0.424	0.505	0.586	0.668	0.749	0.814
25	0.102	0.186	0.271	0.356	0.441	0.525	0.610	0.695	0.780	0.831
26	0.105	0.193	0.281	0.369	0.458	0.546	0.634	0.722	0.803	0.847
27	0.108	0.200	0.292	0.383	0.475	0.566	0.658	0.749	0.819	0.864
28	0.112	0.207	0.302	0.397	0.492	0.586	0.681	0.776	0.834	0.881
29	0.115	0.214	0.312	0.410	0.508	0.607	0.705	0.800	0.849	0.898
30	0.119	0.220	0.322	0.424	0.525	0.627	0.729	0.814	0.864	0.915
31	0.122	0.227	0.332	0.437	0.542	0.647	0.753	0.827	0.880	0.932
32	0.125	0.234	0.342	0.451	0.559	0.668	0.776	0.841	0.895	0.949
33	0.129	0.241	0.353	0.464	0.576	0.688	0.798	0.854	0.910	0.966
34	0.132	0.247	0.363	0.478	0.593	0.708	0.810	0.868	0.925	0.983
35	0.136	0.254	0.373	0.492	0.610	0.729	0.822	0.881	0.941	1.000
36	0.139	0.261	0.383	0.505	0.627	0.749	0.834	0.895	0.956	1.017

Table C-3. Probability of an event occurring during the span of the allowable range

The final step is to calculate ρ for each allowable error. We used the following formula:

$$p_x = .027 \sum_{i=1}^{36} p_{ix}$$

Where: p_{ix} is the probability of a forecast being successful in year *i* given an allowable range of *x*%, and
1/36 = 0.027 is the probability of year *i* being randomly selected as the forecasted year of the event occurring

Table C-4 provides ρ_x, the probability of success for an uninformed guess given an allowable range of *x*%.

Allowable Range	Sensitivity Analysis of Allowable Range										
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
ρ	1.6%	8%	14%	21%	27%	33%	39%	45%	51%	56%	60%

Table C-4. Probability of success for an uniformed guess given various allowable range parameters

C.1. 3 Negative Forecasts

Of the 441 forecasts in the dataset, none of them were negative forecasts. Negative forecasts predict that an event will not occur prior to a certain time frame.

C.1. 4 Test Against a Control with a 30% Allowable Range

The first test we conducted was used to determine whether the eight forecasting methods considered in our study were better than a random guess. We used the probability of success for a random guess described in the previous section as the control probability of success for this test. We conducted a series of binomial tests comparing the number of observed successes and trials against the control probability of success. Our hypothesis test was

$$H_0 : \mu_f \geq \rho_{30\%}$$
$$H_A : \mu_f < \rho_{30\%}$$

Where: μ_f is the observed success rate for forecast methodology *f*, and $\rho_{30\%}$ is the hypothesized probability of success given an allowable range of *30%*

Table C-5 provides the results of the binomial test with $\rho_{30\%}$=21%.

x = 30%; p30% = 21%					
Method	Total	Successes	Failures	Success Rate	Alpha
Expert Sourcing	61	9	52	0.15	0.04
Gaming and Scenarios	133	37	96	0.28	0.83
Models	109	40	69	0.37	1.00
Multiple	7	7	0	1.00	1.00
Qualitative Trend Analysis	37	15	22	0.41	0.99
Quantitative Trend Analysis	80	38	42	0.48	1.00
Source Analysis	8	1	7	0.13	0.37
Total Observations	435				

Table C-5. Results of binomial test given a 30% allowable range

Where:

- "Method" describes the forecasting methodology being tested,
- "Total" is the total number of observed forecasts using the specified methodology,
- "Successes" is the number of successes for the observed methodology,
- "Failures" is the number of failures for the observed methodology,
- "Success Rate" is the value of (Successes/Total),
- "Alpha" is the resulting *p*-value from the hypothesis test. High values of alpha support the hypothesis that the specified methodology is at least as good as a random guess and that low values of alpha support the alternative hypothesis,
- "Total Observations" is the total number of forecasts in the sample population.

C.1. 5 Results of Fisher's Exact Test in Comparing Trend Methods to Other Methods

Both qualitative and quantitative trend analysis demonstrated substantially higher success rates than did the other forecasting methodologies (we excluded "multiple" from previous analyses due to its the small sample size of six forecasts). We used Fisher's Exact Test to determine whether we could draw any statistically significant conclusions concerning these two methods relative to the three forecasting methodologies with large sample sizes (expert analysis, expert sourcing, and source analysis). The following hypothesis test was used:

$$H_0 : \mu_t \le \mu_s$$
$$H_A : \mu_t > \mu_s$$

Where: μ_t is the observed success rate of a trend method, and
μ_s is the observed success rate of the other methods

We present each test in its contingency table set-up. For comparison of the qualitative trend assessment methodology:

	Success	Failure	Total
Qualitative	15	21	36
Gaming and Scenarios	40	88	128
Total	55	109	164

Alternative = Greater Than
p -value = .1658

	Success	Failure	Total
Qualitative	15	21	36
Models	41	66	107
Total	56	87	143

Alternative = Greater Than
p -value = .4342

	Success	Failure	Total
Qualitative	15	21	36
Quantitative	38	40	78
Total	53	61	114

Alternative = Greater Than
p -value = .8167

	Success	Failure	Total
Qualitative	15	21	36
Expert sourcing	10	48	58
Total	25	69	94

Alternative = Greater Than
p -value = .0094

Figure C-6. Results of Fishers Exact Test for qualitative trend analysis methods

For comparison of the quantitative trend methodology:

	Success	Failure	Total
Quantitative	38	40	78
Gaming and scenarios	40	88	128
Total	78	128	206

Alternative = Greater Than
p -value = .0093

	Success	Failure	Total
Quantitative	38	40	78
Models	41	66	107
Total	79	106	185

Alternative = Greater Than

p -value = .1035

	Success	Failure	Total
Quantitative	38	40	78
Qualitative	15	21	36
Total	53	61	114

Alternative = Greater Than

p -value = .3093

	Success	Failure	Total
Quantitative	38	40	78
Expert Sourcing	10	48	58
Total	48	88	136

Alternative = Greater Than

p -value = .0001

Figure C-7. Results of Fishers Exact Test for quantitative trend analysis methods

These results indicate that there is insufficient evidence to suggest the qualitative trend analysis method is better than the two other methods investigated. There is, however, sufficient evidence to suggest the quantitative method is better than the two methods most commonly used.

C.1.6 Description of Data Set for Temporal Analysis

We calculate our temporal error statistics using the following formulas:

$$STFE = date_O - date_F$$

$$TFE = \left| date_O - date_F \right|$$

Where: *STFE* is the signed temporal error,
TFE is the unsigned temporal error,
date$_F$ is the date of the forecasted event, and
date$_O$ is the date the event actually occurred

A forecast which predicts an event will occur before it actually does is considered an optimistic forecast (STFE is positive) while a forecast which predicts and event will occur later than it actually does is considered a pessimistic forecast (STFE is negative). When taken across a specific treatment for an attribute, the average STFE corresponds to the mean error (ME) and the average TFE corresponds to the mean absolute error (MAE).

Only those forecasts whose predicted events actually occurred can have a temporal error. Those forecasts whose events have not yet occurred cannot have a temporal error and are therefore excluded from the temporal analysis phase of the study.

C.1. 7 Results of Tukey-Kramer HSD (TKHSD) Multiple Pair-wise Comparisons

An omnibus test of the means, such as the ANOVA, does not indicate which treatments within the attributes have different means if differences exist. It only tells you if differences do exist. It is possible, however, to conduct a pair-wise comparison between each of the treatments within an attribute to determine specific differences. Had the data satisfied the normality and equal variance criteria, the TKHSD test could have been applied to determine which treatments were different and how they are different. Our revised hypothesis test for this analysis is:

$$H_0 : \mu_i > \mu_j \quad \forall i \neq j$$

$$H_A : \mu_i \leq \mu_j \quad \forall i \neq j$$

Where: μ_i and μ_j are the observed average temporal errors for forecast attribute values *i* and *j*, and *i* and *j* = 1 to *n*

In this analysis, we conduct $_nC_k$ combinations of pair-wise comparisons for the statistic of interest. We subtract the means from each other and compare the results to the test statistic $q_{\alpha,n,v}$, which is the Studentized range at α-level of significance for *n* degrees of freedom for the first sample and *v* degrees of freedom for the second sample. TKHSD accommodates for unequal sample sizes and returns an adjusted *p*-value that accounts for the total number of comparisons to be made. R returns results from the TKHSD by providing the difference between the two means, a confidence interval for the difference of the two means, and an adjusted *p*-value for each comparison. As a visual check, if the confidence interval includes 0 within its boundaries, then there is not statistical evidence that the two means are different. If the difference between the two terms is negative, that indicates the first sample may have a smaller mean than does the second sample.

Treatment	Difference	Lower Bound	Upper Bound	Adjusted p-value
Multiple-Expert Sourcing	-7.5840336	-16.466012	1.2979453	0.0484274
Qualitative Trend Analysis-Expert Sourcing	-3.0939542	-8.91418	2.7262714	0.5098851
Quantitative Trend Analysis-Expert Sourcing	-4.5277149	-10.053271	0.9978414	0.0673246
Source Analysis-Expert Sourcing	-5.4411765	-14.832816	3.9504633	0.4012373
Models-Gaming and Scenarios	-1.954714	-5.52427	1.6148417	0.4726355
Multiple-Gaming and Scenarios	-4.3631961	-12.269519	3.5431266	0.4628555
Qualitative Trend Analysis-Gaming and Scenarios	0.1268832	-4.055869	4.3096355	0.9999999
Quantitative Trend Analysis-Gaming and Scenarios	-1.3068774	-5.068813	2.4550586	0.8879642
Source Analysis-Gaming and Scenarios	-2.220339	-10.695201	6.2545232	0.9697379
Multiple-Models	-2.4084821	-10.281981	5.4650166	0.9363639
Qualitative Trend Analysis-Models	2.0815972	-2.038774	6.2019686	0.5720572
Quantitative Trend Analysis-Models	0.6478365	-3.044616	4.3402892	0.996347
Source Analysis-Models	-0.265625	-8.709873	8.1786235	0.9999998
Qualitative Trend Analysis-Multiple	4.4900794	-3.679723	12.6598818	0.4680599
Quantitative Trend Analysis-Multiple	3.0563187	-4.906245	11.0188822	0.8313115

Source Analysis-Multiple	2.1428571	-8.860485	13.1461995	0.9935385
Quantitative Trend Analysis-Qualitative Trend	-1.4337607	-5.721872	2.8543502	0.9050323
Source Analysis-Qualitative Trend Analysis	-2.3472222	-11.068405	6.3739601	0.9654376
Source Analysis-Quantitative Trend Analysis	-0.9134615	-9.440816	7.6138925	0.9997796

Table C-8. Results of TKHSD test based on forecast methodology

Treatment	Difference	Lower Bound	Upper Bound	Adjusted p-value
Medium-term-Long-term	-1.506564	-4.391306	1.378178	0.2758321
Short-term-Long-term	-3.256564	-7.93479	1.421663	0.1032275
Short-term-Medium-term	-1.75	-6.958467	3.458467	0.584926

Table C-9. Results of TKHSD based on timeframe

We conducted the TKHSD test on the TFE for the forecasts in spite of the violations of the equal variance and the normalized data requirements. While we ran the test for multiple confidence intervals, we provide the results (in both tabular and graphical formats) for the 99% confidence interval only.

We note that all comparisons are significant and therefore we could have concluded the errors associated with the timeframes come from different distributions had we satisfied the criteria (normality and equal variance). Because we fail to satisfy these criteria however, we draw no conclusions with respect to this test.

C.1. 7 Results of Pair-wise Comparison of Quantitative Trend Analysis

Given that our data does not conform to requirements to use standard parametric and non-parametric tests for differences in means (neither normal nor *i.i.d.*), we developed a test to determine if there was a statistically significant difference between a subset of means in our sample data set. For this test, we compare the quantitative trend analysis methods against all other methods. Our hypothesis test is:

$$H_0 : \mu_Q \geq \mu_E$$

$$H_A : \mu_Q < \mu_E$$

Where: μ_Q is the mean of the temporal forecast error for quantitative trend analysis,

μ_E is the mean of the temporal forecast error for all other methods

Our test consists of finding the 95% confidence interval for each of the methods in question, subtracting the quantitative CI from the opinion CI, and then observing if the value zero is included in the resulting interval. If it is, then there is insufficient evidence to reject the null hypothesis. Because these data comparisons are not paired, are not normal, have unequal sample sizes, and are not *i.i.d*; standard methods for developing confidence intervals do not apply – we cannot assume 3 standard deviations include 95% of the data (normality) nor is there an appropriate T-statistic that accounts for 95% of the data. Therefore, we developed a method for creating a confidence interval from which to draw conclusions for our hypothesis test.

We use the standard confidence interval formula as the basis of our test. The standard normal confidence interval is derived from the following formula where $Z_a$is the Z-value returned from the standard normal tables.

$$CI_\alpha = \left(\bar{x} - Z_\alpha \frac{s}{\sqrt{n}} \;\; , \quad \bar{x} + Z_\alpha \frac{s}{\sqrt{n}} \right)$$

We derived the CI from the 95th percentile of the data. We modify the standard CI equation by replacing the Z_a*s term above with the value of the 95th percentile for the given method under investigation.

Results of our test indicate that there is strong statistical evidence that quantitative trend analysis method is larger than that of models. Additionally, because the lower limit for the three remaining methods was positive – Expert Sourcing, Gaming and Scenarios, and Qualitative Trend Analysis – there is evidence the distribution of the TFE for the quantitative trend analysis method is smaller that these three methods. Each of the three tests was conducted at a level of significance of .05.

Forecast Methodology	CI LB	CI UB	Difference LB	Difference UB
Expert sourcing	6.78	13.32	2.53	6.74
Gaming and Scenarios	5.09	8.39	0.84	1.81
Models	3.76	6.27	-0.49	-0.31
Multiple	0.4	4.31	-3.85	-2.27
Qualitative trend analysis	4.41	10.19	0.16	3.61
Source analysis	1.95	7.05	-2.3	0.47
Quantitative trend analysis	4.25	6.58		

Table C-10. Results of Confidence Interval Test Base on 95th percentile of data

C.1. 8 Results of Regression Analysis

We evaluated several regression models to determine if we could develop a predictive model of forecast accuracy. All assessed models were linear since we could not discern an underlying non-linear distribution of the data. Figure D-12 provides a distribution of the data with respect to forecasting method and comparing length of forecast to error.

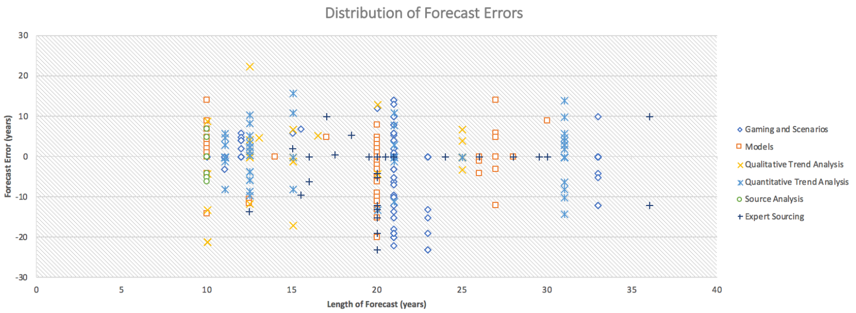


Figure C-11. STFE by time frame

R has a feature that allows you to specify your data set, develop a regression model and then build the best regression model given the data available. It iterates through various combinations of variables seeking the best regression model for the data available. There are two procedures for finding the best model:

- Forward step: it builds the model from the simplest to most complex and stops when it no longer achieves improvement, and
- Reverse step: it builds the model from the most complex to the smallest and stops when it no longer achieves improvement by removing variables.

The literature for R recommend conducting both the forward and reverse paths to ensure both methods converge on a single model.

Before conducting the regression tests we observed that our attributes consisted of two types of data: categorical and numerical. In noisy data sets, regression models may benefit by transforming numerical data. Therefore, we conducted a series of tests that that did forward and reverse passes on our data applying various combinations of transformations on the numerical attribute (time frame years) and on the response variable (STFE).

We explored all combinations up to four variables. To filter out the ill-fitting models, we established the criteria that a model had to produce a *p*–value of .05 or smaller and an adjusted R^2 value of .8 or better. No model produced any results that satisfied the criteria.