International Business Research; Vol. 15, No. 5; 2022
ISSN 1913-9004 E-ISSN 1913-9012
Published by Canadian Center of Science and Education

A Conditioned Forecasting Model: *A-priori* Screening Validation Testing

Frank Heilig[1] & Edward J. Lusk[2]

[1] Strategic Risk-Management, Volkswagen Leasing GmbH, Braunschweig, Germany

[2] Emeritus: The Wharton School, [Dept. Statistics], The University of Pennsylvania, USA & School of Business and Economics [Dep, Accounting], State University of New York: Plattsburgh, USA

Correspondence: Edward Lusk, The Wharton School [Statistics], University of Pennsylvania, Philadelphia, PA. 19104-1686 USA.

Received: March 14, 2022 Accepted: April 14, 2022 Online Published: April 18, 2022

doi:10.5539/ibr.v15n5p63 URL: https://doi.org/10.5539/ibr.v15n5p63

Abstract

Context The forecasting literature over the last three decades documents that judgmental conditions on the performance of the forecasting model *are often used to rationalize the acceptance of the intel from a forecasting model* that will be used in creating an action-plan. However, rarely are these judgmental-conditioning protocols recorded as they should be to intelligently process the interactions of the conditioning protocols with possible adjustments made in the forecasts.

Focus In this research report, we will offer, four judgmental conditioning aspects that are not infrequently used by managers of forecasting divisions. Specifically, the acceptance contingencies of a forecasting model under evaluation scrutiny are: (i) The desired magnitude of the Median benchmarked Precision is in evidence, (ii) The Holdback is in the (1-FPE) Confidence Interval, (iii) The Pearson Product Moment Correlation-Null of the residuals is not rejected, and (iv) The Autocorrelation-Null of the residuals is not rejected. Each of these four conditioning aspects will be evaluated for two standard models typically in the panoply of forecasters: The Two-Parameter [Intercept & Slope] Linear OLS-Regression & the ARIMA(0, 2, 2)/Holt models. The measure of interest for ALL of the selected inferential analyses is: ***How often do selections among these conditioning aspects result in the forecasting model being rejected as informing the decision-making process? Results*** Surprisingly, the range of ***Failures*** for the conditions tested ranged *grosso modo* in the interval:{40% to 80%}depending on the nature of the Conditions. These implications are discussed.

Keywords: OLS-regression & Holt forecasting models

1. Introduction

1.1 Overview

Forecasting the likelihood of the events in the near future is a survival-skill for every organism and collection of organisms. Included therein, of course, are firms in the economic market-place that are traded in active global-exchanges such as: the NYSE, The Shanghai Stock Exchange, and The NASDAQ. In this regard, the forecasting-model's projections compared to the nature of the residual ***profile*** after fitting the model using prior data-Panels could be used to create judgmental adjustments to these forecasting projections to better align the forecasts to the needs of the decision under consideration. Thus, *a-priori* it behooves managers of forecasting divisions or forecasting out-source-links to offer likely conditions on the forecasting-model(s) that they find useful in making the decision to judgmentally adjust the forecasts of their Forecasting Models [FM]. This would fit well in the *Learning-Loop*© of the Balanced Scorecard [BSC™]See Kaplan & Norton (1992). Simply, if there are validity-screens that are recorded, tracked, and evaluated over time, as is usual in the Balanced Scorecard, then this likely will create an effective Learning Loop, the profile of which, often will suggest the more effective forecasting judgmental adjustments.

Attentive experiential consultation observations over decades and the literature on judgmental adjustments over the last five decades, suggest that individual forecasters use standard models to create short- or intermediate-run likelihood projections and then adjust these model-generated projections based upon selected "elements of the actual forecasting profile". Consistent with the literature, in particular the research reports of: Clancey (1983);

Sanders & Ritzman (1992); O'Connor, Remus & Griggs (1993); Adya (2000), Mohammad, Anvari & Saberi (2013) and Adya & Lusk (2016) and in addition to the conversational feedback of our long-time colleagues, who has been forecasters for decades, noted [*liberally paraphrasing*]:

> *'We use simple forecasting models; usually, three are used for every forecasting project: Regression, Moving Average or Exponential Smoothers. How they "work" is easy to understand. So, I can look at their forecasts and their operational-profile and see if the models sort of conform to the assumptions of the models. If these models drift-off here and there I can figure out how to adjust the forecasts to "whip-the-projections" back in line—sort of. I never fuss around to find a better model as: (i) all of the upstream decision-makers need [really would like to have] the forecasts "yesterday", (ii) fussing about the nature of the development data always leads to collecting more data—time that we never have, worse yet, (iii) more data or longer Panels mathematically leads to more precision—i.e., more narrow or exclusive Confidence Intervals that will often error on the side of rejecting otherwise acceptable decisions vis-à-vis "normal or usual" Confidence Intervals.* ***Short-Story****: If I continue to play around with verifying all the assumptions underlying the models, the Senior Managers will be tempted to outsource my forecasting group!!! Also, really, if there was a forecasting model that worked all the time—do they need ME and MY group. So, they expect me to get it right—how I do that is my Midas-touch or Retention-hook so I can pay my mortgage and send my kids to college.'*

1.2 Research Focus

The specific context for our examination is based upon the following insights collect over the years from both the Academic- and the Practice-*Milieux*:

Forecasting Experimental Conditions: ***Failure Screens***

We will offer four FM-conditions that, in practice, seem to be used to suggest to the forecaster that there may be a need to create judgmental-adjustments to the forecasts as generated. The judgmental adjustments will usually be calibrated given the decision-conditions under examination. In this case, we will refer to these as ***Failure Screens***—

> *this just indicates that there is likely a "violation" of the imposed judgmental FM-condition and this usually suggests that the FM did not perform as expected or "desired" and so may rationalize a modification in the forecasts as generated.*

The following conditions will constitute ***Failure Screens*** of the forecasting model that would possibly rationalize judgmental-modifications to better practically align the forecasts to the decision-maker's reality:

1. If the Median Benchmarked Precision of the 95% Confidence Interval is **>=25%**, or
2. If the Holdback is **NOT in** the 95% Confidence Interval, or
3. If the Pearson Product Moment [PPM] Correlation of the Residuals with the Time Index is such that the **Null IS rejected**, or
4. If the Autocorrelation of the Residuals is such that the **Null IS rejected**.

These four ***Failure Screens*** are generalizations; the computational parameters needed to make the calculations will be dealt with *anon*.

1.3 Research Overview

Using the Two-Parameter [Intercept & Slope] Linear OLS-Regression [OLSR] and the ARIMA(0, 2, 2)/Holt models, we will examine the effect of the above four conditioning ***Failure Screens*** on the ***Failure-rate*** of the accrued Panels. In this regard, we will:

1. Provide an illustration of certain aspects of the residuals-test for structure using a Bloomberg-Panel to elucidate the operational issue in a forecasting context,
2. Review the literature on judgmental forecasting that forms the basis of judgmental modification of forecasts,
3. Use Panels randomly sampled from (i) the Bloomberg™ Market Navigation Terminals [BBT] for GAAP-accounting data for firms currently traded on active exchanges, noted as: ***BBT:Data*** and (ii) Economic-Panels randomly taken from the M-Competition [Makridakis et al. (1982)] noted as: ***MComp:Data*** to determine if there is a Failure-effect depending on the nature of these two Panels.

Point of Emphasis We are NOT intending to suggest judgmental adjustments to the forecasts when there were, in fact, judgmental conditions that were in fact violated. ***We are only interested, in this preliminary analysis, in the frequency of the time that the Failure Screens come into play using real data.***

2. Gestalt of the Credentialing a Forecasting-Model

2.1 Graphical Context

Given that we have detailed the association measures, let us offer an illustrative graphical where the process-judgement ***Failure Screens*** are detailed: Figure 1.

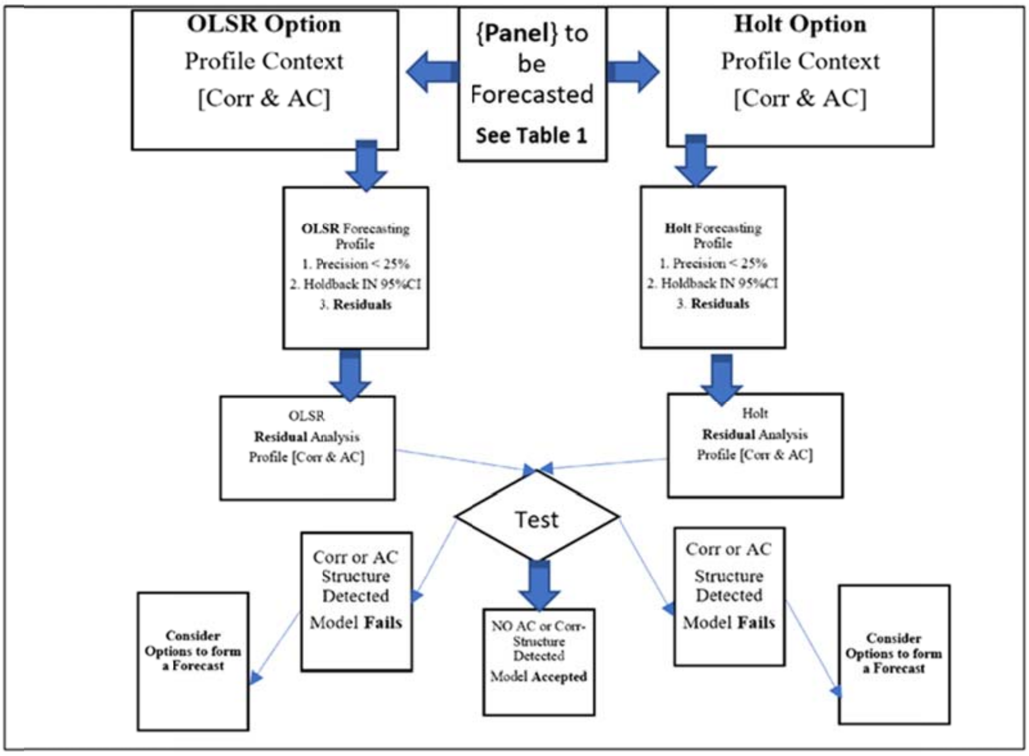

Figure 1. Overview Forecasting Model Testing for Acceptance or Failure ***Source: Authors***

Discussion of Figure 1

First, the selected Panel is forecasted using two models: OLSR & Holt. This usually depends on the initial profile of the Correlation and the Autocorrelation of the selected Panel. The results profile, after the fitting, are used to determine the ***FM-Failure-screening conditions***. The following Forecasting Model Failure Screens are used for ***both*** forecasting models as follows:

1. ***FM-Failure Screen [1]*** Is the (1-FPE) Confidence Interval Precision of the forecasting interval benchmarked by the Median of the series >= 25%? ***Computational Note***: We have elected to use the 95%CIs of the Fixed-Effects version of the construction of the CIs. The FM-fitting uses the first (n-1) Panel-points; whereas, the Median is for the full-Panel. This allows the last data-point to be the Holdback. The relative measure of 25% is a decision-maker election. We had no opinion on what Experts would have used as a relative Precision-benchmarking condition; thus, we used 25% as in a FM-analysis of a random-sample of 150 time-series of varying Panel-sizes ***using the 90th-Perecentile as a Trimming screen*** for the Median-benchmark percentages for the 95%CI-precision, the Median relative percentage of that trimmed-series was approximately 25%. ***Thus***, if the Median Benchmarked Precision **is >=25%**, this would suggest that the CIs are too-wide and the analyst may be prudent to question the utility of the forecasting model,

2. ***FM-Failure Screen [2]*** IF the Holdback—the next Panel point after the last point used in fitting the series—is OUT of the 95%CI of the fitted series for the FM, then the analyst may be prudent to question the utility of the FM to capture the likelihood of the past as a measure of the future, and

ASSUMED that the FM-model FAILS to be accepted as there was residual structure requiring that another model is considered.

2.2 Illustration of the Model Assessment Protocol

Following we will detail the results of the computations that are used to render an opinion of the nature of the forecasting model. Assume that the task is to produce the *one-period-ahead* forecast of the *Current Ratio* [CR] for *Accenture*, *Inc*. listed on the NYSE [*ACN*] using data in Table 1.

Table 1. Accenture PLC: NYSE[ACN] 2005 :2017 Current Ratio [Balance Sheet[Std]], Inc Panel: n=13, Fitting n=12, Point t_{13} is the holdback

1.3749	1.2457	1.1449	1.3375	1.4615	1.4562	1.4508
1.5523	1.4513	1.4592	1.2602	1.3488	t_{13}=1.2314	

As context, the ***initial*** pre-forecast profile is the PPM: Correlation and the Autocorrelation profiles and their related p-values. In this case,

ACN: PPM: Correlation[31.2%] : p-value[0.324]

Computational Clarification Recall, the Correlation is Pearson Product Moment [PPM] Association of the ACN-panel, n=12, ***with*** the matched Time-Index{1,- - - 12).

ACN: Autocorrelation[n=6]: [Average:29.7%] : p-value[Average:0.120**].**

Computational Clarification Recall, Autocorrelation computes an autocorrelation-value at each lag and for each Time-lag there is a related p-value. The p-value computation that we are using is that of Ljung & Box (1978) [L-BQ(k)] as programmed in SAS:JMPv.13. In this regard, we used the L-BQ(k) p-values for the truncated lags k= 1, - - -, n/2.

Discussion In this case, as there is reasonable evidence of structure—i.e., the generating processes are not random—the two models are tested: The OLS:Regression & The ARIMA (0, 2, 2)/Holt. ***Point of Information*** Any set of FM could have been tested. We have selected the OLSR & Holt as they are most often used in the time series context. Assume that the ***OLSR is the first tested***.

2.2.1 OLSR-Profile

Failure Screen(I) Relative Precision >=25% for the one period ahead forecast. So as to create comparative-intel over-time it is often the case that the precision is unitized using the Median of the dataset. Then, the Relative Precision Profiling Screen is computed. In this case, the relative precision for ACN is:

Relative Precision ≡ [[[Upper95%CI – Lower95%CI]/2] / Median]

Relative Precision: [[[(1.747 – 1.141]/2] / 1.3749] = 22.1%

Profile Status: For the OLSR, the Precision Failure Screen of >=25% is ***rejected***. Thus, the OLSR: Failure Screen [1] is **Not Founded**.

For Failure Screen [2], the Holdback of t_{13}=1.2314 is IN the OLSR 95%CI. Thus, the OLSR: Failure Screen [2] is **Not Founded**].

The next stage is to examine the p-value of the **residuals** of the OLSR-model. Profile of the OLSR residuals for ACN is:

ACN: PPM:Correlation: [n=12: **≈ 0**] : p-value[**≈ <1**]

For Failure Screen [3], the p-value of the PPM:Correlation is >=0.25. Thus, the OLSR: Failure Screen [3] is **Not Founded**].

and

ACN: Autocorrelation: [n=12/2: Average: **28.1%**]: p-value[Average:**0.188**].

For Failure Screen [4], Autocorrelation Average is <0.25. Thus, the OLSR: Failure Screen [4] **Is Founded**].

In this case, the OLSR fails as there is reasonable evidence that there is Autocorrelation in the OLS-residuals. True, the p-value of the autocorrelation of the residuals could be interpreted as slightly favoring the Null but nonetheless, it has a p-value <0.25, ***which is our assumed frontier***.

Thus, the next testing phase is the Holt-arm.

2.2.2 Holt-Profile

Holt Relative Precision, in this case, the relative precision for ACN is:

Relative Precision: [(**1.612 – 1.080]/2**) / **1.3749**] = **19.3%**

Also, the holdback of 1.2314 is in the 95%CI.

Profile Status: For the Holt, the Precision Failure Screen of >=25% is rejected. Thus, the Holt: Failure Screen [1] is **Not Founded**.

For Failure Screen [2], the holdback of t_{13}=1.2314 is IN the Holt 95%CI. Thus, the Holt: Failure Screen [2] is **Not Founded**.

The next stage is to examine the residuals of the Holt-model. Profile of the Holt-***residuals*** **for ACN is**:

ACN: PPM:Correlation:[n=10: **49.2%**] : p-value[**0.148**]

For Failure Screen [3], the p-value of the PPM:Correlation is <0.25. Thus, the Holt: Failure Screen [3] is **Founded**.

and

ACN: Autocorrelation: [n=10/2: Average: 11.2%] : p-value[Average:0.92].

For Failure Screen [4], the p-value of the Autocorrelation Average is >0.25. Thus, the Holt: Failure Screen [4] **is NOT founded**.

In this case, the Holt fails as there is reasonable evidence that there is PPM:Correlation in the Holt-residuals as the p-value of 0.148 indicates. Thus, also in the Holt case, there is evidence that there is structure in the residuals.

2.2.3 ACN: Summary

The initial indication for the ACN-Panel was that there was initial structure in the ACN-Panel. Indication: Correlation[31.2%] & p-value [0.324] as well as Autocorrelation[Average: 29.7%] & p-value[Average: 0.120]. Thus, both model selections were tested. However, the OLSR model failed as even though the PPM:Correlation of the OLSR-residuals was effectively zero and thus the p-value was about 1.0, for these OLSR-residuals there was evidence of Autocorrelation that suggested the existence of structure; recall the Autocorrelation p-value was **<0.25** suggesting structure in the OLSR-residuals. ***Thus, the OLSR-arm fails as after the initial OLSR-fit these OLSR-residuals exhibited AC-residual structure and so the fit was not effective.*** For the Holt-arm the Holt-residuals produced a PPM:Correlation p-value of <**0.25** and thus also for the Holt-fit there was structure in the residuals; ***thus, the Holt-arm also failed***. This ACN [Table 1] illustration thus produces a conundrum of sorts regarding selection of the forecasting model. Simply, both the OLSR & the Holt-Model leaves structure in the Residuals and thus there could likely be a more effective and valid forecasting model. ***Point of Information*** When the selected models fail this is just an indication that the fitting assumptions of the models are at variance with the data. ***This is very common***. Sometimes, transforming the data using the Box-Cox set of transforms can better align the model-assumptions with the data. See: (Zhou, Zhu & Sun (2017) and (McInerney, Thyer, Kavetski, Lerat & Kuczera (2017)). However, as a personal observation, transforming the data to effect useful alignment with the model and then to articulate the possible managerial actions given the forecasting results in the transformed data-space is fraught with difficulties. Another possibility is to use instrumental-integrate *via* two-stage models—called two-stage OLSR-models. See**:** Chen & Luo (2019). This tack, which is essentially Forecasting-*Voodoo*, requires highly advanced modeling skills. In our experience employing such instrumental analyses is very rare in practice.

2.3 Study Design

The question of interest, heretofore not yet posed in peer-reviewed sources is: ***How often do empirically collected Expert-opinions that are used to suggest judgmental-forecasting-conditioning Failure Screens create tacit failures in the forecasting process***. This is an ***ex-ante*** context—meaning that the ***Failure Screens*** are created <u>before</u> the forecasts are created. Then, and only then, based upon the actual forecasting profile created after passing the Panel through the forecasting filter, will the profile of the Forecasting ***Failure Screens*** be evaluated to determine IF the FM will be used, and if so, what will the judgmental adjustments that are likely to better align the forecasts. This is an important question as there is long-standing evidence that judgmental factors are rarely used *ex-ante* and are most often created ***ex-post***. For example, Collopy & Armstrong (1992), leveraging previously reported judgmental adjustment results in particular the studies of: Lawrence, Edmundson, O’Connor (1986) and Armstrong (1988), created a Rule-Based Forecasting Expert System [RBF] of combining forecasts due to the research report of Clemens (1989) that is predominately *ex-post* judgmentally based. The

RBF-model makes aligning adjustments based upon the character of the Panel being forecasted—i.e., *ex-post* guidelines. Our conditions are posted before the Panel is profiled—i.e., *ex-ante*.

With this as context, we will detail the experimental design to answer the above question as to:

How often do the four ex-ante Failure screens produce Failure-alerts?

2.4 The Elemental Aspects of the Inferential Design

Following we will suggest the inferential parameters that will offer the sample-size to create meaningful inferences, and the two datasets that will be used as blocking-variables for the ***Failure Screens***. ***Point of Clarification*** In this context, by using the Nature of the Data-Sets as Blocking-variables, we mean we will use the Holt Model for the BBT:Data & then for the MComp:Data to test if there are ***Failure Screen*** differences for the Holt Model that relate to the BBT:Data vis- à-vis those of the MComp:Data. Then we will perform the same test for the OLSR-Model.

FPE & FNE Credibility of the Inferential Framework In the context of this probl ématique, we are interested in three measures. The ***percentage of time*** for the ***Failure Screens*** that:

1. The Median benchmarked relative precision for the 95%Confidence Interval **is >= 25%**,
2. The Panel-Holdback value is **Out** of the forecasting 95%Confidence Interval, and
3. The Residuals have p-values **< 0.25**.

In forming these tests of proportions, we used a set of "standard" factors: FPE[90%[CV=1.645]] & FNE[80%[CV=0.842]] as a guide to determining the number of firms to accrue for the tests proposed. In this case, we assumed that Population[A] has a proportion of H%A and the other Population[B] has proportion of H%B. To form the sample accrual information in this two-population context, we used the following standard two sampled populations Test of Proportions sample size formula: See (Wang & Chow, 2007):

$$\text{Sample size} = \ [1.645 + 0.842]^2 \ \times [\frac{[H\%A*[1-H\%A]+[H\%B*[1-H\%B]}{[Abs[H\%A-H\%B]]^2}]$$

In this case, to initialize the computations, we used as a typical proportion-set for our context: [90% v. 75%]. We selected [90% & 75%] as they give identical results with their binary-partner of [10% & 25%]. Also the effect test of 15% ABS[10% – 25%] seems reasonable given the likely prevalence of 10% as expressed by the experts. This gives a boundary range of a sample size of 77 per generalized binary testing partition.

In our study-frame, we have: One accrual set of 21 BBT-Panels for which we randomly sampled 110 Firm&Account panels. [Appendix A] Further, we also sampled 80 series from the 181 M-Competition series-set used by C&A. ***We accrued 102 Firm&Account Panels and 73 M-Competition Panels***. The slight shortfall was due to missing data and a few series had 95%CIs for which the lower limit was in the negative quadrant. This can occur when there is anomalously high Panel variation. This would have the tendency to compromise the FPE-Null rejection logic giving an illusion of no difference when in fact in a "standard population" failing to reject the Null may not likely be the case. To control for this, rather than screening for non-Ergodic Panel-profiles, we simply screened-out any cases where the Lower-Limit was in the negative quadrant.

2.5 Accounting Variable Set for the BBT-accruals

Using Fraser & Ormiston (2013) as the reference guide for each BBT-firm, we selected four (4) Income Statement variables: **{*Gross Profit; Operating Income; Earnings for the Common Shareholders; Shares for Diluted Earnings per Share*}**, and four (4) Balance Sheet Statement variables: **{*Current Assets; Other Assets & Deferred Charges; Current Liabilities; Current Ratio*}**. As indicated we randomly sampled from the 21-Firms and 8-GAAP Account variables and arrived at 102-Panels. For the M-Competition the final accrual of 73 Panels is noted in Appendix B.

2.6 Firm Datasets of GAAP-Audited Accounts from the Balance Sheet and the Income Statements

[***GAAP:Data***] These datasets are audited under the assumption that management has in place a System of Internal Control over Financial Reporting [ICoFR]. The PCAOB requires that the management's system of ICoFR is audited and if there are issues in conformity with the current PCAOB audit rules a report is required where it is noted that the ICoFR is: *Adequate* or *Deficiency*[*Design*] or *Deficiency*[*Operation*] or *Significant Deficiency* or *Material Weakness* depending, of course, on the nature of the ICoFR-issues discovered during the certified audit. [See Note 1.] In addition, it is important to remember that the GAAP-rules are subject to management's interpretation; the scope of the boundaries of the application of these GAAP rules for recording the results of economic activity offer significant latitude. We tested the initial set of 25 accrual firms and eliminated any firm that had two (or more) Deficiencies of either type or one (one more) Significant Deficiency during the last five years of the accrual frame to arrive at the 21 firms in Appendix B.

The M:Computation firm Datasets [***MComp-Data***] These datasets report in the main results of economic activity. For example, Collective of Automobiles Manufactured: Total Production [France]; Consumer Expenditure OECD: Total Expenditures, Chemical Wood-Pulp Production [Brazil] & Gasoline Production [USA]. These datasets are also impacted by management's (i) sales protocols, (ii) production protocols, and/or (iii) shipping decisions. It is very interesting that there has been no research on the nature of the generating processes of the BBTs market traded organizations vis- à-vis the economic output of long-standing organizations such as those that were selected for the M:Competition.

3. Inferential Context: Vetting, A-Priori & Exploratory Analyses

3.1 Overview

Recently in the statistical community there have been questions as to the meaning of the p-values that are, more or less, focused on the assurance that can be offered as to the veracity of the linkage between the *a priori* expectations as to the nature of the population from which the samples were selected and the inference of the nature the population as drawn from these samples. This is called vetting the population and gives more assurance to the related p-values. Thus, the first tests that are offered are the population-vetting inferences.

3.2 Vetting

Initially, we offer a PPM-correlational vetting-analysis using the following two variable sets:

(i) $Vetting_A${The p-values of the Correlation with those of the Autocorrelation for the datasets as downloaded—i.e., before any forecasting model fitting.} and

(ii) $Vetting_B${The relative precision of the OLSR-model with that of the Holt-model}.

If it were to be the case that for either test that there were to be a PPM-correlation for which the usual Nulls were not rejected this would strongly suggest that these dataset accruals were likely to be atypical and thus cast doubt on any population inferences resulting from testing.

3.3 Vetting Discussion

It would be belaboring the point to discuss why there should be a high degree of association for these variables. But rather, we will offer why it belies economic norms to have Panels of 13-years even though they are from very different accrual periods—the *MComp* accruals are predominantly the 1950s through the 1970s, whereas the Bloomberg firms [*BBTs*] were accruals from 2005 through 2017— where the firm data would not be associated as suggested in $Vetting_A$ & $Vetting_B$. Despite the very different time periods, economic conditions were likely driving the Production Output and so the Accounting variables of these firms were also sensitized by the prevailing economic generating processes. This usually results in association over the Panels—to wit, given a particular Panel-point there is statistical-intel as to the likely neighborhood of the next Panel-point. This is opposed to the case where the economic generating process is Random—to wit there is no associational intel as to where the next Panel-point will lie given the previous Panel-point. Thus, a latent test underlying $Vetting_A$ and $Vetting_B$ is that there is an association among the Panel Points. As a ***pre-curser test*** of this assumed association, we created 175-Panels: n=13 of Random variables for which we computed the ***p-values for the FPE-Null*** of the PPM-Correlation between these Random-Panels & the Time Index. We tested the difference between the p-value Means and Medians for the {MComp & the BBT; n=175} against those of the Random-Panels with the following result:

MComp & BBTs p-value profile: [Mean:0.12 & Median: 0.0003]

Random Panels p-value profile: [Mean: 0.64 & Median: 0.60]

For the Welch-ANOVA [Welch] as well as the Wilcoxon / Kruskal-Wallis Tests (Rank Sums) [WK-W] the p-value profile of the {MComp & BBTs} versus the p-value profile of the {Random Panels}were sufficiently different that the Welch & WK-W inferential p-value indications were both <0.0001. All of the inferential computations were made using SAS[JMP]v.13[Analysis[Fit Y by X]].This clearly indicates that the association of the Panels assumed to have been driven by economic forces was not likely to have been random as the state of nature. Having this context information, we moved to the Vetting Tests.

Vetting Results

(i) $Vetting_A${The 175 p-values of the Correlation with those of the Autocorrelation for the datasets as downloaded—i.e., before any forecasting model fitting.}

This PPM-correlation was 0.54 the p-value of which was < 0.0001 indicating that the Null of no association would logically not be the state of nature.

(ii) $Vetting_B${The relative 175-precisions of the OLSR-model with those of the Holt-model}.

This PPM-correlation was 0.93 the p-value of which was < 0.0001 indicating that the Null of no association would logically not be the state of nature.

3.4 Implication

The MComp and BBT Panels are very likely to be reasonable accruals of Panels being driven by economic and systemic forces typical in the usual economic- and natural-context. These vetting results then lend credibility to the inferential tests to be used to derive the results of this research report.

4. Inferential Tests of Research Questions of Interest

4.1 Overview

There are two research agendas: The first is the test of an *a-priori* hypothesis the Null of which is:

$H1_{Null}A$ *The overall percentage of* ***failures*** *due to the Residual-Test Conditionals of the selected forecasting models: OLSR or Holt will not be greater than 10%. This will be directionally tested individually and unconditionally for both the OLSR- and the Holt-arms and will thus speak to the Nature of these* ***Failure Screens****.*

Discussion

$H1_{Null}A$ is a directional test for the right-hand side rejection of the FPE-Null. The logic of selecting 10% as the minimum upper-limit for not rejecting a forecasting model—{The OLSR and/or The Holt}—was based upon our discussion with practitioners over the years to the Question:

For the collection of forecasts that you have created over the years how ***often—the percentage of time—****do you feel that you have used a sub-optimal forecasting model—in that the fit left evidence of structure in the residuals—****but nevertheless*** *you continued to use that forecasting model?*

The percentages ranged from 1% to 20% in most cases. We took the practical mid-point of 10%. Recall in Figure 1 we are defining ***Failure*** as ***Structure*** in the Residuals, where there are two measures of ***Structure***: Correlation and Autocorrelation with respect to the residuals. This will be a two-step process for testing for Failure. First, the OLSR-Residuals will be tested by the Correlation of the residuals with the relative time index; IF that p-value is >= 0.25, then these OLSR-residuals are tested using the Autocorrelation function. If that second-stage results in an average p-value >=0.25 then, that condition results in the OLSR-model being accepted as a valid model as there was no actionable evidence of Correlation or Autocorrelation structure in the residuals. However, if there is a p-value that is <0.25 at either stage, then the model will have ***failed*** in that the fit did not remove all the structure in the data.

4.2 Results for $H1_{Null}A$ OLSR-Arm

In this case, accepting 10% as the toggle-rejection point where the actual test-point is a directional test against 10%, we tested the OLSR- and Holt-arms using **ONLY** the Residuals-test. The results are:

4.2.1 OLSR-arm

The number of instances where the Correlation OR the Autocorrelation of the OLSR-residuals was < 25% of the 175 cases tested was 96. This ***Failure***-percentage is 54.9%. The p-value for this FPE tested against 10% is < 0.0001. The empirical likelihood result for the OLSR-arm is given by the 95%CI that is: [47.4% : 62.3%]. Using ONLY the residuals test, this empirical result suggests that the population parameter expectation for the failure of the OLSR- model is in the range 47.4% through 62.3% of the time. The p-value indicates that, assuming that the true population percentage of ***Failures*** is 10%, i.e., the experiential indication of the solicited as Expert opinion, the chance of finding a failure rate of 54.9% in a random sample of 175 forecast experiences would happen ***less than 1 time out of 10,000*** by random sampling chance. **Thus, the p-value of <0.0001 provides clear evidence that the Null should be rejected offering the alternative that the Failure-rate is likely greater than 10%.**

4.2.2 Holt-arm

The number of instances where the Correlation OR the Autocorrelation of the Holt-residuals was < 25% of the 175 cases tested was 66. This ***Failure***-percentage is 37.7%. The p-value for this tested against 10% is < 0.0001.

The empirical likelihood result for the Holt-arm is given by the 95%CI that is: [30.5% : 45.0%]. Using ONLY the residuals test, this empirical result suggests that the population parameter expectation for the failure of the Holt model is in the range 30.5% through 45.0% of the time. The p-value indicates that assuming that the true population percentage of ***Failures*** is 10%, i.e., the experiential indication of the solicited Expert opinion, the chance of finding a failure rate of 37.7% or higher in a random sample of 175 forecast experiences would happen less than 1 time out of 10,000 by random sampling chance. ***Thus, the p-value of <0.0001 provide clear evidence that the Null should be rejected offering the alternative that the Failure-rate is likely greater than 10%.***

The simple and clear result is that the Nulls of the two-arms tested can be rejected. For a clarification: The non-rejection region for a directional test for an FPE of 5% re: 10% is <=13.7%. Thus, as both arms tested are multiples of 13.7%, the inferential evidence is: There is a clear indication that it is not the case that for the OLSR or for the Holt that these models rarely, specifically a population reality of 10% of the time, leave structure that is detected by Autocorrelation or Correlation of the post-fit residuals. The alternative, given the empirical directional evidence, is that more than 10% of the time both the OLSR and the Holt models leave evidence of structure in the fitted residuals measured by either Correlation or Autocorrelation. ***Simple Indication: In that case, both models fail the usual residuals-screening forecasting screening test often used in practice.***

5. Exploratory Analyses

5.1 Overview

There are two interesting derivative hypotheses relative to the ***Failure*** of the OLSR & Holt models. Recall that a forecasting model ***fails*** if after the fitting of the model there is evidence of structure in the residuals as detected by <u>either</u> Correlation or Autocorrelation. These exploratory tests have the following Nulls:

$H1_{Null}B$ *The percentage of failures due to the Residual-Test Conditionals of the selected forecasting models: OLRS and Holt will not differ.*

$H1_{Null}C$ *The percentage of failures due to the Residual-Test Conditionals of the selected forecasting models: OLRS and Holt* ***blocked*** *by the nature of the datasets: BBT:Data or the MComp:Data will not differ.*

Discussion of $H1_{Null}B$

5.1.1 Results for $\mathrm{H1_{Null}B}$

These are the same results tested in $H1_{Null}A$ except the Null is that there is no difference in the Failure frequency between the OLSR v. Holt Models even given that they both have Failure rates >>10%.

Table 2. Failure Rate tests OLSR v. Holt **[SAS[JMPv.13[AnalysisTab]]**

Residuals Conditional **n=175**	Overall Direct Test Profile
Failure: Holt<25%	**37.7%**: *0.0*
Failure: OLSR <25%	**54.9%** : *1.0*
Parametric Test	**0.0012**
Non-Parametric Test	*0.0013*

Discussion

For the profile of Table 2, the Mean is noted in **Bold** and the Median is noted in *Italics*. There is clear inferential evidence from the non-directional parametric Welch-ANOVA as well as the non-directional non-parametric Wilcoxon / Kruskal-Wallis Tests (Rank Sums) tests, the p-values of which were **0.0012** and *0.0013* respectively, that clearly indicate that the ***Failure***-rate for OLSR is not the same as the ***Failure***-rate for the Holt model where: the likely empirical indication in the population of interest is that the ***Failure***-rate for the OLSR-model is the higher than that of the Holt-model.

5.1.2 Results for $\mathrm{H1_{Null}C}$

$H1_{Null}C$ *The percentage of failures of the selected forecasting models: OLRS and Holt* ***blocked*** *by the nature of the datasets: BBT:Data and the MComp:Data will not differ.*

In this case, these tests examine the effect of the datasets on the ***Failure***-rates.

Results for $H1_{Null}C$ for the Residuals-Conditional are presented in Table 3:

Table 3. Tests of the Model Effects vis- à-vis the Data Panels**[SAS[JMPv.13[AnalysisTab]]**

Relative Accrual Blocking Residuals Conditional	Holt **Mean** & *Median*	OLSR **Mean** & *Median*
BBTs[n=102]	**38.2%** : *0.0*	**54.0%** : *1.0*
MComp[n=73]	**37.0%** : *0.0*	**56.2%** : *1.0*
Parametric Test	**0.87**	**0.77**
Non-Parametric Test	*0.87*	*0.77*

Discussion

For $H1_{Null}C$ For the profile of Table 3, there is clear inferential evidence from the non-directional parametric Welch-ANOVA as well as the non-directional non-parametric Wilcoxon / Kruskal-Wallis Tests (Rank Sums) tests controlling for the Data Panels: the BBT and the MComp the p-values of which were **0.87** and *0.77* for both inferential-screens that the likely indication is that the ***Failure***-rate is not affected by the nature of the Panel. ***Simple Indication: Consistent with the vetting indications that the selected datasets do not test in either case to be driven by a Random generating process, the relative Failure-rates of the Holt for both the BBTs- and the MComp-Panels & the relative Failure-rates for the OLSR for both the BBT- and MComp-Panels are clear indications that their Nulls are the likely state of nature. In this case, the nature of the Panel does not likely impact the Failure-rates of either the Holt- or the OLSR-Models.***

6. Exploratory Analyses: Precision Conditional-Screen

6.1 Overview

As relative *Precision* is one of the forecasting evaluation variables, it is of interest to determine if there is a difference in the relative *Precision* between (i) the OLSR- and Holt- Models overall, (ii) then examining the same question ***blocking*** by the Nature of the two Panels, and finally (iii) the same analysis *conditioned by a limit on the magnitude of the relative Precision*. In this testing context, the Null hypotheses to be tested are:

$H2_{Null}A$ *The Relative Precision of the Models: OLSR & Holt are not different overall in their central tendency—i.e.,* ***not blocking*** *by the BBT:Data and the MComp:Data,*

$H2_{Null}B$ *The Relative Precision of the Models: OLSR & Holt are not different in their central tendency* ***blocked*** *by the BBT:Data and the MComp:Data,*

$H2_{Null}C$ *The frequency of the Relative Precision that is <25% for the Models:OLSR & Holt are not different in their central tendency blocked by the BBT:Data and the MComp:Data*

Discussion This is a test of the difference in the OLSR- & the Holt-Model's relativity variability that is the principal driver of the width of the one-period-ahead forecast prediction interval. Interestingly, intensive searching using: *ABI:INFORM*™ and *Business Source Premier*™ failed to retrieve intel as to the expectation for the OLSR-model's relative precision for the one-period ahead forecast *vis- à-vis* that of the Holt-model. In this case, these tests will be initial-empirical information on the relative precision of the these forecasting models. Although these tests could be formed as a Factorial-design with main-effects and interactions, due to Power considerations, as presented above, we have elected to present three testing partitions.

6.1.1 Results for $\mathrm{H2_{Null}A}$

Following is the Profile for robust testing of $H2_{Null}A$. In Table 4 the profile of the Mean is **Bolded** and that of the Median is scripted in *Italics*.

Table 4. Overall Test of the Central Tendency of the Models' Relative Precision**[SAS[JMPv.13[AnalysisTab]]**

Precision Conditional	**Relative Precision Profile**
Holt, n=175	**23.5%** : *17.8%*
OLSR, n=175	**28.2%** : *20.5%*
Parametric Test	**0.085**
Non-Parametric Test	*0.11*

Discussion

Results for $H2_{Null}A$ For the profile of Table 4, and using the suggested p-values of Table 4 there is inferential evidence from the non-directional parametric Welch-ANOVA as well as the non-directional non-parametric

Wilcoxon / Kruskal-Wallis Tests (Rank Sums) tests—not controlling for the nature of the datasets—using the p-values 0.085 and *0.11* that the likely indication is that the relative precision of the OLSR-Model is not the same as that of the Holt-Model. The empirical evidence suggests that the OLSR-Model's relative precision is greater than that of the Holt-Model. However, tacitly very interesting, for both the OLSR- & Holt- Models the 95%CI of the Means contain 25%. Specifically: 95%CI OLSR[24.0% : 32.4%] & 95%CI Holt[20.3%: 26.8%]. Recall, relative Precision **>= 25%** is one of the possible ***Failure*** screens of Figure 1. This rationalizes the logic to test $H2_{Null}C$ which investigates the frequency of the events where the relative precision is **< 25%**—a possible "***Desirable State***" event. ***Simple Indication: There is inferential evidence that there is a difference between the OLSR-Model and the Holt-Model relative to the width of their 95%CIs. The empirical evidence suggests that of the OLSR is wider relative to that of the Holt-Model.***

6.1.2 Results for $\mathrm{H2_{Null}B}$

Following is the Profile for robust testing of $H2_{Null}B$. In Table 5 the profile of the Mean is **Bolded** and that of the Median is scripted in *Italics*.

Table 5. The Relative Precision of the Holt- & the OLSR-Models controlling for the Nature of the Datasets BBT & MComp**[SAS[JMPv.13[AnalysisTab]]**

Relative Accrual Blocking **Relative Precision**	Holt	OLSR
BBTs[n=102]	**22.2%** : *19.6%*	**26.2%** : *22.0%*
MComp[n=73]	**25.4%** : *16.1%*	**31.0%** : *17.1%*
Parametric Test	**0.40**	**0.32**
Non-Parametric Test	*0.31*	*0.25*

Discussion

Results for $H2_{Null}B$ For the profile of Table 5, there is suggestive inferential evidence from the non-directional parametric Welch-ANOVA as well as the non-directional non-parametric Wilcoxon / Kruskal-Wallis Tests (Rank Sums) tests—controlling for the nature of the datasets—using the p-values Holt: [0.40 & 0.31] and the p-values of the OLSR: [0.32 & 0.25] that the likely indication is that the Nature of the Datasets do not affect the Median relative Precision. ***Simply the Nulls are the likely state of nature.***

$H2_{Null}C$ *The frequency of the Relative Precision that is <25% for the Models:OLSR & Holt are not different in their central tendency blocked by the GAAP:Data and the MComp:Data*

In this case, recall that the Failure is a judgmental parameter used by a forecasting group to rate the "confidence or the desirability of a possible projection interval" that can be placed in the forecasting parameters that have created the 95%CI. If the relative Precision is ***<25%*** this is a judgmental screen that can be used as an indication that such a 95%CI has a "***special status as a very meaningful projection interval***". Thus a ***Failure*** in this context is: After the fit the relative Precision of the 95%CI is ***>=25%***. This latter indication suggests that the 95%CI is relatively too-wide to inform the decision-making process and thus it is rated as a Failure as it not sufficiently discriminating to inform the decision-making process.

6.1.3 Results for $\mathrm{H2_{Null}C}$

Results for $H2_{Null}C$ are presented in Table 6.

Table 6. ***Success***-Screen Blocked by the Nature of the Datasets**[SAS[JMPv.13[AnalysisTab]]**

BBT & MComp Blocking **RelativePrecision<25%**	Holt[**64.6%**]	OLSR[**58.9%**]
BBT:Data, n=102	**63.7%,** 1.0	**57.8%,**1,0
MComp:Data, n=73	**65.8%,**1.0	**60.3%,** 1,0
Parametric Test	**0.78**	**0.75**
Non-Parametric Test	*0.78*	*0.75*

Discussion

Point of Information the difference 2.1% ABS[63.7% − 65.8%] and the difference 2.5% ABS[57.8% − 60.3%] have Welch-ANOVA as well as the non-directional non-parametric Wilcoxon / Kruskal-Wallis Tests (Rank Sums) p-values of 0.78 and 0.75 respectively. In Table 6 there is no logical reason to reject the Nulls for either the BBT or the MComp blocking partitions as the p-values are >> 25%. The interesting results are that for both the datasets and also overall only about 60% of the time the relative Precision is in the ***judgmentally desirable- or acceptable-zone***. This suggest that about 40% of the time the 95%CI are *too wide to be useful in the forecasting context*.

6.2 Extension of Inquiry The above results beg two interesting questions. What would the performance profile be for the accrued datasets if we were to assume that ALL four of the following ***Success Screens***[SS] were <u>required</u> to be satisfied to validate the use of a forecasting model:

1. SS[1] The Median relative Precision is **< 25%**, **and**
2. SS[2]The Holdback is **IN** the One-Period-Ahead Fixed-Effects 95%Confidence Interval, **and**
3. SS[3]The Residuals, after the forecasting model fit, have a PPM-correlation with the Time-Index p-value of which is **>=25%**, **and**
4. SS[4]The Residuals, after the forecasting model fit, have an Average Autocorrelation Ljung-Box p-value which is **>=25%**.

In this case, an operative question is:

> *IF the forecaster would use the OLSR* ***<u>or</u>*** *the Holt forecasting model but* ***NOT both*** *ONLY under the condition that ALL four of the above success-conditions were to be founded, what would be the frequency profile of the Holt- & the OLRS-Models?*

A related question would be:

> *IF the forecaster would use* ***BOTH*** *the OLSR- & the Holt-Models and select a model that satisfies ALL four of the above success conditions, what would be the frequency profile of that decision-election?*

As for the first question, the profile is presented in Table 7.

In this testing context, we are profiling the Holt or the OLSR model. These profiles address the question: IF the forecaster only were to employ the Holt ***or*** the OLSR ***but not both*** what would be the chance of finding a forecasting model that could be used?

Table 7. Model ***Success*** Profile Overall & Blocked by BBTs & MComp**[SAS[JMPv.13[AnalysisTab]]**

Models	OverAll, n=175			Holt n=175		OLSR, n=175	
Profile	Mean:Median	95%CIs		Mean:Median	95%CIs	Mean:Median	95%CIs
Holt	33.2% : *0.0%*	[26.1% : **40.2%]**		BBTs: 33.3%:0.0%	[24.0% : 42.6]	BBTs: 23.5%: 0.0%	[15.2% : 31.9%]
OLSR	22.9% : *0.0%*	[16.6% : **29.1%]**		MComp: 32.9%:0.0%	[21.8% : 43.9%]	MComp: 21.9%:0.0%	[12.2% : 31.6%]
p-values	**0.03**/*0.03*	N/A		**0.95**/*0.95*	N/A	**0.80**/*0.80*	N/A

Discussion

Point of Information The two p-values are the Welch-ANOVA as well as the non-directional non-parametric Wilcoxon / Kruskal-Wallis Tests (Rank Sums). Overall the Holt-Model seems to be preferred to the OLSR-Model where the p-values are **0.03** & *0.03*. As additional inferential information, we have offered the dataset-partitions. For the Holt-Model the datasets: BBTs and MComp do not figure as important considerations in electing forecasting models as the p-values for **0.95** & *0.95* for the Holt and **0.80** & *0.80* for the OLSR do not suggest that the respective Nulls are not likely. These results speak to the validation profile that has four conditions as noted above.

The second question asks: Assume that we have two forecasting models: the OLSR & the Holt. If BOTH are employed how often will at least ONE of them profile a useful forecasting model.

This information is profiled in Table 8.

In this testing context, we are profiling the Holt- the OLSR-Models in an ***Either OR*** context. This profiling screens each of the 175-datasets for BOTH the Holt- & and the OLSR-Models and IF either [or both] models satisfy the four-success-screens then that is recorded. Clear is, this screen percentage will be greater than the profile offered in response to the first question as the selection population is the entire VENN-space. In this context, to facilitate the comparison, we will format the presentation is the same manner.

Table 8. Combinations of AND / OR Profiles**[SAS[JMPv.13[AnalysisTab]]**

Models	OverAll, n=175			EitherOR[Holt or OLRS]	
Profile	Mean:Median	95%CIs		Mean:Median	95%CIs
Either/OR	41.7% : 0.0	34.3% : 49.1%		BBTs: 41.2%	31.5% : 50.9%
[redacted]	[redacted]	[redacted]		MComp: 42.5%	30.9% : 54.1%
p-values	N/A	N/A		0.87/0.87	N/A

Discussion

In this case, where both the Holt- and the OLSR-Models in the section purview of the forecasting manager, the possibility of finding an acceptable forecasting model among the two standard choices is “clinically” higher than if only one model is on the menu. This is hardly surprising. Of confirmatory intel, the datasets-partition is not a selection factor.

7. Summary and Outlook

7.1 Summary

There has been a paucity of information generated, reported, and discussed in detail. Following, we offer an “*en bref*” summary that can be used as the take-away of the above investigations.

The ***Take-Aways*** offered without the caveats of excessive-inferential verbiage are:

1. There is ***no*** dramatic effect of the ***nature of the datasets*** relative to selecting a forecasting model,
2. The ***Holt-Model outperforms*** the ***OLSR-Model,***
3. Using just the standard Litmus-test of: ***Structure in the Residuals***, the experts seem to have under-estimated the frequency of the time that they have used “sub-optimal forecasting models”. We found that 10% seems too-low by a multiplicative factor.
4. The best-case likely scenario—the upper limit of the 95%CI for the ***successful*** employment of the ***OLSR-Model*** is around ***30%*** while the upper limit of the 95%CI for the ***Holt-Model*** is around ***40%*** as bolded in Table 7.
5. These results, we suggest, may be contemplated with assurance as the p-values were derived from Vetted-Accrual datasets. We are mentioning this so as to encourage researchers to vet their accrual datasets.
6. ***Alert***: ***Neither of these models the OLSR or the Holt can be expected to inform the forecasting decision-making process.*** *Rationale*: If the best-case scenario of creating a relevant and reliable forecasting model is expected to obtain on the order of less than 50% of the time it is not likely that such an analytic-regime can have operational utility. This study has generated interesting and to some extent information that is not without an action frame that at minimum begs reflection.

7.2 Outlook

These results are clear and also confusing. Forecasters have been modifying or “spinning” the results of forecasting-intel since the days of the Oracle of Delphi. These pervasive judgmental adjustments are likely to be endemic and also perhaps useful to correct the projections that use past data to project into an uncertain dynamic world. The unanswered question is:

Do forecasters who indicate ***ex-ante*** their failure conditions create better judgmental forecasts than forecasters who only make judgmental adjustments ***ex-post***—i.e., sort of flying-blind?

Actually, adhering to the usual rigors of Campbell and Stanley (1963) to assure inferential assurance, perhaps there is no experimental design to address this question. However, there can certainly be Quasi-designs to ask the question:

If forecasters or students charged with executing a forecasting task in an experimental setting were randomly assigned to:

Arm[1]: Ex-ante Forecast Failure Screens are recorded ***by the forecasters*** and used—with documentation—to judgmental-adjust the Model(s) generated forecasts and the final forecasts are reported.

Arm[2]: The forecasting context is encoded using RBF-Rules or some similar Ex-post encoding and these ***Rules are given to the forecasters*** and these Encoded-Rules are used to guide the forecasting process. Then forecasts are made and as recorded—with documentation.

Arm[3]: Here there are no instructions or rules of any kind given to guide the forecasting process. The model forecasts are recorded and also any judgmental adjustments that are made to the model generated forecasts.

Inferential context

The forecasts, the judgmental adjustments made to them, and the accuracy of the forecast considering the holdback will be inferentially evaluated for the three study arms.

In addition, our results suggest that there may be information in profiling the sensitivity of the Relative Median Precision [RMP]. This would give a trade-off curve of ***Failure*** relative to the RMP. Also in this test, an accrual of another Set of Panels from the BBTs would add a vetting dimension to better validate this study and also the utility of the sensitivity curve of the Failure & RMP.

Acknowledgments

Appreciation is due to: Mr. John Conners, Senior Vice President, Financial Counseling, West Coast Region, AYCO for his generous philanthropy which funded the establishment of the John and Diana Conners Finance Trading Lab at the State University of New York College at Plattsburgh and the Bloomberg Terminals that were instrumental in this research. Further thanks are due to: Prof. Dr. H. Wright, *Boston University*: Department of Mathematics and Statistics, the participants of the SUNY: SBE Workshop Series, in particular Prof. Dr. Kameliia Petrova: Dept. of Economics[Statistics] and the reviewers of the *International Business Research* journal for their helpful comments and suggestions.

References

Adya, M. (2000). Corrections to rule-based forecasting: findings from a replication. *International Journal of Forecasting*, *16*, 125-127. https://doi.org/10.1016/S0169-2070(99)00034-5

Adya, M., & Lusk, E. (2016). Time series complexity: The development and validation of a rule-based complexity scoring technique. *Decision Support Systems*, on-line. https://doi.org/10.1016/j.dss.2015.12.009

Armstrong J. S. (1988). Research needs in forecasting, *International J. of Forecasting*, *4*, 449-465. https://doi.org/10.1016/0169-2070(88)90111-2

Campbell, D., & Stanley, J. (1963). *Experimental and Quasi-Experimental Designs for Research*. Rand McNally & Company: ISBN-13: 978-0395307878.

Chen, C., & Luo, R. (2019). Executive marketing background, corporate trademark and brand management. *J. Contemporary Marketing Science*, *2*, 345-367. https://doi.org/10.1108/JCMARS-08-2019-0030

Clancey, W.J. (1983). The epistemology of a rule-based expert system: A framework for explanation. *Artificial Intelligence, 20,* 215-251. https://doi.org/10.1016/0004-3702(83)90008-5

Clemens, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*, 559-583. https://doi.org/10.1016/0169-2070(89)90012-5

Collopy, F., & Armstrong, J. S. (1992). Rule-based forecasting: Development and validation of an expert systems approach to combining time series extrapolations, *Management Science*, *38*, 1394-1414. https://doi.org/10.1287/mnsc.38.10.1394

Fraser, L., & Ormiston, A. (2013). *Understanding financial statements* (10th ed.) Pearson, ISBN-10: 0-13-265506- 3.

Kaplan, R., & Norton, D. (1992). The balanced scorecard: Measures that drive performance. *Harvard Business Review*, *J*(F), 71-79.

Lawrence, M., Edmundson, R., & O'Connor, M. (1986). The accuracy of combining judgments and statistical forecasts. *Management Science*, *32*, 1521-1532. https://doi.org/10.1287/mnsc.32.12.1521

Ljung, G., & Box, G. E. P. (1978). On a measure of a lack of fit in time-series models. *Biometrika*, *65*, 297-303. https://doi.org/10.1093/biomet/65.2.297

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., … Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *International Journal of Forecasting*, *1*, 111-153. https://doi.org/10.1002/for.3980010202

McInerney, D., Thyer, M., Kavetski, D., Lerat, J., & Kuczera, G. (2017). Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. *Water Resources Research*, *53*, 2199-2239. https://doi.org/10.1002/2016WR019168

Mohammad S. M., Anvari, M., & Saberi, M. (2013). Targeting performance measures based on performance prediction, *International Journal of Productivity and Performance Management*, *61*, 46-68. https://doi.org/10.1108/17410401211187507

O'Connor, M., Remus, W., & Griggs, K. (1993). Judgmental forecasting in times of change. *International Journal of Forecasting*, *9*, 163-172. https://doi.org/10.1016/0169-2070(93)90002-5

Sanders, N. R., & Ritzman, L. P. (1992). The need for contextual and technical knowledge in judgmental forecasting, *Journal of Behavioral Decision Making*, *5*, 39-52. https://doi.org/10.1002/bdm.3960050106

Wang, H., & Chow, S. C. (2007). Sample size calculation for comparing proportions. Test for equality: *Wiley encyclopedia of clinical trials*. https://doi.org/10.1002/9780471462422.eoct005

Zhou, J., Zhu, J., & Sun, L-Q. (2017). A class of Box-Cox transformation models for recurrent event data with a terminal event. *Acta Mathematica Sinica* [English Series], *8*, 1048-1060. https://doi.org/10.1007/s10114-017-6221-4

End Note

Note 1, See: <https://www.scribd.com/document/535509490/auditing-standards-audits-after-december-15-2020> Auditing_standards_audits_after_december_15_2020.pdf [Section[AS 1305: Communications About Control Deficiencies in an Audit of Financial Statements[p87]]

Appendix A

Table A. Sampled 21 Firms : Bloomberg

6758JP	ACN	AXE	BA	BAE	CVS	EFX	HSY	JBLU	ILMT	ISIE
ISNA	LUV	RAD	ROK	SIE GR	SNA	SPGI	SWK	UTX	WBA	

Appendix B

Makridakis et al. (1982) M-Competition, n=73) The following numbers are the designations give in the M-Competition. **Bolding** is to aid in identification. The M-Competition datasets are found at: **URL-https://www.journals.elsevier.com/international-journal-of-forecasting**

{**1,4,6,10,14,16,17,18,20,22**,23,29,31,43,45,49,50,51,52,54,**56,57,63,64,65,67,71,73,74,76**,78,80,81,86,87,94,97,99,100,101,**102,103,105,108,110,114,118,120,121,122**,123,124,127,132,138,141,142,143,146,150,**151,153,154,157,161,162,165,170,176,177**,178,179,181.}

Most all of these Panels represent Total Units of Production: For example, Collective of Automobiles Manufactured: Total Production [France]; Consumer Expenditure OECD: Total Expenditures, Chemical Wood-Pulp Production [Brazil] & Gasoline Production [USA]

Copyrights