EJSIT
European Journal of Science, Innovation and Technology
ISSN: 2786-4936
www.ejsit-journal.com
Volume 3 | Number 4 | 2023

Vetting the Makridakis Dataset: Further Indications of the Robustness of the Rule Based Forecasting Model

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, SUNY: Plattsburgh, USA & Chair: International School of Management: Otto-von-Guericke, Magdeburg, Germany [https://www.uni-magdeburg.de/en/]

ABSTRACT

Context: The year 2022 marks the 30th anniversary of Collopy and Armstrong's The Rule Based Forecasting [RBF] Expert Systems Model. Over the last three decades, there has been a plethora of research reports—truly a research Cornucopia—spawned by this very unique, effective, and ground-breaking forecasting system. ***Focus:*** The purpose of this research note is to: (i) Briefly, remind the forecasting community of the excellent pre-model-launch vetting used by Collopy and Armstrong [C&A] to form their RBF-model. Important is: their vetting protocols readily generalize to most modeling domains, and (ii) Offer a "re-vetting" analysis of the *M-Competition* dataset used by C&A that addresses their comment: *"This study also used long calibration series - - -; rule-based forecasting benefits from long series because it uses information about patterns in the data.* ***We do not know how the procedure will perform for short series.***" [p. 1403[Bolding Added]]. ***Results:*** We trimmed selected series from the M-Competition to arrive at 165-series all of which had 13-time series points for the OLS Regression-fit [OLS-R] & three panel-points as holdbacks. We found that: (i) there is evidence that these trimmed-series likely have inferentially differentiable variance profiles compared to the performance profiles reported by C&A, and (ii) despite this, these trimmed-segments did not seem to compromise the C&A's parametrization of the RBF Model in comparison to OLS-R forecasts. Finally, we suggest the need for an extension of the RBF Expert System re: (1-FPE) Confidence Intervals that would further enhance RBF-testing with respect to capture-rates and relative precision.

Keywords: Model Development, Short-Segment Forecasting Effect

INTRODUCTION

Overview

Rule Based Forecasting [RBF] started in the mid-1980s with J. Scott Armstrong, an established and recognized global expert in forecasting, who developed with his Wharton PhD student, Fred Collopy, a revolutionary, and, yes bold, departure from the usual forecasting model-tweaking that was the research predilection of the day. Their inspiration and motivation were founded on the ground-breaking research of Spyros Makridakis. Spyros organized, for the first time, an evaluation of the plethora—*really glut*—of the *en vogue* forecasting models, *circa* the 1980s. This was called the Makridakis Competition [*M-Competition*].

Innovative Model Development

Collopy and Armstrong (1992) [C&A] realized, quite correctly, that one of the requirements of acceptance and utilization of an innovative but, intricate forecasting protocol, was to engender confidence that the developmental stages included logical and effective

* Corresponding Author

vetting-protocols[†]. In this context, we offer the following ***practical definition***, which is certainly generalizable, of vetting in the developmental stages of a model:

> *Vetting is the use of data analytical and judgmental protocols to provide answers to logical questions posed, usually in the developmental-phases of a modeling project, by the model-developers that lead, sometimes to inferential testing and sometimes to judgmental actions taken by the research group, to redesign/correct certain modeling aspects that overall engender confidence in the "veracity" of the model with respect to the logical purpose of the model under development.*

RE-VETTING: THERE IS NO EXPIRATION DATE

Overview

The C&A vetting was intensive as well as extensive, however, a possible vetting-stone—the ***short-series forecasting issue***—was left unturned. C&A (p.1400) offer [***Bolding*** Added]:

"*We also hypothesized that rule-based forecasting would provide more accurate forecasts than the random walk*[‡] *or equal-weights combining when:*

(3) historical data **show stable patterns**, *and*

(4) ***good domain knowledge*** *is available.*

Extrapolating ***trends is risky*** *because if the trend forecast is in the wrong direction, the resulting forecast will be less accurate than the random walk. Also, if the* ***trend changes*** *substantially, it can produce errors larger than from the random walk. - - -. The rule base relies upon* ***the patterns in the data*** *to decide which extrapolation methods to use and how to weight them.*" [p.1400] continuing [- - -]

"*Because rule-based forecasting can incorporate causal information, we expected that it would be more useful for long-range than for short-range forecasts,* **because causal factors have stronger effects in the long term.**" [p. 1401] continuing [- - -]

"*This study also used long calibration series (median of 15 years); rule-based forecasting benefits from long series because it uses information about patterns in the data.* ***We do not know how the procedure will perform for short series.***" [p. 1407].

Discussion

Our take-ways from these C&A-excerpts are: ***Short-series are very sensitive to Relative Uncertainly, Level-shifts, or Trend-repositioning that could: (i) erroneously skew the projection-orientation, or (ii) inappropriately encourage rejecting forecasts from a Panel due to excessive volatility in the early sections of the Panel.***

THE RBF-VETTING FOCUS: INFERENTIAL SCREENING

Overview

C&A (1992); Adya, Armstrong, Collopy & Kennedy (2000); Adya, Collopy, Armstrong & Kennedy (2001); and Adya & Lusk (2013 & 2016) used various sub-samples of the 181-annual series of the M-Competition; note this as: The M[181]-dataset. Now that there is an abundance of statistical-information that has been generated using various sub-samples from

[†] Following are the page-citations where C&A note their Vetting-Protocols used in forming their RBF Model: [P. 1359]; [P. 1395]; [Ps.1395-6]; [P. 1396]; [P. 1398]; [Ps. 1401-2]; [P. 1402]; [P. 1403] & [P. 1407]. We used these vetting references in our courses to instruct students in pre-launch vetting of Data Analytic Models.

[‡] Initially, the Random Walk Forecasting Model [RWM] was proposed by the developers of the Makridakis M-Competition; they called it the Naïve 1 Model See Makridakis, et al (1982): Appendix 2[Model (1) Eq4]. Simply, the RWM-forecast for ALL horizons is the last observed data point that was used for parametrizing other competing forecasting models. To this extent the RWM is a benchmark for the other forecasting models.

the M[181]-series, ***all of which are routinely benchmarked against C&A's performance***, we are obligated, better late than never, to pose the following fundamental re-vetting-question:

> *For the M[181]-dataset is there inferential evidence that more than a few of these Panels were composed of initial segments that exhibited excessive variability, the nature of which, may likely have compromise the quality of forecasts?*

Given this mused observation as our focus, we will test for a possible Short-Panel effect in the M[181].

RESEARCH PROTOCOL: THE VETTING COMPONENTS & THE EXPECTATIONS

Components

Following are the Aspects of the Research Plan:

I. Assume that the M-Competition Panels: M[181] were modified, thus ***resulting in a natural increase in Panel-volatility***; refer to this modification as: M[Mod], next
II. We selected ***two-forecasting models*** from among the four used by C&A in their RBF Model that do not require unique user parametrization for forecasting, then
III. Using various standard forecasting performance measures, sensitive to differences in Panel-***volatility***, we will create forecasting profiles from the M[Mod]-dataset, refer to this as M[Mod[Profiles]], finally
IV. Using experiential-inferential methods, we will compare the C&A RBF[Profiles] with the M[Mod[Profiles]].

Expectation

If the C&A RBF[Profiles] are inferentially outperformed by the M[Mod[Profiles]], this would indicate that: (i) The 99-RBF rules were not effective in scoring the parameters of the RBF Model, or (ii) They, in fact, could have been effective but were not properly executed by C&A. Either would be a negative indication for the RBF Model as it would suggest that these Non-Ergodic[§] segments provided erroneous signals that were not detected or correctly used by C&A.

Given this research context, there are a number of technical elements that are needed to create an inferential-montage to test this expectation. We shall take up detailing these elements following.

Panel Size

The first order of investigative-testing, is to determine IF there are indeed ***such anomalous-segments*** in the early stages of the M[181]-Panels. The simplest modification to the M[181] is to trim the Panels in M[181]. The trimming criterion to create the M[Mod] was motived by the A&L(2016, p. 74) study that found that for parametrization of the RBF Model, the Panels should not be less than 13-Panel points. Thus, we have selected all the Panels of the M[181], the sample-size of which were greater than 15. Finally, we selected all the Panels > n=15 and then trimmed all of these Panels to n = 16. Note these Panels as: *M[Mod[165]];* they are noted in Appendix A. This will generate:

I. The shortest series that are *a priori* consistent with adequate forecasting, and

[§] These ***anomalous-segments*** are often referred to as Non-Ergodic segments. Simply, there is inferential evidence for a sufficiently-long Panel that *the segment under investigation* is likely to test as having a FPE[p-value] for its Null of equality *vis-à-vis* the other Panel-segments tested overall that is most likely to be rejected by a prudent analysist. Simply, a section of the full Panel-data-set does not fit the overall statistical profile of the full Panel. If this is the case, then this is referred to as a strong-test for questioning Ergodicity.

II. These trimmed or shortened series are most likely to have ***volatility issues*** that will test the acuity of the 99-Rules and their application by C&A.

Volatility Testing

The begged inferential test-question is: *Do these M[Mod[165]]-Panels have volatility profiles that test to be not similar to the Panels used by C&A?*

Volatility Context

C&A note: The percentage of Panels that have a Coefficient of Variation [CoV] > 0.2 is 21%. [*Table 1, p.1402*]. In C&A [Appendix A], p.1409, C&A note: "Coefficient of Variation. *The standard deviation divided by the mean for the trend adjusted data in the original units.*" The CoV is a measure of unitized variation; it is usually benchmarked by the Mean of the Panel, sometimes by the Median or, for C&A, by the Mean of the trend-adjusted data. The C&A measure thus assumes the computation of the trend. One presumes that they used their RBF-Model for this computation. For the M[Mod[165]]-dataset, we used the Two-parameter[Intercept & Slope] Linear OLS-Regression[OLS-R]. In this case, we found for the M[Mod[165]]-dataset the percentage of Panels where the CoV was > 0.20 was *39.4%*. The conservative non-directional False Positive Error [FPE] p-value] for this difference of 18.4% [39.4% - 21.0%] is: p <0.0005.

Implication

This FPE[p-value] is very strongly suggestive that the M[Mod[165]]-dataset likely had about double the number of series for which the CoV was > 0.2 than did the C&A Panel-set. This test result strongly suggests that the intuition of C&A is correct; short-series invite Non-Ergodic launching-profiles. This being the case, to compensate for detected early Panel volatility, C&A offer the following advice:

Input by the Analyst Based on Inspection of the Data

> *Irrelevant Early Data. An early portion of a series that is believed to have resulted from a substantially different process. For example, the start-up period for sales of a product would be eliminated once the pattern has stabilized. In truncation a series to eliminate such a period, avoid starting the truncated series with an extreme observation.* [***C&A [Appendix A p. 1409***]]
> *Observations Adjusted IF observations are judged to be irregular based upon domain knowledge, THEN adjust the observations prior to analysis to remove their short*-term effects. [***C&A [Appendix B p. 1409]***]

Forecasting Models Selected

To provide illustrations of these experiential inferential computations, we have selected two forecasting models. The first is the OLS-R Model; the other is its RAE-benchmark, the Random Walk Model [RWM]. The RWM uses, as the forecast for <u>all</u> horizons for M[Mod[165]], the Panel-Point, $x_{13,h}$ where h :{1, 2, - - -, 165}. The RWM was initially used in the M-Competition, therein it was called: The *Naïve 1*.

Caveats, Preliminaries, and Context for the Experiential Inference Vetting Protocol

For C&A's RAE- & APE-profiles [See Appendix C] exact inferential-testing is impossible as C&A only report their summary-ERROR measures—***Medians and Means***— of their RAE- & APE-profiles. Thus, the correct standard inferential Design of Experiments [DOE]-protocol of a two-samples-test for their FPE-Nulls: [Median]- or [Mean] ***is not possible***. In this case, we offer, following, our experiential inference-judgments.

Inferential Context

We have *a priori* information from C&A (1992) on the extensive and the intensive vetting protocols that were used to form the RBF Expert Modeling System. In addition, we have the results of their testing and additionally the testing of others who have used the RBF and its various versions. This is valuable conditioning-intel. Finally, we are using the RWM as a benchmark for the OLS-R. C&A also used these two models, in addition to two others. In this case, we used this *a priori* intel to form the testing Null[H_o] of our study as:

Null[H_o]: [The ***Value*** of the C&A[Central Tendency: Error Measure] ***IS Equal to or Greater than*** **[≥]** The [Central Tendency: Error Measure] derived from the Sampled Population: M[H&L[165]]

where: The Central Tendency is: The Arithmetic Mean[Error Measure] or The Geometric Mean[Error Measure] or The Median;[Error Measure], and the Error Measures are verily: The Relative Absolute Error [RAE] and Absolute Percentage Error [APE]. These error measures are presented in detail in Appendix C, finally, when C&A have used the Geometric Mean, the *ln*-transformation is needed (see Carvalho, 2009).

Discussion

In vetting the *Nature of the Intel* offered by C&A, we will use the Null[H_o] as the test-feature for our inferential conclusions. Thus, there needs to be two-stages for the testing as the Null[H_o] is bi-formatted. The first test addresses the ***IS Greater Than*** **[>]** ***condition***. As this is usually, the domain of the False Negative Error, that we are not electing to address, we will use a "conditioning" aspect for the first-stage of the Null[H_o] test. Simply, if The ***Value*** of the C&A[Central Tendency: Error Measure] ***IS Greater Than*** **[>]** The [Central Tendency: Error Measure] derived from the Sampled Population: M[H&L[165], then the FPE[p-value for the **[=]** condition] will be, by definition, >50%. This value will mean that the Null[H_o] is logically not rejected. If the **[>] condition** is not the case *for the set of realizations*, then the test will be the standard test against the realizations for the **[=]** ***condition***. This will be illustrated *anon* in *Section 6.3* when the Mean tests are discussed.

The first testing aspect broached is our inferential scoring for the relationship: ***The Medians*** reported by C&A and the Medians calculated using M[Mod[165]]. Then following, we will offer an scoring protocol for the inferential testing of ***The Means*** reported by C&A and those calculated using M[Mod[165]].

MEDIAN INFERENTIAL VETTING PROTOCOLS

Overview

For the ***Median Error tests***, we will use the Interquartile Range [IQR] endpoints as the following conceptual anchors of the vetting measure:

IQR[Error Measure] ≡ [PERCENTILE.INC(H&L,0.25)] through PERCENTILE.INC(H&L,0.75]

The IQR is a standard range of descriptive importance; for example, it is used in the *Outlier Box-Plot*™ to form the "Whiskers" for screening Outliers in <u>all</u> the *SAS*™ [*JMP*™] versions. *JMP*v.13 offers:

> *The box-plot has lines that extend from each end, sometimes called whiskers. The whiskers extend from the ends of the box to the outermost data point that falls within the distances computed as follows:*
>
> *25^{th} Percentile - 1.5*(IQR)*
> *75^{th} Percentile + 1.5*(IQR)*

Thus, we propose the following Median-Screening Inferential Protocol [MSIP[Error Measures]] that will be used to score the Median performance profiles of the RBF [C&A]

benchmarked by those of the OLS-R Models [H&L]. The [MSIP[Error Measures]] has five-Error Median scoring-zones ***that are generated from the H&L-dataset***; two on the Left Hand Side [LHS] and two on the Right Hand Side [RHS] of the IQR. For the LHS, we have:

←[Lower than the Whisker-Zone]**B**[←Whisker-Zone →]**C**[←IQR-Zone→] R1

where; Points {B & C} are the frontier-points ***between*** the three-Error- zones, Point C is at the 25th Percentile, Point B is at the LHS Whisker-edge→ [*25th Percentile - 1.5*(IQR)], and any point < Point-B falls outside the Whisker-Zone on the LHS that usually extends to about zero*.

Note: For the [MSIP[Error Measures]] the LHS & RHS-zones are symmetrically oriented. In this sense, given the *a priori* intel re: C&A vetting-protocols and their reported results, we offer the MSIP as a directional inferential screen—to wit: One-tailed. ***Summary***: The entire LHS & RHS span the symmetric probability space on either side of the Median Error. We used this space to form our scoring codex for the Medians. In this sense, our Median-scoring in R1 is one-tailed in the p-value sense; also, we tried to form the inferential Median scoring in R1 to coordinate with the p-value scoring that we used for the Means and also for the 95%Confidence Intervals [95%CI] to be discussed following.

Scoring Codex

We have coded the following vetting indications for the [MSIP[Error Measures]] as:

IF the C&A [Median[Error]] is > Point [C] THEN, it is In →H&A [*IQR*[*ErrorRange*]]: Vetting Indication: Equal : [The C&A [Median[Error]] & The H&L [Median[Error]]] are *not likely to be different*; *Summary Indication* The Null[H_o] would not test to have a p-value that would suggest rejecting the Null[H_o]. *Coding: [RBF [Median*[Error]*]] = OLS-R [Median*[Error]*]]; END : ELSE IF*

the C&A [Median[Error] is > Point-B THEN, it is In → H&A [*LHS:Whisker Zone*]: Vetting Indication: Likely : [The C&A [Median[Error]] is likely < The [H&L [Median[Error]]]. *Summary Indication* The RBF Model *likely outperforms* The OLS-R Model relative to their Median[Error] Profiles; in this case, the Null[H_o] would test to have a p-value that would <u>likely</u> suggest rejection. *Coding:* [*RBF* [*Median*[Error]] < *OLS-R* [*Median*[Error]]], *END : ELSE IF*

the C&A[Median[Error]] is < Point-B THEN, it is In → H&A [*Lower than the LHS Whisker-Zone*]: Vetting Indication: Clearly : [The C&A [Median[Error]] is clearly << H&L [Median[Error]]] *Summary Indication* The RBF Model *clearly outperforms* The OLS-R Model relative to their Median[Error] Profiles; in this case, the Null[H_o] would test to have a p-value that would <u>clearly</u> suggest rejection. *Coding:* [*RBF* [*Median*[Error]] << *OLS-R* [*Median*[Error]]]. *END*

Conservative Directional Option

Given the nature of this Median vetting-context, and the fact that the Error-Points {B & C} are critical decision-points, and we have not included their point-values in the MSIP-Median[Error]-codex, we offer the following Zones of *Inexactitude* or *Incertitude* [ZI] for the Points {B & C} so as to better calibrate the vetting-intel for the MSIP[Error]. In this calibration, we have elected to form <u>conservative</u> benchmarking screens—meaning the scoring elections favor *failing to reject* Null[H_o]. After discussions with our colleagues, we have decided to construct a zone around these frontiers-points of a *distance of 15% or ±7.5%*. If the C&A-Median[Error] is IN a particular ZI, we will designate the vetting evaluation with the symbol "≈" thus indicating incertitude in scoring the profile relationships between: The RBF & The OLS-R Models; in this case, we will opt for the Scoring-Zone that is closest to the H&L [Median]—the more directionally <u>conservative</u> election. For example, assume that the MSIP[Error] frontier-point-B was 57.6%. Thus, the ZI would be:

[57.6% –[57.6%*[7.5%]]/2 : 57.6% +[57.6%*[7.5%]]/2]

The B-Zone of Incertitude: [ZI] is: [*53.28% : 61.92%*]

Thus, a small portion of the scoring mass of the B[ZI] extends into the [Lower than the Whisker-Zone]; specifically, 7.5%[4.32%/57.6%]. In this context, if the C&A [Median] were to have been 53.37%, it would have fallen into the [Lower than the Whisker-Zone] which would have constituted a stronger judgmental argument for rejecting the Null[H_o] than if the C&A [Median] were to have been in the [Whisker-Zone]. Thus, conservatively as the C&A[Median [53.37%]] falls in the B[ZI] it will be scored as IN the B-Zone or Whisker-Zone; this designation will be noted as: [RBF [Median] B[<≈] OLS-R [Median]]. This indicates the conservative directional election that locates the C&A-Median in the Whisker-Zone rather than Lower than the LHS of the Whisker-Zone. *Summary*: IF the C&A-Median Parameter were to have been IN a [MSIP[Error]-[ZI]], the analyst is encouraged to make the conservative election as detailed following:

I. Assume that the C&A [Median] is IN the B[ZI] then : [RBF:Median [B] <≈ OLS-R:Median] Indication: Conservatively, the RBF Model *likely <u>will</u> outperform* The OLS-R Model relative to their Median Error Profiles; in this case, these Medians would test to have a p-value that likely <u>will</u> suggest rejecting the [H_o]NULL. Note this indication is slightly less in favor of rejecting the Null than if the C&A[Median] were to have been 100% in the [Lower than the Whisker-Zone] *but* more in favor of rejecting the Null than if the C&A[Median] were to have been 100% in the [Whisker-Zone]-Zone.
II. Assume that the C&A[Median] is IN the C[ZI] then : [RBF:Median [C] ≈ OLS-R:Median]. Indication: Conservatively, the RBF Model *may not likely outperform* the OLS-R Model relative to their Median Profiles; in this case, these Medians, would test to have a p-value that <u>may</u> <u>likely</u> suggest not rejecting the [H_o]NULL. Note this indication is slightly more in favor of rejecting the Null that if the C&A[Median] were to have been 100% in the C-Zone *but* less in favor of rejecting the Null than if the C&A[Median] were to have been 100% in the [Whisker-Zone].
III. Any value Lower than the LHS of the B[ZI]-Zone indicates that the [RBF:Median << OLS-R:Median]. Indication: The RBF Model *clearly outperforms* the OLS-R Model relative to their Median Profiles; in this case, these Medians, would <u>certainly</u> test to have a p-value that would strongly suggest rejecting the [H_o]NULL.

MEAN INFERENTIAL VETTING PROTOCOLS

Overview

The ***Mean vetting tests*** are rather complicated as C&A report both Arithmetic[AM] and Geometric[GM] Means. As these Means are reported by C&A and using the H&L-dataset, we can compute: (i) the M[Mod[165[[AM]] & the M[Mod[165[[GM]], (ii) related directional <u>judgmental</u> p-values, as well as, (iii) the related H&L [95%CIs] for the sampled Means. This will provide two vetting-measures. For the Means, we can compute a p-value for the <u>judgmental</u> FPE: thus, *possibly* rejecting the [H_o]NULL suggesting that: [C&A's Reported Means are less than the Population Means estimated from the H&L-dataset]. In addition, we can use the 95%CIs created from H&L-dataset to vet the reported Means reported by C&A. Of course, these will be judgmental-configurations, logical we hope to argue, but not inferentially valid in a traditional generalizable-sense over all testing domains. This, of course, is the case, as noted above for the Medians, due to the fact the C&A only report the summary indications. ***First*** consider the judgmental inferential p-value context for the AM & GM. Then, we will treat 95%CIs.

The C&A Means: [Geometric [GM] & Arithmetic [AM]]

For the Mean analyses for the RAE and the APE, C&A used both the GM & the AM for reporting purposes. The reason for using the GM is not detailed in C&A; however, the GM is often used in reporting economic Panel-data such as the RAE & the APE so is not uncommon. For vetting the C&A [Means]: Arithmetic [AM] & Geometric [GM], we used the AM & the GM for the M[Mod[165]]-Panels. This Mean[Error]-Screening Inferential Protocol [[G&A]MESIP] protocol encoded in Table 3 is experientially calibrated by the authors as follows:

Table 3: Two Stage Triage Taxonomy: G&AMESIP Experiential directional p-values

Locational-p-value	**[≈>0 : < 1.0*E-10*]**	IZ	**[≥1.0E-10 : <1.0*E-7*]**	IZ	**[≥ 1.0*E-7* : <1.0*E-4*]**	IZ	**[≥ 1.0*E-4* : <0.1]**	IZ	≥ 0.1	*
Inferential p-value	≤0.0001	→	0.001	→	0.01	→	0.1	→	>0.15	

*We truncated Table 3 at: [≥0.1 [>0.15]] as this was the largest directional p-value that we found for vetting the RBF Model. In a spanning set for the Left Hand Side [LHS], this will actually go to a directional p-value of 50% where the C&A RBF-parameter and that of the OLS-R are the same. Also, ***ALERT*** We are using the standard USA-Scientific Notation: 1.0E-7. In the global context, some platforms use: 0,1E-7 which translates in Excel[USA] to: 1.0E-8. Also, in Excel-script 1.0E-7 ≡ 1.00E-07.

Logistics of Table 3 Three Stages are needed for extracting the inferential directional p-value so as to evaluate the [H_o]NULL.

Stage I: *Pre-analysis Logistic* The analyst first checks IF the C&A[Mean] is **[>]** the M[H&L[165]Mean]. If so, then the p-value is > 50% and thus, the [H_o]NULL is not rejected; this then is the analytical-termination point.

Stage 2: *Locational Logistic* If, on the other hand, the C&A[Mean] is **[≤]** the M[H&L[165]Mean], then the analyst computes the standard ***one-tailed p-value*** for the **[=]** test: [The ***Value*** of the C&A[Mean] v. The [***Mean***] derived from the Sampled Population: M[H&L[165]Mean]. This is called the ***Locational p-value***.

Stage 3: *Inferential Logistic* Then the analyst determines for Row [1] of Table 3, the Column where this ***Locational p-value*** is positioned. Finally, the ***Inferential p-value*** to be used in evaluating the [H_o]NULL is found in Row[2] of the Column where the ***Locational p-value*** is positioned.

Computational Illustration: [Table 4, Col[8]]

For example,

Stage 1, The C&A[RBF]:Mean for the APE was: 0.063 and The H&L[OLS-R]:Mean was 0.125. As 0.063 is < than 0.125, we can move to the **[=]** test condition of the Null[H_o].

Stage 2 The following script will create the ***Locational p-value***:

t_{cal} = [M[Mod[Mean]] – C&A[Mean]] / [STDEVA[M[Mod]] / $\sqrt{n}$]

Locational p-value = (T.DIST.2T(t_{cal},(n–1)))/2

In this case, given the parameters: C&A[0.063] & H&L[0.125] & the following computations are:

t_{cal} = [M[Mod[0.125]] – C&A[0.063]] / [STDEVA[Mod[0.18761] / $\sqrt{165}$] = 4.25

Locational p-value[Row1 Table 3] = (T.DIST.2T(4.25,(165–1)))/2 = 1.79E-5

Searching the first row in Table 3 for wherein 1.79E-5 [0.0000179]is located, we arrive at: 1.79E-5 ⊂ [≥ 1.0E-7 : < 1.0E-4].

Stage 3 Thus, the desired section of Table 3 is: [≥ 1.0E-7 : < 1.0E-4], i.e.,—[*Row 1, Column 4*]. In this case, the ***Inferential p-value***: [Table 3[*Row 2 Column 4*]]—is ***0.01.***

Inference Indication Finding two datasets with the following profile: APE[Dataset H&L[0.126] & Dataset C&A[0.063]] suggests that one may reject the FPE:Null:[H_o]: hypothesis that [C&A[APE]] is >= H&L[APE]] when the sampling realizations are: H&L[0.126] & C&A[0.063]. This is the meaning of the ***Inferential p-value***, 0.01, that is encoded in Table 3[Row[Inferential p-value]].

The p-value Locational Incertitude-Zone

As was the case for the Median [ZI], we will, for consistency, form a ***Locational p-value [ZI] protocol***. The shaded frontier-markers in Table 3 are these ***Locational p-value***-ZI zones. If a ***Locational p-value*** were to be in a Shaded-ZI, opting for conservativism, the operant p-value will be the higher contiguous experiential ***Inferential p-value***. For illustration, we will create the ***Locational p-value***: [ZI] for [1.0E-7:] that is at the ***Locational*** junction of the ***Inferential p-values*** of [0.001 ∩ 0.01]. In this case, we are also using a symmetric-zone of 15% of the frontier-junction points. Thus, after the computations, one arrives at the following:

Locational p-value[ZI]: [1.0*E-7*]→[1.0*E-7*) ± [(1.0*E-7*) × 15%/2]] = [9.250E-08 : 1.075E-07]

Assume the Mean reported by C&A created a *Locational p-value* of: 9.27E-8; this is IN the above H&L ***Locational p-value*** [ZI] that bridges the two ***Inferential p-value*** zones: [[0.001]-Zone[7.5%] & [0.01]-Zone[92.5%]. In this case, the C&A [Mean] p-value of: 9.27E-08 is conservatively recorded as 0.01 rather than 0.001. Conservatively, this suggests that the Null [*Ho*] of the vetting-test is rejected at an FPE of 0.01 even though some of the probability mass was in the 0.001-Zone but just barely. This is useful judgmental-intel. Consider now the GM p-value judgmental protocol.

Vetting the Geometric Means [GMs] for the RAEs reported by C&A

Following is the computational locational protocol for the GM; however, it is basically the same as the ***Locational p-value*** computation for the Arithmetic Means [AMs] if one eliminates the *ln*-transformation from the following protocol. We note this as the [GMError p-value Protocol]: [GMEp-vP].

I. Following Carvalho (2009), compute the Natural Log [*ln*] for all the RAEs in the H&L[165]-dataset,
II. Using this *ln*-transformed dataset, compute: the AVERAGE, the Standard Error: [STDEVA / $\sqrt{165}$], and the *ln*(C&A:GM[RAE]),
III. Form the t_{df} as: [ABS[AVERAGE – *ln*(C&A:GM[RAE]]] / Standard Error,
IV. Compute the ***Locational p-value*** as: =(T.DIST.2T(t_{df},164))/2,
V. Finally, using Table 3 find the ***Locational p-value***, record the corresponding ***Inferential p-value***, evaluate the Null[Ho], and discuss its implication.

As noted above, in addition to the above p-values, we are able to compute the 95% CIs for the AM & the GM. This gives the opportunity to form enhanced *confirmatory* vetting measures by using the 95%CIs in conjunction with the ***Inferential p-values*** of the Means.

Judgmental 95%CI: Suggested Computations

In addition to these GM or AM p-value indications, a second vetting measure that we will use is the 95%CIs of the GM or the AM of the H&L-dataset. They are provided so that one could judgmentally compare the C&A [Mean] to these correct M[Mod[165[95%CIs]]] as a companion vetting-indication of the ***Inferential p-values***. Additionally, we have created a scaling to modify the judgmental-inference using a 95%Confidence Interval [IZ]. ***Alert***: There are TWO values in the [95% CI]. For example:

The LHS of the 95% CI≡ [Average – Precision]

The RHS of the 95% CI≡ [Average + Precision]

For the C&A vetting, we are ***only interested*** in the LHS of the [95% CI]. Additional, for the LHS we will create a 95% CI [ZI]. The reason for this is that if the C&A[Mean] is > the H&L[Mean] then there is no further testing as the p-value will be > 50% and the testing is terminated. Also, recall that if C&A[Mean[Error]] is = to the H&L[Mean[Error]], then the [H_o]NULL will be tested. This then creates a conditional limitation on the confirmatory testing using the 95%CIs to the following interval: C&A[Mean[Error]] < the H&L[Mean[Error]]. With this condition, the inferential codex for the LHS of the [95% CI] will be:

The [Lower ZI]-Component of the LHS of the 95% CI[ZI]≡ [Average – Precision] × δ_{Mln}

The [Upper ZI]-Component of the LHS of the 95% CI[ZI]≡ [Average – Precision]× δ_{Max}

where: δ_{Mln} is (1 – 15%/2) and δ_{Max} is (1 + 15%/2). This is consistent with the interval of 15% of the Incertitude-Zone [ZI] that we are using for the Means and the Median.

Specifically, around the end-points of the M[Mod[165[95%CIs]]], we have again created a symmetric 95%CI-Incertitude-Zone of 15% of the values of the LHS-end points-values of the [95% CI [ZI]]. The location of the C&A [Mean] relative to these 95%CI [ZI], will offer a conditioning indication for the ***Inferential p-values***. Specifically, there are three judgmental conditions:

I. If the C&A[Mean[Error]] is < the [Lower ZI]-Component of the LHS-point of the 95% CI [ZI], then this is a strong indication that the RBF outperforms the OLS-R; this is noted as: RBF<<OLS-R,
II. If the C&A Mean is IN [Inclusive] the LHS of the Lower 95% CI [ZI], then, conservatively, this is a possible indication that the RBF[Mean] may outperform the OLS-R [Mean]; this is noted as: RBF≈OLS-R[95%CI[ZI]], and
III. If the C&A Mean is > The [Upper ZI]-Component of the LHS of the 95% CI [ZI], then this is a possible indication that the OLS-R and the RBF are not sufficiently different; this is notes as: RBF=OLS-R[95%CI[ZI]].

This information should be used in conjunction with the ***Inferential p-value*** information from Table 3 to arrive an overall judgmental assessment of the relationship of the RBF [Mean] *vis-à-vis* the OLS-R [Mean].

The Arithmetic Mean

C&A also report the Arithmetic Mean for the APE-measure. In this case, as mentioned above, all of the vetting computations discussed above for the Geometric Mean to create the vetting-intel for the ***Inferential p-values*** and the 95%CI are the same EXCEPTING for taking the *ln*-transformation and the related re-castings.

THE VETTING OF C&A'S PERFORMANCE PROFILES

Overview

Given the above experiential inferential computational context, we will now examine the Means & Medians of the Actual Data created by M[Mod[165]] for the trimmed datasets *vis-à-vis* the summary profile presented by C&A in their Table 3 [p. 1405]. Recall, this is a vetting test of the effect of short series, that we tested and found to have increased volatility compared to the volatility reported by C&A as well as the Quality of the 99-Rules and Wisdom of their application. These profiles are presented in Table 4 following:

Table 4: Error Profiles C&A v. M[Mod[165]]

	Median RAE		Geometric Means RAE		Median APE		Arithmetic Mean APE	
HHs	$HBx_{k=14}$	$HBx_{k=16}$	$HBx_{k=14}$	$HBx_{k=16}$	$HBx_{k=14}$	$HBx_{k=16}$	$HBx_{k=14}$	$HBx_{k=16}$
EW[C&A]	0.70	*N/A*	0.69	*N/A*	4.3%	*N/A*	5.6%	*N/A*
RBF[C&A]	**0.63**	***0.57***	**0.67**	***0.59***	**3.2%**	***7.6%***	**6.3%**	***13.2%***
M[Mod[165]]	**1.06**	**0.79**	**1.01**	**0.82**	**6.7%**	**13.1%**	**12.5%**	**21.1%**
Inf-p-value	N/A	N/A	0.01	0.01	N/A	N/A	0.01	**0.001**
95%CIsZI	N/A	N/A	[0.79:0.92]	[0.65:0.76]	N/A	N/A	[8.9%:10.3%]	[15.9%:18.5%]
Point[C]ZI	[0.56:0.66]	[0.51:0.59]	N/A	N/A	[3.22%:3.8%]	[5.6%:6.5%]	N/A	N/A
Inference	**RBF≈OLS-R**	**RBF≈OLS-R**	**RBF<<OLS-R**	**RBF<<OLS-R**	**RBF<OLS-R**	**RBF= OLS-R**	**RBF<<OLS-R**	**RBF<<OLS-R**

Discussion: The Codex of Table 4

Overall, the evaluation-foci are: For the Geometric & Arithmetic Means, the vetting-intel is created by the p-values from Table 3 and the related scored 95%CIs, while for the Medians the vetting is created by the IQR-measure; R1. In addition, the key vetting elements are: The ***RAE*** is the Relative Absolute Error, the ***APE*** is the Absolute Percentage Error, the HB-Row presents the two-holdbacks tested: The First Holdback: $HBx_{k=14}$, and the Third-Holdback: $HBx_{k=16}$. ***Point of Information*** The actual holdbacks of C&A are variable with respect to the number of actual Panel points as they used the full-Panels from the M-Competition; whereas, M[Mod[165] are trimmed series from the M-Competition Panels. The ***Equal Weights Model*** [***EW]*** [See[C&A fn.4 [**p. 1404**]]] is another forecasting protocol used by C&A to benchmark their RBF Model. *The EW-profiles are shaded and are presented only as context—no vetting tests are made for their EW-values.* In the ***Inf-p-value***-Row, we report the ***Inferential p-values*** taken from Table 3 for the ***Geometric Means*** [GM] & the ***Arithmetic Means*** [AM]. In the ***95%CIsZI***-Row, we report the 95%Confidence Interval ZI. In the ***PointCZI-Row***, we report the Median ZI-anchored at the 25th Percentile-Point for the Medians of the M[Mod[165]]. Finally, in the ***Inference-Row*** we present ***our assessment*** of the RBF *vis-à-vis* the OLS-R Models. ***HoldBack Projections*** Finally, C&A did not report any information regarding the performance profile for the third holdback. H&L used only three holdbacks to produce their performance profiles. To resolve this overlap-disconnect, we created estimated projections for the C&A data at: $HBx_{k=16}$.as follows.

Estimations for the [$\mathbf{HBx_{k=16}}$]

This illustration only considers the RAEs reported by C&A; their **First** HB[RAE]-Value was: 63% and their ***Sixth*** was: 48%. We interpolated that C&A's ***Third*** Median RAE for [$HBx_{k=16}$] likely would have been:

63% – [2 × [(63% – 48%)/5]] = 57% or 0.57 or by symmetry

48% + [2 × [(63% – 48%)/5]] = 57% or 0.57.

Illustrations of the Selected Computations

At this point, as reinforcements to the information presented in Table 4, we detail our analyses for the: RAE [Medians], RAE [Geometric Mean] and APE [Arithmetic Mean] for the First Holdback: $HBx_{k=14}$,

H&L: Analysis RAE[Medians] First Holdback: $HBx_{k=14}$ Column 2 Shaded in Table 4

The H&L Median Profile was: IQR = [25th Percentile = 0.609583, Median = 1.062562, 75th Percentile = 1.559441]; thus, the IQR was: 0.95 [1.56 – 0.61]. The C&A Median reported was: 0.63. The LHS ZI in this case is Point C as the C&A[RAE[Median]] is 0.63 and is closest

to the LHS-edge Point C of the IQR of 0.609583. In this case, the ZI [Point-C] in Table 4 is reported as: [0.56: 0.66]→ [0.609583×(1-0.075) = 0.563864 : 0.609583×(1+0.075)=0.655302]. The C&A [Median[0.63]] is IN this C[ZI]; this indicates to us that there is suggestive evidence *not to assume* that the Median of H&L and that of C&A may test to be inferentially different at a p-value that most analysts would rationalize rejecting the NULL. *Summary Indication* To our conservative experiential judgement, it seems that inferential prudence would suggest that, with respect to the Median [RAE], that the Median-profile of C&A's dataset and that of H&L may not be inferentially sufficiently different to support the assertion that the RBF outperforms the OLS-R on the RAE-measure. *Inference-Codex:* [RBF [RAE:Median]≈ OLS-R [RAE:Median]].

H&L: Analysis RAE[GMeans] First Holdback: $HBx_{k=14}$ Column 4 Shaded in Table 4

The first test addresses the **[>]** case. C&A report the RAE[GM[0.67]] and H&L report the RAE[GM[1.01]. As 0.67 is **[<]** 1.01, we move to the p-value [=] test. In this case, using the Carvalho (2016) script for the Geometric Mean, we will use the Natural log-transformation to create the needed intel. Specifically, the *ln*[0.67] = –0.40047757 is the required transformation. The GM[M[Mod[165][RAE]]] is 1.013606515; thus, taking the *ln*s of the GM-dataset, we have the *ln*-transformed Average as: 0.01351478 [ln[1.013606515]] and the related StError of the H&L dataset is: 0.08755861. The t_{calc} = [[ABS[0.01351478 – [–0.40047757]]/ 0.0875586]] = 4.7282. In this case, the ***Locational p-value*** is: T.DIST.2T(4.7282,164)/2= 0.00000243 or 2.43E-06. Using the Inferential p-value codex of Table 3, we have the inferential indication of **0.01** as 2.43E-06 is IN [≥1.0E-7 : <1.0E-4]. Further, 2.43E-06 is not in any ZI.. Thus, the next step is to compute the 95%CIs. The computation of the 95%CIs also requires the *ln*-transformation of the H&L-REA dataset. Thus, the following computations are made:

$[t_{df=165-1}]$ =T.INV.2T(5%,164)=1.97453458, Precision[95%CI]= [1.97453458 × 0.08755861] = 0.172886024.

Finally,

Lower Limit 95%CI = EXP[0.01351478 – 0.172886024] → *0.852680*

Upper Limit 95%CI = EXP[0.01351478 + 0.172886024] → *1.204905*.

Finally, the 95%CI[ZI] for the LHS of the Lower RAE[95%CIs] is: [0.788729: 0.916631] and the C&A[GM] reported is: 0.67 and is *not in* the 95%CI[ZI] as 0.67 < 0.79.

Summary Indication In this case, as the p-value is 0.01 and the C&A[GM[0.67]] is lower than the LHS of the 95% ZI point of 0.79, this seems to us given these two indications that the RBF Model most *likely* outperforms the OLS-R with respect to the Geometric Mean. *Inference-Codex: RBF* [*RAE:GM*] << *OLS-R* [*RAE:GM*]. Thus, the RBF seems demonstrably different than the OLS-R re: Geometric Mean[RAE].

H&L: Analysis APE[Arithmetic Means] First Holdback: $HBx_{k=14}$ Column 8 Shaded in Table 4

For this measure, it is not required to *ln*-transform the H&L-dataset. Thus, the following computations are made:

The first test addresses the **[>]** case. C&A report the APE[AM[0.063] and H&L report the APE[AM[0.125]. As 0.063 is **[<]** 0.125, we move to the p-value [=] test. The t_{calc} = [ABS[0.12481776 – 0.063]/0.01460603] = 4.2324. in this case, the p-value is =T.DIST.2T(4.2324,164)/2= 0.0000192 or 1.92E-05. The Inferential p-value indication is **0.01** as it is IN [1.0E-7 : 1.0E-4]. Further, 1.92E-05 is not in any ZI. ***Interesting Computational Issue*** *Note that these numbers are slightly different than those used in the illustrative example in 6.3 Computational Illustration. This is due to the number of decimal-places that we used in these computations. Unfortunately, the only way to control for this is to have a Rule of how many decimal-places are required.*

Thus, the next step is to compute the 95%CIs. The computation of the 95%CIs does not require the *ln*-transformation of the H&L-dataset. Thus, the following computations are made: Next is the 95%CIs. The computations are:

AVERAGE=0.12481776, STDEVA=0.18761785, $[t_{df=165-1}]$

=T.INV.2T(5%,164)=1.97453458, StDError = 0.18761785/$\sqrt{165}$ = **0.01460603**,

Precision[95%CI]= [1.97453458×0.01460603] = 0.02884011

Lower Limit 95%CI = [0.12481776 – 0.02884011] → ***0.095978***

Upper Limit 95%CI = [0.12481776 + .02884011] → ***0.153658***.

The APE reported by C&A is 6.3% and is **not in** the LHS: ZI of [8.9% : 10.3%] and thus this is confirmatory evidence of the p-value of **0.01**. ***Summary Indication*** In this case, p-value **of 0.01 and** the fact that the C&A [Mean] of 6.3% **is not** in the 95% CI ZI on the LHS, it seems to us that the RBF Model ***clearly*** outperforms the OLS-R with respect to the APE. ***Inference-Codex: RBF* [*APE:Mean*] << *OLS-R* [*APE:Mean*].** Thus, RBF seems demonstrably different than the OLS-R re: The Arithmetic Mean[APE[].

Summary of Vetting Results: Table 4

The vetting results presented in Table 4 are very clear. The 165 Short or Trimmed Panels used by H&L to provide forecasts using the OLS-R Model, the forecasts of which were benchmarked using C&A's RAE & APE measures, indicated that there is likely no question that the RBF Model used by the C&A, was ***not outperformed by the OLS-R Model***. ***Rationale***: To rationalize this assertion, we will Profile, *en bref*, the eight parameters of Table 4 three, of which, were detailed above.

RAE[Median] Col[2]

Inference-Codex:* [*RBF*[*RAE:Median*] ≈ [*OLS-R*[*RAE:Median*]]] *suggesting that the RBF Model is likely not demonstrably or sufficiently different than the OLS-R re: RAE using the Medians as the measure. The Key index was the fact that the Median of the RBF[0.63] was IN the OLS-R Median:Point[C]ZI. Specifically, **C&A[RBF[0.63]] > LHS[ZI]: H&L[Point C[0.56:0.66]].**

RAE[Median] Col[3]

Inference-Codex:* [*RBF*[*RAE:Median*] ≈ [*OLS-R*[*RAE:Median*]]] *suggesting that the RBF Model is likely not demonstrably or sufficiently different than the OLS-R re: RAE using the Medians as the measure. The Key index was the fact that the Median of the RBF[0.57] was IN the OLS-R Median:Point[C]ZI. Specifically, C&A[**RBF[0.57] > LHS[ZI]: H&L[Point C[0.51:0.59]].**

RAE[GM] Col[4]

Inference-Codex:* [*RBF*[*RAE:GM*] << [*OLS-R* [*RAE:GM*]]*: The RBF seems demonstrably different than the OLS-R re: Geometric Mean. The Key indices were: (i) the GM[p-value] was 0.01, and (ii) the 95%CI LHS[ZI] from the H&L-dataset was [0.79:0.92] while the C&A[GM] was 0.67. Thus, the C&A[GM] of 0.67 is not in the 95%CI ZI on the Lower LHS. These two relationships were consistent indications that the RBF demonstrably outperformed the OLS-R using the GM as the measure.

RAE[GM] Col[5]

Inference-Codex:* [*RBF*[*RAE:GM*] << [*OLS-R* [*RAE:GM*]]*: The RBF seems demonstrably different than the OLS-R re: Geometric Mean. The Key indices were: (i) the GM[p-value] was 0.01, and (ii) the 95%CI LHS[ZI] from the H&L-dataset was [0.65:0.76] while the C&A[GM] was 0.59. Thus, the C&A[GM] of 0.59 is not in the 95%CI ZI on the LHS. These two relationships were consistent indications that the RBF demonstrably outperformed the OLS-R using the GM as the measure.

APE[Median] Col[6]

Inference-Codex:* [*RBF*[*APE:Median*] < [*OLS-R*[*APE:Median*]]] *suggesting that the RBF Model may be marginally or suggestive different than the OLS-R re: APE using the Medians as the measure. The Key index was the fact that the APE-Median of the RBF was just barely *not In* the LHS of Point[C}ZI ***on the Lower LHS***. Specifically, **RBF[3.20%]** < LHS:ZI[**3.22%: 3.8%].**

APE[Median] Col[7]

Inference-Codex:* [*RBF*[*APE:Median*] = [*OLS-R*[*APE:Median*]]] *suggesting that the RBF Model is not likely to have out-performed the OLS-R re: APE using the Medians as the measure. The Key index was the fact that the APE-Median of the RBF was outside the RHS of the of the Point-C ZI. Specifically, **RBF[7.6%]** >>**RHS:Point C: ZI[5.6%:6.5%]** and so the RBF Median is relatively closer to the Median of the OLS-R.

APE[AM] Col[8]

Inference-Codex:* [*RBF*[*APE:AM*] << [*OLS-R*[*APE:AM*]]*: The RBF seems demonstrably different than the OLS-R re: Arithmetic Mean. The Key indices were: (i) the AM[p-value] was 0.01, and (ii) the 95%CI LHS[ZI] from the M[Mod[165]-dataset was [8.9%:10.3%] while the C&A[AM] was 6.3% and so was not IN the 95%CI ZI on the Lower LHS. These two relationships were consistent indications that the RBF demonstrably outperformed the OLS-R using the AM as the measure.

APE[AM] Col[9]

Inference-Codex:* [*RBF*[*APE:AM*] << [*OLS-R*[*APE:AM*][]*: The RBF seems demonstrably different than the OLS-R re: Arithmetic Mean. The Key indices were: (i) the AM[p-value] was 0.001, and (ii) the 95%CI LHS[ZI] from the M[Mod[165]-dataset was [15.9%:18.5%] while the C&A[AM] was 13.2% and so was not IN the 95%CI ZI on the Lower LHS. These two relationships were consistent indications that the RBF demonstrably outperformed the OLS-R using the AM as the measure.

OVERALL TAKE-AWAYS

Referencing the following two questions:

I. *What seems to be the inferential likelihood that that the initial segments of the M[181], as duly profiled by the M[Mod[165]], had a different CoV-volatility profile than the dataset used by C&A?*
II. *If it seems to be the case that the initial segments of the M[181] test to have higher overall CoV re: the dataset used by C&A, is there evidence that this was not taken into account by the Rules of the RBF system as applied by C&A?*

For Question 1, indeed C&A's intuition was on target. Early segments of the M[181]-dataset exhibited about ***twice the number*** of series with CoV-profiles that were >0.2 than were reported by C&A for those that they used from the full-M[181]-dataset. ***Implication*** With such demonstrated early segment volatility, it is not unexpected that the C&A-dataset would likely pose orientation and analytic challenges and issues in the forecasting domain some of which could result in:

> ***Relative Uncertainly, Level-shifts, or Trend-repositioning that could (i) erroneously skew the projection-orientation, or (ii) modify the "true Ergodic nature of the variation of the full-Panel***.

Thus, it is wise to be attentive to "the early stages of" time series Panels because, ***as expected, they have tested to have a higher propensity for volatility when compared to CoV of the full-Panel***. Thus, there are real analytical issues that these early segment sections could create erroneous "launching-pads" that would compromise forecasting acuity. This leads to the next question.

As for question 2, *given that there is evidence that the early segments of the Panels could have posed issues in forming effective forecasts, the silver-bullet against these Non-Ergodic or troublesome early segments of panels, is of course the 99-Rules of the RBF Model* ***and*** *their careful application in forming the parameterization of the RBF Model. Given the vetting results presented in Table 4, there is convincing evidence that the RBF Model, parameterized by C&A, provided effective developmental guidance. Simply, the OLS-R Model, that one would presume would be sensitive to the volatility of these early segments,* ***did not outperform*** *the RBF Model as presented by the C&A[RAE]- & C&A[APE]-profiles.* ***Rationale*** In summary, the vetting evidence, a blend of experiential informed judgmental vetting conjecture, is:

I. There seems to be different judgmental inferential performance profiles depending on the parameters tested. The RBF [Medians] are, in the overview, not noticeably different as between the RBF & the OLS-R. However, for the Means: Arithmetic and Geometric, there seems to be reason to argue for the RBF outperforming the OLS-R. Perhaps this is to be expected as three of the four component models used to combine the forecasts in the RBF Model: The OLS-R, The ARIMA(0, 2, 2)/Holt & Brown(1959) have as drivers OLS-parameter-scaling as their modality. However, the performance profile is nonetheless, that the OLS-R ***did not likely outperform*** the RBF Model given the C&A's RAE & APE-profiles.
II. Ignoring our judgmental inference dimension, all of the eight-central tendency Median and Mean information in Table 4 reported by C&A [$HBx_{k=14}$] or estimated for C&A [$HBx_{k=16}$] are indicative of the dominance of the RBF re: those reported by H&L for the OLS-R,
III. Referencing Table 4, there are no instances where the OLS-R Model seemed inferentially more effective than the RBF Model. The strict Bernoulli-Coin-Flip-Chance for this result, conservatively only using the actual values reported for [$HBx_{k=14}$], is: p-value <0.063 [$50\%^4$]. This is suggestively confirmatory with the above indications.
IV. Overall, the above then speaks to the quality of the 99-Rules. Tedious, certainly, but effective in teasing-out and correcting for volatility [CoV]-issues in the early sections of the Panels used in creating the forecasts produced by the RBF Model.

This, we offer, is solid vetting evidence that even in the presence of a clear increase of CoV in the early segments of series that the RBF Model, and its various versions, deserves its reputation of an effective forecasting model.

OUTLOOK

One of the interesting *lacuna* for the *RBF Model* is that there are NO [1-FPE] Confidence Intervals for the RBF-Model versions. The reason for this is simple: There are "currently" three Basic models: The Random Walk, The OLS-R, and the Holt Models. These are combined judgmentally depending upon the experiential-base of those creating the forecasts considering the particular RBF-Rule base that is used. Thus, there is NO statistically valid protocol that can determine the actual population configuration, the sampling of which, is able to form a sampling theory-estimate for [1-FPE] Confidence Intervals. There have been a few interesting studies of the 22-models used in the M-Competition, where many suggested Confidence Intervals were profiled. See Makridakis, Hibon, Lusk & Belhadjali (1987). However, these are empirical analyses of large datasets specific to models used in the M-Competition—*not the RBF Model*. *These are, indeed, valid efforts*. However, it would also be a valuable effort to partition large ***RBF-datasets*** that have used ONE of the RBF Models and then create empirical Confidence Intervals blocked by the actual RBF Model. Then precision and capture-rates could be used to further refine the RBF Models and possibly lead to theoretical [1-FPE]Confidence Intervals. This theoretical tact would certainly consider combining the [1-FPE]Confidence

Intervals of: The OLS-R & The ARIMA(0,2,2)/Holt model for forecasting projections not outside the time-series interval as these are the only valid OLS-R projections.

REFERENCES

Adya, M., Armstrong, J.S., Collopy, F. & Kennedy, M. (2000). An application of Rule-based Forecasting to a situation lacking domain knowledge. *International Journal of Forecasting*, *16*, 477–484.

Adya, M., Collopy, F., Armstrong, J.S. & Kennedy, M. (2001). Automatic identification of time series features for Rule-based Forecasting. *International Journal of Forecasting*, *17*, 143–157.

Adya, M. & Lusk, E. (2013). Rule Based Forecasting [RBF] - Improving efficacy of judgmental forecasts using simplified expert rules. *International Research Journal of Applied Finance*, *4*, 1006-1024.

Adya, M. & Lusk, E. (2016). Time series complexity: The development and validation of a Rule-Based complexity scoring technique. *Decision Support Systems.* http://dx.doi.org/10.1016/j.dss.2015.12.009

Armstrong, J.S. & Collopy, F. (1992). The selection of error measures for generalizing about forecasting methods: empirical comparisons, *International Journal of Forecasting*, *8*, 69–80.

Brown, R.G. (1959). *Statistical Forecasting for Inventory Control*. McGraw-Hill, New York.

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.

de Carvalho, M. (2016). Mean, what do you mean?. *The American Statistician*, *70*(3), 270-274.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, H., Lewandowski & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, *1*, 111–153.

Makridakis, S., Hibon, M., Lusk, E. & Belhadjali, M. (1987). Confidence intervals: An empirical investigation of the series in the M-Competition. *I.J. Forecasting*, *3*, 489-508.

Theil, H. (1958). *Economic Forecasts and Policy*. North Holland Press, Amsterdam.

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

Secondary References We have presented a number of direct citations from various sources. Some of these citations contained references to literature that these authors used in their research report. Rather than leave these references blank, we have noted them as: *Italicized* & **Bolded**. These ***Secondary References*** may be found in the works herein noted in our references.

APPENDICES

Appendix A

Table 1: M-Competition Series Numbers *NOT* Selected by H&L. n=15

9	88	91	92	98	110	111	152
155	166	167	169	171	178	179	

Discussion There are 181 annual-series in the M-Competition; series 175 had a missing value in the download. *Accounting* [*H&L: Selected 165 + 1 with missing data + 15 series had less than 16 Panel values. Profile for these 15 small panel: Number of Panel-Points: Mean* [*10.8*] *Interval* [*9 : 12*]]

Appendix B

Table 2: Illustrative RAE-values for the first Holdback at t_{14} for 13 Panels

1.21408953	1.12780523	1.02377727	1.46980406	0.667544974	0.71989462	1.01842784
1.06256190	0.86254729	0.96320166	0.82352039	0.868816429	1.30737275	

Appendix C

V3 Relative Absolute Error [RAE] & The Absolute Percentage Error [APE] Note: The RAE & APE are ideal measures as they are scale independent and so can be used where there are magnitude differences. The RAEs, as profiled by the ***Median & Geometric Mean***, are reported in Table 3 [C&A, p.1405]. C&A suggest computing the RAE as a forecasting benchmark as it is an excellent relative measure of forecasting-acuity. C&A note: p. 1402: *The RAE is similar to Theil's U2 as it controls for scale and for the amount of change over the forecasting horizon*; also, see: Theil (1958). In this case, there are two statistical-constructs that need discussion: The ***RAE*** and the ***APE*** as presented following:

The computation of the RAE uses the Absolute Error [AE]:

$$AE_k \equiv ABS(F_k - HB_k)$$

Where: ABS is the Absolute Value Operator, F_k is the OLS-R forecast at time k, and HB_k *is the Holdback at time* k. Uniformly, for our 165-Panels there are three such HB_k: {k: 14, 15 & 16}.

For the H&L-dataset, the AE_k is benchmarked by the Benchmarked Absolute Error [BAE] formed as:

$$BAE_k \equiv ABS(x_{13} - HB_k)$$

Where: x_{13} is last value in our Trimmed Panels used for the OLS-R fit; x_{13} is sometimes called the Random Walk-[RW]-forecast.

Thus, the RAE_k is:

$$RAE_k \equiv [AE_k / BAE_k]$$

XxV4 : The Absolute Percentage Error [APE] as measured by the Median & Mean for the RBF model are reported in Table 3 [C&A, p.1405] and will be vetted using the untransformed original data as this is what C&A must have used. The APE is:

$$APE_k \equiv [ABS(F_k - HB_k) \,/\, HB_k]$$