Transactions on Replication Research

Conceptual Replication DOI: 10.17705/1atrr.00025 ISSN 2473-3458

Replication and Extension of a Forecasting Decision Support System: An Empirical Examination of the Time Series Complexity Scoring Technique

Monica Adya
Department of Management, Marquette University
monica.adya@marquette.edu

Edward J. Lusk
School of Business & Economics, SUNY Plattsburgh
luskej@plattsburgh.edu

Abstract:

This study presents a conceptual replication of Adya and Lusk's (2016) forecasting decision support system (FDSS) that identifies the complexity or simplicity of a time series. Prior studies in forecasting have argued convincingly that the design of FDSS should incorporate the complexity of the forecasting task. Yet, there existed no formal way of determining time series complexity until this FDSS, referred to as the Complexity Scoring Technique (CST). The CST uses characteristics of the time series to trigger 12 rules that score the complexity of a time series and classify it along the binary dimension of Simple or Complex. The CST was originally validated using statistical forecasts of a small set of 54 time series as well as judgmental forecasts from 14 representative participants to confirm that the FDSS successfully distinguished Simple series from Complex ones. In this study, we (a) replicate the CST on a much larger set of data from both statistical and judgmental forecasting methods, and (b) extend and validate the series classification categories from the binary Simple-Complex used in the original CST to Very Simple, Simple, Complex, and Very Complex thus adding an ordinal link between the two previous binary designations. Findings suggest that both the replication and extension of the CST further validate it, thereby greatly enhancing its use in the practice of forecasting. Implications for research and practice are discussed.

Keywords: Decision Support Systems, Forecasting, Complexity

The manuscript was received 11/25/2016 and was with the authors two months for one revision.

1. Introduction

Decision making literature suggests that task complexity influences decision maker's strategies (Payne et al, 1990), information seeking behaviors (Bystrom and Jarvelin, 1995), DSS use (Adya and Lusk, 2016), and performance (Campbell, 1988). In the forecasting literature, a small but consistent set of studies provide similar evidence that the complexity of a time series can challenge the forecasting process and detrimentally influence forecast accuracy (Goodwin and Wright, 1993). For instance, presence of complexity-causing features such as randomness and non-linear trends in time series lead to dysfunctional actions such as overcompensation or confusion (Andreassen and Kraus, 1990). Furthermore, overly difficult time series seem to cause forecasters to ignore cues or to classify important cues as random variations (O'Connor et al, 1993).

The dysfunction created by complexity of forecasting tasks presents an opportunity to examine how Forecasting DSS (FDSS) could be designed to improve human decision processes by selectively focusing analytic activity where it may best conserve the forecaster's time and cognitive resources. This notion emerges from task-technology studies in IS which suggest that the nature of technology support should align with demands of the decision task (Zigurs and Buckland, 1998) i.e. decisions characterized as fuzzy and complex will require support that is different from those characterized as simple. This also positions future work on FDSS within the context of adaptive systems where the complexity of a time series could be used to design systems that restrict or guide forecaster actions. This of course is critical as about 60% of forecasters, in some manner, adjust forecasts from their FDSS and, in the process, often end up compromising forecast accuracy (Sanders and Manrodt, 2003). Complex series might benefit from greater guidance or might require that some features of an FDSS be made "unavailable" to forecasters. Until recently, however, the opportunity to conduct empirical research in these domains was hindered by the lack of a structured and comprehensive mechanism for determining the complexity of a time series.

To address this gap, Adya & Lusk (2016), hereon referred to as A&L, developed and validated an FDSS, the Complexity Scoring Technique (CST). The CST is a rule-based DSS designed to distinguish complex time series from simple ones. Specifically, it uses 12 rules to generate a customized score for individual time series based on its features. For instance, the CST scores a series with a changing slope (a feature) as having higher complexity than a series without this feature. The cumulative score of such features is subsequently used to classify time series along the binary dimension of Simple and Complex. In A&L, the CST was validated on a set of 54 time series across a range of benchmark forecasting methods, e.g. judgmental, Linear Regression, Holt's exponential smoothing. Results confirmed that the CST robustly distinguished Simple from Complex series. From this study, the CST set the stage to tailor the selection of forecasting methods to complexity of a series and to design more nuanced FDSS.

Although A&L made significant strides in addressing this gap in FDSS literature, two issues require further attention. First, the design and validation of the CST was based on a limited set of time series and a small number of data points. A&L had a total of 336 observations for judgmental forecasts and 54 for Regression and Holt's exponential smoothing. It was important, therefore, to conduct "further validation and refinement at many levels" (A&L, p. 81) on a larger data set to establish generalizability of findings. Second, "considering the foundational nature of this work" (p. 76), A&L developed the CST to execute only a binary classification – Simple/Complex – of time series complexity. Whereas, binary classifications are intuitive and can lead to higher proportion of correct classifications, in reality, complexity is a continuum, not a point designation. Using this binary designation runs the risk of misalignment with "forecaster perceptions of the complexity of a time series" (A&L, p. 79). Furthermore, a binary classification limits the precision with which confidence intervals (in effect uncertainty) could be defined around forecasts. To this end, the development of a finer intra-category classification that provides an ordinal set of additional classifications between the two binary ones became a critical aspect of this replication. These extended categories not only enable the validation of the original A&L categories but also enhance the utility of the CST for forecasters when dealing with series that are more intermediate in nature. To extend this binary classification would not only address the above limitations but would be a crucial second level of validation of the mechanism underlying the CST. As such, this study presents a replication and extension of A&L. Specifically, we:

a. replicate the findings of A&L on a larger set of time series and greatly expand the set of judgmental data used in this vetting, and

b. extend the CST by using four ordinal classifications of time series – Very Simple, Simple, Complex, and Very Complex – thus expanding the binary classification developed in A&L. In doing so, we further validate A&L's findings.

In the next section, we briefly summarize the CST and its underlying rule-base for time series classification. This section also rationalizes the need for an FDSS that incorporates complexity considerations. We then describe a new time series classification schema that is tested for internal consistency over the previous binary scoring partition of A&L. The empirical data sets used in the replication and extension are described next along with the proffered hypotheses. Results of the replication and extension are presented thereafter. The paper concludes with a focused discussion of implications for FDSS design.

1. Background

1.1 The Origins of CST: A History of Replications and Extensions

The CST benefits from a strong history of replications and extensions. Its origins lie in the design of Rule-Based Forecasting (RBF) FDSS designed by Collopy and Armstrong (1992). These authors (hereon referred to as C&A), developed an FDSS that combined forecasts from four statistical forecasting methods based on 18 features of time series. This RBF FDSS, an expert system consisting of 99 "IF…THEN…" rules, used forecasting best practices identified by experts to combine forecasts. Eighteen features of time series were used to weight forecasts from these methods, yielding combined forecasts that were customized according to characteristics of the series. The rule below from C&A is representative of how the RBF rules function:

> **RULE 45:** *Unstable Recent Trend*. IF there is an *unstable recent trend*, THEN add 20% to the weight on **Random Walk** and subtract it from **Brown's** and **Holt's**.[1]

In essence, C&A offered, the largest set of features that could be used to characterize time series. However, only about half of these features were programmable and could be identified automatically by embedded modules within RBF. The remaining features had to be determined visually by forecasters, essentially by relying on their experience and information processing capabilities.

This issue was addressed by Adya et al, (2001) who extended RBF by developing various automated routines to detect most time series features that were manually identified in C&As RBF. The design of this Automated RBF (ARBF) set the stage for further replication and validation of RBF on a much larger data set of 3003 time series in Adya et al, (2000) as well as an extension that involved a simpler set of forecasting rules (Adya & Lusk, 2013a). The subject of our study, the CST developed by A&L, has benefited from this history of replications and extensions and presents the next enhancement to ARBF i.e., it systematically builds on the features developed and validated as part of the RBF and ARBF studies. Furthermore, it relies on rules that were built in the original RBF study and validated over the last two decades.

1.2 Overview of A&L's Complexity Scoring Technique (CST)

Whereas typically there should be little need for describing the original study in the conduct of a replication but, because the design, development, and validation of the CST was complex, we begin with a brief overview and background of the CST to inform the context of our study. The CST is a rule-based system that generates a customized score for individual time series based on its features. A&L relied on 14 features[2] of time series to develop and validate 12 rules for identifying complexity of a time series. These rules are presented in Table A in the Appendix. Each series begins with a complexity score of 0. As a feature of a time series is identified, its complexity score is adjusted. For instance, if a series was flagged as having an anomaly between its long-term and short-term trends, its score was reduced by 15 using the following rule.

[1] Rule-numbers are presented as originally designated in C&A. Phrases in italics represent time series features or traits as defined and used in C&A. Here the Random Walk was one of the Models used in Makridakis et al, (1982); the Random Walk is the projection of the last observed value as the forecast value for all the relevant forecasting horizons under examination.

[2] Features include (a) instability causing features such as suspicious patterns in historical data, unstable recent trend, and changing basic trend, (b) uncertainty causing features such as discord between the direction of basic and recent trends, and (c) other features such as domain knowledge, significance of trend, presence of cycles, and number of observations.

Complexity Rule 2: IF the *Direction of Basic* and *Recent* Trends differ OR they agree but differ from *Causal Forces*, THEN add -15 to the Complexity score.

In essence, any complexity causing features lower the complexity score of the time series whereas an alleviating feature would increase it. As CST scoring rules most often reduce the score relative to the baseline score, most series had a negative complexity score. Series with the lowest negative score were, then judged to be the most complex. Scores for time series in the data set ranged 45 units from -40 to +5. Using a variety of measures, A&L defined series with a complexity score of lower than -10 as Complex while -10 or higher were defined as Simple.

1.2.1 Validation of the CST

The CST evolved over two phases and was developed and validated using 126 times series – 74 of these series were used for development and refinement of the rule base while 54 were used for validation. Two sets of validations were conducted:

Model Forecasts:

a. Forecasts were generated for all series, Simple and Complex, using benchmark methods that are well-accepted in forecasting research – Random Walk, Linear Regression, Holt's exponential smoothing/ARIMA(0,2,2), the RBF DSS.
   1.
b. Forecast errors were calculated using two well-established measures recommended in Armstrong and Collopy (1992) – Relative Absolute Errors (RAEs) and Absolute Percentage Errors (APEs). For purposes of validation, the RAEs were given greater emphasis for reasons elaborated in A&L.
c. It was hypothesized logically, referencing the standard forecasting literature, that errors on the RAE would be higher for series determined by the CST to be Complex as opposed to those determined to be Simple. These hypotheses were tested and strongly confirmed.

Judgmental Forecasting:

a. As part of controlled experiments, participants who were well-trained in forecasting in an academic setting were asked to generate forecasts for series classified as Simple and Complex by the CST.

b. The same error measures, RAEs and APEs, were used to measure forecasting accuracy.

c. The same hypothesis was applied for the model forecasts and validations confirmed that errors for forecasts generated judgmentally by participants for Complex series were higher than those for Simple ones.

Example of a research question and hypothesis format ("Research Qs and Hypotheses" in the ribbon).

1.2.2 Contributions of the CST

Although conceptually intuitive and logically defensible, the CST brings much needed value to the FDSS literature. First, it positions future work on FDSS within the context of adaptive systems where features that increase the complexity of a series could be used to design FDSS that restrict or guide forecaster actions. We cannot make recommendations with certainty as the CST has only been recently developed and, to date, no empirical extensions have been reported on it in the literature. Second, the CST makes possible the examination of forecaster behaviors related to FDSS use, particularly regarding how forecasters cope with complexity. While there is some evidence in the forecasting literature that certain features of time series lead to improved use or misuse of FDSS, this literature is preliminary and requires a common framework upon which this body of research might be developed. The CST provides one such approach to managing the dysfunction of complexity but requires further validation.

2. Methodology and Data

Our study is a necessary extension, replication, and validation of the CST offered by A&L. First, we extended the CST to classify time series along four gradations – Very Simple, Simple, Complex, and Very Complex. This elaborated classification was deemed necessary not merely to assess the generalizability of the CST but also to allow FDSS researchers to nuance its integration with other FDSS. The inclusion of additional categories of complexity is critical as it addresses the questions begged by binary only designations: What is the effect of moving off one of the designation points? In other words, how might we deal with intermediate series that are not quite Complex and not quite Simple. While the binary categorizations of A&L are certainly useful for the forecaster as an initial threshold, it is important to consider the intra-category series to offer a rich and articulated decision aid. Second, using these four categories, we replicated the CST on a significantly larger sample of time series as well as judgmental forecasts. These procedures and related data sources are described next.

2.1 Intra-category Classifications – The Elaborated CST

Recall that the CST was designed to classify series with scores between the ranges of +5 and -10 as Simple and those lower than or equal to -15 as Complex. For the extension, we selected relative mid-points along the scoring ranges to give reasonable inferential power over all the four classification categories. This assumes that the scoring line is Cartesian. This mid-range classification would also provide relative balance in the number of sample points and so enhancing the power of the Wilcoxon test and related Multiple Comparison Tests [MCT] (see JMP, 2006). The following intra-category classifications were *a priori* or initially formed and remained unchanged after the preliminary testing:

- Very Complex Series (VC) = Complexity Scores [<= –45 to –20]
- Less Complex Series (LC) = Complexity Scores [–15]
- Less Simple Series (LS) = Complexity Scores [–10 & –5]
- Very Simple Series (VS) = Complexity Score [=> 0]

The sample sizes for these groupings are reported in Table 1 below.

2.2 Intra-category Classifications – The Elaborated CST

Validations of these expanded classifications were conducted in congruence with the protocols used by A&L –i.e., validations were conducted on (a) forecasts from judgmental forecasts generated by participants and (b) forecasts from well-accepted statistical benchmarks. Results from these validations are discussed next.

2.2.1 Judgmental Forecasts – Participant Profile and Treatment

The extended CST was validated with 180 participants who were asked to generate forecasts for time series that were randomly assigned to them. Experiments were conducted at Otto-von-Guericke [OVG] Universität Magdeburg, Germany as well as in Armenia, China, and Leuphana Universität of Lüneburg, Germany where one of the authors was involved in teaching forecasting courses. To maintain consistency with A&L protocols, we selected 66 time series from the same data set as used in their study – i.e., from the M-competition data (Makridakis et al, 1982). All of these 66 series were selected randomly by the authors. Participants were not aware of the complexity of the series that they were assigned in the experimental design. Incidentally, the author executing the experiments was also unaware of the classification of these series. These 66 series and their URL are noted in Table B in the Appendix.

These judgmental validations were conducted over a period of ten years. During this time, we kept track of all the series used in the delivery of these courses. *A posteriori,* we classified these target of opportunity series as to their complexity using the CST complexity scoring protocol. Therefore, this was a truly unbiased assignment as, for more than 85% of the series, there was no complexity information available at the time of the assignment.

About 80% of the judgmental forecasts were generated by participants enrolled in a graduate course on Business Forecasting at the OVG. The general profile of the participants was constant as most of them were

enrolled in the Master's Program in Economics[3] at OVG. All participants were from equivalent Master's programs with the exception of one group at the Lüneburg Universität where the students were undergraduates in a Decision-Making Analytics program in the School of Management. The profile of the students in these courses is presented in Table 1 below:

TABLE 1: Participant Profiles for Judgmental Validation of the CST

Participant	Age Median: Range	Experience+	Performance*
OVG	23:[19-37]	Internships: 34% Prior Work Experience: 6%	Pass: 93% Low Pass: 6% Fail: 1%
Summers	22:[18-24]	Internships: 12% Prior Work Experience:4%	Pass: 99% Low Pass: 1%

*Due to confidentiality considerations, we were not permitted to collect overall GPA for the OVG students. However, we were permitted to report in the aggregate the grades achieved in the forecasting courses. In the regard, we scaled the various diverse grading systems to the USA note metric [A through C]: ***Pass[P]***, D ***Low Pass[LP]*** and Failure [**F**], all failures, there were 3 over the accrual set, were excluded from the dataset].

We tested to determine if there were differences between OVG and the summer groups and over the two grade groups. Their respective p-values were 0.45 and 0.63 respectively suggesting that the Null of no difference was the likely case. Participants were asked to produce 1- to 6-period ahead forecasts for each of the four assigned series, yielding 4,320 forecasts, (180 students, each of whom on average was given four series and generated 6 forecasts: 180×4×6).

2.2.2 Statistical Methods – Data and Validations

In addition to judgmental forecasts for Very Simple to Very Complex series, forecasts were also generated for well-accepted benchmark methods, similar to those used by A&L. Specifically, forecasts were generated using (a) OLS two-parameter Linear Regression, and (b) Holt/ARIMA (0,2,2) two-parameter linear exponential smoothing. For statistical methods, there were [390 x 2 or 780 forecasts for 65 series; series number 36 was eliminated as it exhibited Holt Inversion issues. Specifically:

1. **OLS:** Regression of the form [$Y_t = \alpha + \beta \times t;\ t: [1, n]$], noted as [**R**], and
2. **Holt/ARIMA (0,2,2)**: Exponential smoothing method also known as the Holt Model [**H**]

In total, between all methods, 5,100 forecasts were used – 4,320 for judgmental forecasts and the balance 780 [5,100 – 4,320] from formal statistical methods: Regression and Holt.

2.2.3 Forecast Accuracy Measures and Hypotheses for Empirical Validation of the Elaborated CST

Relative Absolute Error (RAE), a well-accepted measure of forecast accuracy, was used to evaluate the forecasts and thereby, the complexity classifications. This same measure was used in A&L. For consistency, the RAEs were winsorized[4], as was done in A&L and also by C&A. We used the same hypotheses as in A&L with the underlying assumption – *that a full validation of the CST will require that the forecasts for Simple series will be more accurate on the RAE measure than for the Complex ones*.

The hypotheses for this study are developed along two dimensions (a) replication of the original A&L classification, and (b) a validation of the extended classification schema to Very Simple-Very Complex, as described earlier. The hypotheses are tri-conditional, considering the three independent forecast generating methods.

2.3 Replication of the CST

When facing complex tasks, forecasters become conditioned to relying on compensatory decision processes such as frugality in use of cognitive resources and simplifying tasks by eliminating alternatives and processing limited information (Payne, 1976). A&L examined this by asking forecasters to produce

[3] An English language program in the Faculty of Economics & International Studies.
[4] Winsorizing is a common practice in forecasting where if, for a forecast, the RAE < 0.01 then it takes the value RAE = 0.01 and if RAE >10 then it takes the value RAE = 10.

structured judgmental forecasts for series classified as Simple or Complex by the CST. Their tests confirmed that complexity detrimentally impacted the accuracy of judgmental forecasts. Considering that the conditions of our replication were not different from those in A&L, the following hypothesis is proposed:

Hypothesis 1: Median RAEs for *Judgmental* forecasts will be higher for Complex series when compared to those for Simple series.

For the statistical methods, A&L provided consistent evidence across multiple benchmark methods that, for both 1-ahead and 6-ahead forecasts, Complex series are more challenging to forecast than Simple ones. In particular, they found that the forecast accuracy for both OLS Linear Regression and Holt's Exponential Smoothing were higher for Simple time series as opposed to Complex ones. In keeping with those findings, we propose the following:

Hypothesis 2: Median RAEs for *OLS Regression* forecasts will be higher for Complex series when compared to those for Simple series.

Hypothesis 3: Median RAEs for *Holt/ARIMA forecasts* will be higher for Complex series when compared to those for Simple series.

2.4 Extension to Inter-Category Classifications

We extrapolate the hypotheses from A&L to the expanded complexity classifications and posit that forecast accuracy will progressively get worse from series classified as Very Simple to those classified as Very Complex. This aspect of the study is most crucial to CST because if we find no order in the forecast accuracy across the four categories, it would call to question the fundamental validity of the CST. In contrast, if this order were to be confirmed, this second level CST validation would reaffirm the potential impact of CST on FDSS research. Accordingly, the following set of hypotheses were developed, essentially building from H1-H3:

Hypothesis 4: Median RAEs for *Judgmental Forecasts* will follow the pattern: VC > LC > LS > VS.

Hypothesis 5: Median RAEs for *OLS Regression* will follow the pattern VC > LC > LS > VS.

Hypothesis 6: Median RAEs for *Holt/ARIMA* forecasts will follow the pattern VC > LC > LS > VS.

3. Results

The results for all six hypotheses are presented in Table 2 below. For robustness, to test H1-H3, we used three inferential measures: Wilcoxon/Kruskal-Wallis Rank Sums, the Median Test (points above the median), and the van de Waerden test (normal quantiles) platforms in SAS/JMP, v13. This allows us to report the highest, i.e. most conservative, p-value relative to rejecting the usual inferential null. For H1-H3, for the series classified using the original CST, the highest p-value of the three sets of forecasts was such that one can justify rejecting the nulls with high assurance. For all three test trials, the Complex series has a RAE profile significantly higher than that of the Simple series and in two cases, as well as overall, the Median for the Complex series arm is >1.0. This represents strong evidence that the original CST classification in A&L is reliably homomorphic and robust along the binary scoring partition over the three generating processes - Judgment, Regression, and Holt/ARIMA. Table 3 juxtaposes original results from A&L with those from this replication. These comparisons confirm that with the larger judgmental and statistical sample, spanning multiple data collection phases, CST robustly distinguishes between simple and complex time series.

To test H4-H6, we used the non-parametric ANOVA Wilcoxon Rank Sum Test, n >2 and subsequently used the non-parametric multiple comparison test based upon the Wilcoxon Test as found in *SAS/JMP, v13*. While there is no clear inferential test that flows naturally from this dataset as the elements are not likely independent realizations, what seems reasonable is to test the percentage of multiple comparison p-values that are less than the directional hypothesized p-values of 0.05. In this case, we computed the directional 95% confidence interval (CI) for this percentage. Our *rejection* of the null of no effect will be if the lower limit of the 95% CI excludes 50%. This is a strong test because if the null is the reality, then number of low p-values will also be low - on the order of 5%. Conversely, if the lower limit of the above test excludes 50%, admittedly an optimistically high or conservative value, there can be suggestive evidence that the multiple comparison test separation is indicative of structural differences.

Table 2: Results Showing Median RAE for Hypotheses H1-H6							
Test Groups	*Hypotheses H1 – H3*			*Hypotheses H4-H6*			
	Complex: Median, n	Simple: Median, n	Directional p-values*	Very Complex: Median, n	Less Complex: Median, n	Less Simple: Median, n	Very Simple: Median; n
Judgmental Forecasts (JF)	1.20 n = 1902	0.72 n = 2,418	<0.0001	1.26 n= 1440	1.0 n= 462	0.77 n= 1,716	0.61 n= 702
Regression (R)	1.14 n = 156	0.75 n = 234	<0.02	0.87 n= 102	1.41 n= 54	0.80 n= 144	0.72 n= 90
ARIMA/Holt (H)	0.80 n = 156	0.55 n = 234	<0.0001	0.85 n= 102	0.78 n= 54	0.60 n= 144	0.36 n= 90
Overall: JF, R, H	1.16 n = 2,214	0.71 n = 2,886	<0.0001	N/A	N/A	N/A	N/A

*Overall: n = 5,100 or 850 trials each of which produced six (6) forecasts. The p-value reported are for the highest, most conservative respecting the Null, of the three robustness tests.

Table 3: Comparative Median RAEs from Original A&L Study				
Test Groups	***Original Study***		***This Replication***	
	Complex: Median, n	**Simple: Median, n**	**Complex: Median, n**	**Simple: Median, n**
Judgmental Forecasts	1.22 n=168	0.61 n=168	1.20 n=1902	0.72 n=2418
Regression	1.25 n=22	0.53 n=32	1.14 n=156	0.75 n=234
ARIMA/Holt	0.78 n=22	0.36 n=32	0.80 n=156	0.55 n=234

The p-values of the RAE in profile are presented in Table 4:

Table 4: P-values for the Multiple-Comparison Test	
	Overall
Mean	0.045
Median	<0.0001
Mode	<0.0001
Range	[<0.0001 to 0.86]

The number of directional p-values for the multiple comparison tests, derived from Table 2, are in total 18 - three independent methods (Judgment, Regression, Holt/ARIMA) and 6 inter-method contrasts. For these 18 comparisons, 16 were less than the cut-off point p-value of 0.05. The lower limit of the 95% CI is 76.7% [for 88.9% or 16/18] which is significantly above the 50% cut-off, suggesting that there is intra-group separation and internal consistency of the cut-off points. These results are a strong rejection of the null of no directional association for the Median RAE, implying that for the expanded CST one can confidently reject the null of no effect on an inter-group binary comparison. Simply, the integrity of the CST is vetted and it can likely be relied on to provide nuanced classifications of time series classifications that can be useful for FDSS design as well as informing judgmental forecasting best practices.

4. Summary and Implications

We set out to confirm and extend the validity and, as such, the practical utility of the FDSS developed by A&L. This study, then, contributes in two important ways:

(a) The CST complexity categories were expanded and results confirmed that the classifications retained our original inference as presented in A&L - i.e., to distinguish Complex series from Simple ones, potentially spawning a new stream of research. This opens the CST from a simplistic binary partition to four ordered categories from Very Complex to Very Simple.

(b) A more robust replication of CST is presented, tested, and found to be consistent. The CST was replicated and validated across three independent and inherently different datasets of a much

larger magnitude than used in A&L. This robustness is critical as it expands the domain of activity from that of only judgmental methods such as RBF to a wider set of models that are the common fare in forecasting: the Linear Regression and ARIMA models.

The elaborated CST is likely to be a critically valuable system for identifying the nature of time series used for decision-making and fine-tuning the use of forecasting methodologies around these gradations. Most directly, when the decision-maker is engaged in a forecasting task the input of which is a time series, the first question we recommend addressing is:

What is the nature of the complexity with which the decision maker is dealing?

Two critical and poignant observations can be made here regarding the inferences rationalized by the above results:

a. If the informational time stream falls into the ***Very Complex*** category, then the historical data of the time series will likely provide little information that will be useful for a reasonable forecast. *Thus, the forecaster may conserve valuable time by not trying to outperform the Random Walk method that extrapolates the last historical period to the future*.

b. However, where the elaborated CST classification suggests that the time series falls in the class of ***Very Simple***, the decision maker can expect to outperform Random Walk projections of the recent past by *using standard statistical models* that do not require judgmental inputs and, once again, conserve valuable decision-making time.

c. Given the intermediate ranges offered by the new categories, forecasters can use their experience and analytical protocols to improve the acuity of FDSS. This opens opportunities for further design and testing of FDSS that are fine-tuned to identify models and that may be used to create FDSS to improve forecasting accuracy while conserving the resources employed in the forecasting endeavor.

4.1 Implications for FDSS Design

Through this extension and replication of the A&L CST, we have significantly expanded the operational utility of the work of A&L by extending the empirical validation. Over time one would suppose that this would create a competitive advantage while conserving resources as the above points suggest. These are indeed, significant breakthroughs in the FDSS domain as this opens vast new opportunities for DSS development (see Adya & Lusk, 2013b). The current CST system relies on the decision-maker to identify features of the time series – i.e., it assumes that features of the time series have been characterized. In the current world of streaming and big data, for a decision-maker to code time series features is manifestly impractical. This begs the next step - that is to integrate the elaborated CST system presented above into an expanded FDSS that captures the time series at its source/initial engagement and uses automate feature identification routines, such as those found in Adya et al, (2001). These time series characterizations can then feed into a forecasting system, such as C&As RBF that can select forecasting methods based on complexity as well as features of the time series. An exciting application of such an integrated FDSS might be to examine application of restriction and guidance (Silver, 1990) in light of the range of Very Simple to Very Complex time series tasks. The elaborated CST has opened the door to this new stream of necessary DSS development.

In summary, in managerial domains, the dysfunction created by complexity is often manifested in the form of poor decision-making and /or the dissipation of scarce resources. In the case of forecasting, it implies inaccuracies that can have significant implications for firms in terms of lost revenue through excessive inventory or gross miss-estimations of demand. A&Ls work in realty underlies the functionality of most DSS and has a direct role in FDSS design, particularly with regard to task-technology fit (Goodhue and Thompson, 1995). We learn from Table 3 that Very Simple and Very Complex series beg the question of the time commitment needed to form Judgmental forecasts; as simple push-button models of Regression and Holt/ARIMA seem appropriate to the task in the former case and, in the latter case, the Random Walk (Naïve Method) is the model of choice. In summary, we offer that Tables 2 and 3 can form a viable screening protocol for making the behavioral coping decision as to how to deal with time series complexity.

References

Adya, M. & Lusk, E. (2016). Development and validation of a Rule-based time series complexity scoring technique to support design of adaptive forecasting DSS. *Decision Support Systems,* 83, 70-82.

Adya, M. & Lusk, E. (2013a). Rule Based Forecasting [RBF] - Improving efficacy of judgmental forecasts using simplified expert rules. *International Research Journal of Applied Finance*, 4(8), 1006-1024.

Adya, M & Lusk, E.J. (2013b) Designing effective forecasting Decision Support Systems: Aligning task complexity and technology support. In Chiang, J. (Ed) *Decision Support Systems* (pp. 173-196). InTech.

Adya, M., Lusk, E., & Belhadjali, M. (2009). Decomposition as a complex-skill acquisition strategy in management education: A case study in business forecasting. *Decision Sciences Journal of Innovative Education,* 7(1), 9-37.

Adya, M., F. Collopy, J.S. Armstrong, and M. Kennedy (2001). Automatic identification of time series features for rule-based forecasting. *International Journal of Forecasting,* 17(2), 143-157.

Adya, M. (2000). Corrections to rule-based forecasting: Findings from a replication. *International Journal of Forecasting,* 16(1), 125-127.

Adya, M., Collopy, F., Armstrong, J.S. & Kennedy, M. (2000). An application of rule-based forecasting to a situation lacking domain knowledge. *International Journal of Forecasting* 16(4), 477-484.

Andreassen, P. B. & Kraus, S. J. (1990). Judgmental extrapolation and the salience of change. *Journal of Forecasting,* 9(4), 347-372.

Bystrom, K. & Jarvelin, K. (1995). Task complexity affects information seeking and use. *Information Processing & Management,* 31(2), 191-213.

Campbell, D. (1988). Task complexity: A review and analysis. *Academy of Management Review*, 13(1), 40-52.

Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS quarterly*, 19(2), 213-236.

Goodwin, P. & Wright, G. (1993). Improving judgmental time series forecasting: A review of the guidance provided by research. *International Journal of Forecasting*, 9(2), 147-161.

JMP: SAS.2006. *Statistical Discovery: Statistics and Graphics Guide JMP 6*, Cary, NC, USA: The SAS Institute.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, 1(2), 111-1153.

O'Connor, M., Remus, W., & Griggs, K. K. (1993). Judgmental forecasting in times of change. *International Journal of Forecasting*, 9(2), 163-172.

Payne, J.W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis, *Organizational Behavior and Human Decision Processes*, 16(2), 366-387.

Payne, J.W., Bettman, J.R., & Johnson, E.J. (1990). The adaptive decision maker: Effort and accuracy in choice. In R. Hogarth, *Insights in Decision Making* (pp. 129-153). Chicago: University of Chicago Press.

Sanders, N. & Manrodt, K. B. (2003) Forecasting software in practice: Use, satisfaction, and performance. *Interfaces,* 33(5), 90 – 93.

Silver, M.S. (1990). Decision support systems: Directed and non-directed change, *Information Sysetms Research*, 1(1), 47-70.

Zigurs, I. & Buckland, B. K. (1998). A theory of task/technology fit and group support systems effectiveness. *MIS quarterly*, 22(3)**,** 313-334.

Appendix A: CST Rules and Series

Table A: CST Rules (Adya & Lusk, 2016)	
Traits	Related Complexity Scoring Rules
Complexity of Underlying Signal	*Levels of complexity may vary from stationary through linear trend, non-linear trend to no trend.* *CRule 1:* IF *Causal Forces* are Unknown, THEN add -5 to the Complexity score. *CRule 5*: IF *Basic Trend* is not significant (Regression T-Stat <2.0), THEN add -5 to the Complexity score. *CRule 9:* IF the *Functional Form* of a series is additive, THEN add -5 to the Complexity score. *CRule 12:* IF a *Number of Observations* in a series < 13, THEN add -5 to the Complexity score.
Level of Noise around the underlying signal	*CRule 2*: IF *Direction of Basic* and *Recent* Trends differ OR they agree but differ from *Causal Forces*, THEN add -15 to the Complexity score. *CRule 4:* IF Series is *Suspicious*, THEN add -10 to the Complexity score. *CRule 8:* IF the *Basic Trend* of a series is changing, THEN add -15 to the Complexity Score. *CRule 11:* IF the *Coefficient of Variation* about the Trend > 0.9, THEN add +5 to the Complexity score.
Stability around underlying signal	*CRule 3*: IF *Recent Trend* is unstable, THEN add -20 to the Complexity score. *CRule 6:* IF there is a *Level Discontinuity*, THEN add -5 to the Complexity Score. *CRule 7:* IF a series is *Near a Previous Extreme* AND *Cycles* are present, THEN add +10 to the Complexity score. *CRule 10*: IF the *Recent Run is Not Long* THEN add -5 to the Complexity score.

Table B: Series Used in the creating the forecasts for the Three Forecasting Groups*										
4	5	7	8	14	15	17	18	24	27	28
34	35	36	37	38	44	45	47	48	53	54
55	57	58	64	67	68	74	76	77	78	84
87	88	94	96	97	98	104	105	107	108	114
117	118	124	127	128	134	136	137	138	144	147
148	154	157	158	164	167	168	174	175	177	178

*These series [including holdbacks] can be downloaded at: www.forecasters.org. There are 181 time series at www.forecasters.org. A&L used 72 for the AL-CST in the development phase and 54 in the holdback testing phase. In the empirical validation we will use 66 series. For purposes of continued testing the set of series not used in one of the three testing protocols follow the modular repeating set: 1: then {9, 10, 11};- - -; {179, 180 , 181} or there are 1 + 3×18 =55 series that could be used to extend the validation testing. The bolded series were the series most used in comparison to the Median number of series forecasts generated which was 36. There were 25 of these often-used series.

About the Authors

Monica Adya is a Professor and Chair of Management at Marquette University. She received her PhD in Management Information Systems from The Weatherhead School of Management, Case Western Reserve University. Monica's research spans knowledge-based decision support systems with applications to business forecasting. Her work has been published in *Communications of the Association of Information Systems*, *Decision Support Systems*, *Journal of Innovative Education*, *Human Resource Management*, *Information Technology & People*, *International Journal of Forecasting*, *Information Systems Research*, and *Journal of Global Information Management,* among others. Monica has been co-investigator on several research grants including those from 3M Foundation and Naval Surface Warfare Center.

Edward J. Lusk is currently Professor of Accounting, the State University of New York [SUNY], College of Economics and Business, Plattsburgh, NY, USA and Emeritus: The Department of Statistics, The Wharton School, The University of Pennsylvania, Philadelphia, PA. USA. From 2001 to 2006 he held the Chair in Business Administration at the Otto-von-Guericke University, Magdeburg Germany. He has also taught in China at the ShanXi University of Finance and Economics, Taiyuan and the Chulalongkorn University Bangkok, Thailand. He has published more than 225 articles in peer reviewed journals and texts including: *The Journal of Political Economy*, *Statistics and Probability*, *The American Statistician, The New England Journal of Medicine, The Journal of Marketing Research, The Journal of the Academy of Management, Medical Care, The Public Opinion Quarterly, Omega, The Accounting Review*, *Management Science*, *The Journal of Forecasting, Environmental Quality Management, The Journal of the Operational Research Society*, *Gender, Work and Organizations*, and *Decision Support Systems*.

Copyright © 2018 by the Association for Information Systems. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and full citation on the first page. Copyright for components of this work owned by others than the Association for Information Systems must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or fee. Request permission to publish from: AIS Administrative Office, P.O. Box 2712 Atlanta, GA, 30301-2712 Attn: Reprints or via e-mail from **ais@aisnet.org**.