3-1-2016

Development and Validation of a Rule-based Time Series Complexity Scoring Technique to Support Design of Adaptive Forecasting DSS

Monica Adya
Marquette University, monica.adya@marquette.edu

Edward J. Lusk
State University of New York at Plattsburgh

Accepted version. *Decision Support Systems*, Vol. 83 (March 2016): 70–82. DOI.

Development and Validation of a Rule-Based Time Series Complexity Scoring Technique to Support Design of Adaptive Forecasting DSS

Monica Adya
College of Business, Marquette University
Milwaukee, WI
Edward J. Lusk
The State University of New York (SUNY) at Plattsburgh
School of Business and Economics,
Plattsburgh, NY
Emeritus, Department of Statistics
The Wharton School, University of Pennsylvania,
Philadelphia, PA

Abstract: Evidence from forecasting research gives reason to believe that understanding time series complexity can enable design of adaptive forecasting decision support systems (FDSSs) to positively support forecasting behaviors and accuracy of outcomes. Yet, such FDSS design capabilities have not been formally explored because there exists no systematic approach to identifying series complexity. This study describes the

Decision Support Systems, Vol 83 (March 2016): pg. 70-82. DOI.

development and validation of a rule-based complexity scoring technique (CST) that generates a complexity score for time series using 12 rules that rely on 14 features of series. The rule-based schema was developed on 74 series and validated on 52 holdback series using well-accepted forecasting methods as benchmarks. A supporting experimental validation was conducted with 14 participants who generated 336 structured judgmental forecasts for sets of series classified as simple or complex by the CST. Benchmark comparisons validated the CST by confirming, as hypothesized, that forecasting accuracy was lower for series scored by the technique as complex when compared to the accuracy of those scored as simple. The study concludes with a comprehensive framework for design of FDSS that can integrate the CST to adaptively support forecasters under varied conditions of series complexity. The framework is founded on the concepts of restrictiveness and guidance and offers specific recommendations on how these elements can be built in FDSS to support complexity.

Keywords: Benchmark forecasting, Forecasting decision support systems, Structured judgment, Forecasting, Time series, Rule-based Forecasting

1. Introduction

Judgmental forecasting has become an increasingly appreciated approach and, in the process, has "undergone a significant transformation."[1,pg. 493] In particular, best practices have emerged around structuring and formalizing the use of judgment through integration with statistical methods. The practitioner community also has an extensive history with judgmental forecasting. For instance, in a survey of 240 US firms, only 11% used forecasting decision support systems (FDSSs) and, within this sub-group, over 60% judgmentally adjusted software-generated forecasts.[2] Although best practices around judgmental techniques have been rapidly accumulating, many aspects still require further research. In this study, we address one such aspect — time series complexity for decision support and FDSS design.

Alignment between DSS capabilities and task support needs can improve DSS utilization, decision maker performance, and thereby task outcomes.[3,4] This body of research, which often classifies tasks on a continuum from simple to fuzzy (complex), provides support for design of adaptive DSS. Adaptive systems can support judgment by presenting and processing information in ways that adjust to task

context and characteristics, cognitive needs of forecasters, and patterns of information use[5,6] thereby debiasing the decision process.[7] The design of such adaptive systems for forecasting, however, has remained unexplored as there exists no formal way of characterizing the complexity of forecasting tasks.

This study presents the design and validation of a rule-based complexity scoring technique (CST) that relies on a tested and validated set of time series, features, and rules. To this end, the CST is validated using (a) forecasts from benchmark methods on a set of holdback series and (b) experiments with 14 forecasters who rendered 336 structured judgmental forecasts on series scored by the CST as simple or complex. The study concludes with the development of a framework for design of adaptive forecasting decision support systems (AFDSSs) that can respond to forecasting task complexity. This DSS framework is built upon elements of restrictiveness and guidance[8] to limit harmful actions and improve forecaster efficacy under complexity. It must be noted that this study does not justify a theory. Rather, it is positioned in the design science paradigm and seeks to develop capabilities around the design of an IT artifact for series complexity,[9] a task deemed difficult for reasons discussed in later sections. As such, the CST is expected to be refined over multiple design cycles.

2. Background and motivations

2.1. Adaptive DSS

Adaptive DSS (ADSSs) have been defined as systems that aid "decision making judgments by adapting support to the high-level cognitive needs of the users, tasks characteristics, and decision contexts" (pg. 299).[5] Numerous studies have conceptualized ADSSs that support problem formulation, interpretation of the dynamic problem space, and final decision outcomes in response to environments that are known to change within a single decision or across multiple decisions.[5] Piramuthu & Shaw[10] for instance, suggest that an ADSS must have a learning component that can incrementally renew its knowledge base through continuous feedback from the environment. Others have proposed that ADSS must adapt to users' personalities[11] and decision support needs. ADSS can deliver a range

Decision Support Systems, Vol 83 (March 2016): pg. 70-82. DOI.

of problem-solving tools and interfaces that can be invoked by users based on decision context. Decision makers' problem space could also be made more flexible by providing drill-down capabilities into the data,[12] especially as DSSs become integrated with big data. Finally, flexibility can relate to evaluation of decision outcomes. ADSS could be self-evaluative[10] based on internal feedback mechanisms, such as neural networks, or could evaluate users by providing feedback based on decision optimality.

Although numerous forecasting studies have hinted at the need to align forecasting tasks with FDSS capabilities,[4] few have addressed the design and benefits of adaptive systems. Authors in[13] discuss preliminary benefits for ADSS in the domain of water and weather forecasting. Similarly,[12,14] address the need to adapt DSS display, data, and models to the nature of time series being forecast. However, beyond these preliminary indications, insights into design and use of adaptive FDSS are limited as there exists no formal framework around which to conceptualize such systems.

Our review of ADSS indicates that such aids can, and should, be designed to adapt to three primary sources of knowledge: the problem domain,[15] the user,[16] and its own knowledge-base.[10] Although the three elements are interlinked, the focus of our study is on the first i.e. the problem domain. Specifically, our proposals for design of an adaptive FDSS are formulated on understanding time series complexity such that an FDSS could be designed to adaptively support forecasters based on task complexity. A preliminary link between time series complexity and DSS capabilities was established in[17] which found that use of a simple DSS improved forecaster performance in turbulent and complex markets. The challenge, however, is that our understanding of DSS design characteristics, as they relate to time series complexity, is quite dispersed and very few mechanisms currently exist to comprehensively identify series complexity. This is an effective point of departure for our study for which the central issue is the need for identification of series complexity as a necessary pre-condition to framing adaptive FDSS.

2.2. Time series complexity

Drawing parallels from general decision making literature which finds that task-related complexity influences decision makers' strategies,[18,19] information processing behaviors,[20] DSS use,[17] and decision outcomes and performance,[19,21] one may logically suppose that *complexity of a time series* will have similar effects in forecasting. However, the lack of a well-defined and validated approach to identifying series complexity has limited our understanding of the implications of complexity for design and use of FDSS, related research, and forecasting practice.

A small but consistent body of forecasting literature, however, provides useful insights into how and why time series complexity might impact forecast outcomes. Information seeking and processing vary with complexity of cues embedded in the task.[20] Simple tasks require processing of fewer cues and, as such, place lower demands on decision makers' cognitive resources. In contrast, complex tasks cause decision makers to conserve cognitive resources by processing fewer cues.[22] Features of time series (such as direction of trend, presence of variability) are *task cues* that can potentially condition forecaster behavior and performance, evidence for which does exist in the forecasting literature. For example, non-linear trend[23] and the presence of randomness[24] introduce systematic bias in the forecasting process. Forecasters also tend to dampen both increasing and decreasing trends[25] and are particularly confused by the latter[26] or by series with no perceptible trends.[27] The presence of complex seasonal and cyclical patterns seems to bemuse forecasters, leading to lower forecast accuracy.[28] The aggregate impact of these characteristics creates effects, such as sub-optimal use of knowledge, similar to those observed with complex tasks. In other words, these features interject challenges in the forecasting process.

When facing complex tasks, forecasters may become conditioned into unwittingly relying on compensatory decision processes. They become frugal with cognitive resources and simplify the task by eliminating alternatives and processing limited information.[29] In low complexity domains, however, they arrive at correct decision strategies expeditiously and consistently.[18,19] While

Decision Support Systems, Vol 83 (March 2016): pg. 70-82. DOI.

some studies have found such compensatory processes to result in time-savings without noticeable loss in decision accuracy,[30] they produce inconsistent results.[18,31] Such effects may lead forecasters to overlook useful cues or to classify cues as random variations.[26] Forecasters also employ different heuristics for trended and un-trended series where their approach to the former, often considered more difficult, is influenced by the extent of correlation between cues.[27] Such anchoring is illustrative of compensatory practices.

Outcomes for complex tasks have largely been examined in terms of the cumulative effect of task cues, related decision strategies,[32] and use of decision aids.[33] Findings from numerous domains such as auditing[34] and consumer choice[35] confirm that task complexity results in lower decision accuracy. There are, however, some indications that expertise and environmental factors can reverse these effects. Skill[19] and motivation[36] of decision makers can stimulate them in difficult situations, potentially improving outcomes. Similar contradictory effects are evident in forecasting where some studies find experts to be better at applying domain knowledge[1] while others find novice forecasters to be as accurate as experts.[37] Outcomes also improve when DSSs fit task needs. Although little direct evidence is available within the context of complexity, judgmental forecasters do benefit from use of FDSS[38] and by the manner in which the forecasting task is presented.[1] For instance, FDSSs improve forecast accuracy by increasing the slope of analysts' forecasts while decreasing variation[39] and by reducing inconsistencies in outcomes, underscoring decision makers' tendency to smooth to expectations.[40]

The discussions above highlight the confounding processes that underlie complex forecasting tasks. Formalizing these findings for improved research and practice, however, requires simple forecasting tasks to be distinguishable from complex ones. The lack of protocols to create such distinctions warrants development of a complexity scoring technique that can provide a common base from which to study effects of complexity. The next section describes one such protocol.

3. Features: the context for time series complexity

Most forecasting studies have focused on a small set of features when characterizing time series. Studying combined effects of an expanded set of series features on accuracy, while challenging, is necessary as features rarely exist in isolation and may have compensatory, degenerative, or worse yet, random interactions. An underlying process that produces a stable two-parameter linear trended series, for instance, may be exponentially confounded by the level of variation generated by contextual event-instabilities. Yet, by focusing on overall non-event segmented trend effects, the impact of these additional features may be overlooked, possibly to the detriment of forecast accuracy. A complexity schema based on a more inclusive set of features could suggest decision strategies and FDSS capabilities that align better with the task at hand. Use of expanded feature sets for complexity classification is also consistent with the call by,[41] hereon referred to as G&W, to develop a formal characterization of time series to aid judgmental forecasting and "draw firm practical conclusions from research in this area" (p. 151).

G&W suggest a comprehensive definition of time series complexity along three feature categories: (1) complexity of the underlying signal including seasonality, cycles, and trends; (2) level of noise within which the structured signal may be buried; and (3) instability of the underlying signal captured in sudden changes such as level discontinuities. This provides a useful platform upon which to propose a feature-based complexity schema. To do so, a well-established and validated set of features capturing the range of series characteristics is necessary. We identified such a feature set in,[42] hereafter referred to as C&A.

3.1. The Rule-Based Forecasting feature set

C&A generated the most extensive and well-validated set of time series features published in peer reviewed literature. Their study presented the Rule-Based Forecasting (RBF) system, an FDSS that relies on 18 features of time series to combine forecasts from four accepted forecasting methods: Random Walk (Naïve 1), OLS Linear

Regression, Holt's two parameter exponential smoothing (ARIMA [0,0,2]), and Brown's exponential smoothing. These initial set of C&A features were validated and extended in studies such as,[43,44,45] thereby establishing strong theoretical and empirical foundation over two decades.

RBF rules relied extensively on forecasters' domain knowledge and, as recommended in the empirical literature, were designed to allow forecasters to integrate this knowledge as input to the forecasting process. RBF features encompassed all three trait categories proposed in the G&W framework. For instance, features such as trend, seasonality, and presence of general cycles correspond to underlying signal. Traits such as variation around trend, changing trend, and suspicious pattern align with noise around the underlying signal. Finally, features such as outliers, level discontinuities, and unusual last observations capture instabilities underlying the generating process. RBF features, then, empirically captured what was conceptually proposed in G&W. Of these 18 features, our study uses 14 (see Appendix A) that, *a priori*, were deemed essential for developing a robust CST. [c] Next, we describe the development of the CST and its validation using both holdback and structured judgmental forecasts.

4. Development and validation of the rule-based CST

4.1. Overview of CST development process

For development and validation of the CST, we relied on the data and rules developed for the RBF system and presented in C&A and.[46] Three elements were culled from these two sources: (i) 126 time series from M-competition data[48] as used in these studies, (ii) feature codings for each of these series, and (iii) error measures for forecasts from two methods, RBF and Combining A[d] for each series. To this end, the 126 series and related meta-data provided the critical "wind tunnel" data for benchmark comparisons[49] (p. 279). Seventy four of the 126 series were quasi-randomly selected for development and refinement of the CST (*development data set*). Series ending in 2,

3, 5, and 6 were in this group. The remaining 52 series (ending in 4, 7, and 8) were held back to validate the CST (*holdback data set*).

Using data from[46] and C&A was beneficial for several reasons. First, C&A had coded each of the 126 series along the 18 features, thus providing a validated set of feature codings. Inter-rater reliability between the authors was high at 89%,[46, p. 1403] and differences were reconciled to yield a consistent set of series characterizations. Second, the feature codings were validated in several extensions.[43,44,45] Third, forecast errors for RBF and Combining A provided *a priori* validated benchmarks for refinement and sensitivity analysis of the CST during development. *The assumption that complex series will have lower forecast accuracy than simple ones formed the logical basis for calibrations and directional hypothesis formation and testing.*

The CST evolved over two phases. In the *Development Phase*, rules for coding complexity were derived and refined using the 74 series in the development data set. The final forecasting error measures presented in later sections for this data set were generated only after a theoretically defensible rule set was identified. Rule refinements, discussed later, were conducted on this same series set. Upon completion of development, the *Validation Phase* was executed, wherein the CST was tested on the 52 holdback series. The final CST was a rule-based scoring schema that adjusts the score of a series based on its features and generates a customized complexity score for each series. Such rules could be easily integrated into any FDSS or be applied judgmentally by forecasters. Generating an aggregate score for time series may appear to contradict the extensive body of research that supports decomposition of forecasting tasks. However, we view the use of aggregate score as a precursor to forecast generation i.e. the complexity score can signal the difficulty of the series and signal the features that contribute to this complexity while decomposition is a subsequent step to deal with complexity when generating forecasts.

CST scoring is weighted, dynamic, and independent. It is *weighted* as each feature contributes to the score but some do so more than others. It is *dynamic* because the score is modified incrementally as series features are identified. As such, the score for a series with more features will have more modifications, logically

Decision Support Systems, Vol 83 (March 2016): pg. 70-82. DOI.

leading to a higher complexity score. Finally, it is *independent* as there is no starting point in the rule set i.e., as long as the same features are identified, two processes will arrive at the same score irrespective of which feature is considered first. These rules, presented in Appendix B, are discussed next.

4.2. Determination and application of the CST rules

A variety of approaches are available to infer rules. One could statistically infer rules from actual *data*, or generate rules from practitioner/expert *surveys*, *literature review*, and *collaborative scoring*, or from *existing rule sets* within the same domain. In using C&A's RBF as our developmental framework, we used the last approach to obtain initial rules for the CST from the RBF rule set. C&A presented RBF as an expert system consisting of 99 "IF…THEN…" rules that use judgment to combine forecasts from four statistical forecasting methods. Their 18 time series features were used to weight forecasts from these methods, yielding combined forecasts customized according to characteristics of the series. The rule below from C&A is representative of how the RBF rules were structured:

> **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.[e]

C&A relied on protocol analyses of experts, evidence from empirical literature, and comparison of forecasts from multiple benchmark methods to develop, refine, and validate RBF rules. These rules were subjected to subsequent validation in several studies.[43,44,45] RBF rules, then, captured forecasting best practices in a robust knowledge base and were a fitting starting point for identifying CST rules. The following approach was applied to generate CST rules:

4.2.1. RBF rules related to Random Walk

An initial set of CST rules were derived by adapting all RBF rules that shift weight to Random Walk from other component forecasting methods. Typically, this occurs under conditions of instability or uncertainty.[42] For instance, when causal forces are unknown or are known but conflict with basic and recent trend, RBF rules flag the

series as uncertain and shift emphasis to the Random Walk while reducing weights from other component methods. More significantly, the magnitude of such shifts varies by the nature of instability or uncertainty. For example, signal-related uncertainty, e.g. changing basic trend, leads to a greater shift towards the Random Walk as opposed to structural instabilities such as level discontinuity. RBF rules were converted to CST rules as follows. *For every original RBF rule that increased the weight on Random Walk and reduced from other component methods, a new complexity rule was created. For each CST rule, the complexity of a time series was reduced or incremented by 5*. For instance:

> ***Original RBF Rule 40***: IF *Causal Forces* are unknown, THEN add 5% to the weight on Random Walk and subtract it from that on Regression Trend estimate.

was modified to:

> ***Complexity Rule 1***: IF *Causal Forces* are unknown, THEN add – 5 [minus 5] to the Complexity Score of the series.[f]

Incrementing by 5 was judgmentally determined and may well have been 1 or 10. However, increments of 5 generated sufficient variation across series to facilitate separation of time series into simple and complex for later experimental validations. Furthermore, using consistent adjustments of 5 across all rules prevented unintentional biasing of the scoring system. Such equal weighting further enhances robustness of the schema through its uniform application while supporting the Occam's razor principle of simple over complex methods. Other scoring weights were tested but yielded outcomes that did not optimize on validations and, as such, were discarded. Specific results for these can be made available as necessary.

C&A developed rules for a *short model*, which generates 1-ahead forecasts, and a *long model*, which generates 6-ahead forecasts. Interim forecasts are produced by blending these two models using a set of rules (#s 97, 98, and 99). The two models are identical with regard to the features of interest in this study. Similarly, separate rules were developed for forecasting levels and trends. These rules mostly differed with regard to the weights assigned to component methods. As such, we did not develop separate rules for

short and long models or for level and trend forecasts, particularly as complexity was expected to affect both these estimates similarly. Appendix B presents the eight (CRules 1–8) complexity rules generated. Rules are organized around the three trait categories proposed by G&W.

4.2.2. Identification of additional complexity rules

The eight rules exposed gaps related to features known to contribute to uncertain or unstable conditions — functional form, short recent run (recent run not long), and coefficient of variation about the trend. This gap is further highlighted by the fact that C&A used these features in their rule set, not to assign weights to component methods but rather to transform the original time series or their forecasts. In doing so, C&A demonstrated these features to have implications for uncertainty or instability. As such, three additional rules were developed as follows.

a. Functional form

According to the transformation literature, specifically the Box–Cox family of transformations, the functional form of a series is multiplicative or additive based on trend-related motion. C&A also adopted this binary assignment. Essentially, two conditions identify a series as multiplicative:

- Sectional variation differences: The variation of the series is, by a particular section of the overall time series, related to the trend or level of the series. In the continuous, as opposed to discrete, case this suggests that there is a functional, dynamic, link between a series trend and its variation.
- Trajectory changes: The other condition for a multiplicative series is rapid growth or decay of the series. Consistent with C&A, we hesitate to use the term *exponential* growth or decay as sometimes this suggests testing for an exponential fit to rationalize the transformation.

Series that do not match the above conditions are considered additive by default, suggesting that forces act on the series in a way that they produce constant motion in either direction. C&A recommend the log (*ln*) transformations for multiplicative series as such

transformation damps the trajectory of the series, allowing for easier feature detection.

For illustration, Fig. 1 below presents two series and their log transformations. Series 106 was coded as additive by C&A as the conditions for an *ln* transformation were not evident for this series. Log transforming that series does not modify it sufficiently to improve feature identification or forecasting process, that is to say it maintains the same "noisy" profile. In contrast, series 86 was judged by C&A to be a multiplicative with a trajectory change around time period 9, making it a candidate for *ln* transformation. The transformation levels off the series, thereby simplifying feature identification, particularly with respect to causal forces underlying the generating process. Series coded as having an additive functional form were identified as more complex to forecast. As such, CRule 9 was inferred as:

> ***Complexity Rule 9***: If the *Functional Form* of a series is additive THEN add − 5 to the Complexity Score of a series.

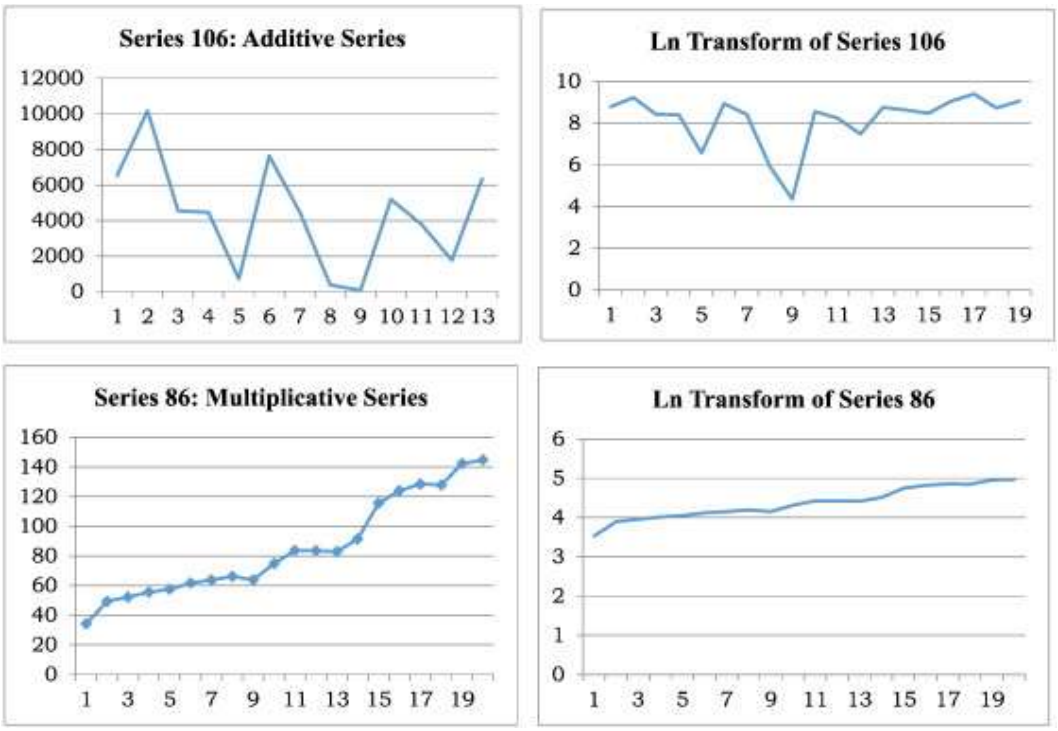

Fig. 1. Original and log transformation of additive and multiplicative series.

b. Long recent run

C&A identify a series as having a *long recent run* if the period-to-period movement for the last six observations is in the same direction. A long recent run suggests recent stability in the trajectory of the series without which historical patterns are not strong enough for accurate extrapolation. Although C&A's Rule 44 related to long recent run does not modify weights for the Random Walk, their empirical evidence suggested that a series lacking a long recent run would be more complex to forecast compared to one that has recent stability. The following rule was developed in response to this argument.

> ***Complexity Rule 10***: If the *Recent Run is Not Long* THEN add − 5 to the Complexity Score.

c. Variation about the trend

Coefficient of Variation (CV) about the trend represents standard deviation of the series divided by mean of its linear regression trend. The CV is used in RBF to estimate parameters for Brown's exponential smoothing and not to assign weights to component methods. C&A suggest that when there is a high degree of variation about the trend (CV > 0.9),[g] uncertainty is high. Considering this, one might infer that any rule developed for this feature should be designed to flag a series as being more complex to forecast. In other words, *if a rule for CV existed* in RBF, it might read as:

> IF the *Coefficient of Variation about the Trend* > 0.9, THEN increase the weight on Random Walk and reduced it from Linear Regression, Holt's and Brown's.

This, in fact, appears contrary to judgmental processes that would suggest fitting trend lines to simplify, maybe with satisficing outcomes, the forecasting process. Shifting the weight to Random Walk, in this case, would reproduce the erratic pattern reflected in the underlying series rather than projecting the underlying trend of the series, which may be more important for a series with high variation. As such, a rule opposite to the above would be more suited:

> IF the *Coefficient of Variation about the Trend* > 0.9, THEN *reduce* weight on Random Walk and shift it to Linear Regression, Holt's and Brown's Exponential Smoothing.

thereby suggesting the following complexity rule:

> ***Complexity Rule 11***: If *Coefficient of Variation about the Trend* > 0.9 THEN add + 5 to the Complexity score of the series.

This rule may initially seem anomalous as it reduces the complexity score when a series has high trend-related variation. However, the underlying assumption relates to the manner in which processing of extreme variation needs to be simplified by allocating weight to the trended methods and reducing it from Random Walk. This supports the satisficing adjustments that become necessary when there is no clear evidence that detailed feature decompositions will be effective in improving forecast outcomes.

d. Number of observations

Short series have insufficient observations to capture historical patterns needed to understand the series and, as such, are associated with higher forecast errors when compared to longer series. To identify the threshold that separates short and long series, we split the 72 series in the development data set along median number of observations and calculated error measures for short series, i.e. series with number of observations below the median, and long series i.e. those above. Specifically, Relative Absolute Errors (RAEs)[h] for RBF and Combining A were generated only for 6-period-ahead forecasts as effects were expected to be more pronounced for longer horizons than for shorter ones. The split along the number of observations was then iteratively lowered from the median until there was a marked lowering of the p-values for the error measures. This occurred at number of observations of < 13. As illustrated in Table 1 below, the p-values for the differences were relatively high for short and long series for the median split. However, for series with less than 13 observations, the error significantly increased as compared series with 13 or more observations. Based on this, the following complexity rule was developed:

Complexity Rule 12: If the *Number of Observations* in a series is < 13, THEN add − 5 to the Complexity score.

Table 1. Results from calibration on number of observations.*

	RBF 6-ahead RAE		Combining A — 6 yr. RAE	
	Median split	# of obs. < 13	Median split	# of obs. < 13
Short series	0.59	0.93	0.71	0.85
Long series	0.64	0.57	0.72	0.71
p-Values*	0.80	0.02	0.94	0.09

*The p-values reported are directional from Wilcoxon/Kruskal–Wallis Rank Sum test.

5. Results from development and validation

5.1. Results from development data set

The development data set series (72) were scored using the 12 CST rules. Based on the feature profile, an aggregate complexity score was generated for each series. An initial score of 0 was assigned to each series. The presence of a feature adjusted the score as prescribed by the rule. For instance, if an anomaly exists between basic and recent trends, 5 would be *deducted* from its complexity score (CRule 2). As rules most often subtracted from the score, most series had a negative complexity score. Series with the lowest negative scores were, then, most complex. Complexity scores ranged 45 units, from − 40 to + 5.[i] Scores for the 72 series demonstrated reasonable symmetry as there were no box-plot outliers for the scores, i.e., no values outside the ± 1.5 Tukey-whiskered inter-quartile range produced using the SAS/JMP v.10.2. Mean complexity score was − 8.4 and the median, − 5, further supporting the relative symmetry and internal validity as major outliers or marked asymmetry are concerns for most calibrations.

5.1.1. Partitioning development series by complexity

It is not the intent of the CST to prescribe a series as simple or complex but rather to generate a complexity score for each series. Such categorizations are domain-specific and require further research. However, some classification was necessary in order to determine

effectiveness of the CST and validate it using benchmark forecasting methods. For this, a simple partitioning of the series into two categories, simple and complex, was reasonable considering the foundational nature of this work. Although both mean and median could be useful in creating such partitions, the mean was preferred for several reasons. First, using the median as a classifier may bias validations based on median error measures (e.g. median RAEs). Second, as the complexity scores are reasonably symmetric, there are likely no classification differences whether using the mean or the median complexity score as a partition. Finally, in the sample of 126 series, only about 20 series had more than two instability causing features, 100 series had low uncertainty (CV < 0.2), and 22 had unknown casual forces. As such, a larger number of series were expected to be simple. Using the median for partitioning would split the sample equally and artificially create groupings at odds with the population profile. The mean rounded to the next whole unit of 5, then, was a better criterion. Specifically, *series with a complexity score equal to or higher than − 10 was coded as simple while those lower were coded as complex*. Given this threshold, in the development data, 23 series were classified as complex and 49 as simple, essentially yielding a 1/3–2/3 split as shown in Table 2.[i] This split in favor of simple series is consistent with empirical results from studies such as.[42,48]

Table 2. Profile of simple and complex series from development data set.

Complexity scores	Complex (n = 23)	Simple (n = 49)	p-Value for the difference
Mean	− 22.8	− 1.6	< 0.0001[Welch test]
Median	− 20.0	0.0	< 0.0001 [Wilcoxon Rank]
95% conf. interval	[− 25.9 to − 19.7]	[− 3.1 to − 0.2]	No CI overlap

Table 2 shows simple and complex series to be nearly symmetric as the mean complexity scores are close to the medians. Again, for the complex-simple split, there were no box-plot outliers. Additionally the 95% parametric confidence intervals are placed in the Cartesian coordinate space with a separation that is a multiple of the average widths of the intervals. In sum, all measures suggest that simple and complex series are significantly, and meaningfully, distinct.

Decision Support Systems, Vol 83 (March 2016): pg. 70-82. DOI.

5.1.2. Error measures for the development series

Armstrong & Collopy[50] recommend the use of multiple independent error measures to evaluate performance of forecasting techniques. In consideration, our results were assessed using two measures — RAEs and Absolute Percentage Errors (APEs). RAEs are the only measures used and reported in[50] and have evolved as definitive measures for judging forecasting models. Although RAEs are sufficient to provide validation for the CST,[51] following the best practice of using multiple error measures for completeness, results were also evaluated using APEs. The APE is inadequate as a sole measure for evaluating forecasting effectiveness as a low APE and a high RAE will usually disqualify the forecasting model under consideration. Following this, the RAE was used to assess effectiveness of the CST while the APEs provided secondary level of validation. As such, APEs are reported in Appendix C, Table A.

Benchmark comparisons were conducted with RBF and Combining A as RAEs and APEs were available for these methods from.[46] In C&A, both methods outperformed other benchmark methods. Furthermore, RBF outperformed Combining A. *It was expected that if the CST had captured complexity with good precision, forecast errors for complex series from RBF and Combining A would be higher than those for simple ones.* Tables 3 summarizes error measures for the 72 development series for RAEs. Table A in Appendix C provides results for APEs. Following recommendations from,[53] all error measures were winsorized using the following replacements: if RAE or APE is < 0.01 ➔ 0.01 or if RAE or APE is > 10 ➔ 10 for all *h*. Additionally, Wilcoxon Kruskal–Wallis Rank Sum test Chi^2 version for inference was used because outliers and asymmetries are still possible even though winsorizing bounds the data [0.01 and 10]. All measures reported were medians of winsorized errors for 1- and 6-ahead forecast horizons. Finally, for p-values, all tests consistent with the *a-priori* directional effects are one-tailed and shaded in Table 3 and Table A in Appendix C. Unshaded p-values are two-tailed.

Decision Support Systems, Vol 83 (March 2016): pg. 70-82. DOI.

Table 3. Median RAEs for development data set.*

Benchmark methods	1-Period ahead			6-Period ahead		
	Complex (n = 23)	Simple (n = 49)	p-Value	Complex (n = 23)	Simple (n = 49)	p-Value
Combining A	1.08	0.60	0.0421	1.08	0.65	<0.0001
RBF	0.85	0.33	0.0026	0.92	0.41	<0.0001

*All p-values are directional one-tailed tests.

Results showed lower forecast accuracy for complex series on both RBF and Combining A for 1- and 6-year ahead forecasts. These results are compelling as neither of the benchmark methods are pure judgment and, as such, are free from human bias and inefficiencies derived from complex tasks. The more immediate interpretation of the results, however, is that on the development series, CST rules generated a classification of simple and complex time series tasks that produce the expected accuracy profiles.

5.2. Results from validation data set

Next, effectiveness of the CST was assessed on the 54 series held back as the validation data set. The 12 CST rules were applied to this set *with no modifications*. Additionally, the same cutoffs as used in the development data set were used to segregate simple series from complex, i.e., series with complexity score between + 5 and − 10 were categorized as simple while those lower than or equal to − 15 were coded as complex. Using these parameters, 22 series in the test sample were classified as complex and the remaining 32 as simple. This 1/3–2/3 split is consistent with the development data set.

Benchmark comparisons for the validation data set were conducted across a larger set of methods. First, similar to the development data set, both RBF and Combining A were part of the benchmarks. Second, established forecasting methods, specifically the Random Walk (or Naïve), Holts' exponential smoothing, and OLS Linear Regression models were added for a more robust validation. There is no support for the belief that the Random Walk produces a consistent directional split for the APE. However, the same cannot be said for the Linear Regression or Holt's as these models are two parameter models and, unlike the Random Walk, are parameterized from the entire data set and not merely from the last observation. As such, it is plausible that performance of these benchmarks could differ

with respect to complexity. To explore this aspect, we proffer following hypotheses for additional non-RBF validations.

> **H1.** Median RAEs for RBF forecasts will be higher for complex series as compared to simple series on 1- and 6-period-ahead horizons.
> **H2.** Median RAEs for forecasts from Combining A will be higher for complex series as compared to simple series on 1- and 6-period-ahead horizons.
> **H3.** Median RAEs for forecasts from OLS Regression will be higher for complex series as compared to simple series on 1- and 6-period-ahead horizons.
> **H4.** Median RAEs for forecasts from Holt's exponential smoothing will be higher for complex series as compared to simple series on 1- and 6-period-ahead horizons.

Table 4 summarizes findings related to the above hypotheses. Table B Table 4 summarizes findings related to the above hypotheses. Table B in Appendix C provides related hypotheses and results for APEs. Again, results are consistent with the a-priori directional expectations. For H1, there is strong and consistent evidence that for both horizons, complex series are more challenging to forecast, even when using an extensive knowledge-based system such as RBF or a composite of methods (Combining A). All results presented in the tables confirm that the CST provides a robust and sensitive schema for scoring the complexity of time series.

Table 4. Median RAEs for holdback series on all benchmarks.*

Horizons	Benchmark methods	Complex series (n = 22)	Simple series (n = 32)	p-Values
All horizons	Random Walk	N/A	N/A	N/A
	Linear Regression	1.25	0.53	.0004
	Holt's	0.78	0.36	< .0001
1-Period horizons	Random Walk	N/A	N/A	N/A
	Linear Regression	1.85	0.90	0.058
	Holt's	1.18	0.19	0.017
	Combining A	0.79	0.61	0.121
	RBF	0.99	0.47	0.006
6-Period horizons	Random Walk	N/A	N/A	N/A
	Linear Regression	1.28	0.42	0.017
	Holt's	0.80	0.27	0.025
	Combining A	0.66	0.65	0.427
	RBF	0.86	0.43	0.045

Decision Support Systems, Vol 83 (March 2016): pg. 70-82. DOI.

*As obtained from Collopy [46] and C&A.

For H2, H3 ; H4, related to Combining A, OLS Regression and Holt's, hypotheses were only developed for RAEs as it is our primary error measure. However, all APE results are presented for completeness in Table B (Appendix C). Results are as expected across all three benchmarks. Median RAEs for all confirm that forecast accuracy for complex series is worse than for simple ones. These forecasting methods are each unique in terms of the underlying generating processes, not merely from each other but also from RBF. Specifically, both the individual models such as OLS, Holt's, and Random Walk as well as combined models i.e., Combining A and RBF, provide independent confirmatory evidence for effectiveness of the CST as well as evidence that complexity impacts forecasting practice. Overall, results for the RAEs are definitive — the CST produces a usable technique for scoring series complexity based on the general expectation that simple series are less demanding than complex ones**.**

6. Judgmental validation of CST — preliminary evidence

The CST was further validated using an experiment that asked forecasters to produce *structured judgmental forecasts* for series classified by the CST as simple or complex. The process was structured as participants judgmentally applied knowledge from RBF to generate forecasts for assigned series. As such, the forecasts were not generated using pure judgment but rather by blending judgment with statistical methods, a best practice supported by the judgmental forecasting community. For simplicity, this approach is referred to as "judgmental" hereon. Note that forecasters were not asked to assess series complexity, only to generate forecasts. In fact, they were unaware of any complexity classifications.

The experiment was conducted with 14 participants and was designed to address the question — *Do judgmental forecasts for simple and complex series*, *as scored by the CST*, *follow the hypothesized pattern i.e., lower accuracy for complex series*? The intent was not to provide insights into judgmental forecasting of complex series but to provide alternate confirmatory evidence

validating the CST. The study was conducted after parameters of the CST were finalized and results were confirmed on the validation data set. Participants were advanced undergraduate students enrolled in a Business Forecasting course in Germany. They were trained in general forecasting knowledge and best practices, as captured in a simplified set of rules and features from RBF, and component forecasting methods in RBF. The average age was 22 and the gender mix was about 1/3 females and 2/3 males, typical for the gender mix in the program. All participants had the Excel™ and statistical skills to complete the experimental task.

Twelve series from the validation data set described in previous sections were randomly selected, six each from the complex (series 14, 27, 28, 37, 48, 177) and simple (series 54, 64, 104, 134, 138, 144)[k] sets. Table 5 provides the complexity profile for these series. The 12 series were quasi-randomly assigned to the 14 participants such that each participant received two simple and two complex series. They were to produce 1- to 6-period ahead forecasts for each assigned series, yielding 336 [14 × 4 × 6] forecasts. Series assignments were adjusted to provide a 50/50 allocation of simple and complex series across the group. Both series allocations and forecast generation were conducted on the last day of the course.

Table 5. Complexity profiles for series used for judgmental validation.

Complex series		**Simple series**	
Series #	**Complexity score**	**Series #**	**Complexity score**
14	− 25	54	− 5
27	− 20	64	− 5
28	− 30	104	− 5
37	− 40	134	− 5
48	− 20	138	0
177	− 25	144	− 10

Series assignments were controlled for order effects by first giving complex series to seven participants and simple to the remaining. Once these initial forecasts were delivered, the order was reversed. This created two test groups: Group I (n = 7): [complex, simple] and Group II (n = 7): [simple, complex], producing a total of 336 forecasts, 168 each for simple and complex series. Controls for order effects were also factored. If, for example, Participant 1 was

paired for group work during the course with Participant 2, Participant 1 received [simple, complex] and Participant 2 received [complex, simple].

All participants used a Visual Basic™/Excel based DSS to aid the forecasting process. This DSS is available from the authors without restriction on use. As has been practice in this course, participants had dedicated time to apply knowledge learned through the course to produce forecasts in the classroom, a computer lab. To do so, 2½ h was dedicated in the morning session followed by a mandatory break and a second session of 2½ h. Additional time was offered but was not used by any participant.

6.1. Results from structured judgmental validation

Forecast accuracy of participants was evaluated using winsorized Median RAEs and Median APEs as for earlier validations. For inference purposes, we used the Wilcoxon/Kruskal–Wallis Rank Sum Test, specifically the Chi^2 version as programmed in SAS/JMP, v.10. Grade effects were also tested to determine whether students scoring in the top half of the assigned grades were of different caliber than students those in the lower half. Forecast errors, and errors by order and grade effects, are reported in Table 6 along with appropriate p-values.

Table 6. Median RAE and APE for judgmental forecasting results.

Series blocks	Judgmental forecasts		Order effects test		Grade effects test	
	Complex series	**Simple series**	**Simple: Complex**	**Complex: Simple**	**< Median**	**> Median**
Median RAE	1.22	0.61	0.860	0.831	0.796	0.898
p-Value	p-Value < 0.0001		p-Value 0.1799		p-Value 0.5824	
Median APE	0.119	0.106	0.114	0.109	0.108	0.114
p-Value	p-Value 0.04625		p-Value 0.3390		p-Value 0.6286	

*The sample size for each test block is 168.

Table 6 provides confirmatory evidence on effectiveness of the CST based on judgmental forecasting. Tangentially, the results also

Decision Support Systems, Vol 83 (March 2016): pg. 70-82. DOI.

provide preliminary evidence on effects of complexity on judgment as median RAEs for complex series are nearly twice those for simple series. Interestingly, using the Wilcoxon Signed-Rank as a directional test for Median RAE of 1.0 for the population, a Median RAE of 1.22 for complex series suggests that, when forecasting complex series, participants *did worse* than if they had just used Random Walk to produce forecasts. The related test of APE shows similarly significant results though the separation between error measures for simple and complex series is less profound as compared to RAEs. Table 6 also shows no evidence of order or grade effects. Finally, each participant was given the option to select any *one* series the difficulty of which was such that they felt the least confident in their forecasts. All series identified were complex. The p-value of this is < 0.0001, confirming that even for the recently trained, complexity is both recognizable and challenging.

7. Implications for FDSS and judgmental forecasting

Although the CST is preliminary, it is a crucial first step. Its implications are numerous, in particular for design of FDSS and for research in judgmental forecasting. These are summarized in Table 7 below.

Table 7. Summary of research opportunities related to CST and complexity.

ID	Research need	Domain
1	How will presentation of complexity information to forecasters influence their forecasting strategies and process?	FDSS Judgment
2	What adjustments need to be made to the CST to allow for short period (quarterly, monthly, weekly, hourly) data?	CST
3	What decomposition strategies are most suited to simple and complex time series tasks?	Judgment Forecasting process
4	To what extent do informative and suggestive guidance benefit and enhance forecaster strategies and mental models?	Judgment FDSS
5	In what ways do interface characteristics enhance or harm forecaster effectiveness on simple as opposed to complex time series tasks?	FDSS Judgment
6	What design and human factors must be considered for optimally identifying and presenting time series features to	FDSS Judgment

ID	Research need	Domain
	forecasters? For instance, could big-data analytics be used to develop and visualize time series features?	
7	What sort of guidance and feedback are most beneficial for simple and complex tasks?	FDSS
8	How do these forms of guidance influence forecaster mental models and strategies?	Judgment
9	Can specifying confidence intervals for simple and complex tasks in FDSS design direct forecasters towards better adjustment practices?	FDSS
10	How does prolonged use of confidence intervals for simple and complex tasks modify adjustment behaviors?	Judgment
11	How does judgmental adjustment of simple series impact forecast accuracy as opposed to similar adjustment of complex series?	Judgment
12	Are FDSSs uniformly useful for supporting simple and complex tasks? If not, what capabilities are necessary for optimally supporting both?	FDSS
13	Do forecaster perceptions of the complexity of a time series align with those suggested by the CST?	CST
14	What additional rules might improve the efficacy of the CST?	CST
15	What additional features and feature combinations might improve the efficacy of the CST?	CST
16	Can the CST be delivered as effectively with fewer rules and features?	CST
17	Can integrating magnitude of features (e.g. level discontinuity) enhance CST efficacy?	CST

7.1. A framework for adaptive forecasting decision support systems

This section elaborates on DSS enrichments possible through integration of CST, specifically the design of adaptive FDSS (AFDSS) that responds to time series complexity. Our intent is not to provide a technical design of AFDSS components, as has been done in.[5,10] Rather, considering the scope of this study, we focus on how the CST could feed into specific components of AFDSS. In essence, we specify the broad frameworks proposed by earlier studies to the context of FDSS.

Fig. 2 presents a conceptualization of FDSS built upon four well-established phases of decision making and support — (i) problem recognition, (ii) solution formulation and rationalization of the proposed solution, (iii) implementing actions from alternative sets, and (iv) evaluating the realized outcomes [52]. These elements explicitly integrate forecasters' organizational, domain, and technical expertise with FDSS use and outcomes. The model suggests that the forecasters' cognitive mapping shapes, and is shaped by, their interpretation and knowledge of forecasting tasks. This determines how forecasters interact with the task, data, and analytical models when approaching the solution space. Forecasters' domain knowledge and conceptual decomposition paradigm coupled with FDSS guidance play a crucial role in evaluating and selecting alternatives. Finally, in an ideal design, forecasters' mental models can mature through active reflection on outcomes, FDSS feedback, and reformulation of the problem domain as necessary. The next few sections elaborate on the left side of this figure i.e., how the CST can enhance this experience by enabling adaptiveness in FDSS.

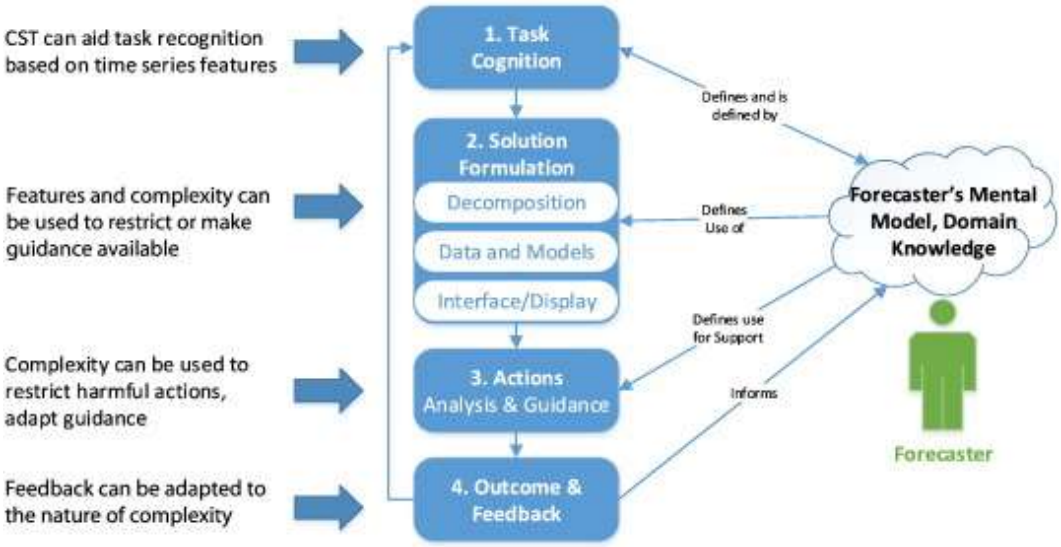

Fig. 2. A conceptual model for adaptive FDSS design.

7.1.1. Automate task cognition

The CST offers a framework for a feature-based approach to task cognition. As a time series is input into the FDSS, automated feature detection routines, such as those described in C&A and,[44] can

categorize time series based on complexity. This information about series complexity and its drivers could be made available to forecasters to allow them to draw upon relevant knowledge and strategies for forecast generation. This, however, raises interesting questions about forecasters' response to complexity-related information in early stages of the forecasting process. On the one hand, such information may enable forecasters to focus cognitive resources on relevant factors, but on the other, it may bias the judgmental processes, as through unwarranted observer effects.[47] In executing such studies, then, care must be taken against biasing effects that run the risk of removing the forecaster's expertise from the process, something a well-designed DSS should prevent. Specific FDSS design elements that can positively focus cognitive resources and de-bias the process require exploration and testing. After determining a series to be more likely complex or simple, the FDSS could use the underlying series information to provide guidance on possible actions. For instance, it is empirically shown that in light of changing basic trend, forecasters often place more emphasis on smoothing methods such as Holt's and Brown's. This guidance can be made available to forecasters. Steering the DSS design process from conditioning to helpful guidance is the goal — a challenge for designers.[9]

7.1.2. Restrict or expand solution formulation based on time series complexity

Time series profiles can be used to adaptively restrict or expand forecasters' cognitive model during solution formulation. Restrictiveness is the "degree to which, and the manner in which, a DSS limits its users' decision-making process to a subset of all possible processes."[8] (pg. 52). The following three aspects of the solution space can be adapted to complexity (see[12] for an excellent review of DSS restrictiveness):

> *Support task decomposition according to complexity*: Our working memories are limited[53] and, as such, complex tasks broken into simple "chunks" are more effectively executed when compared to tasks not simplified thus.[45] Decomposition is found to improve performance over unaided and intuitive judgment[54] by breaking down a complex task into sets of easier tasks that

Decision Support Systems, Vol 83 (March 2016): pg. 70-82. DOI.

are more accurately executed than the holistic task.[12] Cognitive and information overload can be controlled by providing greater structure to the environment[55] through decomposition strategies that simplify the domain.

Although decomposition can be argued as being restrictive when its use is forced upon the decision maker,[12] most often, an FDSS user may not focus on the benefits of task decomposition nor recognize how to proceed with it. To this end, we suggest that decomposition be implemented in both restrictive and decisional guidance mode. Specifically, we use the framework by[45] who suggest that decomposition can be applied at three levels: *transformation* of problem space using characteristics of the forecasting task and domain; *simplification* of process, i.e., decomposing and understanding components of the forecasting process from problem formulation to forecast use;[56,57] and *decomposition for method selection* i.e., applying forecasting knowledge and rules to selecting fitting methods.

Transformation should be a restrictive feature in FDSS. The decomposition of time series into its features, when combined with effective displays, can enhance forecaster's ability to recognize meaningful patterns as opposed to random ones. So should be the case for *simplification* which could restrict early convergence on use of specific forecasting methods without adequate analysis and problem formulation. Finally, *method selection* could be implemented as decisional guidance. Users may be prompted with forecasts from multiple relevant methods, e.g. using RBF rules, to consider use of alternative methods and combining. Furthermore, suggestive guidance to on how to proceed with method selection and combination could be useful for simpler tasks.

Restrict action on data and models according to task complexity: FDSS can make some processes easy to use while making other, less desirable ones, more challenging. Restrictiveness may be relaxed for simple tasks by increasing the range of available data and models. For instance, forecasters tend to replace missing or erroneous data with their own estimates rather than using estimates from quantitative methods.[12] Such adjustments can be restricted, particularly when series are complex and domain knowledge is weak. Automating and, thereby simplifying, the application of ideal strategies can reduce effort associated with executing the more desirable ones[56] and tendency to make damaging adjustments.[58]

Although restricting the range of models available to forecasters may be unwarranted and frustrating, under certain conditions when some forecasting models consistently underperform, FDSS may be designed to restrict availability of those models, especially for simple series. In contrast, a wider range of methods may be made available for complex series to support combining, which has demonstrated value in enhancing forecasting accuracy. In such cases, the success rate of specific methods on analogical series may improve choice of methods to be combined.

Adapt FDSS display to task complexity: Because simple tasks create lower cognitive strain,[29] performance on such tasks can be improved by increasing user awareness of forecasting cues, such as by displaying features underlying the time series, forecasts from component methods, and the forecasting process generating final outcomes. For instance, making available the long-term trend of a time series improves accuracy as it allows forecasters to overlook distracting patterns and apply knowledge consistently.[59] Because decision makers tend to trade off accuracy in favor of cost efficiency,[60] *informative* and *suggestive* guidance could be displayed for simple series such that the forecaster need not drill down to make satisficing decisions. As simple tasks impose less cognitive strain on forecasters, the processing of such displays will be less intrusive. In contrast, FDSS displays for complex tasks can be restricted because this same information presented to the forecaster can result in greater cognitive overload, strain, and over-reaction. Indeed, in complex task settings, decision makers ignore suggestive advice and focus on informative guidance.[61] To reduce such cognitive overload, information for complex tasks could be made available as layered, drill-down options. Such adaptive support can reduce information overload and related information processing challenges in the context of complex tasks.[62]

7.1.3. Provide in-task guidance for simple tasks and post-task guidance for complex ones

Decisional guidance is "the degree to which, and the manner in which, a DSS guides its users in constructing and executing the decision-making processes by assisting them in choosing and using its operators"[8] (pg. 57). Guidance and feedback promote learning and behavior modification with the assumption that organizational

practices encourage such review. Broadly speaking, guidance can be offered to forecasters at two stages – *during* and *post task execution* – the former being critical to outcome accuracy and the latter as beneficial for fostering learning.[12] Forecasters facing complex tasks do not have the time and cognitive resources to reflect adequately upon the impact of their actions on the forecasting environment[63] and consequently, may fail to consider control actions. However, extensive feedback during execution of complex tasks can worsen information overload and frustrate users. As such, FDSS designers may benefit from focusing on post-execution feedback for complex tasks which improves decision quality[64] and attainment of challenging goals. Holistic learning is possible, for instance, by supporting informative guidance with the ability to drill down to the components.

Simple tasks, in contrast, are cognitively less demanding and do not require the same level of feedback and support as complex tasks. Consequently, in-task feedback may be less detrimental and could be designed to guide the user, for example by displaying features of time series and discussing their impact on forecasts, providing original series contrasted with series that have been cleansed of distracting features such as outliers and irrelevant early data, and by providing guidance in form of rules and relevant methods. As a case in point, RBF rules pertaining to a specific set of features could be displayed such that the user can recognize the knowledge that has gone into generating the forecast.

7.1.4. Adapt outcome-related flexibility based on complexity

Outcomes from FDSS are often adjusted to accommodate forecaster's domain knowledge as well as enhance ownership of outcomes. However, not all such adjustments improve outcomes.[1] Two recommendations are proposed.

> *Restrict where harmful judgment can be applied*: When unrestricted, forecasters are free to apply judgmental adjustments at many levels in the forecasting process such as towards data to be used or excluded, models to be applied or ignored, and changes to decision outcomes, even when undesirable.[12] While, on an average, such adjustments improve accuracy, studies have found specific circumstances in which these can be harmful. For instance, in their examination of over

60,000 forecasts,[65] found that small adjustments, and those that are optimistic, are less likely to improve forecast outcomes. Few studies, however, tie these findings specifically to complexity. In ongoing extensions of this study, participants make smaller, positive adjustments to simple series as opposed to more complex ones. In doing so, they end up harming accuracies of simple tasks as opposed to complex ones. As such, while[65] did not make an explicit link between series complexity and the nature of adjustments, preliminary evidence from our studies suggests the potential. Assuming that such linkage exists, FDSS can restrict harmful adjustment based on complexity drivers and also guide forecasters to specific forecasting processes where adjustments may be beneficial e.g. adjustments to data and models as opposed to outcomes.

Restrict to impose standards and best practices: Restrictions can be applied when organizational best practices and standards need to be supported in the forecasting process. For instance, a critical issue in supply chain forecasting is the bullwhip effect of adjustments as a forecast moves down the supply chain.[66] Overly optimistic and large adjustments for simple series, for example, can continue to get compounded along the supply chain. Embedding practices that constrain the magnitude and directionality of adjustments in FDSS may potentially reduce risks associated with overcompensating for each link in the supply chain. These restraints may be in the form of boundaries or confidence intervals defined by the nature and complexity of series being presented to the forecaster. This is particularly true for complex series where forecasters may overemphasize random patterns in the data or for simple series where forecasters may want to overcompensate for seemingly aggressive forecasts.

7.2. Implications for judgmental forecasting

The use of specific time series features in the CST expands opportunities for studying individual and cumulative effects of series features on information processing behaviors of forecasters and for executing condition analysis.[42,44] Similarly, series complexity should impact adjustment behaviors. We have observed in ongoing studies that judgmental adjustment of complex series seems not to harm forecast accuracy to the same extent as that of simple ones and could, in fact, improve accuracy for complex series. These preliminary

findings need further examination. FDSS studies tend to interlace task needs and technology capabilities in ways that make it a challenge to disengage the two. The CST provides a way of untangling the two and promoting a detailed examination of factors such as trust in forecasts,[67] organizational and individual use of forecasts, and adjustment behaviors under varying conditions.[68]

This study raises relevant, and perhaps troubling, questions about meaningful use of FDSSs and judgment for varying complexity levels. Many studies (e.g.[69]) suggest that DSSs are better for uncertain and complex tasks while human-centric approaches may be preferable for simple but equivocal and ambiguously defined tasks. One might question whether, at some point, complexity cannot be meaningfully dealt with by FDSS and requires greater forecaster intervention instead. In a similar vein, certain forecasting tasks may be simple enough that any judgmental intervention could destabilize accuracy. Might an inverted-U curve relationship exist between task complexity and forecast accuracy where complexity impacts judgmental processes and FDSS effectiveness positively up to a point but eventually, proves detrimental beyond? Moving forward on this frontier may be challenging but necessary to rationalize commitment of resources to support specific methods or FDSS.

7.3. Considerations for enhancing and evolving the CST

The CST is a first index of its kind. Its development is embedded in design science research with the intent of refining the IT artifact[9] to solve pragmatic forecasting problems. As such, it will likely be a launching point for further research leading to refinements of our results. Most significantly, the complexity schema presented here is not defined around a particular domain, presenting numerous opportunities for domain-specific customization. Specifically, the twelve rules presented in Appendix B may benefit from domain-based calibrations, such as by modifying weights on specific rules or removing some rules altogether. For instance, domains that rely on recent consumer trends may find *CRule 10* to be less relevant than more stable domains such as demographic forecasting. Similarly, public-utilities demand forecasting may find level discontinuities to be more destabilizing and prefer to increase the complexity score for that

rule. Forecasters from specific domains may also consider developing and calibrating rules for features prevalent in their industries. For instance, natural gas demand forecasters may prefer to give consideration to outliers as these often represent unusual demand days that providers may want to proactively model rather than suppress.

The features represented in the twelve complexity rules are limited to 14. Since C&A categorization, these have been expanded to 28 features by including features for seasonality and forecast horizon.[52] Future research might explore the role of these additional features in improving precision of the CST and to develop and calibrate related rules. Other features may be considered for exclusion or more sophisticated representation in the schema. For instance, neither the C&A nor the CST rules consider contribution of the *magnitude* of an instability feature towards increasing the complexity of a series. A series with a small level discontinuity, for example, may be easily overlooked, both judgmentally and statistically, as compared to one with larger magnitude. The possibility of moving from a binary feature set to a scaling measure may allow for more contextual application of the CST. Whether this approach leads to significant gains in efficacy remains to be determined.

Are series classified as simple or complex by the CST perceived similarly by forecasters? Currently, our evidence is anecdotal and based on casual observation. For instance, for several participants, we observed extra periods of hesitation and increased eye movement for complex series but not for simple ones. Future research can formally capture such biological interpretations of complexity using techniques from biological sensors such as eye trackers to self-reported measures of difficulty around the series. Finally, the CST is based on RBF that was originally developed, calibrated, and validated on annual time series. While we presume that the CST rules would apply similarly to shorter period data such as monthly or quarterly series, feature and weight calibrations will be necessary for short period series as the underlying generating processes will, in all likelihood, be different.

The CST presented herein is a stable, validated, robust, and fully disclosed technique that invites the possibility of creating AFDSSs that respond to the forecasting environment based on task complexity.

Fully disclosing development of the CST provides opportunities for further validation and refinement at many levels. Implications for judgmental forecasting and AFDSS design are numerous and, as such, the CST seeds a new stream of forecasting research on series complexity and supporting processes.

Acknowledgments

The authors wish to thank Prof. Fred Collopy, *Case Western Reserve*, Cleveland, Ohio, USA for his guidance and for unrestricted access to his RBF datasets. We also thank Prof. J. Scott Armstrong, The *Wharton School*, University of Pennsylvania, for his guidance. Our appreciation also extends to participants at the 2014 International Symposium on Forecasting, Rotterdam and those at SBE Research Workshop at SUNY: Plattsburgh for their discussions of complexity in forecasting. Finally, we thank the reviewers for their rigor and articulation of recommendations that have significantly benefited this research.

Appendix A. Features of time series from C&A used in the CST

Feature categories	C&A features used	Description of feature[a]
Instability[b]	Suspicious pattern	Series that show a substantial change in recent pattern.
	Unstable recent trend	Series that show marked changes in recent trend pattern.
	Recent run not long	The last six period-to-period movements are not in same direction.
	Near a previous extreme	A last observation that is 90% more than the highest or 110% lower than lowest observation.
	Changing basic trend	Underlying trend that is changing over the long run.
	Level discontinuities	Changes in the level of the series (steps)
Uncertainty	Coeff. of variation about the trend > 0.2	Standard deviation divided by the mean for the trend adjusted data.
	Direction of basic trend[c]	The direction of the trend (up or down) as identified by fitting linear regression to the historical series.
	Direction of recent trend	Direction of the trend that results from fitting Holt's exponential smoothing to the historical series.
Trend	Significant basic trend	The t-statistic for linear regression is greater than 2.
Cycles expected	Cycles	Regular movement of the series about the basic trend.

Feature categories	C&A features used	Description of feature[a]
Domain knowledge	Causal forces	The net directional effect of the principal factors acting on the series. *Growth* exerts an upward force. *Decay* exerts a downward force. *Supporting* forces push in direction of historical trend. *Opposing* forces work against the trend. *Regressing* forces work towards a mean. When uncertain, forces should be *Unknown*.
	Functional form	Expected pattern of the trend of the series. Multiplicative and Additive functional forms were considered.
Length of series	Number of observations[d]	Number of observations in the series, not including the holdout data.

[a]Adapted from C&A and Forecasting Principles site — http://forecastingprinciples.com/index.php/features-of-time-series.

[b]Outliers and unusual last observation were additional instability features used in C&A. However, these were not considered in this study as these features were assumed to be adjusted prior to the forecasting process.

[c]Note that uncertainty occurs when the basic and recent trends are not in the same direction.

[d]Not an original C&A feature.

Appendix B. CST rules

Characterizations as in Goodwin & Wright[41]	Complexity rules related to characterizations
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	*There may be sudden changes to a new underlying mean level* (*steps*), *gradual changes to new levels* (*ramps*), *or a trended series might exhibit reversals in trend* etc. *CRule 3*: IF *Recent Trend* is unstable, THEN add − 20 to the Complexity score.

Decision Support Systems, Vol 83 (March 2016): pg. 70-82. DOI.

Characterizations as in Goodwin & Wright[41]	Complexity rules related to characterizations
	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.

Appendix C. Results and discussion from ape comparisons

Table A. Median APEs for development data set.*

Benchmark methods	1–Period ahead			6–Period ahead		
	Complex (n = 23)	Simple (n = 49)	p–Value	Complex (n = 23)	Simple (n = 49)	p–Value
Combining A	0.061	0.030	0.0342	0.26	0.16	0.0834
RBF	0.059	0.018	0.0049	0.24	0.10	0.0083

*All p-values are directional one-tailed tests.

For the validation data set:

Hypothesis.

Median APEs for RBF forecasts will be higher for complex series as compared to simple series on 1- and 6-period-ahead horizons.

Median APEs are directionally consistent for RBF though errors for complex and simple series are not as divergent as they are for median RAEs. The odds for rejecting the APE nulls for H2 are non-trivial and confirm support for the strong complexity differentials using the RAE. Specifically, APEs suggest that only about a third of the time such median results could be randomly drawn from the population where simple and complex errors are not different.

Table B. Median APEs for holdback series on all validity testing benchmarks.*

Horizons	Benchmark methods	Complex series (n = 22)	Simple series (n = 32)	p-Values
All horizons	Random Walk	0.09	0.19	< .0001
	Linear Regression	0.11	0.12	0.080
	Holt's	0.07	0.08	0.557
1-Period horizons	Random Walk	0.03	0.06	0.241
	Linear Regression	0.03	0.05	0.382
	Holt's	0.05	0.03	0.349

Decision Support Systems, Vol 83 (March 2016): pg. 70-82. DOI.

Horizons	Benchmark methods	Complex series (n = 22)	Simple series (n = 32)	p-Values
6-Period horizons	Combining A	0.03	0.04	0.923
	RBF	0.03	0.02	0.307
	Random Walk	0.09	0.31	0.004
	Linear Regression	0.13	0.16	0.316
	Holt's	0.08	0.11	0.397
	Combining A	0.06	0.21	0.012
	RBF	0.05	0.13	0.298

*As obtained from Collopy[46] and C&A.

References

[1]M. Lawrence, P. Goodwin, M. O'Connor, D. Önkal. Judgmental forecasting: a review of progress over the last 25 years. *International Journal of Forecasting*, 22 (3) (2006), pp. 493–518

[2]N.R. Sanders, K.B. Manrodt. Forecasting software in practice: use, satisfaction, and performance. *Interfaces*, 33 (5) (2003), pp. 90–93

[3]D.L. Goodhue, R.L. Thompson. Task-technology fit and individual performance. *MIS Quarterly*, 19 (1) (1995), pp. 213–236

[4]C. Smith, J. Mentzer. Forecasting task-technology fit: the influence of individuals, systems and procedures on forecast performance. *International Journal of Forecasting*, 26 (1) (2010), pp. 144–161

[5]B. Fazlollahi, M.A.M.A. Parikh, S. Verma. Adaptive decision support systems. *Decision Support Systems*, 20 (4) (1997), pp. 297–315

[6]D.M. Lamberti, W.A. Wallace. Intelligent interface design: an empirical assessment of knowledge presentation in expert systems. *MIS Quarterly*, 14 (1) (1990), pp. 279–311

[7]G. Bhandari, K. Hassanein. An agent-based debiasing framework for investment decision-support systems. *Behaviour and Information Technology*, 31 (5) (2012), pp. 495–507

[8]M.J. Silver. Decision support systems: directed and non-directed change. *Information Systems Research*, 1 (1) (1990), pp. 47–70

[9]A. Hevner, S. Chatterjee. *Design Research in Information Systems: Theory and Practice*. Springer, New York, NY (2010)

[10]S. Piramuthu, M.J. Shaw. Learning-enhanced adaptive DSS: a design science perspective. *Information Technology and Management*, 10 (1) (2009), pp. 41–54

[11]P. Paranagama, F. Burstein, D. Arnott. ADAPTOR: a personality-based adaptive DSS generator, in syst. Sciences. *Proceedings of the Thirty-First Hawaii International Conference5* (1998), pp. 54–63

[12]R. Fildes, P. Goodwin, M. Lawrence. The design features of forecasting support systems and their effectiveness. *Decision Support Systems*, 42 (1) (2006), pp. 351–361

[13]J.W. Labadie, D.G. Fontane, J.H. Lee, I.H. Ko. Decision support system for adaptive river basin management: application to the Geum River basin, Korea. *Water International*, 32 (3) (2007), pp. 397–415

[14]M. Lawrence, W. Sim. Prototyping a financial DSS. *Omega*, 27 (4) (1999), pp. 445–450

[15]C.W. Holsapple, R. Pakath, V.S. Jacob, J.S. Zaveri. Learning by problem processors: adaptive decision support systems. *Decision Support Systems*, 10 (2) (1993), pp. 85–108

[16]B.L. Dos Santos, C.W. Holsapple. A framework for designing adaptive DSS interfaces. *Decision Support Systems*, 5 (1) (1989), pp. 1–11

[17]D. Arnott, P. O'Donnell. A note on an experimental study of DSS and forecasting exponential growth. *Decision Support Systems*, 45 (1) (2008), pp. 180–186

[18]J.W. Payne, J.R. Bettman, E.J. Johnson. *The Adaptive Decision Maker*. Cambridge University Press, Cambridge (1993)

[19]S.E. Bonner. A model of the effects of audit task complexity. *Accounting, Organizations and Society*, 19 (3) (1994), pp. 213–234

[20]K. Byström, K. Järvelin. Task complexity affects information seeking and use. *Information Processing and Management*, 31 (2) (1995), pp. 191–213

[21]D.J. Campbell. Task complexity: a review and analysis. *The Academy of Management Review*, 13 (1) (1988), pp. 40–52

[22]C. Speier, I. Vessey, J.S. Valacich. The effects of interruptions, task complexity, and information presentation on computer-supported decision-making performance. *Decision Sciences*, 34 (4) (2003), pp. 771–797

[23]H. Timmers, W.J. Wagenaar. Inverse statistics and misperception of exponential growth. *Perception and Psychophysics*, 21 (6) (1977), pp. 558–562

[24]P.B. Andreassen, S.J. Kraus. Judgmental extrapolation and the salience of change. *Journal of Forecasting*, 9 (4) (1990), pp. 347–372

[25]M. Lawrence, S. Makridakis. Factors affecting judgmental forecasts and confidence intervals. *Organizational Behavior and Human Decision Processes*, 43 (2) (1989), pp. 172–187

[26]M. O'Connor, W. Remus, K. Griggs. Going up–going down: how good are people at forecasting trends and changes in trends? *Journal of Forecasting*, 16 (3) (1997), pp. 165–176

[27]F. Bolger, N. Harvey. Context-sensitive heuristics in statistical reasoning. *The Quarterly Journal of Experimental Psychology*, 46 (4) (1993), pp. 779–811

[28]N. Harvey, F. Bolger. Graphs versus tables: effects of data presentation format on judgmental forecasting. *International Journal of Forecasting*, 12 (1) (1996), pp. 119–137

[29]J.W. Payne. Task complexity and contingent processing in decision making: an information search and protocol analysis. *Organizational Behavior and Human Performance*, 16 (2) (1976), pp. 366–387

[30]L. Paquette, T. Kida. The effect of decision strategy and task complexity on decision performance. *Organizational Behavior and Human Decision Processes*, 41 (1) (1988), pp. 128–142

[31]J.R. Bettman, E.J. Johnson, J.W. Payne. A componential analysis of cognitive effort in choice. *Organizational Behavior and Human Decision Processes*, 45 (1) (1990), pp. 111–139

[32]J.E. Russo, B.A. Dosher. Strategies for multiattribute binary choice. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 9 (4) (1983), pp. 676–696

[33]R. Webby, M. O'Connor, B. Edmundson. Forecasting support systems for the incorporation of event information: an empirical investigation. *International Journal of Forecasting*, 21 (3) (2005), pp. 411–423

[34]C. Dowling, S. Leech. Audit support systems and decision aids: current practice and opportunities for future research. *International Journal of Accounting Information Systems*, 8 (2) (2007), pp. 92–116

[35]J. Swait, W. Adamowicz. The influence of task complexity on consumer choice: a latent class model of decision strategy switching. *Journal of Consumer Research*, 28 (1) (2001), pp. 135–148

[36]E.A. Locke, G.P. Latham. Building a practically useful theory of goal setting and task motivation: a 35-year odyssey. *The American Psychologist*, 57 (9) (2002), pp. 705–717

[37]D. Önkal, G. Muradoglu. Effects of task format on probabilistic forecasting of stock prices. *International Journal of Forecasting*, 12 (1) (1996), pp. 9–24

[38]W. Tych, D.J. Pedregal, P.C. Young, J. Davies. An unobserved component model for multi-rate forecasting of telephone call demand: the design of a forecasting support system. *International Journal of Forecasting*, 18 (4) (2002), pp. 673–695

[39]S.M. Whitecotton. The effects of experience and a decision aid on the slope, scatter, and bias of earnings forecasts. *Organizational Behavior and Human Decision Processes*, 66 (1) (1996), pp. 111–121

[40]H. Moskowitz, R.K. Sarin. Improving the consistency of conditional probability assessments for forecasting and decision making. *Management Science*, 29 (6) (1983), pp. 735–749

[41]P. Goodwin, G. Wright. Improving judgment time series forecasting: a review of guidance provided by research. *International Journal of Forecasting*, 9 (2) (1993), pp. 147–161

[42]F. Collopy, J.S. Armstrong. Rule-based forecasting: development and validation of an expert systems approach to combining time series extrapolations. *Management Science*, 38 (10) (1992), pp. 1394–1414

[43]M. Adya, J.S. Armstrong, F. Collopy, M. Kennedy. An application of Rule-based Forecasting to a situation lacking domain knowledge. *International Journal of Forecasting*, 16 (4) (2000), pp. 477–484

[44]M. Adya, F. Collopy, J.S. Armstrong, M. Kennedy. Automatic identification of time series features for Rule-based Forecasting. *International Journal of Forecasting*, 17 (2) (2001), pp. 143–157

[45]M. Adya, E.J. Lusk, M. Belhadjali. Decomposition as a complex-skill acquisition strategy in management education: a case study in business forecasting. *Decision Sciences Journal of Innovative Education*, 7 (1) (2009), pp. 9–36

[46]F. Collopy. *Rule-based Forecasting: Development and Validation of an Expert Systems Approach to Time-series Extrapolation*. (Doctoral dissertation) University of Pennsylvania. Ann Arbor (1989)

[47]H.A. Simon. *The Sciences of the Artificial*. The MIT Press, Cambridge, MA (1996)

[48]S. Makridakis, A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, R. Winkler. The accuracy of extrapolation (time series) methods: results of a forecasting competition. *Journal of Forecasting*, 1 (2) (1982), pp. 111–153

[49]J.S. Armstrong, M. Adya, F. Collopy. Rule-based Forecasting: using judgment in time series extrapolation. J.S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Kluwer Academic Publishers, Norwell, MA (2001), pp. 259–282

[50]J.S. Armstrong, F. Collopy. The selection of error measures for generalizing about forecasting methods: empirical comparisons. *International Journal of Forecasting*, 8 (1) (1992), pp. 69–80

[51]S. Morlidge. Do forecasting methods reduce avoidable error? Evidence from forecasting competitions. *Foresight: The International Journal of Applied Forecasting*, 32 (2014), pp. 34–39

[52]J.P. Shim, M. Warkentin, J.F. Courtney, D.J. Power, R. Sharda, C. Carlsson. Past, present, and future of decision support technology. *Decision Support Systems*, 33 (2) (2002), pp. 111–126

[53]A. Baddeley. Is working memory working? The fifteenth Bartlett lecture. *Quarterly Journal of Educational Psychology*, 44 (1) (1992), pp. 1–31

[54]D.G. MacGregor. Decomposition for judgmental forecasting and estimation. J.S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Kluwer Academic Publishers, Norwell, MA (2001), pp. 107–123

[55]F.J. Lee, J.R. Anderson. Does learning a complex task have to be complex? A study in learning decomposition. *Cognitive Psychology*, 42 (3) (2001), pp. 267–316

[56]J.S. Armstrong. Extrapolation of time-series and cross-sectional data. J.S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Kluwer Academic Publishers, Norwell, MA (2001), pp. 217–243

[57]P. Todd, I. Benbasat. Evaluating the impact of DSS, cognitive effort, and incentives on strategy selection. *Information Systems Research*, 10 (4) (1999), pp. 356–374

[58]P. Goodwin. Improving the voluntary integration of statistical forecasts and judgment. *International Journal of Forecasting*, 16 (1) (2000), pp. 85–99

[59]E. Welch, S. Bretschneider, J. Rohrbaugh. Accuracy of judgmental extrapolation of time series data: characteristics, causes, and remediation strategies for forecasting. *International Journal of Forecasting*, 14 (1) (1998), pp. 95–110

[60]I. Vessey. The effect of information presentation on decision making: a cost–benefit analysis. *Information Management*, 27 (2) (1994), pp. 103–119

[61]A.R. Montazemi, F. Wang, S. Nainar, C.K. Bart. On the effectiveness of decisional guidance. *Decision Support Systems*, 18 (2) (1996), pp. 181–198

[62]B. Xiao, I. Benbasat. E-commerce product recommendation agents: use, characteristics, and impact. *MIS Quarterly*, 13 (1) (2007), pp. 137–209

[63]J.D. Sterman. *Misperceptions of Feedback in Dynamic Decision Making*. Springer, Berlin, Heidelberg (1989), pp. 21–31

[64]D.T. Singh. Incorporating cognitive aids into decision support systems: the case of the strategy execution process. *Decision Support Systems*, 24 (2) (1998), pp. 145–163

[65]R. Fildes, P. Goodwin, M. Lawrence, K. Nikolopoulos. Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, 25 (1) (2009), pp. 3–23

[66]F. Chen, Z. Drezner, J.K. Ryan, D. Simchi-Levi. Quantifying the bullwhip effect in a simple supply chain: the impact of forecasting, lead times, and information. *Management Science*, 46 (3) (2000), pp. 436–443

[67]P. Goodwin, M. Sinan Gönül, D. Önkal. Antecedents and effects of trust in forecasting advice. *International Journal of Forecasting*, 29 (2) (2013), pp. 354–366

[68]S. Asimakopoulos, A. Dix. Forecasting support systems technologies-in-practice: a model of adoption and use for product forecasting. *International Journal of Forecasting*, 29 (2) (2013), pp. 322–336

[69]M.H. Zack. The role of decision support systems in an indeterminate world. *Decision Support Systems*, 43 (4) (2007), pp. 1664–1674

Corresponding author. Tel.: + 1 414 288 7526.

[a]Both authors have contributed equally to this paper.

[b]Tel.: + 1 518 564 4190; fax: + 1 518 564 3183.

[c]The 14 features include one (number of observations) that was not represented in the C&A feature set. Features such as outliers, unusual last observation, and irrelevant early data were not considered for the taxonomy as these are corrected in the original series before being processed for forecasting (e.g. irrelevant early data are truncated).We expected these corrective processes to continue to be used as best practices when deemed fitting.

[d]Combining A averages forecasts from the methods in Typical Method-five proposed in.[51] The five methods are single exponential smoothing, adaptive response rate exponential smoothing, automatic AEP filtering, Holt's exponential smoothing, and Brown's linear exponential smoothing (see [50]).

[e]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;[51] the Random Walk is the projection of the last observed value as the forecast value for all the relevant forecasting horizons under examination.

[f]In the development of the complexity scoring we decided to initiate each series with a score of 0. When a feature increased complexity, a negative value was added to the complexity score. When complexity is decreased, a positive value was added.

[g]To ascertain the validity of this cutoff, we computed the median CV values for the 72 series in the development data. This median was 0.865 which, rounded off, is the same as the threshold validated in prior RBF studies.

[h]These measures are discussed in later sections.

[i]Other scoring variations using constant increments were considered. Specifically, we experimented with using (a) an initial score of 100 and reducing the score as complexity increased, (b) two scores (simplicity and complexity) for each series, and (c) an initial scores of − 100 and adding as complexity increased. In the end, we found that starting

with a base of 0 yielded the simplest calibration. Possibly, other scoring schema could be considered in future replications.

[j]We did not test this mean classification relative to a possible median classification because of concerns that it would compromise the final testing of the CST. As such, we worked from the features so as to preserve the dataset at the development stage as a valid initial test of the classification. The final validity check, however, was expected to be the holdback test.

[k]Series numbers are those assigned by C&A.