4-2001

Automatic Identification of Time Series Features for Rule-Based Forecasting

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

Fred Collopy
Case Western Reserve University

J. Scott Armstrong
University of Pennsylvania

Miles Kennedy
Case Western Reserve University

Automatic Identification of Time Series Features for Rule-Based Forecasting

By Monica Adya, Fred Collopy, J. Scott Armstrong, and Miles Kennedy

Rule-based forecasting (RBF) is an expert system that uses features of time series to select and weight extrapolation techniques. Thus, it is dependent upon the identification of features of the time series. Judgmental coding of these features is expensive and the reliability of the ratings is modest. We developed and automated heuristics to detect six features that had previously been judgmentally identified in RBF: outliers, level shifts, change in basic trend, unstable recent trend, unusual last observation, and functional form. These heuristics rely on simple statistics such as first differences and regression estimates. In general, there was agreement between automated and judgmental codings for all features other than functional form. Heuristic coding was more sensitive than judgment and consequently, identified more series with a certain feature than judgmental coding. We compared forecast accuracy using automated codings with that using judgmental codings across 122 series. Forecasts were produced for six horizons, resulting in a total of 732 forecasts. Accuracy for 30% of the 122 annual time series was similar to that reported for RBF. For the remaining series, there were as many that did better with automated feature detection as there were that did worse. In other words, the use of automated feature detection heuristics reduced the costs of using RBF without negatively affecting forecast accuracy.

1. Introduction

Rule-based forecasting (RBF) is an expert system developed by Collopy and Armstrong (1992a) (hereon referred to as C&A). Its original version consisted of 99 rules that combine forecasts from four simple extrapolation methods (random walk, linear regression, Holt's exponential smoothing, and Brown's exponential smoothing). RBF relies on the identification of up to 28 features of time series to weight forecasts from these four methods. Based on empirical comparisons conducted on 36 time series, C&A concluded that RBF provided more accurate forecasts than could be obtained from an equal-weights combination of the four methods.

Rule-based forecasting is based on the premise that the features of time series can be reliably identified. Eight of these features are identified by analytical procedures coded in RBF while the rest rely on an analyst's knowledge of the domain or visual inspection of plots. Judgmental identification of these remaining features is time-consuming, relies on scarce and

expensive expertise, and has only a modest inter-rater reliability.

The identification of time series features has already been automated in one study. Vokurka, Flores and Pearce (1996) automated three features from RBF: irrelevant early data, outliers, and functional form. Their extension differed in a number of ways from C&A. They allowed for user interventions at several points in the forecasting process. The base methods — simple exponential smoothing, Gardner's damped trend exponential smoothing (Gardner, 1999), and classical decomposition — differed from those used in RBF. Finally, they used only a subset of the features identified in C&A. Their results were similar to those reported in C&A. Forecasts were, in general, more accurate than those from equal-weights as well as from a random walk.

Improved reliability of feature identification may improve accuracy (Stewart, 2001). In this study, we developed and validated heuristics for the identification of six features used in RBF. We expected that automating the feature detection process would reduce the inconsistencies in feature coding that result from differences in the experiences, abilities, and biases of expert coders. From a practical standpoint, automatic identification is less expensive because it automates time-consuming judgments. This is important for coding large data sets. From a research point of view, it should also aid replication and extension.

The next section describes the features used in RBF to characterize time series. We then discuss the six feature detectors. Judgmental and automatic codings are used to produce rule-based forecasts. These forecasts are compared with those from common benchmark methods. The paper concludes with an evaluation of the forecast accuracies from judgmental and automatic codings.

2. Features of Rule-Based Forecasting

Forecasting experts report that they base their selection of a method in part on patterns in the data. In a survey of forecasters (Yokum & Armstrong, 1995), 72% of the 319 respondents agreed that 'experts can, by examining a time series and its characteristics, improve the accuracy of forecasts by selecting the best among available extrapolation methods'. Only 12% disagreed, the rest being undecided. Historically, forecasters have characterized time series on broad patterns that represent the level, trend, seasonal variation, and uncertainty. A variety of features have been used to characterize each of these. To develop a comprehensive list of the conditions to describe historical time series, Collopy and Armstrong (1992b) examined the literature, surveyed experts, and conducted protocol sessions with forecasting experts.

Armstrong, Adya and Collopy (2001) describe 28 features of time series. Table 1 summarizes these features. C&A used 18 of these features to describe statistical characteristics

of the historical data and domain knowledge about future events. Of these, eight were identified by statistical procedures contained within the RBF rules. For instance, the direction of the basic trend is obtained by fitting a linear regression to the historical data and that for the recent trend is determined by fitting Holt's exponential smoothing model to the same data. Six of the remaining features such as level discontinuity, changing basic trend, unusual last observations, and unstable recent trends were identified by the features analyst based on visual inspection of the series. The remaining features relied on the analyst's knowledge of the domain, including information about the expected functional form, cycles, whether the series represents a start-up, and the causal forces impacting the series.

3. Automatic Feature Detectors

C&A relied on the judgment of its authors for the identification of features. Agreement on these codings ranged from 75 to 100% and averaged 89%. Coding a single series took about 5 minutes. This was a deterrent for further enhancement and validation of RBF across large samples of time series. In this paper, we develop heuristics for the identification of some features of RBF.

We determined that not all features of RBF could be automated. Features that relied on domain knowledge could be better identified by domain experts based on their expectations about the future. Once identified, domain-based features seldom change over time. While we did attempt to identify one domain-based feature, functional form, for the most part we focused our efforts on (a) instability features, and (b) features that were originally identified by viewing plots of the time series. These include level discontinuities, unusual last observation, changing basic trend, and unstable recent trend. In addition, we simplified the procedures for one feature, outlier, that was already automated in C&A.

The detection of features is a sequential process. The order of detection was partly motivated by the procedural requirements for RBF. In addition, through a process of trial and error, we found that certain features were better detected if other instabilities were not presented. For instance, a change in slope was easier to identify if no outliers and level discontinuities were present, because the heuristic for changing trend relies on a good fit for the regression line. Similarly, RBF procedures rely on the identification of the functional form of a series before any models can be fitted.

Our automatic features detector proceeds in the following sequence. Firstly, the functional form of a series is determined. Then, instability features are examined by first looking for outliers and level discontinuities. If an outlier is detected, its value is replaced by the average

of the adjacent points. If a level discontinuity is encountered, the historical data before the level discontinuity is equalized with the current level. For instance, if at t_{20}, the level were 60 and at t_{21} it increased to 100, then the level from t_1 to t_{19} would be adjusted to 100. This adjustment for level discontinuities is only used to aid detection of the remaining features since several of them rely on a good regression fit. For producing final forecasts, the level for t_0 to t_{19} reverts back to its original. Adjustments to outliers, however, are retained for the remaining procedures.

After this, the last observation is examined to determine if it is unusual (an outlier). If so, it is replaced by the average of the previous observation and a forecast obtained from fitting a regression line through data points up to $t - 1$. Finally, possible slope changes and instabilities in the recent trends are identified by fitting regression lines and examining the direction and magnitudes of the slopes.[1]

4. Development of the Heuristics

C&A used 126 time series from the M-Competition data (Makridakis et al., 1982) to develop, refine, and validate RBF. For this study, we used 122 of these 126 time series to develop and test the automatic feature detector. The remaining four series had regressing causal forces, indicating that the forces acting on these series tended to regress to a mean. Since we have not encountered any regressing series in our further work with RBF, these series were not considered in this study. For testing and refining the heuristics, the 122 series were split into two subsamples — the development sample and the validation sample. Series whose identifiers ended with 6, 5, 3, and 2 (a total of 70 series) were used to develop and refine the procedures. The remaining 52 series, with identifiers ending in 7, 4, and 8, were used to test effectiveness of these heuristics. The series were normalized so that historical values were between 0 and 100 before being used for feature detection.

For each feature, we used statistical tests to identify a heuristic. For example, we tested linear regression residuals to determine trend-based discontinuities, such as a change in basic trend or an unstable recent trend. Similarly, we examined first and second differences in a series to identify outliers and level discontinuities. We tested the heuristics on the development sample by comparing them with the judgmental coding implemented in C&A. Our objective was to minimize the disagreements between judgmental and heuristic codings. If disagreement was high, we refined the parameters of the heuristics. Typically these parameters related to the magnitude of first or second differences (as in outliers and level discontinuities), or to the standard deviation in a series (as in unstable recent trends), or to the magnitude of residuals from a linear regression fit. If adjusting these parameters did not produce improvements, we

considered alternative heuristics. When there was a reasonable agreement between judgmental and heuristic codings on the development sample for a particular feature, we retained the heuristic. We did not have a threshold acceptance level for the classifications on the development sample but used our judgment on what constituted a reasonable agreement. (Identification of such acceptance thresholds was not possible without a very large sample of time series. For instance, for the 70 series in the development sample, the judgmental coders had identified only five series with a changing basic trend. Consequently, we were calibrating our heuristics to these five series in the development sample.)

We typically developed a two-step approach for the identification of most features. The first step screened all observations for the likely occurrence of an instability. Where an instability was suspected, a second test was used to identify the nature of the instability. For instance, in the detection of an outlier, the heuristic first compared each data point in the series to determine where the deviation existed. However, such a deviation could signal either an outlier or a level discontinuity. Therefore, the second test examined the nature of the deviation to clearly identify if it were an outlier or a level discontinuity. The first test was intended to bring attention to any suspicious patterns in the data and the second test was used to confirm the nature of this pattern. We next describe and illustrate the specific heuristics for feature identification.

4.1. Functional Form

The functional form of a series represents the pattern of growth in the trend of the series. C&A shows that it is an area in which domain knowledge and historical data each play a role. The two forms used in RBF are multiplicative (exponential growth or decay) and additive (linear). While examining results from C&A's judgmental coding, we found that the identification of functional form led to substantial forecast errors on certain series. These were primarily start-up series with only about 8 to 10 historical observations available. Often these series were characterized as multiplicative; however, when the exponential growth of the early years was extrapolated, it produced large forecast errors since the rate of growth for the early years could not be sustained.

For the automatic identification, we assumed all series to be growing multiplicatively except when the series was (a) a start-up, (b) a short series (defined as a series with fewer than eight observations), (c) expressed in percentage terms, (d) contained negative observations, or (e) had a growth rate that appeared to be unsustainable in the long run. We assumed an annual growth rate of 20% or more to be unsustainable. We made this assumption since, in our experience, most business and economic series are multiplicative (Armstrong, 1985).

4.2. Outliers

Outliers are isolated observations that deviate substantially from the pattern in the rest of the series. Such deviations can be due to an unusual, non-recurring event or, sometimes, simply due to mistakes during data transcriptions.

We considered a large second difference to be indicative of the occurrence of an instability. We then used a regression trend line to specify the nature of this instability. A large second difference at any point in the historical data suggests the presence of an outlier or a level shift immediately preceding it. The pattern of the regression residuals from this point to the end of the series, however, confirms the nature of this instability. Specifically, the presence of an abrupt though temporary increase or decrease in residuals at a point before the large second difference indicates the presence of an outlier. This change in residuals should revert back to approximately the same range as before the outlier. On detection, outliers are replaced by the average of the preceding and subsequent data points.

Fig. 1 illustrates outlier detection. In that series, an outlier occurs at t_{40}. A large second difference is indicated at t_{41}. A regression line is fitted through from t_0 to t_{38} to avoid fitting the regression line to an extreme point. The parameters from this fit are used to predict the remaining time periods (i.e. from t_{39} to t_{47}). This fitted line indicates a large and temporary change in residuals at the point of the outlier (i.e. at point t_{40}). Notice that within the next two periods, the pattern of residuals reverts back to the same range as before the occurrence of the outlier.

4.3. Level Discontinuities

Level discontinuities are defined here as permanent shifts in the level of a series. For instance, sales may abruptly increase as a result of an increase in plant capacity that was provided to meet latent demand.

The initial screening for level discontinuities is the same as for outliers; both are identified by the presence of large second differences. The point before a large second difference is a possible discontinuity. Just as for outliers, a regression trend line is fit up to the point where this large difference occurs and residuals are obtained for the remaining periods. Because a level change is assumed to have lasting effects, both the magnitude and the direction of residuals must be examined. Two conditions must then be satisfied for qualifying an instability as a level discontinuity. Firstly, the large increase or decrease in residuals should be sustained for at least three periods after the indicated point of discontinuity. Residuals for these three points should be of similar magnitude. Secondly, the direction of the residuals after the discontinuity point should

be opposite to that of the periods before the discontinuity. If the residuals are positive before the level shift, they should now become negative.

The identification of a level discontinuity is illustrated in Fig. 2. A level shift occurs at t_9. This discontinuity is indicated by the presence of a large second difference at t_{10}. Notice that the second differences also indicate a relatively large change at t_9 and t_{11}. Our interest though is in the largest difference and that occurs at t_{10} here. A regression line is run through t_0 to t_7 and residuals are produced for the periods t_8 to t_{12}. The residuals decrease abruptly at t_9 from positive to negative and this negative pattern of residuals is maintained till the end of the series.

4.4. Unusual Last Observation

An unusual last observation occurs when the last data point deviates substantially from the previous pattern. The detection of this instability is important because it has a strong effect on the level and trend estimates of most extrapolation methods. An unusual last observation can be regarded as a special case of outliers. This feature is detected using first differences. If the first difference for the last data point, t_n, is greater than three standard deviations from t_{n-3}, then an unusual observation exists at the last data point. An unusual last observation is replaced by the average of its original value and the forecast from regression.

4.5. Changing Basic Trend

A change in the basic trend of a series is identified by comparing the slope in the early part of the historical data with that in the more recent past. If there is a large difference in slopes, a change in the basic trend could have occurred.

Fig. 3 illustrates the concepts behind this heuristic. In this series, the basic trend is changing although the change is masked by the instability in the trend of the series. A regression line is fitted on the first third and the last third of the series that is from point t_1 to t_5 and from t_{10} to t_{13}. The fit indicates a significant difference between the slopes as is evidenced from the fit lines. Results from this procedure, however, may be biased if either of the two regression lines runs through extreme points or if the number of observations is low, as in this example. A second test is then conducted by fitting another pair of regressions — this time in the first half and the second half of the series, from t_1 to t_7 and from t_8 to t_{13}. If the second test also indicates the presence of a slope change, then the slope change is confirmed. Both of the conditions must be met to classify the series as one with a changing basic trend.

4.6. Unstable Recent Trend

Occasionally, short-term trends become unstable, making it difficult to estimate parameters for short-term exponential smoothing methods. For the feature detector, we developed two tests to identify an unstable recent trend. Firstly, we examine the standard deviation of residuals for the recent trend. For annual series, we estimate that the last 20% of the series would be an appropriate representation of the recent trend. If this were less than five data points, we used a default value of 5. If the standard deviation of residuals for the recent trend was beyond a threshold, the series was assumed to be unstable in the recent period.

In the second test, we compare standard deviation of the residuals from the recent trend with that for the previous trend. If residual for the recent trend is greater than that for the first half of the series by a certain threshold, then the recent trend is determined to be unstable.

Fig. 4 illustrates application of the unstable recent trend heuristic using a series for which the trend is fluctuating and the series is noisy. The standard deviation of the residuals for this series is greater than 5.0, which is the threshold determined for the first part of this test. If this condition is not met, then a regression trend line is fit from *t* to *t* and another from *t* to *t* .If the residuals for the second half of the series are greater than 2.5 standard deviations from that of the first half, the series would pass the second test of the heuristic and thus be classified as a series with an unstable recent trend.

5. Comparison of Classifications from Heuristic and Judgmental Codings

We compared the heuristic identification of time series features on 122 series from C&A with judgmental classifications from C&A. Tables 2 and 3 provide details of this comparison for the development and validation samples. In these tables, 'judgment' indicates features identified by the judgmental codings, 'automatic' indicates features identified by the heuritistics. 'Both' indicates series for which there was agreement between the judgmental feature detection heuristics. For instance, in the development sample the heuristics identified a total of 27 changing basic trends (slope changes). The experts identified 24. The two approaches agreed on 14 codings ('both'). Another 10 series that were coded judgmentally as having slope changes were missed by the heuristics. The heuristics identified 15 series as having a slope change that the coders did not.

There was wide variation in the agreement between expert and automated codings. The highest agreement occurred on the coding of changing basic trend. On the other hand, codings of functional form differed greatly. Of the 122 series used in this study, 24 were coded as having an additive functional form by the automated procedures. While the experts had coded 22 series as additive, there was no agreement between the automatic and judgmental codings.

Of the eight series identified judgmentally as having a level discontinuity, the heuristic identified three. The heuristic identified six additional series as having discontinuities. An examination of plots for those series with differences between judgmental and automatic identification indicated that these were often noisy series. Sometimes the nature of the series was confounded by the presence of other features. For instance, in series 46, the level discontinuity identified judgmentally occurred over several time periods. Consequently, instead of being identified as a level discontinuity, the feature detector identified a slope change. Differences at the 'automatic' coding indicated the presence of a small level discontinuity that might have been overlooked at the time of visual inspection. At other times, the sensitivity of the test appears to have identified a slight deviation in an otherwise smooth series as a level discontinuity.

The heuristic for identification of unusual last observation agreed with all three series coded judgmentally. The heuristic identified nine additional series that were not identified. Eight of these 12 observations were in the development sample and four in the validation sample. The test appears to be sensitive to small deviations that were not noticed by experts.

Slope changes were identified judgmentally for 41 of the 122 series. The heuristic identified 24 of these but missed 17. An additional 15 series were identified by the heuristic as having a slope change. On examination of plots, it appeared in several cases that the judgmental process had identified slope changes in series that were relatively smooth (e.g. 16, 168). In other cases, what appeared to be a slope change in the recent periods was rectified as a result of adjustment of previously identified features (e.g. in series 53 adjustment for an unusual last observation removes a slope change). Conversely, 'automatic' coding differences were influenced by detection and adjustment of features such as outliers and level discontinuities.

Seventeen of the 26 series identified by the experts as having an unstable recent trend were also identified automatically as such. Twelve more series were identified by automatic procedures but not by judgment. Patterns in 'judgment' and 'automatic' differences were similar to those for slope change detection.

6. Validation and Testing Procedures

As expected, the differences between judgmental and automatic codings were more serious in the specification of functional form. Therefore, we compared the forecast accuracy of heuristic and judgmental codings to gain a better understanding of the impact of automating the feature identification process. We expected that the gains made in terms of cost savings and reliability might compensate for a small decline in forecasting accuracy.

We integrated the heuristics with the original RBF rules into an expanded version of RBF, RBF(A). We then produced forecasts for 122 of the 126 time series used by C&A and compared their accuracy to those from RBF as reported in C&A. RBF was originally calibrated on 36 series, then tested on validation sample V1 which contained 18 series. V1 was then combined with the calibration sample and RBF was re-calibrated. This procedure was repeated with V2 which contained 36 series. The final validation of RBF was then done only on 36 series in sample V3. We present our validations of RBF(A) on the same samples used in C&A — V1 to V3. Forecasts were produced for 1 to 6-ahead horizons, resulting in 732 forecasts across all 122 series. Furthermore, as in C&A, we used equal-weights and random walk for comparisons. If RBF(A) suffered no serious loss in accuracy as compared to RBF and these benchmarks, then the case for using automated feature detection would be fairly strong. However, if RBF(A) declined in performance, then the heuristics would need to be subjected to further validations.

While implementing RBF(A), 10 corrections were made to the original rule-base as presented in C&A. These corrections are described in Adya (2000) and the corrected rules are available on the web site hops.wharton.upenn.edu/ forecast. Consequently, original feature codings from C&A were rerun on the updated version of RBF. Results from this run indicated no improvements in accuracy on validation sample V3 from C&A. Future references to RBF in the paper apply to this corrected version of RBF.

We used multiple error measures for assessing the performance of RBF(A) as recommended by Armstrong and Collopy (1992). One of these, the relative absolute errors (RAEs), relates the performance of a method to that of the random walk. Armstrong and Collopy (1992) found that both mean and median RAEs and absolute percentage errors (APEs) were reliable and had good construct validity. The geometric mean of RAEs (GMRAEs) and mean APEs (MAPEs) are sensitive to the impact of parameter changes but do not provide sufficient outlier protection. Median RAEs and APEs are relatively less sensitive to small changes. The insensitivity is, however, valuable in protecting against outliers. Consequently, in this study, all four measures are reported. Both RAEs and APEs are computed for each horizon in each series. Cumulative RAEs and cumulative APEs summarize performance across horizons. Geometric mean of RAEs (GMRAEs), median RAEs (MdRAEs), mean APEs (MAPEs), and median APEs (MdAPEs) are used to summarize across series.

Table 4 presents summary results using multiple error measures for the 1-ahead, 6-ahead, and cumulative forecasts for the three validation samples.

Over all the horizons, RBF(A) performed about as well as RBF. For validation samples V1 and V3, GMRAEs for RBF(A) improved over RBF for all horizons. MdAPEs indicate equal or

better performance as summarized in Table 5.

Table 6 summarizes the forecast accuracy of judgmental and automatic identification on series in which their features differ. Since the samples are small, the conclusions here are only suggestive.

6.1. Changing Basic Trend

For series where automatic identification did not identify a change in basic trend but judgement did, forecast accuracy suffered. When the detectors identified a series that judgmental coding missed, forecast accuracy for the short periods improved slightly.

6.2. Unstable Recent Trend

For series where the heuristics identified an unstable recent trend, typically the 1-ahead forecasts were better than those for judgment. Judgmental coding yielded same or better forecasts for the long horizon.

6.3. Functional Form

Forecasts for series that were coded as additive by the automatic feature detectors but multiplicative by experts were at least as good as those reported for RBF on the 1-ahead, 6-ahead, and cumulative horizons. Similar gains were not observed for series that were coded as multiplicative by the heuristics and as additive by the experts. This raises questions about the sensitivity and completeness of the functional form heuristic which must be subject to further examination. There is also an argument that functional form is essentially a domain-based feature that does not benefit from identification of patterns in the historical data.

Considering all features, there were 78 series where heuristic and judgmental codings differed on one or more features. For almost 30% of the series, judgmental and heuristic codings yielded similar error measures for 1-ahead, 6-ahead, and cumulative horizons. Of the remaining series, there were as many series that did better with automated feature detection as there were that did worse. Decreased accuracy in some series were offset by gains in others. In general, then, the introduction of automated heuristics for feature detection in RBF did not reduce forecast accuracy.

7. Some Issues Related to Automatic Feature Identification

The detection and correct of outliers, level discontinuities, and unusual last observations occasionally either induced another feature or removed other features that had been identified in

the judgmental codings. For instance, in series 56, a level discontinuity was identified. When the heuristic corrected for this, it resulted in highlighting a slope change that had otherwise been masked by the more prominent level discontinuity.

It is possible that when features were judgmentally coded, judges were unable to anticipate the effects of corrections or adjustments that the rules would make in response to other features. Therefore, the automatic process may be beneificial over the more wholistic judgment employed by experts in C&A. We do not know how it would compare to a more decomposed judgment.

A major challenge was to apply the heuristic to series that differed in terms of variation and noise. For instance, the heuristics should be able to identify a changing basic trend or a level discontinuity for series that are smooth as well as those that are noisy. We used different threshold levels. For instance, a noisy series would have to meet more rigorous threshold requirements.

Our heuristics were developed on 70 time series and were tested on 52. Further work is required to validate them with larger samples of data. In particular, further analysis is required to better understand the causes for and the effects of differences in judgmental and automatic coding of the functional form.

8. Conclusion

Automatic feature identification significantly reduced the costs of forecasting large data sets with no appreciable loss in forecast accuracy in this study of 732 forecasts. The most significant contribution of this study is that it has automated forecasting decisions that are time consuming. Analysts are required to code only the domain-based features that typically take under a minute to identify. Moreover, these features tend to stay stable over longer periods of time thereby further reducing costs of recoding. Automating feature detection has also introduced consistency and reliability into the forecasting process. The added reliability could contribute to further efforts in the validation and refinement of rule-based forecasting. More importantly, it is now possible to apply RBF to a large number of series, such as the M3-IJF competition (Adya, Armstrong, Collopy and Kennedy, 2000).

Acknowledgements

The authors would like to thank Guisseppi Forgionne, Leonard Tashman, William Remus, and two anonymous reviewers for their valuable comments on earlier versions of the paper.

Biographies

Monica Adya is an Assistant Professor at DePaul University. She received her Ph.D. in Information Systems from Case Western Reserve University in 1996. Her research interests relate to the application of decision support systems and Artificial Intelligence to various business functions, particularly for forecasting and prediction. Recently, she has also examined the use of intelligent DSS in supporting health care issues, defense contracting, and digital libraries. Monica has published in *Information Systems Research*, *The International Journal of Forecasting* (*IJF*), and *The Journal of Forecasting*.

Fred Collopy is an Associate professor of Information Systems in the Weatherhead School of Management at Case Western Reserve University. He received his Ph.D. in Decision Sciences from the Wharton School of the University of Pennsylvania in 1990. He has done extensive research in forecasting, including the development of rule-based forecasting. He has also published on objective setting in organizations, time perception, and visual programming. His research has been published in leading academic and practitioner journals including *Management Science*, *Journal of Marketing Research*, *International Journal of Forecasting*, *Journal of Forecasting*, *Chief Executive*, and *Interfaces*. He is a member of the editorial boards of the *International Journal of Forecasting*, and of *Accounting*, *Management*, *and Information Technologies*.

J. Scott Armstrong (Ph.D. from MIT, 1968) is Professor of Marketing at the Wharton School, University of Pennsylvania, where he has been since 1968. He has also taught in Switzerland, Sweden, New Zealand, South Africa, Thailand, Argentina, Japan, and other countries. He was a founder and editor of the *Journal of Forecasting*, the *International Journal of Forecasting*, and of the International Symposium on Forecasting. A study in 1989 ranked him among the top 15 marketing professors in the US. In another study, he was the second most prolific Wharton faculty member during the 1988–1993 period. In 1996, he was selected as one of the first six 'Honorary Fellows' by the International Institute of Forecasters. His book, *Principles of Forecasting*: *A Handbook for Researchers and Practitioners*, will be published by Kluwer in 2001. He is the creator of the Forecasting Principles website (hops.wharton.upenn.edu/forecast).

Miles Kennedy is an Associate Professor in the Weatherhead School of Management's Information Systems Department at Case Western Reserve University. He received his Ph.D. from the London School of Economics. His research interests center around the use of information systems to amplify the intelligence of individuals and organizations. For several years, his major focus has been on the Causeway project — software that helps decision makers

create (potentially very large) categorical decision tables that are guaranteed to be complete and consistent: the system also supports rule induction and the creation of point scoring schemes.

References

Adya, M. (2000). Corrections to rule-based forecasting: findings from a replication. *International Journal of Forecasting* 16(1), 125–128.

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

Armstrong, J. S. (1985). *Long-Range Forecasting*: *From Crystal Ball to Computer*, John Wiley, New York.

Armstrong, J. S., Adya, M., & Collopy, F. (2001). Rule-based forecasting: using domain knowledge in time series extrapolation. In: Armstrong, J. S. (Ed.), Principles of Forecasting: A Handbook for Researchers and Practitioners, Kluwer Academic, Norwell, MA.

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

Collopy, F., & Armstrong, J. S. (1992a). Rule-based forecasting: development and validation of an expert systems approach to combining time series extrapolations. *Management Science 38*, 1392–1414.

Collopy, F., & Armstrong, J. S. (1992b). Expert opinions about extrapolation and the mystery of the overlooked discontinuities. *International Journal of Forecasting 8*, 575–582.

Gardner, E. S. (1999). Note: rule-based forecasting vs. damped trend exponential smoothing. *Management Science 45*(8), 1169–1176.

Makridakis, S., Anderson, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: results of a forecasting competition. *Journal of Forecasting 1*, 111–153.

Stewart, T. (2001). Improving reliability of judgmental forecasts. In: Armstrong, J. S. (Ed.), Principles of Forecasting: A Handbook for Researchers and Practitioners, Kluwer Academic, Norwell, MA.

Vokurka, R. J., Flores, B. E., & Pearce, S. L. (1996). Automatic feature identification and graphical support in rule-based forecasting: a comparison. *International Journal of Forecasting 12*, 495–512.

Yokum, T., & Armstrong, J. S. (1995). Beyond accuracy: comparison of criteria used to select

forecasting methods. *International Journal of Forecasting 11*, 591–597.

Appendix

Figure 1: Detection of an Outlier: An Example

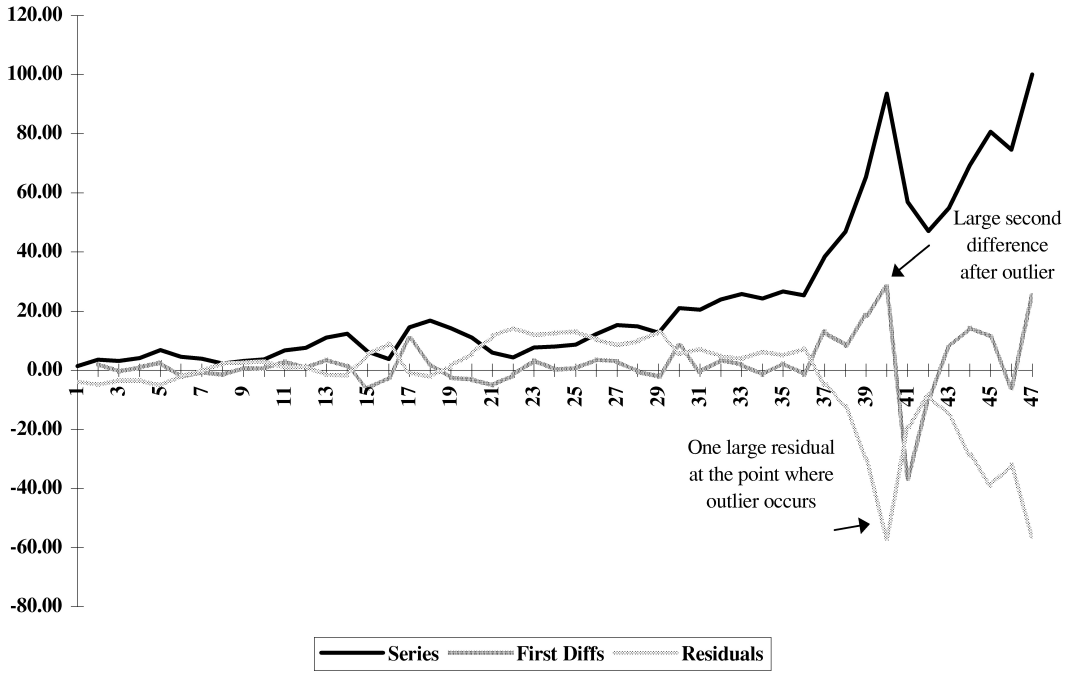

Figure 2: Detecting a Level Discontinuity: An Example

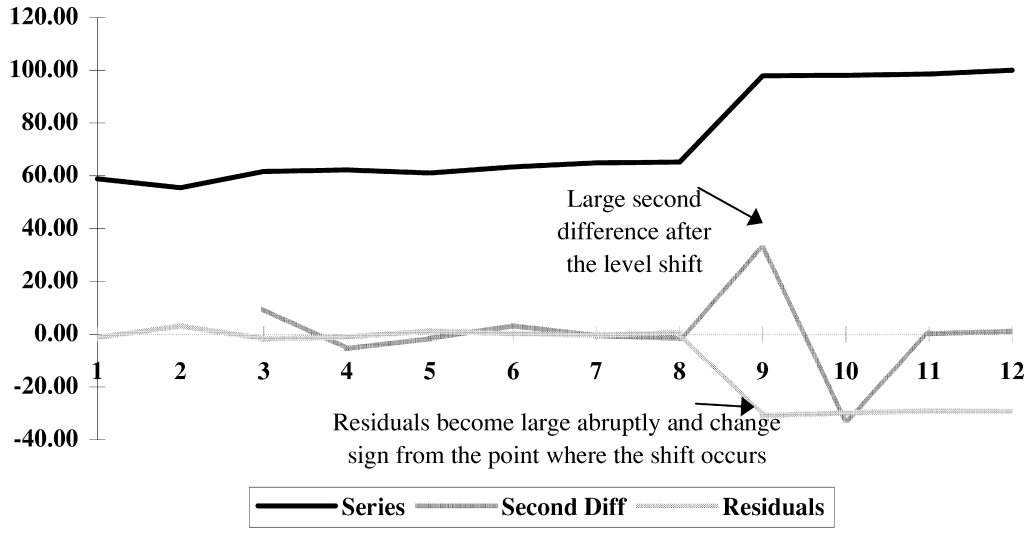

Figure 3: Detecting a Changing Basic Trend: An Example

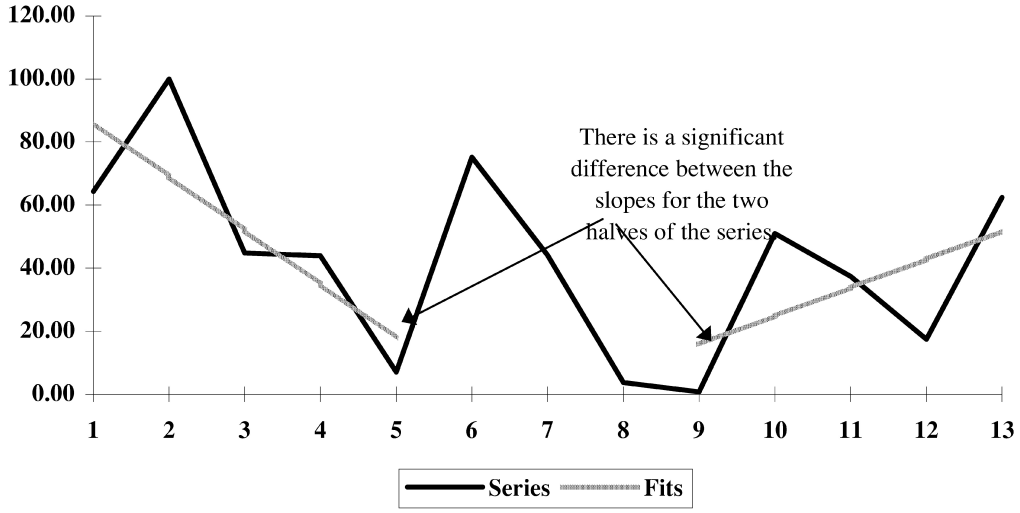

Figure 4: Detecting an Unstable Recent Trend: An Example

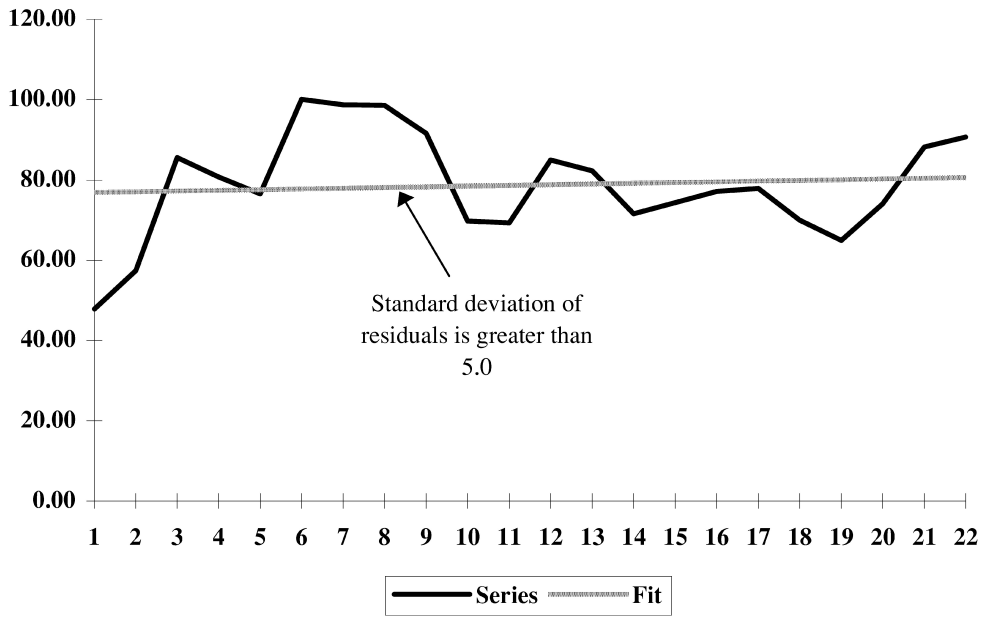

Table 1: Rule-based Forecasting Relies on 28 Time Series Features

Domain knowledge	Historical data	
Causal forces	Types of data	Uncertainty
Growth	Only positive values possible	Coefficient of variation about trend>0.2
Decay	Bounded (e.g. percentages, asymptotes)	Basic and recent trends differ
Supporting	Missing observations	
Opposing		
Regressing	Level	Instability
Unknown	Biased	Irrelevant early data
Functional form	Trend	Suspicious pattern
Multiplicative	Direction of basic trend	Unstable recent trend
Additive	Direction of recent trend	Outliers present
Cycles expected	Significant basic trend ($t > 2$)	Recent run not long
Forecast horizon	Length of series	Near a previous extreme
Subject to events	Number of observations	Changing basic trend
Start-up series	Time interval (e.g. annual)	Level discontinuities
Related to other series	Seasonality	Last observation unusual
	Seasonality present	

Table 2: Agreement of Judgmental and Heuristic Coding: Development Sample

Features	Codings		Agreement	Differences	
	Judgment	Automatic		Judgment	Automatic
Additive functional form	8	14	0	8	14
Outliers	–	5	–	–	–
Level discontinuity	5	8	3	2	5
Unusual last observation	3	8	3	0	5
Changing basic trend	24	27	14	10	13
Unstable recent trend	17	17	11	6	6

Table 3: Agreement of Judgmental and Heuristic Coding: Validation Sample

Features	Codings		Agreement	Differences	
	Judgment	Automatic		Judgment	Automatic
Additive functional form	14	10	0	14	10
Outliers	–[a]	11	11	–[a]	–[a]
Level discontinuity	3	1	0	3	1
Unusual last observation	0	4	0	0	4
Changing basic trend	17	15	10	7	5
Unstable recent trend	10	12	6	4	6

Table 4: Ex-ante Forecast Errors for Validation Samples V1-V3

Extrapolation procedure	MdRAE			GMRAE			MdAPE			MAPE		
	1-year	6-year	Cum	1-year	6-year	Cum	1-year	6-year	Cum	1-year	6-year	Cum
Validation sample V1												
Equal weights	0.68	0.66	0.77	0.77	0.69	0.79	3.79	20.57	13.95	8.20	24.50	15.97
RBF	0.58	0.61	0.77	0.45	0.58	0.66	1.89	14.37	12.23	8.71	24.10	15.67
RBF(A)	0.55	0.61	0.54	0.42	0.50	0.60	2.11	13.18	9.18	6.70	21.64	13.85
Validation sample V2												
Equal weights	0.78	0.55	0.60	0.94	0.58	0.61	3.56	17.62	11.23	7.46	28.48	15.98
RBF	0.46	0.71	0.61	0.42	0.47	0.47	1.28	9.21	7.54	4.95	25.42	13.72
RBF(A)	0.62	0.46	0.64	0.43	0.43	0.44	1.55	9.84	7.56	5.25	23.40	13.38
Validation sample V3												
Equal weights	0.82	0.63	0.69	0.95	0.76	0.77	4.69	19.29	10.69	8.62	32.56	18.58
RBF	0.58	0.58	0.60	0.57	0.62	0.68	2.98	14.71	12.22	5.74	14.71	12.22
RBF(A)	0.56	0.51	0.61	0.49	0.46	0.60	2.49	12.91	10.26	6.17	18.56	12.87

Table 5: Ex-ante Forecast Errors for Extrapolation Procedures, Median APEs[a]

Extrapolation procedure	1-Year ahead forecasts				6-Year ahead forecasts				Cumulative forecasts			
	V1 (18)	V2 (35)	V3 (35)	Wtd avg	V1 (18)	V2 (35)	V3 (35)	Wtd avg	V1 (18)	V2 (35)	V3 (35)	Wtd avg
Random walk	6.05	5.23	5.61	5.55	30.09	25.16	25.39	26.26	23.83	19.75	22.99	21.87
Equal weights	3.79	3.56	4.69	4.06	20.57	18.00	19.29	19.04	15.75	14.57	10.69	13.27
RBF	2.80	3.10	3.20	3.08	13.00	9.10	14.20	11.93	–	–	–	–
RBF(A)	2.11	2.10	2.83	2.39	15.39	10.18	13.89	12.72	11.82	9.00	9.92	9.94

[a] Cumulative forecasts from RBF were not available for comparison.

Table 6: Comparison of Automatic with Judgmental Coding Series with Different Codings: Relative Absolute Errors

Feature	Total different	1-Ahead RBF(A)			6-Ahead RBF(A)			Cumulative RBF(A)		
		Better	Same	Worse	Better	Same	Worse	Better	Same	Worse
Outliers										
Automatic	16	5	6	5	3	5	8	4	7	5
Judgment		–	–	–	–	–	–	–	–	–
Unusual last										
Automatic	9	2	1	6	3	2	4	2	2	5
Judgment	–	–	–	–	–	–	–	–	–	–
Level discontinuities										
Automatic	6	1	4	1	1	3	2	1	4	1
Judgment	5	2	2	1	3	–	2	3	1	1
Changing basic										
Automatic	15	6	6	3	5	3	7	5	6	4
Judgment	12	4	1	7	4	1	7	4	1	7
Unstable recent trend										
Automatic	12	8	3	1	3	4	5	4	5	3
Judgment	6	1	3	3	3	4	–	3	3	1
Functional form										
Automatic	18	11	4	3	8	3	7	8	4	6
Judgment	24	14	3	7	5	4	15	10	3	11