A COMPARISON OF MISSING-DATA PROCEDURES FOR ARIMA TIME-SERIES ANALYSIS

WAYNE F. VELICER
University of Rhode Island

SUZANNE M. COLBY
Brown University

Missing data are a common practical problem for longitudinal designs. Time-series analysis is a longitudinal method that involves a large number of observations on a single unit. Four different missing-data methods (deletion, mean substitution, mean of adjacent observations, and maximum likelihood estimation) were evaluated. Computer-generated time-series data of length 100 were generated for 50 different conditions representing five levels of autocorrelation, two levels of slope, and five levels of proportion of missing data. Methods were compared with respect to the accuracy of estimation for four parameters (level, error variance, degree of autocorrelation, and slope). The choice of method had a major impact on the analysis. The maximum likelihood very accurately estimated all four parameters under all conditions tested. The mean of the series was the least accurate approach. Statistical methods such as the maximum likelihood procedure represent a superior approach to missing data.

Keywords: *missing data; ARIMA models; time-series analysis; autocorrelation*

Missing data are a common practical problem in applied research. During the past two decades, advances in statistical theory combined with the availability of high-speed computers has resulted in the development and dissemination of new methods for analyzing incomplete data. Despite these ad-

Grants CA27821, CA63045, CA71356, and CA50087 from the National Cancer Institute supported this work. Correspondence concerning this article should be addressed to Wayne F. Velicer, Cancer Prevention Research Center, 2 Chafee Road, University of Rhode Island, Kingston, RI 02881-0808; e-mail: velicer@uri.edu. Further information can be obtained at the Cancer Prevention Research Center's Web site (http://www.uri.edu/research/cprc).

Educational and Psychological Measurement, Vol. 65 No. 4, August 2005 596-615
DOI: 10.1177/0013164404272502

vances, traditional ad hoc procedures continue to be widely used. This is partly the result of inertia—researchers tend to use familiar procedures—and partly due to a lack of awareness of the extent to which the new statistical procedures represent a significant improvement on the ad hoc methods. The purpose of this article is to evaluate four procedures, one statistical and three ad hoc, applied to different degrees of missing data in time-series analysis.

Time-series analysis can be viewed as an exemplar of the new methods developed to analyze longitudinal data. Time-series analysis is used to analyze data that consist of repeated observations on a single experimental unit. An interrupted time-series analysis includes an intervention and may be conceptualized as the reverse of an analysis-of-variance (ANOVA) design: Rather than taking one observation each on many individual subjects, one takes many observations on a single subject. One major advantage of using time-series analysis is that it allows a researcher to look at a pattern of change over time rather than at a single discrete point in time. Like other longitudinal methods, time-series analysis requires that attention be paid to the temporal structure of the data. Repeated observations on the same unit cannot be assumed to be statistically independent. Dependency is the extent to which an observation at time t is predicted by the observation at time $t - 1$. The terms *dependency* and *autocorrelation* are used interchangeably. All longitudinal procedures must address the issue of dependency.

Time-series analysis can be viewed as an excellent method for studying the impact of missing data. The method involves the estimate of four basic parameters, two similar to cross-sectional procedures and two unique to longitudinal methods. The estimations of the level of a series and the error variance of a series are directly analogous to estimating the mean and error variance in cross-sectional designs. The estimations of the slope, or change over time, and the dependency represent the unique aspects of a longitudinal method.

Missing data are also a particularly critical problem in longitudinal research. Studies that are characterized by repeated observations on the same experimental unit, particularly when the experimental unit is a person, are especially likely to have some observations missing (Laird, 1988). The likelihood of missing observations increases as the number of observations required increases. The minimum number of observations required for adequate power in the baseline (or control) phase of a time series ranges between 20 and 50 and may be much higher, depending on the effect size. In a simple design, one would prefer the same or a similar number in the postintervention or treatment phase. Obtaining complete data at regular intervals for 40 to 100 observations is difficult. The causes of missing data are often beyond the control of experimenters. For example, if a study requires that a participant come to the research site for every observation, missing data may result on weekends if appointments cannot be scheduled, or on any days when unforeseen events (e.g., illness, lack of transportation) preclude attendance.

It is reasonable to assume that the greater the number of observations required on a single experimental unit, the more likely it is that some observations will be missing. Because the implementation of time-series analysis assumes complete data, it is necessary to account for the missing data as part of any analysis.

This study examines the effects of using different techniques for handling missing data on time-series analysis. Available techniques vary widely in terms of their ease of implementation and theoretical appropriateness. Most of the more sophisticated approaches (e.g., maximum likelihood estimation for state-space models) have been developed within the field of econometrics, leading some to question their applicability to behavioral research (Rankin & Marsh, 1985). There are very few examples of the use of the new statistical procedures available. Although the mathematical superiority of the procedures has been demonstrated, the practical advantages remain unknown. It is unclear how large an improvement they represent over the well-known and easily implemented ad hoc procedures. There is also no guidance available about the conditions under which the use of the ad hoc procedures in previous research may have distorted the conclusions. An empirical comparison of missing-data techniques can provide a basis for researchers to choose the most appropriate and practical methods available to them. It also can guide a reappraisal of the degree of distortion that might have been introduced by ad hoc procedures in the published literature. Procedures that are superior in the context of time-series analysis are likely to be superior in other types of longitudinal data analysis.

Time-Series Analysis

Crosbie (1993), Velicer and Colby (1997), and Velicer and Fava (2003) provided recent reviews of the use of time-series analysis in behavioral research. Box and Jenkins (1970) developed the autoregressive integrated moving-averages (ARIMA) models to mathematically represent the dependence in data (also see Glass, Willson, & Gottman, 1975; Gottman, 1981; McCleary & Hay, 1980). If dependency were ignored, an estimate of error variance would be incorrect. If dependency is positive, apparent variability in data is decreased, and the probability of a Type I error increases. If dependency is negative, apparent variability in data is increased, and the probability of a Type II error increases. Time-series analysis takes dependency into account and therefore yields accurate parameter estimates and significance tests.

ARIMA models have three parameters: p represents the order of the autoregressive component, d represents the amount of differencing necessary to remove any cyclicity present in the series, and q represents the order of the moving-averages component. The numeric values of these three parame-

ters specify the type of ARIMA model that best fits a series. Among the many methods available for model identification, the most common method (Glass et al., 1975) consists of examining the patterns of autocorrelations and partial autocorrelations in data. Once an ARIMA model has been identified, the data are transformed to meet the assumptions of the general linear model.

For the purposes of this study, the type of ARIMA model was held constant while the degree of dependency, the slope, and the proportion of missing data in the series were manipulated. Even though model identification may be considered by some to be a central issue in time-series analysis, it is not essential to time-series analysis. Because methods of model identification (e.g., Glass et al., 1975) are subjective and have been shown to be unreliable (Velicer & Colby, 1997; Velicer & Harrop, 1983), several methods of time-series analysis without model identification have been proposed. All of these methods use an estimated or generalized transformation matrix rather than a transformation specific to the ARIMA model that underlies a series. Velicer and McDonald (1984, 1991) suggested analyzing a higher order (5, 0, 0) autoregressive model for all series. Simonton (1977) assumed that a simple (1, 0, 0) autoregressive model is appropriate for all cases. Algina and Swaminathan (1979) used a sample estimation of the transformation matrix. For a discussion of these approaches' respective merits, see Velicer and Fava (2003).

Parameter Estimates and the ARIMA (1, 0, 0)

To limit the scope of this study, a single ARIMA model was selected. The first-order autoregressive ARIMA model was selected because it is the most commonly encountered model in the behavioral sciences. Glass et al. (1975) stated that higher order models are unusual in the behavioral sciences. In a study of couple interaction during marital counseling, all of the 98 series could be represented by either (0, 0, 0) or (1, 0, 0) models (Revenstorf, Kessler, Schindler, Hahlweg, & Bluemner, 1980). Also, in a reanalysis of 70 clinical series published in journal articles over a 4-year period, Marsh and Shibano (1984) found that 40% of the series could be described as (0, 0, 0) models, and 48% could be described as autoregressive models with one or two terms. More relevant is the fact that the use of a (1, 0, 0) model proved satisfactory with 80% of the series tested. To determine characteristics of time-series data for a simulation study, Rankin and Marsh (1985) examined baseline data obtained from the Elderly Support Project at the School of Social Service Administration at the University of Chicago. Of the 16 series that met their criteria (complete data and a minimum of 30 data points), 11 were identified as (0, 0, 0) models, 4 as (1, 0, 0), and 1 as a (0, 0, 1) model. As a result, data for Rankin and Marsh's simulation study were generated to fit a first-order autoregressive model.

Four parameters describe first-order autoregressive ARIMA models: level (L), error variance (σ^2), slope (S), and autocorrelation (ϕ). The level and variance represent parameters that are analogous to cross-sectional statistical methods. The slope and autocorrelation parameters represent parameters unique to longitudinal designs.

L is an estimate of the level of a series at $t = 0$. This value is the intercept of the best fitting straight line through the plotted observations of a series. When the series is level (i.e., the slope of this line is equal to zero), L is the mean of the series. If the series is not level, its mean is determined in part by the number of observations (N) in the series. Therefore, the intercept of the series (L) is used because it is independent of the length of the series.

Error variance (σ^2) estimates the chance variation that remains in a time series once the series has been transformed to remove the dependency from the data. The residual error variance must be uncorrelated (i.e., not significantly different from a series of random errors), with a mean of zero. It is calculated by dividing the error sum of squares by the number of observations in the series.

The slope (S) of a series estimates the change of the series over time. If the slope of the series is zero, the series is stable across time. A positive slope indicates that the series is increasing, and a negative slope indicates that the series is decreasing.

The autocorrelation (ϕ) estimates the degree of dependency in the data. For a (1, 0, 0) model, the value of ϕ must lie within the "bounds of stationarity" (Box & Jenkins, 1970; Dixon, 1988), that is, from –1.00 to 1.00. In this case, ϕ is analogous to a correlation coefficient. When $\phi = 0$, there is no dependency in the data. When $\phi = 1.00$ or –1.00, behavior is considered to be perfectly predictable. Negative and positive autocorrelation are distinguished by the direction in which a subject's behavior deviates from one time point to the next. For example, ϕ is positive when a subject's behavior deviates in the same direction at time t as it did at time $t - 1$. Phi is negative when a subject's behavior deviates in the opposite direction at time t than it did at time $t - 1$.

Missing Data: Assumptions

The choice of an appropriate method for handling missing data depends in part on the reason why the data are missing, or the missing-data mechanism (Little & Rubin, 1987; Schafer, 1997; Schafer & Graham, 2002). Rubin (1976) developed the following classifications:

- Data are missing completely at random (MCAR) if the observations with any missing values are a random subsample of the full sample. In this case, the missing-data mechanism is unrelated to the model and is therefore ignorable.

- Data are missing at random (MAR) if the pattern for a variable is not a function of its observed values, but it may be a function of other values in the model. For example, a response for annual household income on a survey may be missing for several reasons. One reason is that a respondent may not know his or her household income. The missing-data mechanism may be a function of a respondent's age (e.g., very young respondents often do not know their families' incomes) but not a function of the respondent's household income. MAR is a less stringent assumption than MCAR.
- Values are classified as nonignorable if they are systematically missing from the data set (i.e., the missingness is a function of the values that are missing). In time series, this might mean that the missing data occur in patterns or are related to the numeric values of the series (i.e., when the number of drinks consumed is very high, a subject may not record the amount).

In this study, the eliminated observations are assumed to be MCAR. Data that are MAR or systematically missing from a data set may be more typical of most missing-data problems. Data that meet the MCAR assumption also meet the MAR assumption. The MCAR condition is the most basic missing-data condition and therefore is a logical starting point. A subsequent article will evaluate the performance of alternative methods when the MAR assumption is violated.

Methods of Handling Missing Data

Four methods of handling missing data were evaluated in this study. The first three are ad hoc procedures that have been extensively used and have the advantage of being easy to implement.

The first ad hoc procedure for handling missing data is deletion, that is, eliminating the observations that are missing from the series and then analyzing the condensed series as if it were the same as an original shorter series. This is analogous to complete case analysis, which is often used to handle missing data in other types of univariate and multivariate analyses. The costs associated with using deletion include that (a) the method will always decrease the sample size and (b) the method can lead to biased parameter estimates if the data are not MCAR. This procedure does preserve order information in longitudinal data. However, in time-series analysis, the procedure will result in violations of the assumption that there is an equal time interval between observations. When the missing data are eliminated from the series, the result is equivalent to a series with irregular time intervals. When data are collected at irregular time intervals but treated as though they were collected at regular intervals, the autocorrelation coefficients may be inaccurate, and the estimation of the slope parameter may be affected. Rankin and Marsh (1985), on the basis of the results of their simulation study that examined the effects of missing data on time-series analysis, concluded that once more

than 20% of the data are missing, a series deviates from the complete data pattern. Rankin and Marsh also concluded that this approach is not detrimental when less than 20% of a series is missing.

The second ad hoc technique involves substituting the mean of the series for the missing value. Substituting the mean of the entire time series is comparable to the typical mean imputation used in other analyses. In this case, the mean would be obtained from all of the nonmissing observations in a series, and that value would be imputed for each missing observation. This method ignores the order of the observations. With respect to the estimation of dependency, imputing the mean values may inappropriately smooth a series with negative autocorrelation. It is also likely to provide inaccurate estimates when there is a nonzero slope in the series. The appeal of this method is simplicity and its widespread use.

The third ad hoc technique involves substituting the mean of the adjacent observations for any missing data. This method takes into account the order of the observations. The mean of the adjacent observations should be more accurate when the slope of a series is not zero. Imputing the mean of the adjacent observations may also be more accurate than the mean of the entire series when ϕ (autocorrelation) is positive. However, when ϕ is negative, using the mean of the adjacent observations may artificially smooth the series and mask the amount of autocorrelation actually present in the behavior of interest.

The fourth method, maximum likelihood estimation, is the most theoretically justified method for handling missing data in time-series designs. The maximum likelihood algorithm (Jones, 1980), available in SAS (SAS Institute, 1988), was selected for inclusion in this study. Of all the statistical methods, it was the fastest and easiest to implement. This is related to the expectation maximization algorithm, a general iterative algorithm for maximum likelihood estimation in missing-data problems. The algorithm had originally been applied to a limited and distinct set of problems. With each new application, the generality of the algorithm's underlying principle became more apparent (Baum, Petrie, Soules, & Weiss, 1970; Beale & Little, 1975; Hartley, 1958). Dempster, Laird, and Rubin (1977) demonstrated the generality of what they named the expectation maximization algorithm and provided a broad range of examples of its applications. Because the initiation of this study, new missing-data methods have become more generally available. Specifically, multiple-imputation procedures represent another approach that should be evaluated in future studies (Graham, Cumsille, & Elek-Fisk, 2003; Schafer, 1997; Schafer & Graham, 2002). It was fully expected that the maximum likelihood technique would be the most accurate approach to handling missing data.

Previous Research

Although many techniques are available for handling missing data for other types of experimental designs, the special complications presented by autocorrelation in time-series analysis have received limited attention in the domain of applied behavioral research. To date, only one simulation study has been published that examined the effect of missing data on behavioral time-series analysis (Rankin & Marsh, 1985). This study used the deletion approach to handling the missing data. The fields of econometrics and engineering have produced more in-depth treatment of this topic (Harvey & Pierse, 1984; Kohn & Ansley, 1986). However, this research is based on general state-space models developed by Kalman (1960) rather than Box and Jenkins's (1970) ARIMA models, the dominant approach in the behavioral sciences.

Study Design

The purpose of this study was to compare the performance of different methods of handling missing data in time-series analysis under conditions in which all assumptions are met. The study focused on computer-generated data that are normally distributed, with the data MCAR and the time-series model correctly identified. Fifty different types of series (10 samples of each) representing different degrees of dependency ($\phi = -.80, -.40, .00, .40$, and .80), different types of slope ($S = 0$ or a positive slope of 15°), and different proportions of missing data points (10%, 20%, 30%, or 40%) were generated. Four different techniques for handling missing data (deletion, mean of series, mean of adjacent series, and maximum likelihood estimation) were then used on each series, and results were compared for each technique. It was expected that the maximum likelihood procedure would perform best overall, but the degree of accuracy was unknown. It was further expected that the mean of the series would perform worst, because this method completely ignores the longitudinal nature of the data. It was unclear under what conditions and to what degree the different methods of handling missing data would result in any practical differences. Effect-size estimates were used to provide an indication of the degree of practical importance.

Method

Data Generation

A comparison of missing-data techniques could have been done either using well-known data sets or simulated data. The decision to simulate data

was made because factors that may affect the outcome could be systematically manipulated. In addition, the use of simulated data provided population parameter values (criterion values) against which estimates could be compared. Simulation studies can be viewed as validity studies, whereas the use of existing data sets can be viewed as reliability studies.

Time-series data were generated using a FORTRAN computer program that was adapted from a program originally developed by Padia (1975) and revised by Harrop and Velicer (1985). Initially, 10 series were generated, representing all possible combinations of five levels of dependency (ϕ) and two levels of slope (*S*). Ten replications (samples) of each of the 10 series were then generated. Ten replications were chosen on the basis of Harrop and Velicer (1990) and some preliminary testing. The preliminary runs involved 5, 10, and 20 replications. Both 10 and 20 replications yielded estimates that were more accurate and stable than 5 replications. In addition, there was little or no improvement in the estimates when using 20 replications as opposed to 10. It should be noted that simulation studies sometimes use thousands of replications, but those studies are focusing on the evaluation of probability values rather than parameter estimation and only include a few exemplars rather than performance across the broad number of conditions in this study.

The 100 complete data sets (i.e., 10 replications of each of the 10 series) served as the basis for the missing-data estimation conditions. All series fit an ARIMA first-order autoregressive (1, 0, 0) model. The mean of the random component of all series was 0.0, and the variance was 1.00.

Independent Variables Manipulated

*Slope (*S*).* Half of the series generated had a slope of zero; the other half had a positive slope of 15°. Because the series consisted of 100 data points, a slope of 15° was sufficient to test the effects of slope on the effectiveness of the missing-data methods. Also, there was no reason to expect that the effects of a negative slope would be any different than the effects of a positive slope, so negative slopes were not investigated in this study.

Degree of dependency (ϕ). Series were generated with five different levels of autocorrelation ($\phi = -.80, -.40, .00, .40$, and .80). The values were chosen to represent moderate and severe levels of autocorrelation that may be encountered in time-series designs. It was considered necessary to include negative and positive values of ϕ, because most techniques should affect these series differently. The fifth level was the white-noise model ($\phi = .00$). (When $\phi = 0.00$ and the slope is zero, the time-series model defaults to an ANOVA model, and techniques appropriate for ANOVA may be used.

However, an ANOVA would ignore the temporal pattern of change in an interrupted time-series study.)

Proportion of missing data. In Rankin and Marsh's (1985) investigation of the effects of missing data (with no replacement), six different percentages were deleted: 10%, 20%, 30%, 40%, 50%, and 60%. The authors concluded that the higher the percentage of missing data, the poorer the model's overall fit, with poor fit beginning to occur when the percentage is more than 20%. Because a negative impact began to occur at 20% missing data, this study limited the percentages to 10%, 20%, 30%, and 40%. The complete case condition (0% missing) was also included to provide a validity check.

Data were randomly eliminated from each of the 100 original series in the four different proportions. This resulted in a total of 50 conditions, or 500 series (10 replications each of the original 10 complete data sets; the same 10 data sets with 10% of observations randomly eliminated; and the same 10 data sets with 20%, 30%, and 40% of observations randomly eliminated).

Techniques for handling missing data. Four different missing-data techniques were used: (a) deletion, involving no estimate of the missing data, with the series condensed and analyzed as a shorter series; (b) mean of series, whereby the mean of the series was imputed in place of the missing observations; (c) mean of the adjacent observations, whereby the mean of the adjacent observations was imputed in place of the missing observations; and (d) maximum likelihood estimation, whereby Jones's (1980) algorithm provided the maximum likelihood estimates for the parameters of the series.

Dependent Variables

Four dependent variables were obtained from each time-series analysis, which corresponded to the four parameters of the model.

*Level (*L*)*. Estimates of level were also obtained for each series to assess the effects of missing-data techniques on this parameter. This parameter is crucial for interrupted time-series analysis, in which change in level is one of the two parameters, which measures intervention effects. For this reason, it is important to have an accurate measure of baseline level. The population value of *L* was 0.0.

Error variance (σ^2). Minimum residual error variance was obtained by dividing the sum of the squared error at the last iteration by the number of data points in the series (usually 100). The number of data points had to be adjusted by the missing-data technique used. For example, when deletion

was used and 40% of the data were missing, the number of data points in the series was 60. Error variance was used as a dependent variable because it is critical in the calculation of significance tests, when testing intervention effects (e.g., in an interrupted time-series design). The population value for error variance was 1.00.

*Slope (*S*)*. This is one of the two parameters critical for longitudinal designs. In an intervention study, change in slope is the other parameter that measures intervention effects. Estimates of slope were obtained for each series to test whether missing-data methods affect the accuracy of estimating this parameter.

Autocorrelation parameter (ϕ). This is the other parameter that is critical for longitudinal designs. An estimate of ϕ was obtained for each series to see how missing-data techniques affect this parameter.

Analyses

Time-series analyses were performed using SAS/ETS Version 6, Proc ARIMA (SAS Institute, 1988) on an IBM 4381 mainframe computer. SAS uses a nonlinear algorithm for its solution. Analyses used the true model identification and default values for starting estimates and stopping criterion. The maximum number of iterations was set at 50. The conditional least squares method of estimation was used.

Results

Separate repeated-measures ANOVAs were used for each of the four dependent variables: error variance, ϕ, level, and slope. In the cases of ϕ and slope, which are independent as well as dependent variables, separate ANOVAs were performed for each level of the independent variable. Each analysis met the assumptions for a repeated-measures ANOVA. Effect size was estimated by η^2. As a guide to interpretation, Cohen (1988) classified an effect size of $\eta^2 = .01$ as a "small" effect size, an effect size of $\eta^2 = .06$ as a "medium" effect size, and an effect size of $\eta^2 = .14$ as a "large" effect size. To simplify the interpretation of the results and focus on the most important finding, only results that were both statistically significant and had an effect size of $\eta^2 = .03$ or higher are reported.

Level

A $2 \times 5 \times 5 \times 4$ (Slope $\times$ ϕ $\times$ Percentage Missing $\times$ Technique) ANOVA was used to compare the estimates of level associated with using each of the

four techniques for handling missing data, under all of the experimental conditions. Slope and ϕ were between-groups factors, and percentage missing and technique were within-groups factors in this analysis. The correct value for all series is 0.00.

There was one significant three-way interaction: technique by percentage missing by slope, $F(12, 1{,}080) = 101.56$, $p < .001$, $\eta^2 = .081$. There were three two-way interactions to interpret: slope by percentage missing, $F(4, 360) = 56.48$, $p < .001$, $\eta^2 = .024$; slope by technique, $F(3, 270) = 479.82$, $p < .001$, $\eta^2 = .157$; and technique by percentage missing, $F(12, 1{,}080) = 102.88$, $p < .001$, $\eta^2 = .082$. There were three main effects to interpret: slope, $F(1, 90) = 25.93$, $p < .001$, $\eta^2 = .069$; percentage missing, $F(4, 360) = 66.47$, $p < .001$, $\eta^2 = .029$; and technique, $F(3, 270) = 488.04$, $p < .001$, $\eta^2 = .160$.

Simple effects tests were used to interpret the results. It was found that the interaction between technique and percentage missing was significant only for the $S = 15°$ condition. When $S = 0°$, all of the missing-data techniques performed equally well, and none of their level estimates was significantly different from those based on complete data. For $S = 15°$, using the mean of the series for missing data resulted in inflated estimates of level. This approach produced inaccurate level estimates when only 10% of the data were missing ($M = 1.2$) and got worse when more data were missing. When 40% of the data were missing, the mean of the series estimated level was 5.0. The other major finding was that both maximum likelihood and deletion resulted in accurate estimates of level under all conditions. The performance of the mean of adjacent scores was also generally acceptable.

Error Variance

A $2 \times 5 \times 5 \times 4$ (Slope $\times$ ϕ $\times$ Percentage Missing $\times$ Technique) ANOVA was used to examine differences in estimation of variance in the series. The percentage of data missing from the series and the technique for handling missing data were within-groups (i.e., repeated-measures) factors; ϕ and slope were between-groups factors. The correct value for all series is 1.00.

There was one significant three-way interaction with a large effect size: technique by percentage missing by slope, $F(12, 1{,}080) = 1{,}797.54$, $p < .001$, $\eta^2 = .121$. There were three two-way interactions to interpret: percentage missing by slope, $F(4, 360) = 1{,}327.76$, $p < .001$, $\eta^2 = .043$; technique by slope, $F(3, 270) = 5{,}944.72$, $p < .001$, $\eta^2 = .263$; and percentage missing by technique, $F(12, 1{,}080) = 1{,}854.24$, $p < .001$, $\eta^2 = .125$. There were three main effects to interpret: slope, $F(1, 90) = 2{,}623.51$, $p < .001$, $\eta^2 = .093$; percentage missing, $F(4, 360) = 1{,}721.68$, $p < .001$, $\eta^2 = .056$; and technique, $F(3, 270) = 6{,}062.99$, $p < .001$, $\eta^2 = .268$.

Simple effects tests found that the interaction between technique and percentage missing was significant only when $S = 15°$. When $S = 0°$, all the miss-

ing-data techniques performed equally well, and none of their level estimates was different from those based on complete data. There was one major finding and one secondary finding.

The major finding involved the mean of the series when the time series had a 15° slope. The estimates of variance were extremely high, ranging from 7.6 when 10% of the data were missing to 24.4 when 40% of the data were missing (criterion = 1.00). By way of contrast, the maximum likelihood approach yielded accurate variance estimates at all levels of missing data.

In addition, the severe negative autocorrelation condition ($\phi = -.80$) presented special problems for the deletion and mean of the adjacent observations methods. When $\phi = -.80$, these techniques substantially overestimated the error variance, with mean of the adjacent observations significantly less accurate than deletion. The effect size was much smaller than the problems observed for the mean of the series, with the overestimates generally in the 2.0 to 3.0 range (compared with the criterion of 1.00). At every other level of ϕ, deletion and mean of the adjacent observations yielded accurate variance estimates.

Slope

Separate $5 \times 5 \times 4$ ($\phi \times$ Percentage Missing $\times$ Technique) ANOVAs were conducted for each value of slope used in the study (i.e., 0° and 15°). Phi was a between-groups factor, and percentage missing and technique were within-groups factors.

S = *0*°. When $S = 0°$, there were no significant effects in the analysis. All techniques accurately estimated slope to be 0°, regardless of the level of ϕ or the percentage of missing data.

S = *15*°. One two-way interaction with a large effect size was interpreted: technique by percentage missing, $F(12, 540) = 1{,}339.06$, $p < .001$, $\eta^2 = .347$. There was one main effect to interpret: technique, $F(4, 135) = 6{,}153.54$, $p < .001$, $\eta^2 = .575$. Figure 1 illustrates this result.

The same pattern was observed at all four levels of missing data (10% missing, 20% missing, 30% missing, and 40% missing) and all levels of ϕ ($\phi = -.80, -.40, .00, .40$, and .80). Maximum likelihood and mean of the adjacent observations both resulted in accurate estimates of slope. The use of deletion led to significant overestimates of slope. Using the mean of the series significantly underestimated slope. Both of these problems in slope estimation became more exaggerated as the percentage of missing data increased.

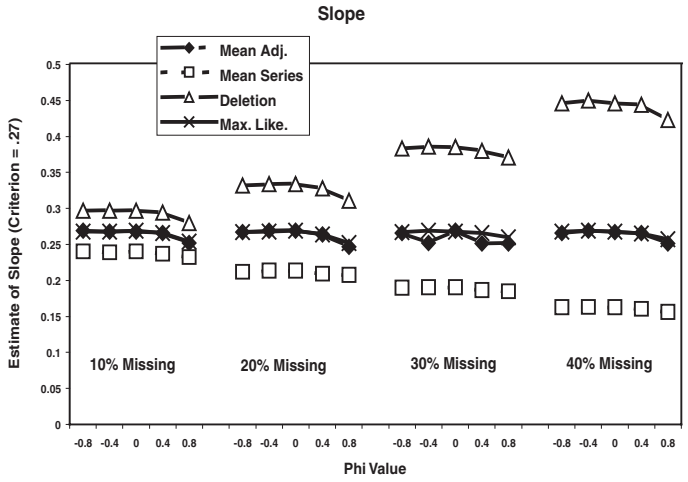

Figure 1. Mean estimate slope as a function of percentage missing data and ϕ for slope present (criterion = .27) conditions.

Dependence

Because ϕ was an independent variable in this study, five separate 2 × 5 × 4 (Slope × Percentage Missing × Technique) ANOVAs were done for each level of ϕ (–.80, –.40, .00, .40, and .80). In this design, slope was a between-groups factor; percentage missing and technique were both repeated-measures (i.e., within-groups) factors. The results were the most exaggerated for the ϕ = –.80 condition.

Phi = –.80. One two-way interaction, slope by technique, $F(3, 54) = 92.50$, $p < .001$, $\eta^2 = .050$, and all of the main effects were interpreted: slope, $F(1, 18) = 19.52$, $p < .001$, $\eta^2 = .029$; percentage missing, $F(4, 72) = 192.11$, $p < .001$, $\eta^2 = .311$; and technique, $F(3, 54) = 601.37$, $p < .001$, $\eta^2 = .324$. Figure 2 illustrates the results.

The maximum likelihood approach was the only method that yielded accurate estimates of ϕ under all conditions. All other techniques generally underestimated ϕ (i.e., produced estimates that were closer to zero than the correct value of –.80). Estimation was slightly less accurate when $S = 15°$ as opposed to 0°. Using mean of the adjacent observations produced the least accurate estimates of ϕ; these estimates got substantially worse at each increasing level of missing data. At 40% missing data, this approach estimated ϕ to be moderately positive, at approximately .20 instead of the high negative value that was correct.

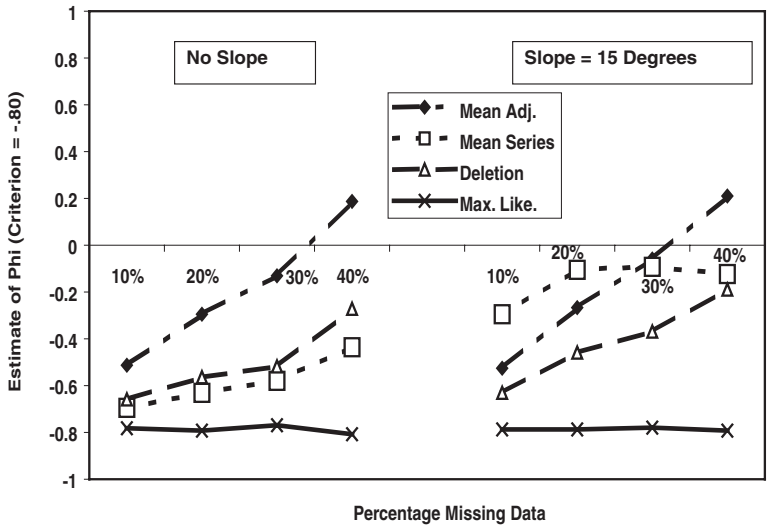

Figure 2. Mean estimates of ϕ as a function of percentage data missing and slope of series for the high negative dependency (criterion = –.80) condition.

Phi = –.40. One two-way interaction, percentage missing by technique, $F(12, 216) = 38.36$, $p < .001$, $\eta^2 = .146$, was interpreted. Two of the main effects were interpreted: percentage missing, $F(4, 72) = 68.30$, $p < .001$, $\eta^2 = .199$; and technique, $F(3, 54) = 176.37$, $p < .001$, $\eta^2 = .261$. The same pattern of results occurred as was observed with –.80. Maximum likelihood was the only accurate method. All other techniques generally underestimated ϕ. Mean of adjacent observations was the most inaccurate approach. Estimation was slightly worse when $S = 15°$.

Phi = .00. One two-way interaction was interpreted: percentage missing by technique, $F(12, 216) = 29.96$, $p < .001$, $\eta^2 = .145$. Two main effects were interpreted: percentage missing, $F(4, 72) = 8.89$, $p < .001$, $\eta^2 = .073$; and technique, $F(3, 54) = 161.38$, $p < .001$, $\eta^2 = .269$. The incorrect estimates all involved using the mean of the adjacent observations. This method led to severe overestimates of ϕ (.20 to .40 instead of the criterion value of .00). Estimation got substantially worse as the percentage of missing data increased. The other three methods produced generally accurate results.

Phi = .40. Two two-way interactions were interpreted: technique by slope, $F(3, 54) = 35.41$, $p < .001$, $\eta^2 = .067$; and technique by percentage missing, $F(12, 216) = 32.64$, $p < .001$, $\eta^2 = .141$. One main effect was interpreted: technique, $F(3, 54) = 188.54$, $p < .001$, $\eta^2 = .355$. The mean of the adjacent observations was inaccurate under all conditions, overestimating ϕ.

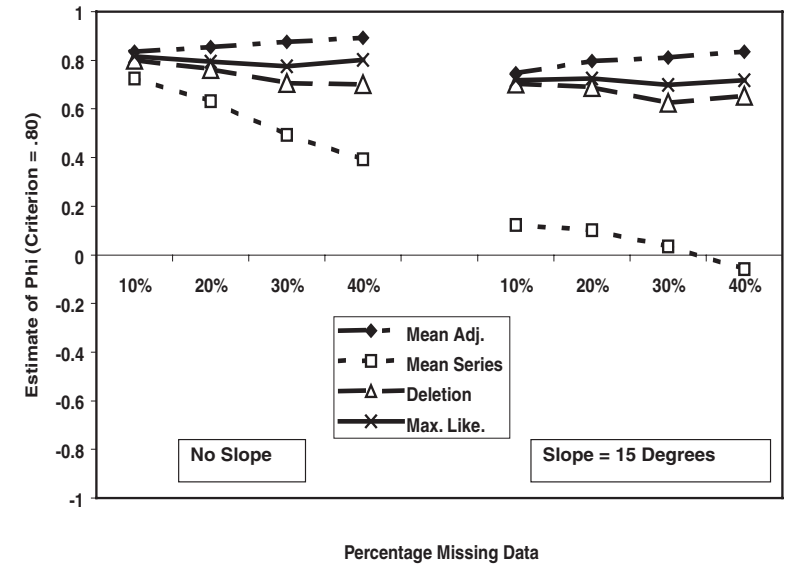

Figure 3. Mean estimates of ϕ as a function of percentage data missing and slope of series for the high positive dependency (criterion = .80) condition.

At 40% missing data, this approach led to average estimates of ϕ ranging from .59 to .62, compared with the correct value of .40. Using mean of the series resulted in underestimates of ϕ when $S = 15°$. Maximum likelihood resulted in accurate estimates of ϕ under all conditions. Deletion was generally accurate.

Phi = .80. Two two-way interactions were interpreted: technique by slope, $F(3, 54) = 149.46$, $p < .001$, $\eta^2 = .091$; and percentage missing by technique, $F(12, 216) = 78.35$, $p < .001$, $\eta^2 = .133$. Three main effects were interpreted: slope, $F(1, 18) = 18.33$, $p < .001$, $\eta^2 = .110$; percentage missing, $F(4, 72) = 41.50$, $p < .001$, $\eta^2 = .043$; and technique, $F(3, 54) = 693.21$, $p < .001$, $\eta^2 = .422$. As illustrated in Figure 3, the use of maximum likelihood always resulted in accurate estimates of ϕ. The use of the mean of the series always led to underestimates of ϕ. Estimates of ϕ when using mean of the series were substantially worse when $S = 15°$, with estimates ranging from .12 when 10% of the data were missing to –.06 at 40% missing data. Mean of adjacent observations and deletion were generally accurate.

Discussion

This study compared the accuracy of four methods for handling missing data in time-series analysis. Accuracy was measured by comparing four

parameter estimates (level, error variance, slope, and dependency) for the time-series model with the known population parameters. To evaluate the strengths and the limitations of the four methods, time series were computer generated that manipulated two different levels of slope, five different levels of autocorrelation, and five different levels of percentage of data missing in the series. The effect sizes associated with the differences ranged from large to extremely large ($L = 16$; $\sigma^2 = .268$; $S = .575$; and dependency = .324 [$\phi = -.80$], .261 ($\phi = -.40$], .269 [$\phi = .00$], .355 [$\phi = .40$], and .422 [$\phi = .80$], where $\eta^2 = .14$ was classified as large by Cohen, 1988).

The maximum likelihood procedure for handling missing data outperformed all others. Although this result was expected, the degree of accuracy was very impressive. The method provided accurate estimates of all four parameters. Furthermore, the method provided accurate parameter estimates across all levels of missing data, even when 40% of the data had been randomly eliminated. These results contradict a common statistical "rule of thumb" that suggests that data should be used only if there is less than 10% or 20% missing data. However, this result is limited by the fact that only one sample size ($N = 100$) was considered. The results may be primarily a function of the size of the remaining sample size after deletion.

The maximum likelihood method was so accurate that researchers might consider "planned missingness" studies (i.e., studies in which the series extends over a longer period of time but costs are contained by randomly not assessing the series on a planned percent of the occasions). Alternatively, if assessment is performed on a random sample of the possible occasions, time series could still be used by treating an interval in which no observation occurred as a randomly missing observation. In retrospect, this study should have included more levels of percentage missing data, because the method was so accurate at the 40% missing level that it did not adequately test the limits of the procedure. A minor limitation of the maximum likelihood technique is that estimates of the actual missing data points cannot be obtained in the current version of SAS/ETS. This could be important to a researcher who prefers a visual presentation of the data.

Imputing the mean of the series is an unacceptable method for handling missing data in longitudinal designs. Whenever a slope parameter was introduced into the data, this method led to very inaccurate estimates of all four parameters. The severe overestimates of error variance and level would result in very inaccurate tests of significance. These results reflect the fact that this procedure ignores the ordinal position of the observations. When a slope is introduced, the temporal order becomes extremely important. Phi was moderately underestimated by this method, regardless of the slope of the series. These findings indicate that this method of handling missing data should not be used for longitudinal designs, even when as few as 10% of the observations are missing.

The other two ad hoc methods also produced inaccurate estimates for some of the parameters. The mean of adjacent observations produced reasonable estimates of level and slope. However, the method produced extremely inaccurate estimates of the dependency parameter. This can be explained by considering a specific example. Assume that a large negative dependency is present in the data. That means that an extremely high observation is likely to be followed by an extremely low observation and then another extremely high observation. The mean of adjacent would replace the missing extremely low observation by an extremely high observation, the mean of the two adjacent observations. This would result in an underestimate of the degree of dependency present in the data. In addition, when $\phi = -.80$, mean of the adjacent observations substantially overestimated error variance.

Deletion was generally accurate for the estimation of level and error variance but was inaccurate for the two longitudinal parameters. Deletion led to an overestimate of the slope. This can be explained by the fact that compressing a time series into a shorter interval will require a steeper slope (i.e., the rate of change of the series over time to fit the data). Deletion was also inaccurate for moderate and high degrees of negative dependency. For the estimation of this parameter, deletion will operate in much the same way as the mean of adjacent, placing a high observation next to another high observation by the act of closing up the series, resulting in an underestimate of the dependency in the data. In addition, when $\phi = -.80$, deletion substantially overestimated error variance.

All three ad hoc procedures were inaccurate for estimating the degree of dependency in the data. This accuracy can effect both model identification and tests of significance in an interrupted time-series analysis. Model identification can be an intermediate step in analysis, serving to identify to proper ARIMA model and selecting the transformation matrix needed for testing the significance of the parameters of interests. In other cases, model identification may be the goal of the study. Finding what model best represents a behavior may provide scientists with a better understanding the nature of a behavior of interest (Velicer, 1994; Velicer, Redding, Richmond, Greeley, & Swift, 1992). An inaccurate estimate of the dependency can also affect the test of significance in an interrupted time-series model (Crosbie, 1993). If a large negative dependency is present and is not accounted for in the analysis, the error variance will be overestimated, decreasing the value of the test statistic and potentially resulting in a Type II error. If a large positive dependency is present and is not accounted for in the analysis, the error variance will be underestimated, increasing the value of the test statistic and potentially resulting in a Type I error.

One limitation of this study is that the methods were examined under ideal circumstances. The data were normally distributed. The data were MCAR. The time-series model was accurately estimated. In applied situations, one or

more of these assumptions is likely to be violated. Violations of these assumptions will not improve the performance of the three ad hoc procedures. However, the performance of the maximum likelihood procedure may not be so impressive under less than ideal circumstances. Velicer and Colby (2005) have completed an investigation of the performance of this method under violations and determined that it was very robust. However, that study evaluated only the consistency of the estimator and did not investigate the impact of nonnormality on the estimation of standard errors.

In summary, this study found that the maximum likelihood approach yielded accurate estimates of all four parameters at all levels of missing data. This method should be viewed as the method of choice for time-series studies. The mean of the series resulted in very inaccurate estimates and is not recommended. All three ad hoc procedures produced inaccurate estimates of the dependency parameter. The results of this study also demonstrated that statistical methods for handling missing data represent a substantial improvement over the available ad hoc procedures. The effect sizes involved were generally very large. Modern statistical procedures for handling missing data should be used in all analysis when data are missing.

References

Algina, J., & Swaminathan, H. A. (1979). Alternatives to Simonton's analysis of the interrupted and multiple-group time series designs. *Psychological Bulletin*, *86*, 919-926.

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, *41*, 164-171.

Beale, E. M. L., & Little, R.J.A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society, Series B*, *37*, 129-145.

Box, G. E. P., & Jenkins, G. M. (1970). *Time-series analysis: Forecasting and control*. San Francisco: Holden-Day.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology*, *61*, 966-974.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1-38.

Dixon, W. J. (1988). *BMDP statistical software*. Berkeley: University of California Press.

Glass, G. V, Willson, V. L., & Gottman, J. M. (1975). *Design and analysis of time series experiments*. Boulder: University of Colorado Press.

Gottman, J. M. (1981). *Time series analysis: A comprehensive introduction for social scientists*. New York: Cambridge University Press.

Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka & W. F. Velicer (Eds.), *Research methods in psychology* (pp. 87-114). New York: John Wiley.

Harrop, J. W., & Velicer, W. F. (1985). A comparison of three alternative methods of time series model identification. *Multivariate Behavioral Research*, *20*, 27-44.

Harrop, J. W., & Velicer, W. F. (1990). Computer programs for interrupted time series analysis: II. A quantitative evaluation. *Multivariate Behavioral Research*, *25*, 233-248.

Hartley, H. O. (1958). Maximum likelihood estimation from incomplete data. *Biometrics*, *14*, 174-194.

Harvey, A. C., & Pierse, R. G. (1984). Estimating missing observations in economic time series. *Journal of the American Statistical Association*, *79*, 125-131.

Jones, R. H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics*, *22*, 389-396.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, *82*, 34-45.

Kohn, R., & Ansley, C. F. (1986). Estimation, prediction, and interpolation for ARIMA models with missing data. *Journal of the American Statistical Association*, *81*, 751-761.

Laird, N. M. (1988). Missing data in longitudinal studies. *Statistics in Medicine*, *7*, 305-315.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley.

Marsh, J. C., & Shibano, M. (1984). Issues in the statistical analysis of clinical time-series data. *Social Work Research and Abstracts*, *20*, 7-12.

McCleary, R., & Hay, R. A., Jr. (1980). *Applied time series analysis for the social sciences*. Beverly Hills, CA: Sage.

Padia, W. L. (1975). The consequences of model misidentification in the interrupted time-series experiment. *Dissertation Abstracts International*, *36,* 4875A. (University Microfilms No. 76-3938)

Rankin, E. D., & Marsh, J. C. (1985). Effects of missing data on the statistical analysis of clinical time series. *Social Work Research and Abstracts*, *21*, 13-16.

Revenstorf, D., Kessler, A., Schindler, L., Hahlweg, K., & Bluemner, E. (1980). Time series analysis: Clinical applications evaluating intervention effects in diaries. In O. D. Anderson (Ed.), *Analyzing time series: Proceedings of the international conference held on Guernsey, Channel Islands, October 1979* (pp. 291-312). Amsterdam, the Netherlands: North-Holland.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581-592.

SAS Institute. (1988). *SAS/ETS user's guide, version 6*. Cary, NC: Author.

Schafer, J. L. (1997). *Analysis of incomplete multivariate.* New York: John Wiley.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147-177.

Simonton, D. K. (1977). Cross-sectional time-series experiments: Some suggested statistical analyses. *Psychological Bulletin*, *84*, 489-502.

Velicer, W. F., & Colby, S. M. (1997). Time series analysis for prevention and treatment research. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 211-249). Washington, DC: American Psychological Association.

Velicer, W. F., & Colby, S. M. (2005). Missing data and the general transformation approach to time series analysis. In A. Maydeu-Olivares A. & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift to Roderick P. McDonald.* Hillsdale, NJ: Lawrence Erlbaum.

Velicer, W. F., & Fava, J. L. (2003). Time series analysis. In J. Schinka & W. F. Velicer (Eds.), *Research methods in psychology* (pp. 581-606). New York: John Wiley.

Velicer, W. F., & Harrop, J. W. (1983). The reliability and accuracy of time series model identification. *Evaluation Review*, *7*, 551-560.

Velicer, W. F., & McDonald, R. P. (1984). Time series analysis without model identification. *Multivariate Behavioral Research*, *19*, 33-47.

Velicer, W. F., & McDonald, R. P. (1991). Cross-sectional time series designs: A general transformation approach. *Multivariate Behavioral Research*, *26*, 247-254.

Velicer, W. F., Redding, C. A., Richmond, R. L., Greeley, J., & Swift, W. (1992). A time series investigation of three nicotine regulation models. *Addictive Behaviors*, *17*, 325-345.