Name: Mingyuan Cheng

Title: Effect of filling methods on the forecasting of time series with missing values

Abstract

The Gulf of Alaska Mooring (GAK1) monitoring data set is an irregular time series of temperature and salinity at various depths in the Gulf of Alaska. One approach to analyzing data from an irregular time series is to regularize the series by imputing or filling in missing values. In this project we investigated and compared four methods (denoted as APPROX, SPLINE, LOCF and OMIT) of doing this. Simulation was used to evaluate the performance of each filling method on parameter estimation and forecasting precision for an Autoregressive Integrated Moving Average (ARIMA) model. Simulations showed differences among the four methods in terms of forecast precision and parameter estimate bias. These differences depended on the true values of model parameters as well as on the percentage of data missing. Among the four methods used in this project, the method OMIT performed the best and SPLINE performed the worst. We also illustrate the application of the four methods to forecasting the Gulf of Alaska Mooring (GAK1) monitoring time series, and discuss the results in this project.

Introduction

A time series is a collection of data points obtained through repeated measurement over time. The daily value of the stock market and the quarterly sales volume of a product are all examples of time series data. Time series analysis can provide information about the internal structure of data, such as autocorrelation and seasonal variation, and forecast future trends. Most time series are regular time series, such as data measured at daily, weekly, monthly, quarterly or annual intervals. However, irregular time series, where data are recorded at irregular intervals, often occur, for example in long-term field data collections. This may be due to factors such as weather, funding, and others.

Missing data are a common problem in quantitative research studies such as biology, marine sciences, social sciences, and space sciences (Acock 2005, Baraldi 2009). Missing data may cause large statistical and design problems in research. According to Acock (2005), there are four different types of missing data: missing by definition of the subpopulation, missing completely at random (MCAR), missing at random (MAR), and nonignorable missing data. Missing by definition of the subpopulation happens when some survey participants are excluded from the analysis because they are not in the subpopulation under investigation. The idea of missing completely at random is that thinking of the data set as a large matrix, the missing values are randomly distributed throughout the matrix. Missing at random is a much weaker (but still strong) assumption than MCAR. Nonignorable missing data are the data missing in ways that are neither MAR nor MCAR, but nevertheless are systematic. Two approaches, traditional and modern alternative approaches, are commonly used to work with missing data. The traditional approaches to working with missing data include listwise deletion, pairwise deletion, indicator variable, and mean substitution (Acock 2005). Listwise or case deletion is the most common solution to missing values, it removes all data for a case that has one or more missing values, and the listwise deletion will give unbiased estimates if the missing values are MCAR. Pairwise deletion, unlike listwise deletion, only removes the specific values from the analysis (not the entire case). Indicator variable is a method for dealing with missing data on predictors in a regression analysis. For each predictor with missing data, a dummy variable is created to indicate whether or not data are missing on that predictor. Mean substitution uses the sample mean to replace the missing values. Traditional approaches, especially mean substitution, for working with missing data can lead to biased estimates and may either reduce or exaggerate statistical power and lead to invalid conclusions (Acock 2005). Modern alternative approaches are single imputation, multiple imputation, and full information maximum likelihood estimation (Acock 2005). Single imputation is hot-dock imputation where a missing value is impute from a randomly selected similar record. Multiple imputations are run on the same data set multiple times and the imputed data sets are saved for later analysis. Full information maximum likelihood estimation does not actually impute missing values but uses all the available information to provide maximum likelihood estimation and maximum likelihood estimation has proven to be an excellent method for handling missing data in a wide variety of situations. Investigations of methods for imputing missing values for forecasting in time series include Junninen et al. (2004) and Plaia and Bondi (2006).

Time series forecasting is the process of using a model to generate predictions or forecasts of future events based on known past events. Many models can be used for time series forecasting. Among them is the Autoregressive Integrated Moving Average (ARIMA) model which consists of three parts, a autoregressive (AR) part (future values are estimated based on a weighted sum of past values), a Moving Average (MA) part (a linear regression of the current value of the series against current and previous white noise error terms or random shocks), and an integrated (I) part (where an initial differencing step can be applied to remove any non-stationarity in the signal). ARIMA, introduced by Box and Jenkins (1970), which has been one of the most popular approaches to forecasting and has been used to forecast, for instance, daily peak electricity demand (Asad, 2012), world broadband and mobile telecommunications' penetration (Christodoulos et al. 2010), traffic flow (Williams et al. 2003) and stock price (Pai and Lin 2005).

The data for the Gulf of Alaska Mooring (GAK1) (http://www.ims.uaf.edu/gak1/) time series is one example of an irregular time series. This time series data contains measurements of temperature, salinity and depth profiles obtained from an oceanographic station located at the mouth of Resurrection Bay near Seward, Alaska. The data collection began in 1970, and the time interval of sampling varies from several times per month to a few times per year. One approach to dealing with irregular time series is to regularize them, either by deleting values or adding values at given intervals, in order to create regularly-spaced measurements. The purpose of this project is to determine the best methods to "fill in" missing data in a time series. Of particular interest is identifying which methods result in the most reliable forecasts.

Our objective in this study is to forecast the measurements in the GAK1 data sets using ARIMA models. Because the time series in the data sets are irregular, we worked on identifying an appropriate method for filling in missing values. We performed a simulation study to compare four filling methods on the basis of forecasting accuracy and parameter estimate bias. We considered two additional factors in our simulations, specifically we investigated how the true values of model coefficients and the percentage of observations that were missing affected forecasting accuracy and parameter estimate bias. After simulation, the methods that least affected time series forecasting accuracy in terms of Mean Absolute Deviation (MAD) were used to transform four irregular data sets from the GAK1 time series (from 2000 to 2010) to regular data, then ARIMA models were fitted. Three known observations were forecasted according to the fitted models, and the performance of the filling methods were compared for these time series.

1. Background

1.1. Filling Methods for Time Series Data

Four methods, which we will denote APPROX, SPLINE, LOCF and OMIT, were selected to fill the missing values. These four methods were carried out by the functions *na.approx*, *na.spline*, *na.locf* and *na.omit*, which are provided in the zoo package in R (Kukuyeva 2010; Zeileis et al. 2014). The first three methods belong to modern approaches to interpolating missing values and the last method performs listwise deletion, a traditional approach to working with missing values. Next we describe each of the four methods.

1.1.1. APPROX: This method replaces missing values, denoted by NA, with a linear approximation from surrounding observations. For a value x in the interval (x_0, x_1), the value y along the straight line is given by the equation:

$$y = y_0 + (y_1 - y_0) * \frac{x - x0}{x1 - x0}$$

This is the formula for linear interpolation in the interval (x_0, x_1). For example, suppose we have three observations 1, <u>NA</u>, 2, <u>NA</u>, 10 and want to create two additional observations between the first and the second, and between the second and the third. A linear approximation would result in the series 1, 1.5, 2, 6, 10.

1.1.2. SPLINE: This method replaces the missing values with a cubic spline interpolation. Suppose we have a table of points [x_i, y_i] for i=0, 1,…, n for the function y=f(x), with n+1 points and n intervals between them. The cubic spline curve is a piecewise polynomial curve and is constructed by using a different cubic polynomial curve between each two data points. There is a separate cubic polynomial for each interval, each with its own coefficients:

$$F_i(x)=A_i(x-x_i)^3 + B_i(x-x_i)^2 + C_i(x-x_i) + D_i \qquad \text{for x from } [x_i, x_{i+1}]$$

Together, these polynomial segments are denoted $F(x)$, the spline. Since there are n intervals and four coefficients for each we require a total of 4n parameters to define the spline $F(x)$. We need to find 4n independent linear equations to fix them. Two equations for each interval come from the requirement that the cubic polynomial match the values of the table at both ends of the interval:

$$F_i(x_i)=y_i, F_i(x_{i+1})=y_{i+1}$$

This gives n +1 conditions.

The continuity conditions for the spline, its derivative, and its second derivative are the following:

$$F_{i-1}(x_i) = F_i (x_i), F_{i-1}'(x_i) = F_i'(x_i), F_{i-1}''(x_i) = F_i''(x_i)$$

These conditions apply for i=1,2,…,n-1, resulting in 3(n-1) constraints. So we need two more conditions to completely fix the spline.

$F_0''(x_0) = 0$, $F''_n(x_n) = 0$

With 4n coefficients and 4n linear conditions it is straightforward to work out the equations that determine them.

1.1.3. LOCF : This method is known a Last Observation Carried Forward and replaces each NA value with the most recent non-NA prior to it (if the first observation is NA it is removed by default) (Kukuyeva, 2010; Zeileis et al 2014). For example, one data set with 1, NA, 3, 5, NA, 8 will become 1, 1, 3, 5, 5, 8 when LOCF is used.

1.1.4. OMIT: This method is used just to omit the NA (Kukuyeva 2010; Zeileis et al 2014) and it is the same as the listwise deletion of the traditional approaches. For example, the data set with 1, NA, 3, 5, NA, 8 will become 1, 3, 5, 8 when OMIT is used.

1.2. ARIMA model

The Autoregressive Integrated Moving Average (ARIMA) model may be used to analyze and forecast equally spaced univariate time series data. The model is generally referred to as an ARIMA (p, d, p) where parameters p, d, and q are non-negative integers that refer to the number of autoregressive terms, the number of nonseasonal differences, and the number of lagged errors in the prediction equation. In an ARIMA model, a future value is assumed to be a linear combination of past values and past and present errors. The equation of the ARIMA is:

$$Y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots\ldots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots\ldots\ldots \theta_q \varepsilon_{t-q}$$

where y_t and ε_t are the actual value and random error at time period t, respectively, and ϕ_i (i=1, 2, ….. ,p) and θ_j (j=0, 1, 2, …. , q) are model parameters. The values p and q are integers and are often referred to as orders of the model. Random errors ε_t , are assumed to be independently and identically distributed with a mean of zero and a constant variance of σ^2. When d=0, the ARIMA (1, 0, 1) may be expressed as ARMA (1, 1) and is represented as follows

$$y_t = \theta_0 + \phi_1\, y_{t-1} + \varepsilon_t + \theta_1\, \varepsilon_{t-1}.$$

1.3. Best models for irregular GAK1 time series

Four datasets, temperature at surface, temperature at 250 meters, salinity at surface and salinity at 250 meters, were initially fit using ARIMA models. A variety of orders were investigated including ARIMA (1, 0, 1), ARIMA (0, 1, 0), ARIMA (1, 1, 0), ARIMA (0, 1, 1) and ARIMA (0, 2, 1). The best ARIMA models for all four datasets were selected according to the lowest Akaike Information Criterion (AIC) values. The test results showed that ARIMA (1, 0, 1) had the lowest AIC value and the ARIMA (1, 0, 1) model was used in this project for all four time series and also used for simulations. We note that this selection was determined using the original irregular time series, and therefore may be inaccurate. However, it provides a starting point for our investigation.

2. Simulations

2.1. Simulation Design

Data for simulations were generated under the ARIMA (1, 0, 1) model. Data were generated having different values of model coefficients ϕ and θ. The variance σ^2 was set to one for all simulations. For each simulated data set a given proportion of values were randomly deleted, and then replaced using each of the four filling methods. To evaluate the accuracy of the forecasts, the Mean Absolute Deviation (MAD) was computed and used to compare different models. The calculation of MAD is as follows

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n} |Ai - Fi|$$

where n is the number of forecast values, Ai are the actual observed values and Fi are the forecast values from a fitted model.

The simulation procedure is described in the flowchart (Fig. 1). Briefly, first, the simulations began by fixing the value of the parameters (ϕ and θ) to generate 203 observations using arima.sim in R. Second, varying percentages of the first 200 observations were randomly selected and deleted to create new data sets. These missing values in each new data set were filled in by the four methods, APPROX, SPLINE, LOCF and OMIT. The ARIMA (1, 0, 1) models were fit using the data set with non-missing values. Then each fitted model was used to forecast the remaining 3 values. In addition the values of the estimated parameters ϕ and θ were recorded. Forecast values were used to compute MAD. Parameter estimates were used to estimate bias. Differences in mean MAD among the four filling methods were analyzed using one-way ANOVA. The Tukey Honest Significant Difference method and pairwise *t* tests were conducted to test the MAD differences between filling methods and also between filling methods and non-missing control. Significant bias in estimated parameters was determined with t-tests. All comparisons were made at significance level $\alpha = .05$.

Three factors were varied during the simulation. These three factors and levels of each factor were six levels of percentages of missing observations (5, 20, 30, 40, 50 and 70%), four levels of

1st order AR-MA values (level 1: ϕ =0.1, θ=0.1, level 2: ϕ =0.1, θ=0.9, level 3: ϕ =0.5, θ=0.5 and level 3: $\phi = 0.7$, $\theta = 0.2$), four filling methods (APPROX, SPLINE, LOCF and OMIT). The original simulated data, with no missing values served as a CONTROL. For each combination of factor levels, 1000 such data sets were simulated resulting in 1000 MADs and 1000 estimated AR and estimated MA coefficients.

2.2. Simulation results

To investigate the effect of the percentage of missing values and filling in methods on the forecast of the time series, three simulations were conducted separately for each of the four levels of true AR-MA values (level 1: ϕ =0.1, θ=0.1, level 2: ϕ =0.1, θ=0.9, level 3: ϕ =0.5, θ=0.5 and level 4: ϕ =0.6, θ=0.2) using a 6 (factor percentage) × 5 (factor methods including control) factorial design. Analysis of Variance was used to compare mean MAD among combination of the factors of missing and filling method. The effect of percentage of missing values and filling in methods on the bias of estimated coefficients (ϕ and θ) was similarly investigated using the same factorial design. For each fitted model, bias was estimated as the difference between the estimates AR and MA coefficient and the true values of the coefficients. Analysis of Variance models were fit separately to investigate the effects of percentage of missing + filling method on means of bias of estimated AR and MA coefficients.

2.2.1. The effect of methods and percentage of missing values on the mean MAD and the bias of estimated AR and MA values when the true AR-MA values at level 1

The results for mean MAD:

- When the true AR-MA values were at level 1, both missing percentage (F(5, 29970) =9.1, $p < .001$) and methods (F(4, 29970)= 50.1, $p < .001$) significantly affected the forecast of time series in terms of mean MAD but interaction between two factors was not statistically significant (F(20, 29970) = 0.05, $p > .05$). Post-hoc analyses using Tukey Honest Significant Difference method indicated that for factor methods, only method OMIT had no significant difference compared to no missing CONTROL at $\alpha = .05$ and performed best (Table 1). There are no significantly different between the method APPROX and LOCF (Table 1). The method SPLINE was the worst in all four methods (Table1). For factor percentage, the effect were no significant differences on MAD between 5%, 30%, and 50%, no differences between 5%, 20%, 50% , and 70%, and no differences between 5%, 20%, 40% and 70% (Table 2).

The results for the estimated AR

- When the true AR-MA values were under level 1 ($\phi = 0.1$ and $\theta = 0.1$), simulation indicated that percentage of missing values (F(5, 29970) =2.28, $p = .044$) and methods (F(4, 29970) = 1363, $p < .001$) had a significant effect on the bias but the interaction between two factors (F(20, 29970) = 0.29, $p > .05$) was not statistically significant. Post-hoc analyses using Tukey Honest Significant Difference method indicated that for factor methods, all four methods and CONTROL were biased (Table 3). The estimated AR values from all four methods were more biased than those from CONTROL at $\alpha = .05$ and the values from APPROX and LOCF were the most biased (Table 3). For factor percentage, the estimated AR values were biased at all different percentages of missing values (Table 4). The effect of missing values on the bias of estimated AR values can be divided into two significant different groups. One group was 20%, 40% and 70%, and the other was 5%, 30% and 50% (Table 4).

The results for estimated MA

- When the true AR-MA values were under level 1 ($\phi = 0.1$ and $\theta = 0.1$), simulation indicated that factor percentage (F(5, 29970) = 3.69, $p = .0024$) methods (F(4, 29970) = 11778, $p < .001$) had significant effect on the bias of estimated MA value but interaction between two factors (F(20, 29970) = 0.43, $p > .05$) was not statistically significant. Post-hoc analyses using Tukey Honest Significant Difference method indicated that for factor methods, all four methods and CONTROL had significant different effect on the bias of estimated MA values and the estimated values from method SPLINE was the most biased at $\alpha = .05$ (Table 5). For the factor percentage, the estimated MA values were biased at all percentages of missing values. The effect of percentages on the bias of the estimated MA values can be divided into three groups. The first group was 5%, 30% and 50%, the second group was 5%, 20%, 50% and 70%, and the third group was 5%, 20%, 40% and 70% at $\alpha = .05$ (Table 6).

2.2.2. The effect of methods and percentage of missing values on the mean MAD and the bias of estimated AR and MA values when the true AR-MA values at level 2

The results for mean MAD values

- When the true AR-MA values were at level 2, factor methods (F(4, 29970)= 55.8, $p < .001$) and factor percentage (F(5, 29970) =10.2, $p < .001$) significantly affected the forecast of time series in terms of mean MAD but the interaction was not statistically significant (F(20, 29970) = 0.14, $p > .05$). Post-hoc analyses using Tukey Honest Significant Difference method indicated that for factor methods, all four methods had significant differences compared to no missing CONTROL at $\alpha = .05$ (Table 1). The method SPLINE was the worst; there was no significantly different between the method APPROX and LOCF and no significantly different between the method LOCF and OMIT. For factor percentage, the effect of missing values on the forecast can be divided into two significant different groups. One group was 5%, 30% and 50% and the other was 20%, 40% and 70%, (Table 2).

The results for estimated AR values

- When the true AR-MA values were under level 2 ($\phi = 0.1$ and $\theta = 0.9$), only factor methods affected the bias of estimated AR values (F(4, 29970) = 5408, $p < .001$), factor percentage (F(5, 29970) =1.41, $p = 0.22$) and interaction between two factors (F(20, 29970) = 0.12, $p > .05$) were not statistically significant. Post-hoc analyses using Tukey Honest Significant Difference method indicated that for factor methods, the estimated AR values were biased at all four methods and CONTROL (Table 3). The estimated AR values from all four methods were more biased than those from no missing CONTROL at $\alpha = .05$ and the values from SPLINE were the most biased (Table 3). For factor percentage, there were no significant different effect on the bias of estimated AR values between 5%, 20%, 30%, 40%, 50%, and 70% (Table 4).

The results for estimated MA values

- When the true AR-MA values were under level 2 ($\phi = 0.1$ and $\theta = 0.9$), only factor methods affected the bias of estimated MA values (F(4, 29970) = 11986, $p < .001$), factor percentage (F(5, 29970) =1.65, $p = 0.14$) and interaction between two factors (F(20, 29970) = 0.50, $p > .05$) were not statistically significant. Post-hoc analyses using Tukey Honest Significant Difference method indicated that for factor methods, the estimated

MA values from all four methods and CONTROL were biased at α = .05 (Table 5). All four methods and CONTROL had significant different effect on the bias of estimated MA values and the estimated values from method APPROX and LOCF was the most biased at α = .05 (Table 5). For factor percentage, the estimated MA values from all six proportion of missing values were biased but there were no significant differences between all six proportion of missing values at α = .05 (Table 6).

2.2.3. The effect of methods and percentage of missing values on the mean MAD and the bias of estimated AR and MA values when the true AR-MA values at level 3

The results for mean MAD values

- When the true AR-MA values were at level 3, factor methods significantly affected the forecast of time series in terms of mean MAD (F(4, 29970)= 5.38, $p < .001$) but factor percentage (F(5,29970) = 0.011, $p > .05$) and interaction (F(20, 29970) = 0.46, $p > .05$) were not statistically significant. Post-hoc analyses using Tukey Honest Significant Difference method indicated that three methods (APPROX, LOCF and OMIT) had no significant difference compared to no missing CONTROL at α = .05 (Table 1). The method SPLINE was the worst and significantly different from the other three methods and CONTROL (Table 1).

The results for estimated AR values

- When the true AR-MA values were under level 3 ($\phi = 0.5$ and $\theta = 0.5$), only factor methods affected the bias of estimated AR values (F(4, 29970) = 107.8, $p < .001$) but factor percentage (F(5, 29970) = 0.23, $p = .95$) and the interaction between two factors (F(20, 29970) = 0.009, $p > .05$) were not statistically significant. Post-hoc analyses using Tukey Honest Significant Difference method indicated that for factor methods, the estimated AR values were biased at all four methods and CONTROL at α = .05 (Table 3). The estimated AR values from APPROX and SPLINE were the most biased (Table 3). For factor percentage, there were no significant different effect on the bias of estimated AR values between 5%, 20%, 30%, 40%, 50%, and 70% (Table 4).

The results for estimated MA values

- When the true AR-MA values were under level 3 (ϕ = 0.5 and θ = 0.5), only factor methods affected the bias of estimated MA values (F(4, 29970) = 497.4, *p* < .001) while factor percentage (F(5, 29970) = 0.21, *p* = .96) and the interaction (F(20, 29970) = 0.098, *p* > .05) between two factors were not statistically significant. Post-hoc analyses using Tukey Honest Significant Difference method indicated that for factor methods, all four methods and CONTROL had significant different effect on the bias of the estimated MA values and the estimated values from method LOCF was the most biased at α = .05 (Table 5). For factor percentage, the estimated MA values from all six proportion of missing values were biased but there were no significant differences between all six proportion of missing values at α = .05 (Table 6).

2.2.4. The effect of methods and percentage of missing values on the mean MAD and the bias of estimated AR and MA values when the true AR-MA values at level 4

The simulation results for mean MAD values

- When the true AR-MA values were at level 4, factor percentage (F(5, 29970) =0.961, *p* > .05) and methods (F(4, 29970)= .96, *p* > .05) and interaction between (F(20, 29970 = 0.015, p > .05) did not significantly affect the forecast of time series in terms of mean

MAD. Post-hoc analyses using Tukey Honest Significant Difference method indicated that for factor methods, all four methods had no significant difference compared to no missing CONTROL at α = .05 (Table 1). For factor percentage, there were no significant differences between 5%, 20%, 30%, 40%, 50% and 70%.

The simulation results for estimated AR values

- When the true AR-MA values were under level 4 (ϕ = 0.7 and θ = 0.2), factor percentage (F(4, 29970) = 0.89, *p* > .05), factor methods (F(4, 29970) = 0.49, *p* > .05) and interaction between two factors (F(4, 29970) = 0.0035, *p* > .05) did not affected the bias of estimated AR values (Table 3 - 4) although the estimated AR values were biased.

The simulation results for estimated MA values

- When the true AR-MA values were under level 4 (ϕ = 0.7 and θ = 0.2), only factor method had significantly affected on the bias of estimated MA (F(5, 29970) = 46.7, *p* < .001) while factor percentage (F(4, 29970) = 0.73, *p* > .05), and interaction between two factors (F(4, 29970) = 0.12, *p* > .05) did not affected the bias of estimated MA values. Post-hoc analyses using Tukey Honest Significant Difference method indicated that for factor methods, all four methods and CONTROL had significant effect on the bias of the estimated MA values at α = .05 (Table 5). For factor percentage, the estimated MA values from all six proportion of missing values were biased there were no significant differences between all six proportion of missing values at α = .05 (Table 6).

In summary, the filling in methods and the percentages of missing values significantly affected the forecast of time series in term of MAD values but the effect depended on the internal structure of data sets (i.e. ϕ and θ values). OMIT performed best and SPLINE performed worst in all four methods. When the values of ϕ increased, the effect of the filling methods and the percentages of missing values on the forecast of time series decreased. The estimated AR and MA values generated from all the simulations were biased. Generally, for the estimated AR values, the bias was positive when ϕ value was 0.1 and the bias was negative when ϕ values was 0.5 and 0.7 regardless of the method used and the percentages of missing values. For the estimated MA values, the bias was positive only when the θ value was 0.1 no matter what kinds of methods were used and what percentages of missing values were.

Applications

The GAK1 time series is a data set of temperature and salinity changes with time at different depths (from 0 to 250 meter below surface). However, the data are not a regular monthly time series since no observations were recorded in some months and multiple observations were recorded in other months. In order to forecast the GAK1 time series, the multiple values from the same month were averaged and this resulted in an irregular monthly time series, with a proportion of missing observation of 23.4%. The estimated AR and MA coefficients were approximately $\phi = 0.7$ and $\theta = 0.2$ according the fitted ARIMA (1, 0, 1) models from irregular data sets. Our simulation results suggest that the four filling in methods did not significantly affect the forecast of time series in mean MAD when $\phi = 0.7$ and $\theta = 0.2$. Therefore, we use all four methods to fill in the missing values and to forecast the last true observations in each time series and compare the methods. After filling, ARIMA (1, 0, 1) models were fit to all but the last three values of each series and the Mean Absolute Deviation (MAD) values were calculated using the 3-ahead forecasted values and last three values. The time series plots constructed from the filling data sets showed that the four filling methods, except OMIT, created a regular time series with monthly interval (Figs. 2-5). The MAD values were smaller for the method OMIT except the data set for salinity time series at the surface, indicating OMIT generally performed better than the other three methods (Table 7). The MADs from the data sets at 250 meters were much smaller than those at the surface (Table 7). Therefore, the forecasting of temperature and salinity time series from the 250 meters measurement was better than those from the surface (Table 7). Salinity time series forecasting were better than temperature time series because of smaller MADs (Table 7). These results suggest that the proportion of missing values may not be the only factor to affect the forecast of the time series, the variance of temperature and salinity time series at 250 meters was smaller than the one at the surface and the variance of salinity time series was smaller than the ones of temperature time series. This may be one of the reasons to explain the smaller MADs for the data sets of salinity, and for the data sets of temperature and salinity at the 250 meters blow surface.

In the applications, the MAD values from method SPLINE and APPROX were higher than the ones from the methods LOCF and OMIT in three of four data sets, confirming the results from simulations that effects of methods SPLINE and APPROX on forecast of time series were significantly different than those of methods LOCF and OMIT.

Conclusions

The overall goal of this project was to investigate whether the four filling methods (APPROX, SPLINE, LOCF and OMIT) affected the forecasting of time series, and apply them to the GAK1 time series. Our overall conclusions are as follows:

- The irregular data set can be regularized by filling in the missing values using the four methods tested in this project but the method chosen would depend on the structure of the data set (the true coefficients values of AR and MA) and to a lesser degree on the proportion of missing values.

- APPROX may be used to fill missing values of time series when the true coefficient AR values are greater than 0.5 and 0.7 without significantly affecting the ARIMA forecasting. The mean MAD values from APPROX are the second highest in all four simulations.
- SPLINE is the worst among four methods. Our simulations show that SPLINE can be used in one of the four simulations (only when the true coefficient AR-MA values are at level 4: $\phi = 0.7$ and $\theta = 0.2$). The mean MAD values are always the highest in all four simulations.
- LOCF can be used to fill the missing values when the true coefficient AR-MA values are at level 3 ($\phi = 0.5$ and $\theta = 0.5$) and 4 ($\phi = 0.7$ and $\theta = 0.2$). The mean MAD values are the third highest in all simulations.
- OMIT can be used in three of four simulated conditions in this project and its performance is the best among the four methods. The only simulation that does not perform well is when the true coefficient AR-MA value is at level 2 ($\phi = 0.1$ and $\theta = 0.9$).
- The estimated AR and MA values from ARIMA models are biased for all four methods and CONTROL in the simulations. For the estimated AR values, the bias was positive when ϕ value was 0.1 and the bias was negative when ϕ values was 0.5 and 0.7. For the estimated MA values, the bias was positive only when the θ value was 0.1.
- In the application, the forecasting results from the fitted models show that OMIT has the lowest MAD of the four methods. This confirms the result from the simulations that OMIT is the best method to fill the missing values in this project. The better

performance for forecast of the data sets from the 250 meters below the surface and of data sets from salinity may relate to the internal structure of the data sets.

Forecasting time series data is important because these data often provide the foundation for decision models. Selecting the correct and best models to use to forecast is also important. In this project, the ARIMA model is the only model selected to test the effect of missing observations on the forecasting. In the future, more models should be selected to test and the time series with trend should also be included to determine the effect of missing values on forecasting and compare with the current research.

Acknowledges

I would like to thank my committee co-chairs Dr. Julie McIntyre and Dr. Ronald Barry for their guidance and valuable time, and committee member Dr. Lubov Zeifman and former committee member Dr. Margaret Short for their help. I also thank Dr. Thomas Weingartner from School of Fisheries and Ocean Sciences of University of Alaska Fairbanks for letting me use the GAK1 data sets.

References

1. Acock, A. 2005. Working with missing values, Journal of Marriage and Family 67, 1012-1028.
2. Asad, M. 2012. Finding the best ARIMA model to forecast daily peak electricity demand. Proceedings of the Fifth Annual ASEARC Conference-Looking to the future-Programme and Proceedings, 2-3 February2012, University of Wollongong.
3. Baraldi, A. N., Enders, C. K. 2010. An introduction to modern missing data analyses. Journal of School Psychology 48, 5-37.

4. Box GEP, Jenkins G. Time series analysis, forecasting and control. San Francisco, CA: Holden-Day; 1970.
5. Christodoulos, C., Michalakelis, C., Varoutas, D. 2010. Forecasting with limited data: combining ARIMA and diffusion models. Technological Forecasting & Social Change 77, 558-265.
6. Kukuyeva, I. 2010. Introduction to time series in R. UCLA Department of Statistics Statistical Consulting center. http://scc.stat.ucla.edu/page_attachments/0000/0136/timeseries-10w.pdf
7. Pai, P-F., Lin, C-S. 2005. A hybrid ARIMA and support vector machines model in stock price forecasting. Omega 33, 497-505.
8. Saigal S., Mehrotra D., 2012. Performance comparison of time series data using predictive data mining techniques. Advances in Information Mining, 4 (1), 57-66.
9. Simon Stevenson, (2007) "A comparison of the forecasting ability of ARIMA models", Journal of Property Investment & Finance. 25, .223 – 240.
10. Williams, B. M., Asce, M., Hoel, L. A., Fasce, F. 2003. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. Journal of Transportation Engineering 129, 664-672.
11. Zeileis A., Frothendieck G., Ryan J A., 2014. S3 Infrastructure for regular and irregular time series (‘Z’s ordered observations). http://cran.r-project.org/web/packages/zoo/zoo.pdf

Table 1. Tukey Honest Significant Difference tests show the effect of filling methods on Mean MADs at different true AR-MA values and σ^2 =1

		methods				
Level	AR-MA	APPROX	SPLINE	LOCF	OMIT	Control
1	0.1-0.1	0.84 (b)	0.88(a)	0.83 (b)	0.82 (c)	0.81 (c)
2	0.1-0.9	1.08 (b)	1.16 (a)	1.07 (bc)	1.05 (c)	1.02 (d)
3	0.5-0.5	1.07 (b)	1.14 (a)	1.06 (b)	1.06 (b)	1.04 (b)
4	0.7-0.2	0.04 (a)	0.04 (a)	0.04 (a)	0.04 (a)	0.04 (a)

Note: the MAD values are means from 1000 replications. Different letters show the differences between the methods and the same letters indicate values that are not significant different according to Tukey Honest Significance Difference tests at α = .05.

Table 2. Tukey Honest Significant Difference tests show the effect of the proportion of missing values on Mean MADs at different true AR-MA values and σ^2 =1

		Percentage of missing values					
Level	AR-MA	5%	20%	30%	40%	50%	70%
1	0.1-0.1	0.84 (abc)	0.83 (bc)	0.85 (a)	0.82 (c)	0.85 (ab)	0.83 (bc)
2	0.1-0.9	1.11 (a)	1.04 (b)	1.11 (a)	1.04 (b)	1.11 (a)	1.04 (b)
3	0.5-0.5	0.74 (a)	0.71 (a)	0.74 (a)	0.71 (a)	0.73 (a)	0.70 (a)
4	0.7-0.2	0.04 (a)	0.04 (a)	0.04 (a)	0.04 (a)	0.04 (a)	0.04 (a)

Note: the MAD values are means from 1000 replications. Different letters show the differences between the percentages of missing values and that the same letters indicates values that are not significantly different according to Tukey Honest Significant Different tests at α = .05.

Table 3. Tukey Honest Significant Difference tests show the effect of the filling methods on the bias of estimated AR at different true AR-MA values and $\sigma^2 = 1$

		methods				
Level	AR-MA	APPROX	SPLINE	LOCF	OMIT	Control
1	0.1-0.1	0.27 (a)	0.22 (b)	0.25 (a)	-0.05 (d)	-0.021 (c)
2	0.1-0.9	0.33 (b)	0.36 (a)	0.29 (c)	0.02 (d)	-0.006 (e)
3	0.5-0.5	-0.049 (a)	-0.051 (a)	-0.08 (b)	-0.16 (c)	-0.16 (c)
4	0.7-0.2	-0.67 (a)	-0.67 (a)	-0.67 (a)	-0.67 (a)	-0.67 (a)

Note: the values are means of AR bias from 1000 replications. Different letters show the differences between the methods and that the same letters indicates values that are not significantly different according to Tukey Honest Significant Difference tests at $\alpha = .05$.

Table 4 Tukey Honest Significant Difference tests show the effect of the proportion of missing values on bias of estimated AR at different true AR-MA values and σ^2 =1

		Percentage of missing values					
Level	AR-MA	5%	20%	30%	40%	50%	70%
1	0.1-0.1	0.15 (a)	0.12 (b)	0.14 (a)	0.12 (b)	0.15 (a)	0.12 (b)
2	0.1-0.9	0.20 (a)	0.20 (a)	0.20 (a)	0.20 (a)	0.20 (a)	0.20 (a)
3	0.5-0.5	-0.10 (a)	-0.10 (a)	-0.10 (a)	-0.10 (a)	-0.10 (a)	-0.10 (a)
4	0.7-0.2	-0.67 (a)	-0.67 (a)	-0.67 (a)	-0.67 (a)	-0.67 (a)	-0.67 (a)

Note: the values are means of AR bias from 1000 replications. Different letters show the differences between the percentages and that the same letters indicate values that are not significantly different according to Tukey Honest Significant Different tests at $\alpha = .05$.

Table 5. Tukey Honest Significant Different tests show the effect of the filling methods on bias of MA at different true AR-MA values and $\sigma^2 = 1$

		methods				
Level	AR-MA	APPROX	SPLINE	LOCF	OMIT	Control
1	0.1-0.1	0.84 (b)	0.88 (a)	0.82 (bc)	0.82 (c)	0.91 (c)
2	0.1-0.9	-0.38 (d)	-0.12 (c)	-0.55 (d)	-0.06 (b)	0.008 (a)
3	0.5-0.5	-0.20 (d)	-0.024 (a)	-0.34 (e)	-0.11 (b)	-0.15 (c)
4	0.7-0.2	-0.19 (b)	-0.18 (a)	-0.19 (b)	-0.19 (b)	-0.19 (b)

Note: the values are means of MA bias from 1000 replications. Different letters show the differences between the methods and that the same letters indicate values that are not significantly different according to Tukey Honest Significant Difference tests at $\alpha = .05$.

Table 6. Tukey Honest Significant Difference tests show the effect of the proportion of missing values on estimated MA at different true AR-MA values and σ^2 =1

		Percentage of missing values					
Level	AR-MA	5%	20%	30%	40%	50%	70%
1	0.1-0.1	0.84 (abc)	0.83 (bc)	0.85 (a)	0.82 (c)	0.85 (ab)	083 (bc)
2	0.1-0.9	-0.22 (a)	-0.22 (a))	-0.22 (a))	-0.22 (a)	-0.22 (a)	-0.22 (a)
3	0.5-0.5	-0.16 (a)	-0.17 (a)	-0.16 (a)	-0.17 (a)	-0.16 (a)	-0.17 (a)
4	0.7-0.2	-0.19 (a)	-0.19 (a)	-0.19 (a)	-0.19 (a)	-0.19 (a)	-0.19 (a)

Note: the values are means of MA bias from 1000 replications. Different letters show the differences between the percentages of missing values and the same letters indicate values that are not significantly different according to Tukey Honest Significant Difference tests at $\alpha = .05$.

Table 7. MAD values from four filling methods for GAK1 time series data sets according to 3-ahead forecasting values

data sets	filling methods			
	APPROX	SPLINE	LOCF	OMIT
Temperature time series at surface	2.48	2.67	2.4	2.4
Salinity time series at surface	1.24	1.17	1.14	1.32
Temperature time series at 250 meters	0.68	0.65	0.65	0.65
Salinity time series at 250 meters	0.098	0.10	0.098	0.097

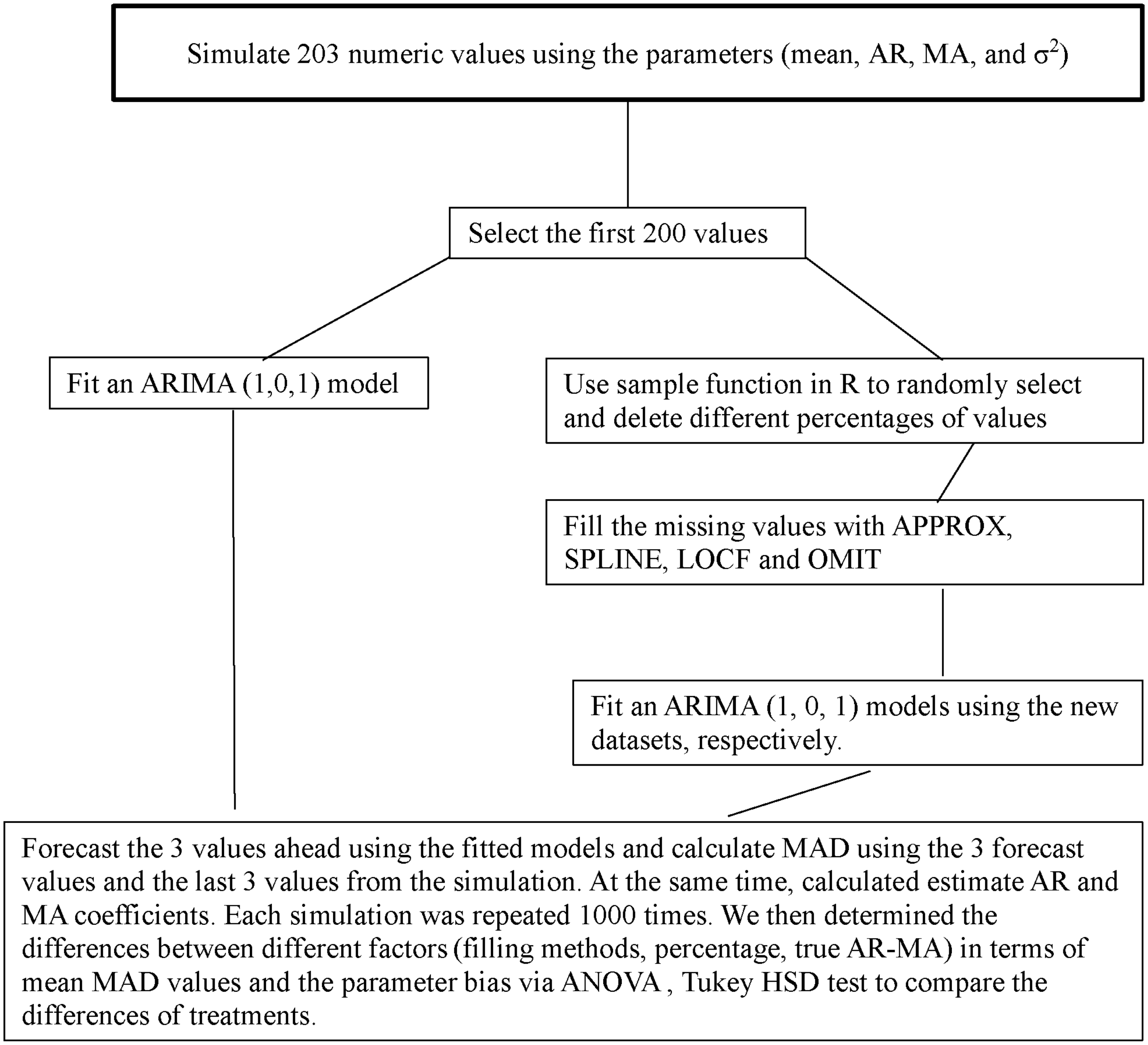
Simulate 203 numeric values using the parameters (mean, AR, MA, and σ^2)
Select the first 200 values
Fit an ARIMA (1,0,1) model
Use sample function in R to randomly select and delete different percentages of values
Fill the missing values with APPROX, SPLINE, LOCF and OMIT
Fit an ARIMA (1, 0, 1) models using the new datasets, respectively.
Forecast the 3 values ahead using the fitted models and calculate MAD using the 3 forecast values and the last 3 values from the simulation. At the same time, calculated estimate AR and MA coefficients. Each simulation was repeated 1000 times. We then determined the differences between different factors (filling methods, percentage, true AR-MA) in terms of mean MAD values and the parameter bias via ANOVA , Tukey HSD test to compare the differences of treatments.

Figure 1. The flowchart shows the procedure of simulation.

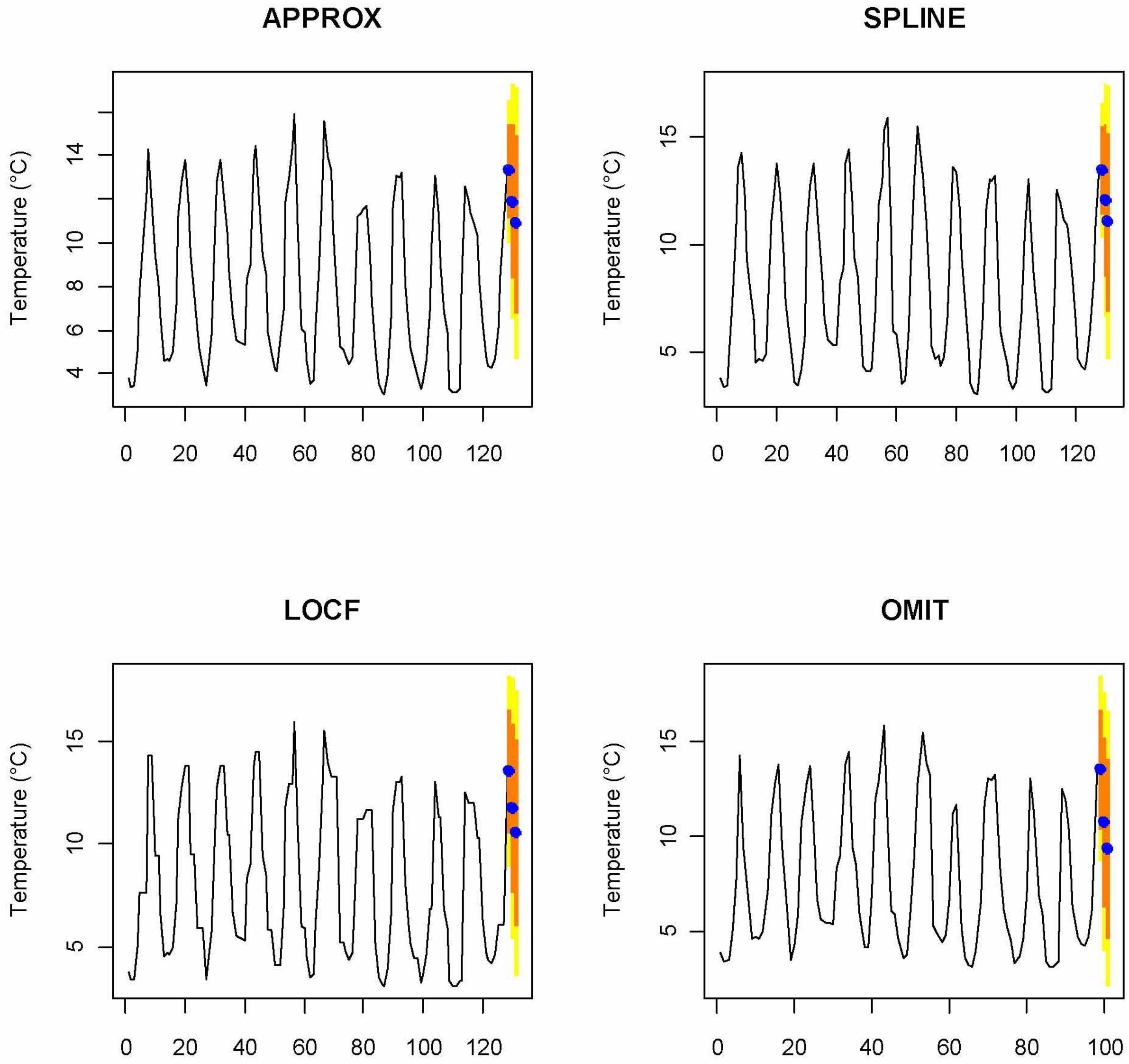

Figure 2. Temperature time series (GAK1 dataset from the surface) filled with four different methods and their 3-ahead forecasts with a 95% confident level. GAK1 time series was transformed into a regular time series by filling it with the average values for the months with multiple observations and filling with APPROX, SPLINE, LOCF and OMIT for the months with missing observations. ARIMA (1, 0, 1) was used to fit the model and to forecast.

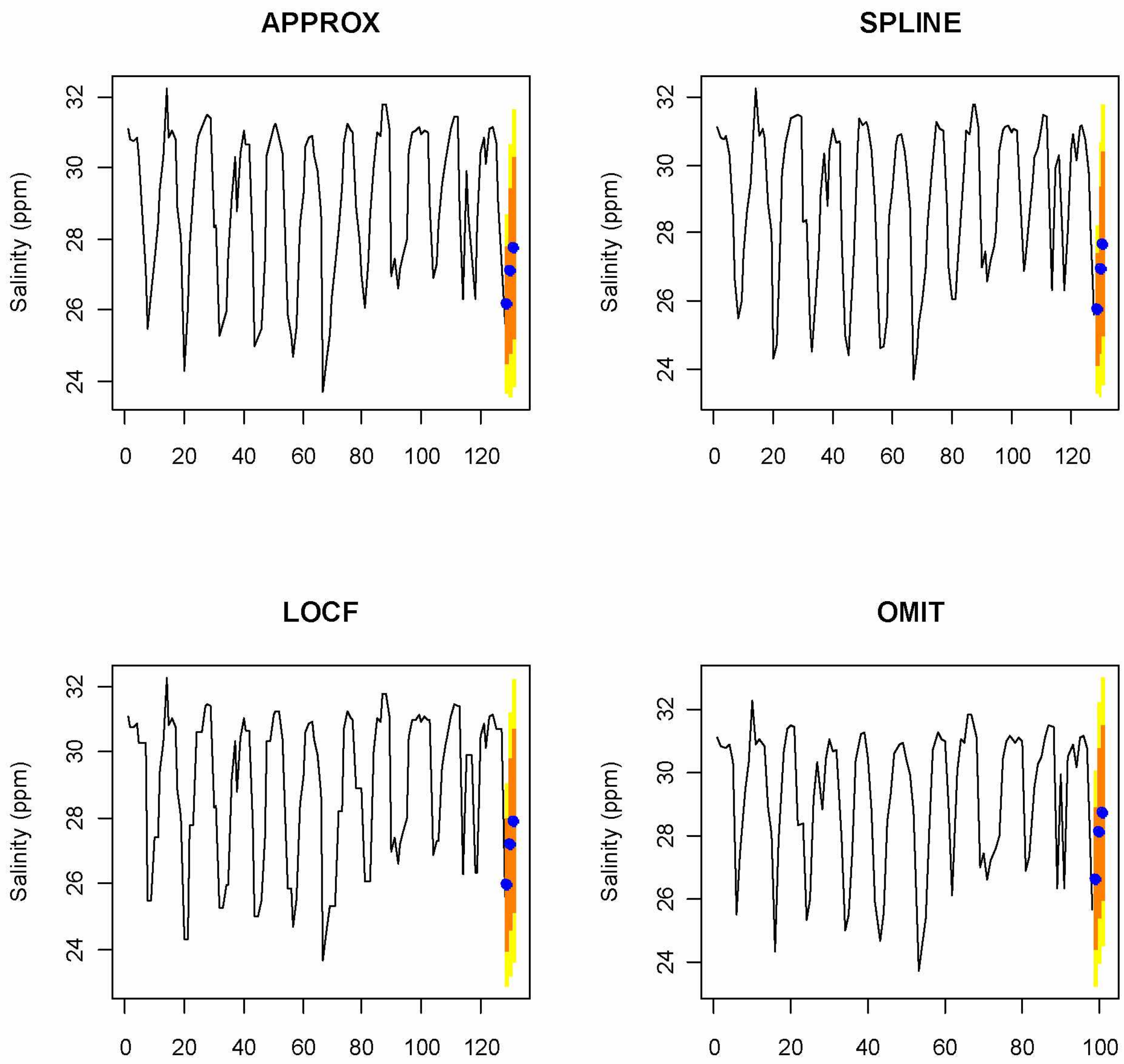

Figure 3. Salinity time series (GAK1 dataset from the surface) filled with four different methods and their 3-ahead forecasts with a 95% confident level. GAK1 time series was transformed into a regular time series by filling it with the average values for the months with multiple observations and filling with APPROX, SPLINE, LOCF and OMIT for the months with missing observations. ARIMA (1, 0, 1) was used to fit the model and to forecast.

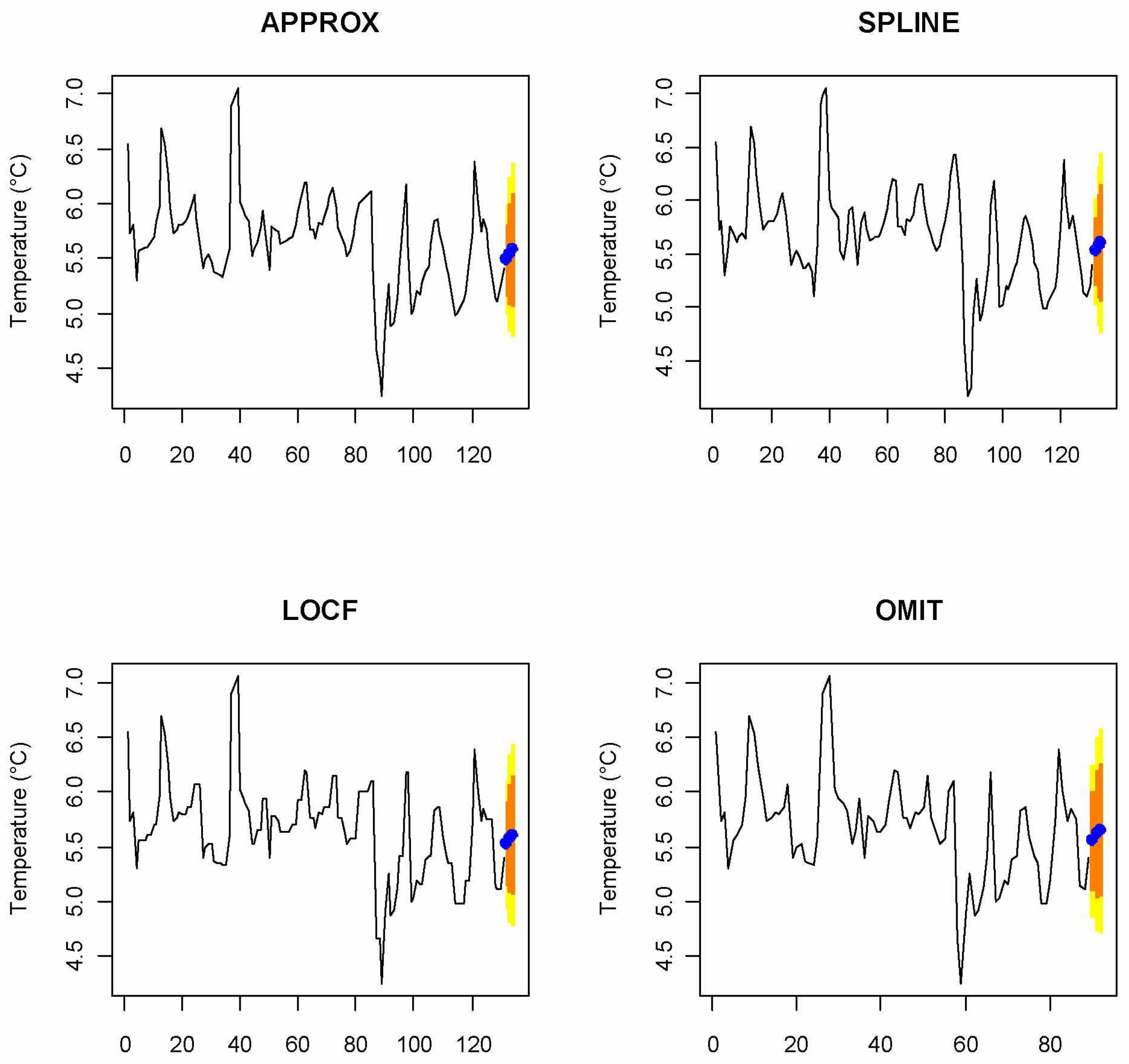

Figure 4. Temperature time series (GAK1 dataset from the depth of 250 meters) filled with four different methods and their 3-ahead forecasts with a 95% confident level. GAK1 time series was transformed into a regular time series by filling with the average values for the months with multiple observations and filling values for the months with missing observations with PPROX, SPLINE, LOCF and OMIT. ARIMA (1, 0, 1) was used to fit the model and to forecast.

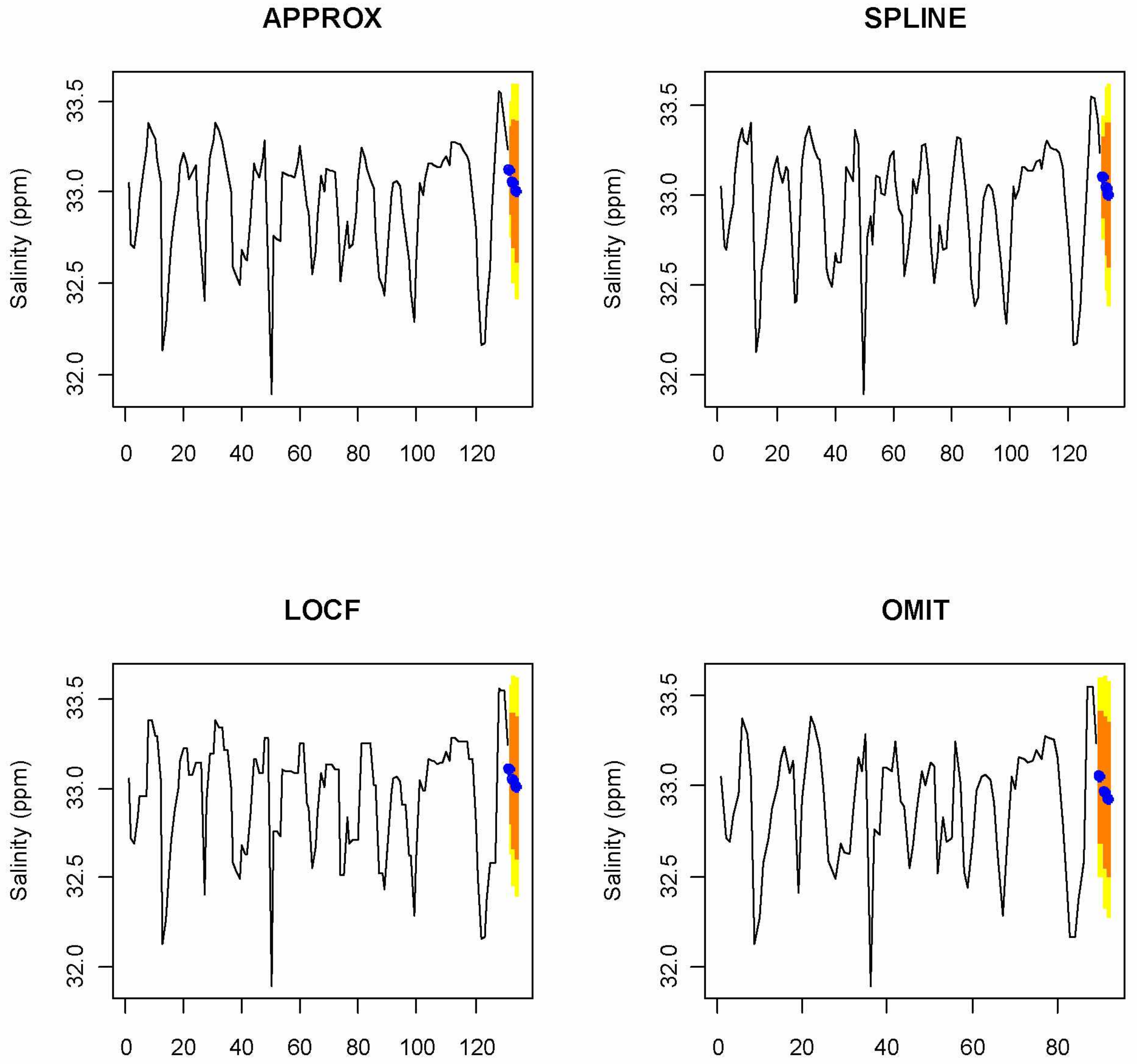

Figure 5. Salinity time series (GAK1 dataset from the depth of 250 meters) filled with four different methods and their 3-ahead forecasts with a 95% confident level. GAK1 time series was transformed into a regular time series by filling it with the average values for the months with multiple observations and filling it for the months with missing observations with APPROX, SPLINE, LOCF and OMIT. ARIMA (1, 0, 1) was used to fit the model and to forecast.