A Comparative Analysis of Seven Methods for the Estimation of Values for Observations Missing from Temperature Climate Data Series

by

Ross Andrew Benton
B.Sc., University of Victoria, 1981

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Geography

We accept this thesis as conforming to the required standard

[redacted]

Dr. S.E. Tuller, Supervisor
(Department of Geography)

[redacted]

Dr. K.O. Niemann, Department Member
(Department of Geography)

[redacted]

Dr. R.R. Davidson, Outside Member
(Department of Mathematics and Statistics)

[redacted]

Dr. S. Nadeau, External Examiner
(Department of Economics)

University of Victoria

Supervisor: Dr. Stanton E. Tuller

Abstract

The advent of low cost electronic data recorders for the collection and compilation of weather data for research and climate database purposes has brought with it the ability to collect vast amounts of relatively inexpensive data from areas where it is uneconomical to maintain manned stations. While these dataloggers are generally very reliable they will inevitably break down or will suffer from sensor failure. The result is a loss of data which must be estimated for many practical applications of the data.

There are several methods available for the estimation of missing climate data. The selection of method is dependent on the nature of the estimated variable. This thesis compares the results of the application of seven different methods of estimating missing mean daily temperature data. All methods are applied to four data sets; one synthetically generated and three from different physioclimatic regions of British Columbia. Ten, twenty, and thirty percent of the data is removed from complete data series and the

estimated data are compared to the original data.

The seven methods used in this thesis can be grouped into three general categories: methods based on the series mean, ordinary least squares regression methods, and time series methods. The means based methods are the difference and ratio methods with the the absolute temperature being used with the latter method. Ordinary least squares regression and polynomial modelling are least squares regression methods. Lagged dependent regressor, autoregressive integrated moving average (ARIMA), and modified Kalman Filter models are used as examples of time series methods. Each of the various methods has advantages and disadvantages both in terms of ease of use and restrictions for application.

Method application results were compared using several criteria. These included mean square error for least squares and time series methods and the Akaike's Information Criteria (AIC) for time series models. Final comparison for all methods was based on a comparison of estimated and actual data for all series. The difference between these two values was calculated for each estimated point and data were grouped into range classes. The distribution of estimated values

within each class was used to compare the methods for each of the data sets.

The results of the application of these seven methods to the different data sets showed no definitive 'best' method while a few methods proved to be significantly better than others. The commonly used difference method provides ease of use and average estimation results relative to the other methods. The methods based on ordinary least squares regression provide moderate to good estimation results with relative ease of use but the data violate presumptions upon which the methods are based. The three time series methods provide mixed results. The ARIMA and Kalman Filter method provide generally poor results and require sophisticated computer software to apply. The functional response of the ARIMA method is dependent on the experience of the modeller and thus further experience with this method may provide better results. The lagged dependent regressor method, on the other hand, provides good estimates with ease of use comparable to the ordinary least squares methods. The lagged dependent regressor method also accounts for the spatially and temporally correlated nature of the temperature data which the ordinary least squares

methods do not.

Examiners:

[redacted]

Dr. S.E. Tuller, Supervisor

(Department of Geography)

[redacted]

Dr. K.O. Niemann, Departmental Member

(Department of Geography)

[redacted]

Dr. R.R. Davidson, Outside Member

(Department of Mathematics and Statistics)

[redacted]

Dr. S. Nadeau, External Examiner

(Department of Economics)

Table of Contents

List of Tables

List of Figures

Availability of Data

All the data sets used for analysis in this thesis are available from the author through the following address.

Ross Benton
Pacific Forestry Centre
Forestry Canada
506 West Burnside Road
Victoria, BC
Canada
V8Z 1M5

Acknowledgements

I wish to thank Dr. S.E. Tuller, Dr. R.R. Davidson, and Dr. K.O. Niemann of the University of Victoria for the time and effort spent in the evaluation and comments required in the development of this thesis. I also wish to thank Dr. R.H. Silversides of the Pacific Forestry Centre, Forestry Canada, for the encouragement, information, and advice he so readily provided.

Dedication

This thesis is dedicated to my family and to the memory of my parents, all of whom played a key role in its completion.

I. Proposition of Problem and Literature Review

Missing observations from a climate data series pose many problems for researchers wishing to use the data. This problem is particularly acute when the series of missing observations occurs during a critical period in experimentation. Periods of missing observations may be single points or include extensive data sets. Large gaps in the data force the researcher to rely on the ability of some form of data interpolation method to project close approximations of the actual environmental variables for a particular time period.

Several mathematical and statistical methods are available for estimating variables to fill in missing observations. The methods which are available range in complexity of application from very simplistic models, requiring little understanding, to extremely complex ones, requiring extensive knowledge of statistical methods and modeling procedures. Of the methods currently in widespread application, most are designed for applications where only a single or a few observations are missing from a particular series. They are not intended for projecting meaningful and realistic values into large gaps in a data series.

A useful missing data completion method should have several attributes. It should be easy to apply and not

have a long list of constraints which greatly limit its applications, be able to project reliable and plausible values into the missing observations, take advantage of data both prior to and following the missing data period(s) which may provide additional information, and be available to most users without a requirement for sophisticated hardware and software computing capabilities.

Review of current literature

A review of literature from the fields of statistics, meteorology, climatology, and economics provides much of the current information on theoretical and applied methods used for completion of missing data in time and/or space related serial data analysis. Many of the methods in practical use for meteorological and climatological application tend to be either simplistic methods which can be readily applied or ones which are extremely complex.

Methods such as the ratio or difference methods (Atmospheric Environment Service, 1982a, 1982b), regression analysis (Findlay, 1985; Kemp et al., 1983; Murphy and Katz, 1985; Flocas et al., 1983; Tabony, 1983), or polynomial curve fitting (Reuter, 1980), provide ease of application, require relatively few data points, and have limited computational requirements.

Others, such as autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), lagged dependent regression, and other time domain forecast methods (Steinberg, 1984) require complex computational tools and a great deal of data (Findlay, 1985).

The two most common methods, used by the Atmospheric Environment Service (AES) of Environment Canada, are the ratio and difference methods (Atmospheric Environment Service, 1982a, 1982b). These two methods are used for filling in missing data points for the calculation of statistics such as thirty year normals and, as such, do not require a high degree of prediction accuracy or resolution. Both methods rely on the availability of a standard station series to be used in conjunction with the incomplete series. A standard station is defined as one with sufficient length (28 years or more) and with climatic and physiographic characteristics similar to the station with the incomplete series (Atmospheric Environment Service, 1982a). Choice of the standard station is made on the basis of the best correlation using Pearson's correlation coefficient (r). The ratio (or quotient) method is calculated using the ratio of the station means from the incomplete and standard station series which is then multiplied onto the standard station value for the particular time period in question. For this

method to be applied to climatological variables that can be either positive or negative, such as temperature, the variable would have to be transformed in some manner since this method produces a model which is assumed to have a zero intercept. In the case of temperature, the data could be transformed into units of °K from units of °C or °F. The ratio method is commonly used for environmental variables such as precipitation or degree days.

The difference method is similar in configuration to the ratio method except that the difference of the two series means is used as an adjustment factor rather than a ratio. One of the advantages of the difference method is that it does not require a transformation of climatological variables which can have either positive or negative values. This property results in this method being used for the estimation of missing temperature values.

Both the ratio and the difference methods are used for completion of data series that contain few missing data points, that cover a long period of time, and rely on a relatively consistent relationship between the two stations. They are not specifically designed to provide the best possible estimates of a particular environmental variable but are used to provide an average value based on long term results. These methods

provide a mechanism by which a shorter term mean value for a particular date or period can be adjusted to emulate a longer term mean.

There are other methods which provide more reliable and realistic values for estimating missing data points. They may incorporate information from various sources such as simultaneous data from several standard stations, and/or past or future values from the same station. This added information conveys more about the nature of conditions at the station during the missing data period than may be obtained from the application of the difference or ratio methods in their basic form.

Simple and multiple regression methods offer a variety of possibilities for application. It is possible to develop a model which incorporates several pieces of relative information which would aid in the accurate estimation of a missing data point using multiple regression. For example, one may wish to incorporate the data from several surrounding stations in order to estimate the value of a missing data point at a particular station. Regression analysis allows for the easy use of transformations of particular variables but does not necessarily require it. This method also allows for a variety of different variables to be incorporated into the model.

Polynomial curve fitting techniques also provide a

variety of potential models including the use of simple transformations of the short series variable so that no other station data are required but also allows the use of second station data for prediction purposes. The order of the polynomial can be determined by several methods, many of which are available in common statistical software packages.

Autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) models are time domain models which incorporate the autoregressive nature of the station variable through time and moving averages. An ARMA or ARIMA model can incorporate model parameters for seasonality as well, if necessary. The seasonality parameters adjust the model for longer term cycles which may otherwise be unaccounted for. These models vary widely in composition and application and their components are often determined by the nature of the variable under investigation.

Thiébaux and Pedder (1987) consider several potential methods for estimating a missing observation (data point) in a spatial or temporal context and suggest smoothing (averaging) or polynomial projection for single data points. They also note that estimation of a series of missing data points requires a more complex approach although they provide no specific

solution. Guiot (1985) offers an extrapolation method for missing data using spectral canonical regression, a technique which combines spectral and regression analysis. While Guiot's technique combines the simplicity of regression with the efficiency of spectral analysis, the method relies on digital filtering and other smoothing techniques to accommodate spectral (frequency domain) analysis and does not take advantage of time domain analysis.

In the field of economics, missing data series are dealt with by filling missing data points with series averages, zero values (Chaghaghi, 1985), or dummy values (Chow and Lin, 1976) which have little or no effect on the variables under study. Least squares estimators (Doran, 1974) are another commonly used method in this field. These methods are also used in other fields and are found in various statistical applications (Dunsmuir, 1981; Burkhardt and Hense, 1985; Barham and Dunstan, 1982; Jones, 1980; Parzen, 1970b; Neave, 1970).

While methods such as averaging, zero value, or dummy value insertion may be effective in dealing with economic and other variables which have non-negative values, it presents problems in dealing with variables which may have values which can readily swing positive or negative. Other problems arise for series where the mean value may not be appropriate as a substitute value

such as those where seasonal or short term effects change the nature of the mean. An example of this problem is when one considers temperature data where the mean value for any given period will change as time progresses through the seasons or through shorter climatic events such as the movements of fronts through an area. The determining factor is the time scale under consideration. While the completion of a data series for economic purposes is often informative, the primary goal of serial data analysis in this field is for future forecasting purposes. Long data series are not always required or available and the data collected often utilize long periods such as weeks, months, quarters, or years. The use of standard economic techniques therefore does not incorporate methodology for dealing with long series of missing data covering relatively short periods of time, such as hourly data, because short term influences in economic variables are rarely of great significance and the data are not collected.

The use of related series for interpolation of data is common practice in economics (Granger and Newbold, 1977; Makridakis and Wheelwright, 1978). This method is also used for interpolation of missing data in a series (Chow and Lin, 1976; Friedman, 1962; Nerlove et al., 1979). Using related serial data sets for model development provides the opportunity to apply knowledge

of how the series relate to each other. In the event that one of the series is incomplete over a particular time period, it may be possible to estimate the missing data portions using the response of the longer series in the same time period.

Methods where related series are used for analysis have been used with mixed success in climatological applications relating sun spot and other solar activity to atmospheric temperature (Zerefos, 1983; Fleer, 1982; Schönwiese, 1978). Perceived failures in these climatic applications may be partially due to a lack of sufficient reliable data. Schönwiese (1978) also points out that extremely long data series are required in order to perform an effective analysis. Van Loon and Labitzke (1988) were finally able to achieve a high degree of association between solar activity and the earth's temperature by incorporating a third variable relating to global wind patterns. The incorporation of two serially dependent data series to predict a third greatly reduced the need for extremely long series for predictive purposes.

Time series analysis has been used in meteorology for several years for forecasting purposes (Murphy and Katz, 1985; Essenwanger, 1976) but has been avoided by such agencies as the Atmospheric Environment Service because of its relative complexity of application and

relative lack of success (Baird and Associates et al., 1986; Steinberg, 1984). Specific types of time series models such as lagged dependent regressor models (Fuller, 1987, 1976; Young, 1984; Harvey, 1981) provide a relatively simple method of estimating missing values in an incomplete series while maintaining the nature of the series in terms of variance and error structure. Specific applications of simple lagged dependent regressor models provide for time space projection using data series from a nearby climate station and past (or future) responses of environmental conditions at the location where data is missing. Users who wish to apply time series techniques to climate data series have three options available to them. These are: forward projection based on observations collected prior to the missing data, backward projections based on observations collected following the missing data, and mixed or recursive methods which incorporate both forward and backward projection. Methods developed for the analysis of missing data and discrete data in serial data sets are often adaptations of forecast methods (Jones, 1981; Seaman and Hutchinson, 1985) or products of models produced on a subset of the data series (Barham and Dunstan, 1982). Jones (in Murphy and Katz, 1985) also suggests a method for estimation of a missing data point based on the Kalman state space model (or Kalman

Filter). Work done by Brubacher and Wilson (1976) on comparison of actual and projected demands for electricity over holiday periods uses time series interpolation techniques. This latter work involves the projection of a single variable over a period of more than a few missing observations and a shorter sampling interval (i.e. hourly observations) is used while in most of the other examples only single data points are being projected and sampling intervals are averaged or totalled into monthly, quarterly, or annual observations. In many cases only a single data series is being used and comparative data series are not available for performing related series methods.

To date, few authors have performed comparative analyses of the various methods for the estimation of missing observations in climate data series. Those, such as Tabony (1983, 1986) and Guiot (1985) who have done some comparison have included very few methods. With the advances in computer hardware and software, there are several new methods available, which may be used with relative ease. A comparison of the current methods available is required to demonstrate the response of the variety of methods available.

The purpose of this study is to evaluate seven methods for the estimation of missing data in temperature series. It will compare the methods

currently in use for this application as well as others which potentially lend themselves to this use. Most of the methods applied are available in many commonly used computer software packages or can be easily adapted for current software.

Temperature is an important environmental variable for many biological studies. It is often a key factor in the growth, development, and survival of living organisms and the ecosystems in which they live. The loss of key periods of temperature information in a given study can potentially result in the failure of the experiment or incorrect interpretation of the collected information. Since atmospheric temperature data can have either positive or negative values and is both temporally and spatially correlated, the methods or the data are modified accordingly for application.

II. Interpolation Methods General Theory

Three general classes of data interpolation methods are presented here for comparison. These are methods based on the series mean values, spatial correlation, and spatial and serial correlation. Different methods within each of these general classifications are also presented for comparative purposes.

The ratio and difference methods are both based on the series mean values. These are the methods currently used by the Atmospheric Environment Service for adjusting climatic normals. Temperature series are used for this study and so it is necessary to convert the series to the Kelvin scale in order to apply the ratio method for predictive purposes.

The remainder of the methods to be compared are based on least squares regression techniques. Two of the methods, ordinary least squares regression and polynomial curve fitting, rely on spatial correlation and the basic premises and constraints of standard regression techniques. The remaining three methods; autoregressive integrated moving averages (ARIMA), lagged dependent regressor (after Fuller, 1976; SAS Institute Inc., 1984), and modified Kalman filtering (after Jones, 1985), are based on standard regression techniques with modifications to allow for the serial

correlation in both observations and errors.

The methods based on serial correlation and regression techniques do not meet many of the assumptions which apply to standard regression methods. The standard assumptions of independent observations and independent error terms do not hold in serially correlated data such as is found in climate data sets. It is necessary to adjust the estimation techniques to account for these differences.

The general theory presented here is a simple summary of that which may be found in general texts and, as such, is limited to the essentials of parameter and variance estimation. Each method is presented with the individual adjustments and criteria necessary for temperature estimation capabilities. General regression theory has been summarized for those methods which use it with unique characteristics of the least squares regression based methods being dealt with on an individual basis.

Ratio Method

The ratio method has been used for many applications because of its relative ease of use. Adjustments to data sets can be readily calculated by hand if necessary using this method. This method has been used for many years for these reasons. The

Atmospheric Environment Service currently uses this method for the estimation of precipitation normals for stations which have insufficient series length to calculate a true thirty year normal. Periods of concurrent data are used for the calculation of the ratio (r).

Standard Formulae:

Estimator of a population mean μ_y;

$$\hat{\mu}_y = r\mu_x$$

where $r = \dfrac{\Sigma y_i/n}{\Sigma x_i/n}$ is the ratio,

y is the dependent variable,

x is the independent variable,

n is the number of observations, and

μ_x is the population mean of the independent variable.

Estimated variance of $(\hat{\mu}_y)$;

$$V(\hat{\mu}_y) = \frac{(N - n)}{(nN)} \cdot \frac{(\Sigma(y_i - rx_i)^2)}{(n - 1)}$$

where r is the ratio,

y is the dependent variable,

x is the independent variable,

N is the population size, and

n is the sample size.

Modification for Missing Data Interpolation

Estimation of the series ratio for y:

$$\hat{y}_t = r * x_t$$

where $\hat{y}_t$ is the estimated value of the dependent variable at time t,

x_t is the value of the independent variable at time t, and

$$r = \frac{\Sigma y_t / n}{\Sigma x_t / n} \quad .$$

Constraints

The ratio method has two basic constraints on the data. These are based on the relationship between two variables. The implied constraints are that the relationship between the dependent variable (y) and the independent variable (x) is linear through the origin and that the variance of the dependent variable is proportional to that of the independent variable.

In this particular application the variables are the temperature responses from the research station and the AES climate stations. The constraint of linearity may be maintained but that of being linear through the origin is violated as temperatures can frequently be below zero in either the Celsius or Fahrenheit scales. The data must be transformed into another scale, such as the Kelvin scale, in order to apply this method.

If such a transformation were to be applied, the resulting ratio is applicable to only the transformed data and thus the data would have to be transformed before any estimation could be attempted. The second constraint, that of variance of the dependent variable being proportional to that of the independent variable, can be readily determined by plotting one against the

other. A linear or fan shaped distribution would indicate this constraint is being maintained.

Difference Method

The difference method is the standard method used by the AES for its calculation of normal temperature values where data is missing from one of its stations or one of its stations has a record of insufficient length for the calculation of a temperature normal. The particular method applied by the AES uses the mean temperatures for each of the dependent and independent stations series for periods of concurrent data and the temperature value for the independent station for the same time as the dependent station missing value is to be predicted.

Standard Formulae:

Difference estimator of a population mean μ_y;

$$\hat{\mu}_y = \bar{y} + (\mu_x - \bar{x}) = \bar{d} + \mu_x$$

where $\bar{y}$ is the sample mean for the dependent variable,

μ_x is the population mean of the independent variable,

$\bar{x}$ is the sample mean for the independent variable, and

$\bar{d}$ is the difference between the sample means for the dependent and independent variables $(\bar{y} - \bar{x})$.

Estimated variance of $\hat{\mu}_y$;

$$V(\hat{\mu}_y) = \frac{(N - n)}{nN} \cdot \frac{\Sigma(d_i - \bar{d})^2}{n - 1}$$

where d_i is the difference between the dependent and independent variables for any one observation $(y_i - x_i)$,

$\bar{d}$ is the difference between the sample means for the dependent and independent variables,

N is the population size for the dependent variable, and

n is the sample size.

Modification for Application to Missing Data Interpolation

Difference estimator of a series y;

$$\hat{y}_t = \bar{d} + x_t$$

where $\hat{y}_t$ is the predicted value of the dependent variable at time t,

x_t is the value for the independent variable at time t.

Constraints

The difference method has two basic constraints which limit its application. The first of these constraints is that the data, when the dependent variable is plotted against the independent variable, should lie on a line that has slope close to one. The second constraint is the variance of the difference between the dependent and the independent values should be minimal.

The ratio and difference methods offer simplicity of use and ease of calculation as their greatest assets. These assets outweigh the relative inaccuracy of predictive capability for estimates of long term values such as normal values when computing capabilities are

limited. The assumption of stationary mean values upon which the use of these two methods are based may tend to hold for long term applications but will not necessarily hold for short term situations where seasonal effects are present.

General Least Squares Regression Methods

While the difference and ratio methods are based on using the mean value of the data for model estimation, most of the other methods presented for comparison are based on the least squares method for model parameter estimation. An alternate method for estimating the parameters in these models is to use maximum likelihood estimation but it was found that the practical difference in estimated parameters using each of these methods was not significant. Only least squares is discussed here since many commonly available computer statistics software packages only offer this computational method.

The parameter and variance estimation methods for the ordinary least squares, the polynomial curve fit, and lagged dependent regressor methods have a common base in least squares regression. As such, the generalized formulation is presented here and specific differences are presented in the discussion of the individual methods.

Standard Formulae:

For the general model:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_i X_i + \epsilon$$

where Y is the dependent variable,

X_i is the i^{th} independent variable,

β_0 is the constant term,

β_i is the i^{th} regression coefficient, and

ϵ is the error term.

Model Parameter Estimation:

$$b = (X'X)^{-1}X'y$$

where y is the vector of values for the dependent variable,

X is the n x k matrix of values for the k independent variables,

$(X'X)$ is the matrix of sum of squares and crossproducts for the independent variables,

$X'y$ is the vector of sum of

crossproducts of the dependent variable with the independent variables, and

b — is the vector of estimated parameter coefficients.

Estimate of Error Variance:

$$s^2 = \frac{(y-Xb)'(y-Xb)}{n - k - 1}$$

where n — is the number of observations, and

k — is the number of independent variables.

Constraints

Least squares regression methods attempt to minimize the mean square error of a data series and much of the performance of an estimated model is dependent on the distribution of its residuals. The residuals are expected to have uniform variance, have a zero mean, and be uncorrelated.

It is not always possible to achieve these constraints in practical application. This is particularly true in the case of climate data sets as the data is correlated both through time and space. It

is thus necessary to account for this known limitation in the application of standard regression methods.

Ordinary Least Squares Regression

Ordinary least squares regression has been used for many years as a method for climate data interpolation. This method allows for the ready incorporation of additional variables which will aid in the estimation of missing data points. Pertinent data from other nearby stations such as coincidental observations of the same or other variables can be introduced into a prospective model.

While regression models are capable of being composed of many variables, one of the optimum goals for the application of this method is to achieve the best estimate of the missing data with the minimum number of estimated parameters (parsimony). Achieving a good balance between the number of estimated parameters and model performance significantly reduces the amount of extra data required for estimation as well as computational requirements. As a consequence it is possible to use one of the many inexpensive hand held calculators which are capable of developing simple linear regression models for predictive purposes.

The example model illustrates an ordinary least squares model which incorporates two independent

factors, X_1 and X_2, used to predict variable Y. While not implicitly stated, it is assumed that all three factors are measured simultaneously. For practical purposes we can assume that X_1 and X_2 represent simultaneous data series from two reference stations while Y is a station with missing data points from its series. Optimally, the series from X_1 and X_2 should be complete over the series being covered by Y.

Example Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

where

Y	is the dependent variable,
β_0	is the constant term,
β_1, β_2	are the regression coefficients,
X_1, X_2	are the independent variables, and
ϵ	is the error term associated with the model.

The estimated model equation is therefore:

$$y = b_0 + b_1 x_1 + b_2 x_2$$

Applications

Since ordinary least squares regression allows for the inclusion of more than one related variable, it is possible to greatly improve the predictive ability of a single predictor model. This is particularly useful for climatic data estimation as related variables can be included into the model.

For example, if temperature from a particular station is to be predicted at a given location and one has temperatures from other stations nearby then it is possible to include the data from more than one other station in a network of related stations to fill a gap. In this way more complex spatial relationships could be incorporated into a model.

Another application could include other related data from the station in which the incomplete series was generated. This situation can often arise as a result of a sensor failure. In this case a temperature data series might be reconstructed from solar radiation, precipitation, and wind data, as well as temperature data from another nearby station.

Constraints

The primary constraints for the application of ordinary least squares regression techniques to climate data interpolation involve the nature of the data series

themselves. Two of the standard assumptions upon which least squares and normal theory are based are violated. The assumption that observations are independent does not hold since they are serially correlated. The other violation of concern is that the error term is not necessarily random due to nonrandom oscillations in the data over time.

Polynomial Curve Fitting

Polynomial modeling using power functions of a single regressor variable and least squares regression is a method that has been used with varying levels of success. In standard application, a single variable is selected and then powers of that variable are used to generate a polynomial model of a specified order.

General Model:

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \ldots + \alpha_k X^k + \epsilon$$

where Y is the dependent variable,

$\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_k$ are coefficients of the powers of X,

X is the independent variable, and

ϵ is the error term.

Constraints

Polynomial models allow for improving the correlation between two variables by means of a simple power transformation of the regressor variable. This can be greatly beneficial in some applications but it has been shown to be of limited use in practical applications on some climatic variables such as temperature.

The selection of order for the polynomial model will determine its practicality of application, particularly over long time periods. Application of this method to interpolation of a temperature series has been shown to produce unrealistic values when extrapolated beyond the reasonable limits of the model (Thiébaux and Pedder, 1987).

Time Series Methods

There are several classes of time series models. This thesis deals with three classes of the many potentially available for applications such as forecasting and backcasting. The three methods dealt with here are the autoregressive integrated moving average (ARIMA) and its relatives, a form of lagged dependent regression, and the modified Kalman Filter.

Autoregressive Moving Average (ARMA) Models

Autoregressive moving average (ARMA) models consist of two primary components which are, in themselves, individual classes of time series models. The two components of this class are the autoregressive and moving average models which are united to combine the advantages of the individual model types, to reduce the shortcomings of each type, and to achieve a more parsimonious model.

ARMA models can be composed of autoregressive (AR) or moving average (MA) parameters or a combination of the two. They can also include other parameters such as seasonal and other trend components. The composition of the ARMA model depends on the nature of the variable being modeled and will determine the types of model parameters required to adequately model its response over time.

Autoregressive Integrated Moving Average (ARIMA) Models

The autoregressive integrated moving average (ARIMA) model is a further development of the ARMA class of models. The three main elements of the ARIMA model are the autoregressive, the moving average, and the differencing components. Each of these individual parts will affect the way in which the model will predict results. Selection of the component parts to be

included or excluded from a given model should be based on the climatological variable which is to be predicted. For example, the inclusion of a high order moving-average component in a temperature series can introduce false harmonics in a series and produce unrealistic values.

It has also been suggested that differencing of a long term climate data series is unnecessary as these series naturally tend toward a long term mean (normal) value (Jones, 1985). Single order differencing must be done for short time period climate data sets in order to achieve stationarity. This is necessary to remove seasonal effects inherent in climate data.

The example model that follows is an ARMA(1,1) model which has 1 autoregressive and 1 moving average factor.

Example Model:

$$Y_t = \alpha_1 Y_{t-1} + \beta_1 \epsilon_{t-1} + \epsilon_t$$

where Y_t is the variable to be estimated at time interval t,

α_1 is the autoregressive coefficient,

Y_{t-1} is the value of the variable at time interval t-1,

β_1 is the moving average coefficient

ϵ_{t-1} is the error term at time t-1, and

ϵ_t is the error term at time t.

Applications

Autoregressive moving average models have been used in economics and engineering applications for several years. This procedure is used in economics for forecasting economic indicators and market trends while for engineering applications it has been applied in systems measurement and control. Since climate is a long term function of the atmospheric system, it follows that this type of model could be well adapted to both long and short term interpolative models.

Due to the recursive nature of model development, the use of computer software is a must for practical application of this procedure. Several packages are available. Most packages are suitable although care must be taken in selection as the algorithms used to optimize in some packages can corrupt the nature of some environmental variables. For example; a common algorithm practice is to pad series data with zero values in order to perform a fast fourier transform. This practice will alter the nature of a temperature

series by forcing its mean value closer to zero (Chatfield and Pepper, 1971).

Estimation and Variance Formulae

Parameter and variance estimation for ARIMA models is a complicated procedure and is well covered by others (Box and Jenkins, 1976; Jones, 1985). Since most available computer software offers this procedure using least squares methods, standard regression hypothesis testing techniques are applicable.

Constraints

The greatest limitations to the use of this modeling method are computing costs and the required learning curve. The sophisticated computer software currently available to perform ARIMA modeling tends to be expensive and requires considerable computer resources and training in order to develop efficient ARIMA models.

The extreme flexibility of this model poses two potential difficulties. Firstly, one must have a knowledge of the effect of each of the three main components (i.e. AR, MA, and I) on the environmental variable series being estimated. Secondly, one must determine what order the model must take to adequately emulate the real series. Efficient application of this

modeling technique requires an adequate knowledge of the series to which it is being applied as well as methods of determining model order.

There are several problems which can occur as a result of selecting the wrong order for any one of the three components of this type of model. The effect of overdifferencing the data, that is to difference the data more than necessary to achieve stationarity, can introduce false harmonics into the series (Gottman, 1981). For most short term climate applications, the level of differencing should be set to one. The effect of too high an order for the moving average component is to introduce false harmonics into a system. This will tend to either dampen or increase natural trend patterns. The selection of order for the autoregressive component model will often be evident by natural patterns (or cycles) in the data. If, for example, a three day pattern occurs then an order three autoregressive term should be used.

Lagged Dependent Regressor Models

There are several types of lagged dependent regression models. Only a simple form of the model is evaluated in this study. This is because the simple method has wider application and can be readily applied without an extensive knowledge of time series modeling

or the need for sophisticated computer software. Therefore, the model presented here represents only one of many possible forms of this model.

While the simple model presented here can be applied using a combination of common statistical software, the use of time series software will produce a superior model. This is due to the fact that time series software will incorporate the serial autocorrelation in the error term while standard statistical software assumes an independently distributed error.

A straightforward generalization of the autoregressive model is one containing an ordinary independent variable as well as a lagged dependent variable.

Basic Model:

$$Y_t = \alpha_0 + \alpha_1 X_t + \beta Y_{t-1} + \epsilon_t$$

where Y_t is the value for Y at time t,

α_0 is a constant term,

α_1, β are model coefficients,

Y_{t-1} is the value of Y at time t-1,

X_t is the value for X at time t, and

ϵ_t is the error term.

In this case, the series for station Y is incomplete and information contained in the series for station Y and the complete series for station X is incorporated to estimate the missing values for series Y. The selection of station X is based on the best correlated fit between series Y and corresponding series for other stations. This model could be readily expanded to incorporate other lags at station Y (Y_{t-2}, Y_{t-3},...) and other series (X_1, X_2, etc.) to further improve the model response.

Practical application of this method is a three step process (after Fuller, 1976). The first step is to perform an instrumental regression using ordinary least squares regression. This step results in the following preliminary model:

$$y_t = b_0 + b_1 x_t + b_2 x_{t-1}$$

where y_t is the predicted value of y at time t,

b_0, b_1, b_2 are estimated model parameters, and

x_t, x_{t-1} are the values of x at times t and t-1.

The second step is to fit an ordinary least squares model using the original values for y_t, the first lag of fitted series from the instrumental regression, and the original x series. This step provides initial estimates of the model parameters.

The final step involves fitting a first order autoregressive process to the estimated residuals from the second step and then using generalized least squares in order to improve the initial estimates of the model parameters. Unlike other time series methods the lagged dependent regressor method does not require the data to be stationary. The use of instrumental regression and the inclusion of a constant term in the final model alleviate the need to assume stationarity.

A simple comparison was performed using the method used throughout the remainder of this thesis and one where the data were made stationary, by first order differencing, and developing models with no constant terms in them. It was found that the difference in predictive performance between the use of the two different methods for generating the model is marginal in practical application and thus instrumental regression technique with no adjustment was used.

Applications

To date most applications for lagged dependent regressor models have been in the fields of economics, physics, and engineering. Since this class of time series model is multivariate it allows for the introduction of spatial factors as well as the temporal factors of a strictly temporal model such as the autoregressive moving average models.

Constraints

This form of model attempts to resolve the observation and error serial correlation problems encountered with ordinary least squares regression. This goal is achieved when used with a time series software package. When applied using standard regression software, this method is still very useful as it includes prior information about the short series as well as incorporating information from a related series.

Modified Kalman Filter

This method for data completion is an adaptation of the Kalman filter process as suggested by Jones (in Murphy and Katz, 1985). Kalman filtering is a recursive method for the development of real time estimation and prediction models used in a statespace framework. This filtering method not only takes into consideration a

predicted future value but also considers future variance in the system.

Application of the standard Kalman filter is a seven step process as described by Jones. He states that, in order to apply this to serial missing data points, only the first two steps of the standard filter need be applied. These two steps involve, first, doing a one step forecast using an ARMA(1,1) model and then calculating its variance. This procedure is repeated for following missing data points with consideration of the variance structure used to maintain the integrity of the data series.

Formulae:

(1) One-step forecast using an ARMA(1,1)

$$Y_{(t+1|t)} = \alpha Y_{(t|t)} + \epsilon_t$$

where $Y_{(t+1|t)}$ is the missing value to be estimated given the previous series values,

α is the regression coefficient, and

$Y_{(t|t)}$ is the value immediately preceding the missing value in the series.

(2) Estimate the variance of the predicted value

$$P_{(t+1|t)} = \alpha^2 P_{(t|t)} + \sigma^2$$

where: $P_{(t+1|t)}$ is the estimated variance at time t+1 given the variance up to and including time t,

$P_{(t|t)}$ is the variance at time t and,

σ^2 is the overall series variance.

Constraints

The modified Kalman Filter is applied by first making the one-step prediction and then calculating its variance. It has several advantages which make it useful for missing data applications. Firstly, it incorporates prior information about the nature of the series by using a first order autoregressive moving average model. It also incorporates the variance of prediction for further predictive purposes. A further advantage of this method is that it can be further modified so that it may either be applied in a forward movement through time (future prediction) or backward (past prediction) so that, in the event of a long series of missing data, the method can be applied starting from both 'ends' of the missing data period to take advantage of the preceding and following data.

The greatest limiting factors to application of this method is the probability calculations required in each of the iterative steps. While computers greatly ease this problem, it does not lend itself well to simple application. This poses several potential problems. Much of the predictive capability is based on a good estimate of the series variance and a series that is adequately represented by an ARMA(1,1) model. These restrictions may hold true for longer time periods but the assumption of the ARMA model parameters may not hold for shorter periods. The ARMA(1,1) model would have to include a differencing parameter of at least one, thus becoming an ARIMA(1,1,1) model, in order to apply this method to periods of a year or less in order to accommodate seasonal variation. The availability of software to perform Kalman filtering directly is somewhat limited at present hindering its ready application. An increasing number of software packages are becoming available that will accommodate simple ARMA modeling and thus enable more users to perform this estimation method.

The only real way to judge the predictive performance of any one particular method in relation to the others is in practical application. The remainder of this thesis deals with the application of each of

these methods to identical data sets. Incomplete data sets used are generated from complete data series with 10, 20, and 30 percent of the data points removed. The predicted values generated are then compared to the original data. In this way it is possible to evaluate the accuracy of prediction and the limits of practical application. Data from three regions of British Columbia representing very different climatic and physiographic regimes are used as well as an artificially generated data series. This latter series represents a control series where all aspects of the series generation are known.

III. Comparative Analysis of Methods

Introduction

In order to perform a balanced comparison of the different interpolation methods each was applied to different sets of data. The data sets consisted of artificially generated temperature data series and actual temperature data series from three different physioclimatic regimes of British Columbia: the coastal, central interior plateau, and the Rocky Mountain trench. The use of a synthetically generated data series provides a controlled series where all sources of input can be accounted for. The three actual data series are used to compare the relative response of each method in practical application. The actual data sets also provide comparison of each method under potentially quite different climatic regimes to allow for evaluation of the portability of methods under different applications.

All statistical analysis, model development, and artificial data series generation, were performed using the SAS Institute's SAS Version 5.18 computer software. The SAS system components used for this thesis were base SAS, SAS/ETS, and SAS/Graph. Base SAS was used for artificial series generation, and the ratio, difference, regression, and polynomial methods. SAS/ETS is the

econometrics and time series package and was used for the ARIMA, lagged dependent regressor, and Kalman Filter methods. SAS/Graph was used to produce all graphic representations of the data and analysis results found in Appendix A.

A) Artificial Data Series

1) Methods

a) Artificial data series generation

In order to compare data completion methods with a controlled data series it was necessary to generate three synthetic temperature series. Three series were generated in order that methods which require more than the one station could be evaluated.

The generation of each data series was accomplished by first generating a white noise series of 365 data points. This was performed using the RANNOR function in SAS. This function generates a pseudo-random number (SAS Institute, 1985). When used in a series application such as this, the series of random numbers will approximate a normal distribution with unit variance.

The second step was to take a series of monthly temperature normals (30 year mean values) and, using

these 12 normals, generate a smoothed function which emulates an annual cycle. This was accomplished by taking 30 year normal data from three AES stations from the same physioclimatic region and generating three series. A sine function was used to smooth the 12 monthly values into 365 daily values. The sine function was generated using the SAS NLIN procedure.

Finally, the pseudo-random number series generated in the first step were added to the sine function series. In this way a known white noise function was added to the pure sine (or annual) function and the basis for all sources of signal variation are known.

The AES station normal data were selected from stations used as part of the actual series data analysis. This provided a basis for comparison of the response of the artificially generated series with actual measured data.

In order to generate the missing segments from the complete data series a random number generation strategy was established which reasonably approximates actual data loss. Tests on a five year data set from Stony Lake, which is used for method comparison later in this thesis, showed that the probability of instrument failure was approximately P=0.282 for any given 28 day period. This interval was selected as regular servicing would generally be performed on a four week basis. The

failures in the data set tested occurred as a result of datalogger and/or sensor failure, lightning strikes, and human error. The station selected for this calculation is in the interior plateau region of the province and represents reasonably current electronic datalogger technology but this general probability of failure may readily be applied to any of the regions since most of the conditions responsible for instrument failure are present in all regions. Other potential causes of failure are present where some conditions, such as lightning, are not as prevalent in a given region and thus the general probability of failure is assumed to be transportable from one region to another. Once the instrument has failed it cannot restart itself until it is serviced. Based on these premises, the data was divided into 28 day segments and any residual was added to the beginning of data series on the assumption that instrument failure is less likely to occur at the beginning of the installation. For each 28 day interval a random number representing the probability of failure was generated using the SAS function RANUNI which generates a uniformly distributed random variate (SAS Institute, 1985). If this number was less than 0.282 then a second random uniform number was generated. This number was restricted in the range from 1 to 28 and represented the day within the given interval on which

the failure occurred. All data from the failure day to the end of the interval was then considered to be missing.

Three copies of the first artificial data series were made and the aforementioned missing data generation strategy was applied to each of the artificial series copies to create data series with ten, twenty, and thirty percent of the original 'observations' missing. Data selection to the appropriate level (ten, twenty, or thirty percent) was controlled by generating only sufficient numbers of missing observations based on the length of the original complete series. The data series were subjected to the aforementioned data removal sequence. If the number of data points slated for removal did not total the set goal of ten, twenty, or thirty percent of the data set then another pass of the data was made. This process continued until the required percentage of data was removed. Slightly more or less than the ten, twenty, or thirty percent of the data may have been removed but this was limited to ±2 percent. This same process of missing data selection was used on the actual data sets used later for comparative analysis. In each case study and for each level of missing data, a new random selection series was used.

Each of the predictive methods is applied to the

individual levels of missing data using the same data series. Two steps are used to compare the effectiveness of the different methods. The first step is the comparison of the mean square error for each of the models. This provides an initial best approximation of the method to apply for interpolation of the missing values. The mean square error is used as the determining factor for order selection in the case of methods such as ordinary least-squares regression, polynomial curve fitting, or the time series methods where models could be developed with varying numbers of parameters. The second step is to compare the values removed from the original complete series with the values predicted by each applied method. This provides a test of the real variance of predicted versus actual values.

b) Model development

Models were developed for each of the 10, 20, and 30 percent levels of missing data using each of the methods under investigation. Model development for each of the seven methods was based on either standard technique using the data available after the generation of missing values or to relatively large contiguous segments of data dependent on the application

limitations of each model.

The artificial series 1 was arbitrarily assigned as the series from which to remove data to simulate missing data episodes. Series 2 and 3 were maintained for relational purposes. Series 3 was used where methods rely on a single relational station for predictive purposes as it was found to have the highest correlation coefficient with series 1.

The models for the ratio and difference methods were developed by obtaining the mean values for the available data. The temperature data were transformed to the Kelvin scale for use with the ratio method.

The ordinary least-squares, polynomial curve fit, and lagged dependent regressor methods also used all available data as a unit set. The type of lagged dependent regressor model used here allowed for the use of all available data. Different configurations of lagged dependent regressor models may not permit this depending on the computer software used for model development.

Due to the nature of the calculations and the limitations of current computer programs, it was necessary to develop individual segment models for each of the other time series methods. This required increased numbers of calculations but reduced the overall complexity of the required model.

The following section presents the results of the applications of each of the seven different methods applied to the artificially generated data series. The models developed using each method and level of missing data along with the mean square error (MSE), where applicable, can be found in appendices B, C, D, and E.

Where more than one model was possible within each method, the model with minimum MSE was selected to use for that method. This was the case for multiple regression, polynomial, and the time series models. The method with the minimum MSE for a given data set was deemed to perform the best for that situation. Additional criteria were used for selecting optimal models for the time series methods. These included minimum values for AIC (Akaike's Information Criterion), white noise variance, and estimated model variance.

The values estimated for the missing data are compared with the original values removed from the data sets. The absolute values of the residuals for the estimated missing values are classified in half degree Celsius (Kelvin) increments and the cumulative percentages are presented in a series of tables. This shows the percentage of predicted values for a given set which fall within set limits and allows comparison between methods. The acceptable limits of error were set at ±3.5°C for this thesis and methods with only a

small number of predictions within these specified bounds are rejected.

The half degree classification categories were established based on the general accuracy of instrumentation and an upper threshold of acceptable error. A residual with an absolute value which exceeded 3.5°C was considered outside usable limits and all values above this limit were grouped into a single category for the purposes of this thesis. The upper residual error limit in actual practice would most likely be set at a lower value.

It is not possible to estimate some of the missing data using some of the time series methods and the SAS software used in this analysis. The ARIMA and Kalman Filter methods are not capable of dealing with non-contiguous data, therefore models were developed for individual segments with 20 or more observations and forecasts were made using the segment models. The data set was then reassembled from the existing data and resulting forecasts. In some cases the missing data are present early in the data set and insufficient data exist to create a model and thus a segment of the data becomes inestimable. The missing data segment may be sufficiently large that forecasts become unrealistic and the estimated data points are dropped. The data are not recoverable as a result and are shown in the tables as

unrecoverable data and the value is given as a percentage of total missing data.

2) Results of analysis using artificial data series

The methods can be compared by examining the distribution of the differences between actual and estimated data. The range of differences is also an important consideration when evaluating the relative performance of any given method with a narrower range of values being preferred. The application of the various methods to the three levels of missing data for the artificial series produces the following results (Table 1).

The polynomial, lagged dependent regressor, and regression methods tend to be better overall estimators of the missing data than the other methods with 10 percent of the original data removed (Table 1). The lagged dependent regressor method has a slight advantage over the regression and polynomial methods in that the range of estimates tends to be slightly narrower.

With 20 percent of the data missing the lagged dependent regressor model performs considerably better than the others (Table 1). Four of the remaining six methods produce comparable results to one another up to the ±2.0°C range but diverge above this range. The

ARIMA and Kalman Filter methods produce quite good estimates with this data set, particularly in the low end ranges, but fail to match the performance of the other methods overall.

The Kalman Filter and ARIMA modeling methods have slightly fewer unrecoverable data points with 20 percent missing data than with 10 percent. This is most likely due to the locations of the missing data segments within the data series. It is possible that the missing data segments in the 10 percent data set come in sequences that produce relatively short segments of contiguous data whereas the 20 percent data may have longer segments of contiguous data. Long contiguous data segments allow for better model estimation for the ARIMA and Kalman Filter methods since trends that should be in the entire data series are more likely to appear in a long sequence than a short segment. These two methods also perform as relative equals up to and including the ±1.5°C range which includes 92-93% of the estimated missing values.

Table 1.

Ranges of Residuals for Missing Values
in Cumulative Percentage Frequency for
the Artificial Series Data Sets

Range [°C]	Pct. Miss.	Estimation Method						
		Ratio	Diff.	Regr.	Poly.	ARIMA	LDR	Kalm.
± 0.5	10	45.9	45.9	37.8	48.6	28.3	43.2	23.4
	20	51.9	51.9	49.4	53.2	51.2	61.0	47.1
	30	18.3	42.6	45.2	31.3	12.1	45.2	14.5
± 1.0	10	75.7	75.7	78.4	81.1	50.0	91.9	40.4
	20	75.3	75.3	79.2	80.5	79.8	87.0	76.5
	30	46.1	70.4	74.8	69.6	25.8	71.3	27.4
± 1.5	10	94.6	94.6	94.6	91.9	82.6	100.0	70.2
	20	89.6	89.6	94.8	92.2	91.7	98.7	92.9
	30	67.0	87.8	91.3	85.2	36.3	93.9	41.0
± 2.0	10	97.3	97.3	100.0	100.0	89.1	100.0	93.6
	20	100.0	100.0	100.0	100.0	95.2	100.0	97.6
	30	80.9	96.5	99.1	97.4	48.4	99.1	58.1
± 2.5	10	99.7	100.0	100.0	100.0	93.5	100.0	93.6
	20	100.0	100.0	100.0	100.0	96.4	100.0	98.8
	30	93.0	99.1	100.0	99.1	57.3	100.0	64.1
± 3.0	10	100.0	100.0		100.0	95.7		97.9
	20	100.0	100.0		100.0	96.4		98.8
	30	98.3	100.0		100.0	62.1		75.2
± 3.5	10	100.0				95.7		97.9
	20	100.0				96.4		96.4
	30	99.1				71.8		78.6
±> 3.5	10	100.0				97.8		97.9
	20	100.0				98.8		98.8
	30	100.0				100.0		100.0
Unre-	10					2.2		2.1
cover-	20					1.2		1.2
able	30					0.0		0.0

The lagged dependent regressor, regression, difference, and polynomial methods are the better performing models with thirty percent of the data missing (Table 1). The former two methods are the best and the regression method has a slight edge over the lagged regressor method with a higher percentage of the estimated data falling within the ±1.0°C range. The ARIMA and Kalman Filter methods have no unrecoverable data in this case but have a much higher percentage of data in the ±>3.5°C range than in either of the other two test data sets. The estimation capabilities of these two methods are poor compared to any of the other methods but tend to be comparable to each other.

The regression and lagged dependent regressor methods produce consistently better results on these artificial data series than does the difference method, which has been the most commonly practiced method of temperature climate data estimation, or the ARIMA or Kalman Filter methods. This is more evident as the percentage of missing data in the sets increases.

The application of the different methods to the artificial data series has separated the methods into three rather distinct categories of predictive capabilities. The regression, lagged dependent regressor model methods and, to a certain extent, the polynomial method were very good estimators of the

missing series data. The difference and ratio methods were average estimators and the ARIMA and Kalman Filter methods tended to be rather poor estimators of the missing temperature data. The quality of estimation for the latter two methods dropped significantly with increasing proportions of missing data. Some of the variation in the predictive capabilities of the ARIMA method can be explained by the quality of fit of the models for any given segment of data. A relatively short data segment may not provide adequate information to establish a good model with which to estimate a long period of missing data that follows. Thus, the estimated values for any given missing data period may not be of high precision when the data segments are reintegrated into a single data set.

In this application it is fairly easy to pick which method one may want to use based on ease of application and available computing capabilities. A better test of the estimation capabilities of each of the different methods is to apply them to actual temperature data sets.

B) Actual Data Series

Data sets from three very different physiographic regions of British Columbia are used which give an

indication of the robustness of the missing data estimation methods. The three data sets represent coastal, central interior plateau, and the Rocky Mountain trench regions of the province (Table 2).

Table 2.

Forestry Canada Weather Station Information

Station	Latitude	Longitude	Elevation
Driftwood Creek	50° 53'N	116° 34'W	1370 m
Green Mountain	49° 03'N	124° 35'W	760 m
Stony Lake	53° 27'N	121° 54'W	980 m

1) Data and methods

a) Stations and data

Data sets consisting of approximately one year's average daily temperature values were collected using automated datalogging equipment capable of recording hourly and daily maximum, minimum, and mean temperature data. These recording stations, installed by Forestry Canada, collected environmental information for research conducted in these regions. Automated system failure and human error resulted in the loss of information in the data collected. Contiguous segments of data are

used here.

Data from Atmospheric Environment Service climate stations within the appropriate physioclimatic regions or, in the case of the interior plateau site, adjacent physioclimatic regions are used for predictive purposes for the ratio, difference, polynomial, regression, and lagged dependent regressor methods which take advantage of this information.

Some differences exist in the method of temperature measurement and reporting between the Forestry Canada and the Atmospheric Environment Service data. The Atmospheric Environment Service reports the daily average value as the midpoint between the observed daily maximum and daily minimum values on glass mercury and alcohol thermometers. The Forestry Canada automated stations scan glass encapsulated thermoresistive sensors every sixty seconds and integrate the data both hourly and daily. The response time of the electronic sensors to reach equilibrium is approximately ten seconds to three minutes while the mercury and alcohol thermometers is approximately five to ten minutes depending on exposure and degree of temperature change. This response time difference means that the electronic sensors are more susceptible to transient extremes that may occur over short time periods thus producing more extreme maximums and minimums. Both the electronic

sensors and Atmospheric Environment Service thermometers were exposed in standard screens at approximately 1.3 meters (standard) height.

The difference in calculation method of the daily mean temperature poses a potential problem in that the method used for reporting by the Atmospheric Environment Service is highly dependent on the range between the maximum and minimum temperature values. This dependency may not necessarily reflect the true mean daily temperature. For the sake of uniformity, the Forestry Canada mean daily temperatures were calculated in the same way used by the Atmospheric Environment Service. This problem may diminish in the future as the Atmospheric Environment Service deploys more automated stations which are capable of producing a true daily mean value. The potential for error exists since the automated stations are capable of measuring more extreme daily ranges between maximum and minimum values, thus influencing the daily mean temperature values and potentially influencing the results of any method used to estimate missing data values when the two different measurement systems are used.

The three different physioclimatic regions used for this study display quite varied daily, weekly and, seasonal responses to air temperature. The Driftwood Creek site, is located in the Purcell Mountains on the

west side of the Rocky Mountain trench near Golden. This site is representative of mountainous regions where valleys play a key role in local climate. The coastal site, Green Mountain, is located near Nanaimo on Vancouver Island. This site is quite representative of a coastal watershed subject to maritime effects. The central interior plateau site, Stony Lake, is located near Prince George. This site, while located at the junction of three large open valleys, is subject to weather systems typical of the general region.

The Driftwood Creek site is located on a kame terrace on the eastern slopes of the Purcell Mountains. There is a slight east-northeast aspect to the exposure with an effective slope of approximately 44 percent to the surrounding terrain but the weather station site is effectively flat. Localized weather and climate patterns are predominantly affected by the north-south flow along the valley.

There were five Atmospheric Environment Service climate stations within the same general physioclimatic region as the Driftwood Creek station. These were Bobbie Burns, Bugaboo Creek Lodge, Glacier National Park Mount Fidelity, Glacier National Park Rogers Pass, and Golden (Table 3). These stations range from 5 to 70 kilometers from the Driftwood Creek station. The former three stations were located on the sides of valleys, as

was the Forestry Canada station, and the the latter two were located in valley bottoms. The close proximity of a large number of AES stations provided a good selection of potential candidates for methods such as ordinary least squares regression which can take advantage of this.

Table 3.

Atmospheric Environment Service Climate Stations Near the Driftwood Creek Station

Station	Latitude	Longitude	Elevation
Bobbie Burns	50° 56'N	116° 56'W	1370 m
Bugaboo Cr. Lodge	50° 45'N	116° 42'W	1494 m
Mount Fidelity	51° 14'N	117° 42'W	1874 m
Rogers Pass	51° 17'N	117° 31'W	1323 m
Golden	51° 18'N	116° 58'W	785 m

The Green Mountain site is located on a mountain side on the eastern side of Vancouver Island at an elevation of 760 meters. The site has a southwest aspect with a slope of 64 percent and is influenced by the localized mountainous terrain and heavily biased by maritime effects due to its proximity to the Pacific Ocean and the Georgia Strait.

Four Atmospheric Environment Service climate stations were used for comparison with the Green

Mountain data series due to their close proximity to the Forestry Canada climate station. These were Cowichan Lake Forestry, Duncan Forestry, Port Alberni Airport, and Nanaimo Airport (Table 4). The AES stations range from 20 to 65 kilometers from the Green Mountain station. The four AES stations were located in relatively open valleys on or near the valley floor in contrast to the Forestry Canada stations which was located on a hillside.

Table 4.

Atmospheric Environment Service Climate Stations Near the Green Mountain Station

Station	Latitude	Longitude	Elevation
Cowichan Forestry	48° 50'N	124° 08'W	177 m
Duncan Forestry	48° 47'N	123° 41'W	6 m
Nanaimo Airport	49° 03'N	123° 52'W	30 m
Port Alberni A.	49° 15'N	124° 50'W	2 m

The Stony Lake site, while located in the western foothills of the Cariboo Mountains, is primarily affected by the westerly flows across the Fraser plateau. The weather station site has a very slight southern exposure in a rolling landscape with relatively low relief.

The four closest Atmospheric Environment Service

climate stations to the Stony Lake site were Prince George Airport, Hixon, Quesnel Airport, and Barkerville (Table 5). These stations range 50 to 80 kilometers from the Stony Lake installation. Unlike the Green Mountain or Driftwood Creek sites, none of the AES stations fall within the same general physioclimatic regions as the Stony Lake station. This is a situation which occurs in much of the less inhabited portions of the province.

Table 5.

Atmospheric Environment Service Climate Stations Near the Stony Lake Station

Station	Latitude	Longitude	Elevation
Prince George A.	53° 53'N	122° 40'W	691 m
Hixon	53° 25'N	122° 35'W	541 m
Quesnel Airport	53° 02'N	122° 31'W	545 m
Barkerville	53° 04'N	121° 31'W	1265 m

The data collected from each of the Forestry Canada automated stations were quality checked to ensure the longest contiguous segment of temperature data available was used for comparative purposes. Three copies were made of the complete data sets for each of the different regions and data was removed from the Forestry Canada data series using the same random number series method

applied to the artificially generated series described in the previous section.

b) Model development

Where several potential models are possible for any given method, the model with the minimum mean square error was selected. Additional criteria, such as the minimum AIC and minimum white noise variance, were used for the selection of optimal time series models. Due to the complexity and range of time series models, particularly with the ARIMA model method, limits were placed on the potential models tested. The limits were based on analysis of the autocorrelation and partial autocorrelation plots which indicated possible components which should be included in the potential model and required data transformations.

The missing data values estimated by the application of the seven different methods are then compared to the corresponding values actually measured and the absolute values of the residuals are classified into ranges and presented in tables. Data listed as unrecoverable is data that could not be estimated using the method applied and based on the available data.

2) Results of analysis using actual data series

a) Driftwood Creek data results

The Driftwood Creek site, located in the Rocky Mountain trench region, has five highly correlated AES stations located in the same physioclimatic region. Two of these stations are within a few kilometers of the experimental site and one of these is located on the same ridge. Location was used as a weighting criterion in addition to correlation for selection of AES stations to use with this site although it was not used in actual model parameterization.

The Bobbie Burns AES station data was used for the ratio, difference, polynomial, and lagged dependent regressor methods which used a single correlated station for predictive purposes. This station proved to have the highest correlation coefficient for all levels of missing data. The ordinary least squares regression method incorporated all the AES stations with the exception of the Bugaboo Creek Lodge station which was an incomplete series in itself.

With ten percent of the data missing from the original data set, the ARIMA and Kalman Filter methods forecast very little in the way of usable results with less than half of the forecasted values falling within

the ±3.5°C acceptable limits whereas the regression, polynomial, and lagged dependent regressor methods offer the best predictive performance (Table 6). The polynomial method performs the best of the three in the range ±2.0°C but has close to 4% of the predicted values falling outside usable limits. The only method where all of the predicted values fall within the set acceptable limits is the regression method.

As with the ten percent missing data, the polynomial method outperforms the other methods with 92% of the predicted values falling within ±2°C of the actual values with twenty percent of the data missing (Table 6). The regression method closely follows and has the advantage of all the predicted data points falling within the usable limits of ±3.5°C.

The regression method outperforms all other methods and is the only method to have all estimated data points within acceptable limits with thirty percent of the data missing (Table 6). The lagged dependent regressor and polynomial methods are comparable at ranges up to ±2.5°C but have 5% and 4%, respectively, of predicted values falling outside acceptable limits.

Table 6.

Ranges of Residuals for Missing Values in Cumulative Percentage Frequency for the Driftwood Creek Data Sets

Range [°C]	Pct. Miss.	Estimation Method						
		Ratio	Diff.	Regr.	Poly.	ARIMA	LDR	Kalm.
± 0.5	10	21.4	21.4	25.0	39.3	3.7	28.6	3.6
	20	16.7	16.7	22.2	38.9	5.7	22.2	5.6
	30	26.8	29.3	32.9	26.8	4.6	28.0	9.8
± 1.0	10	28.6	35.7	45.8	64.3	11.1	39.3	14.3
	20	36.1	38.9	38.9	63.9	14.3	36.1	13.9
	30	50.0	51.2	61.6	53.7	9.2	56.1	17.1
± 1.5	10	50.0	53.6	62.5	89.3	14.8	57.1	17.9
	20	58.3	58.7	61.1	86.1	17.1	57.1	19.4
	30	67.1	68.3	79.5	72.0	16.1	74.4	23.2
± 2.0	10	75.0	75.0	83.3	89.3	22.2	71.4	25.0
	20	77.8	77.8	80.6	91.7	20.0	77.8	19.4
	30	80.5	80.5	84.9	81.7	21.8	86.6	29.3
± 2.5	10	78.6	78.6	95.8	89.3	25.9	89.3	28.6
	20	80.6	80.6	91.7	91.7	28.6	88.9	25.0
	30	85.4	85.4	89.0	86.6	26.4	91.5	37.8
± 3.0	10	82.1	89.3	95.8	92.9	33.3	89.3	39.3
	20	86.1	91.7	94.4	94.4	37.1	88.9	36.1
	30	90.2	90.2	98.6	96.3	35.6	91.5	46.3
± 3.5	10	92.9	92.9	100.0	96.4	37.0	89.3	42.9
	20	94.4	94.4	100.0	97.2	40.0	94.4	38.9
	30	96.3	96.3	100.0	96.3	40.2	95.1	53.7
±> 3.5	10	100.0	100.0		100.0	100.0	100.0	100.0
	20	100.0	100.0		100.0	100.0	100.0	100.0
	30	100.0	100.0		100.0	100.0	100.0	100.0
Unre-	10					0		0
cover-	20					0		0
able	30					0		0

At all levels of missing data the regression method provides the best performance for the Driftwood Creek site in that this is the only method to estimate all values within the limits defined as acceptable. This is followed very closely by the lagged dependent regressor and polynomial methods which provide estimates which are equal to or better than the regression method in many of the lower residual ranges. The polynomial method estimates for all ranges tend to group in the upper and lower extreme ranges with the majority being in the lower ranges. The regression and lagged dependent method estimates tend to group more in the lower and middle ranges with emphasis towards the lower ranges.

b) Green Mountain data results

The Green Mountain data set is representative of a coastal climatic regime. The results from application of the different methods to this data set are similar to those from the Driftwood Creek (Rocky Mountain trench) data.

The Green Mountain site was much higher than the nearby AES stations used for modelling purposes. Since temperature tends to be a linear function of elevation this is not a significant problem. The relationships between this station and the AES stations is more likely

to be influenced by the moderating effect of the large water bodies nearby and the fact the highly correlated AES station are both affected by similar valley exposures.

The Cowichan Lake Forestry AES station had the highest correlation coefficient for those stations in the vicinity of Green Mountain and was used for those methods using a single correlated station. The regression method only used the Cowichan Lake Forestry and Port Alberni Airport AES stations as the remaining stations added little to the predictive capabilities of the model when in the presence of these two stations. These relationships held for all levels of missing data.

The regression method is generally a better estimator of the missing data in the ranges up to ±2.5°C with the polynomial and lagged dependent regressor methods showing similarly good responses when ten percent of the data has been removed (Table 7). Above this range all methods except the ARIMA and Kalman Filter methods are equally good estimators of the missing values.

With twenty percent of the data missing, no real distinction exists between any of the methods other than the ARIMA and Kalman Filter methods which do not perform as well. The lagged dependent regressor method has a slight advantage over the other methods with 1.2% of the

Table 7.

Ranges of Residuals for Missing Values in Cumulative Percentage Frequency for the Green Mountain Data Sets

Range [°C]	Pct. Miss.	Estimation Method						
		Ratio	Diff.	Regr.	Poly.	ARIMA	LDR	Kalm.
± 0.5	10	26.8	26.8	31.7	29.3	2.5	22.0	7.3
	20	26.3	26.3	28.8	23.8	19.5	25.0	12.5
	30	28.7	27.9	22.1	30.3	5.1	30.3	5.8
± 1.0	10	39.0	36.6	41.5	51.2	15.0	39.0	17.1
	20	37.5	37.5	48.8	48.8	27.6	41.3	28.8
	30	45.1	45.9	45.1	52.5	10.1	49.2	12.2
± 1.5	10	63.4	63.4	63.4	61.0	17.5	65.9	24.4
	20	62.5	62.5	61.3	62.5	37.9	62.5	40.0
	30	63.1	63.1	65.6	67.2	15.9	63.1	15.8
± 2.0	10	75.6	75.6	80.5	75.6	27.5	80.5	26.8
	20	71.3	73.8	75.0	75.0	44.8	75.0	51.3
	30	77.0	77.9	77.9	82.0	18.1	81.1	18.7
± 2.5	10	90.2	90.2	92.7	85.4	35.0	90.2	36.6
	20	88.8	87.5	86.3	87.5	51.7	87.5	56.3
	30	92.6	83.6	91.0	92.6	20.3	90.2	20.9
± 3.0	10	97.6	97.6	97.6	97.6	40.0	95.1	41.5
	20	93.8	93.4	93.8	93.8	55.2	93.8	60.0
	30	98.4	87.7	95.9	96.7	23.9	99.2	25.2
± 3.5	10	100.0	100.0	100.0	100.0	45.0	100.0	46.3
	20	97.5	97.5	96.3	96.3	62.1	98.8	62.5
	30	100.0	100.0	100.0	99.2	27.5	99.2	30.2
±> 3.5	10	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	20	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	30	100.0	100.0	100.0	100.0	92.8	100.0	92.8
Unre-	10					0		0
cover-	20					0		0
able	30					7.2		7.2

data being out of usable range whereas the ratio and difference methods have slightly more than 2.5% outside acceptable limits.

The regression, ratio, and difference methods are the only methods which estimate all missing values within acceptable limits with thirty percent of the data removed. The ratio and difference methods produce estimators which are comparable with the other methods in all ranges which suggests that these methods are preferred due to ease of use for most operational applications. The ARIMA and Kalman Filter methods perform poorly with less than 31% of the data falling within acceptable limits.

The Green Mountain data sets, once again, show very little overall difference between the regression, lagged dependent regressor, and polynomial methods over the three levels of missing data. The ratio and difference methods also perform reasonably well with no real advantage going to either method.

c) Stony Lake data results

The Stony Lake data sets are representative of the interior plateau region of British Columbia. The surrounding AES stations are all located in relatively narrow river valleys and tend to be influenced by this

factor whereas the Stony Lake site is at the junction of three much broader valleys. The Stony Lake site was also at a considerably higher elevation than the AES stations with the exception of the Barkerville AES station which is higher than Stony Lake.

The Hixon AES station had the highest correlation with this site with ten and twenty percent of the data missing while the Quesnel Airport AES station had the highest correlation with thirty percent of the data missing. These stations were used for the single correlated station methods at their respective levels of missing data. All four of the surrounding AES stations were used for the ordinary least squares regression method.

The response of the application of the various methods to the Stony Lake data is slightly different from that for the other data sets. Some methods which consistently perform well at the other two stations are relatively not as effective when applied to these data. In addition, the serial correlation methods generally performed better on this data set than with the other data sets. There was a higher level of estimated data outside the acceptable limits in all levels of missing data than was found in the coastal or Rocky Mountain data sets.

The lagged dependent regressor method significantly

outperforms all other methods in estimating the missing values with ten percent of the data removed (Table 8). The polynomial, ratio, and difference methods follow and are fairly comparable to one another. The regression method does not perform well at all for this application.

The lagged dependent regressor method is generally the better method with more of the distribution of estimates tending toward the lower ranges of residuals and a significantly lower percentage of the estimated data falling outside the acceptable limits than any of the other methods with twenty percent of the data missing (Table 8). The performance of the ratio, difference, regression, and polynomial methods were comparable. Consistent with all other data sets, the ARIMA and Kalman Filter methods perform poorly.

Table 8.

Ranges of Residuals for Missing Values in Cumulative Percentage Frequency for the Stony Lake Data Sets

Range [°C]	Pct. Miss.	Estimation Method						
		Ratio	Diff.	Regr.	Poly.	ARIMA	LDR	Kalm.
±0.5	10	11.1	11.1	2.8	19.4	8.6	30.6	13.9
	20	16.0	14.7	16.0	20.0	0.0	25.3	8.0
	30	12.4	13.3	11.5	11.9	9.2	19.5	3.5
± 1.0	10	41.7	41.7	25.0	41.7	14.3	52.8	16.7
	20	30.7	30.7	33.3	34.7	8.1	41.3	10.7
	30	21.2	20.4	29.2	24.6	17.6	34.5	9.7
± 1.5	10	58.3	55.6	58.3	55.6	17.1	75.0	22.2
	20	41.3	42.7	50.7	44.0	18.9	58.7	18.7
	30	34.5	34.5	40.7	34.9	22.7	46.9	18.6
± 2.0	10	72.2	69.4	66.7	72.2	25.7	83.3	30.6
	20	56.0	52.0	60.0	58.7	24.3	68.0	24.0
	30	45.1	44.2	54.9	46.8	31.1	61.9	23.9
± 2.5	10	80.6	80.6	69.4	80.6	34.3	83.3	41.7
	20	69.3	69.3	65.3	69.3	25.7	76.0	28.0
	30	56.6	56.6	65.5	59.5	37.8	70.8	30.1
± 3.0	10	83.3	83.3	86.1	88.9	48.6	86.1	50.0
	20	72.0	72.0	73.3	76.0	35.1	84.0	33.3
	30	66.4	66.4	74.3	65.1	47.1	76.1	36.3
± 3.5	10	88.9	88.9	88.9	88.9	54.3	91.7	61.1
	20	78.7	80.0	84.0	78.7	37.8	86.7	38.7
	30	72.6	72.6	82.3	71.4	49.6	82.3	43.4
±> 3.5	10	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	20	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	30	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Unre-cover-able	10					0		0
	20					0		0
	30					0		0

With thirty percent of the data missing the lagged dependent regressor method outperforms all other methods in virtually all ranges. The regression method is second to the lagged dependent regressor method in quality of estimation in all ranges above ±1.0°C. The ratio, difference and polynomial methods are very similar in their estimation capabilities.

The Stony Lake data set proves to be the most challenging for all methods at all levels because the surrounding AES stations are within slightly different physioclimatic regimes. Subsequently, the performance of the various methods proves to be slightly different than with the other data sets used in this thesis. Unlike the response with other data sets, the regression method tends to produce estimate range distributions which tend to polarize toward the extremes. This is true at all levels of missing data. As a result this method's performance drops in relation to the other methods. The lagged dependent regressor method provides consistently better estimates of the missing data than the other methods at the three levels of missing observations.

IV. Summary and Conclusions

The results cited in this thesis provide a basis for establishing effective and efficient methods for the estimation of missing temperature data. The methods presented offer a cross section of techniques which range from simple to difficult in ease of application with widely varying degrees of operational success. User selection of the most effective method with which to estimate missing values in a temperature data series will ultimately be determined by the amount of data from the incomplete series, additional information from nearby stations, and the availability of computing resources.

The nature of the environmental variable being estimated is a prime consideration. Temperature data are highly correlated spatially and temporally. Standard methods such as the difference method and ordinary least squares regression do not take this into account and thus, resulting estimates may be biased as a consequence.

The data analysis for this thesis shows that while no method was vastly superior in all cases some methods were consistently better than others in a large proportion of the cases. It was possible to rank the effective use of the different methods based on their

overall performance with all the data sets, the different levels of missing data, and ease of use (Table 9).

Table 9.

Method Performance Ratings

Method	Reliant on other station	Adherance to method restrict.	Ease of Use	General Perfor-mance
Ratio	Yes	Moderate	Moderate	Moderate
Difference	Yes	Moderate	Good	Moderate
Regression	Yes	Poor	Moderate	Good
Polynomial	No	Poor	Moderate	Moderate to Good
ARIMA	No	Good	Difficult	Poor
Lagged Dependent Regressor	Yes	Good	Moderate	Good
Kalman Filter	No	Good	Difficult	Poor

The ratio and difference methods, which are in standard use for adjustment of climate data normals, were found to be better than other methods in only a very few cases. These methods generally produced moderately good results which, for the most part, would be adequate for many purposes. The reliance of these

methods on the mean of each series remaining relatively constant compared with the data segments available is perhaps the most limiting factor in their usefulness. While the series mean will drift very little with long series of data, it is not likely that the calculated mean for a short series and a long term mean will necessarily be similar. These methods still have the advantage of application ease since data adjustment can be done easily without the requirement of computer facilities but this advantage is doubtful given the current low cost and easy access to such facilities.

Ordinary least squares regression and polynomial regression methods have the advantage of moderately easy application. These methods are very easy to apply using commonly available computer software and can be done on some handheld calculators without great difficulty. The results of application, as presented in this thesis, show two features. First, the regression method tends to produce distributions of residuals that group toward the smaller (±0.5°C, ±1.0°C) and middle (±1.5°C, ±2.0°C) ranges. Second, the polynomial method tends to produce results that group in the extreme ranges. This tendency toward extremes for the polynomial method produces estimates which are highly desirable and those which are beyond the usable limits specified earlier in this thesis. These two features of the least squares methods

suggest that a further study might include a mixed regression model with several variables and relevant powers of those variables. The development of such a model could be achieved using stepwise regression techniques. Final selection of this type of model should be based on criteria such as minimum mean square error and Mallows' C_p as well as maximum R^2.

The regression based methods generally outperformed other methods but key assumptions upon which standard regression is based do not hold. Observations are serially correlated due to annual and other cycles and thus are not independent. The vector of error terms is also autocorrelated. When more than one station is used there is a spatial correlation and thus variables are not necessarily independent. This would also hold true if other environmental variables, such as solar radiation, were used as variables as they are rarely independent of one another.

While two of the three time series methods did not perform well, and indeed performed very poorly, the lagged dependent regressor method consistently produced very good results. Time series methods are readily adaptable for spatial and temporal applications and take autocorrelation between observations and variables into account.

The application of ARIMA and modified Kalman Filter

(or Statespace) methods was restricted by their general inability to deal with noncontiguous data sets and the necessity for relatively sophisticated computer software. Models must be fitted to individual data segments and several problems arise as a consequence. Modeling individual segments is time consuming. The data segments must be of sufficient length to provide enough observations to produce forecasts. Determination of the required number of observations is based on trends found within the contiguous data segments which may or may not adequately represent the true nature of the series. The results of this thesis indicated that, as a general rule of thumb, at least twenty contiguous daily temperature observations are required to produce reasonable forecasts to bridge a missing data gap when using only one year of data. This rule is limited by the length of missing data to be bridged as the forecasting limits of the model are determined by the number of model parameters and length of data series used to develop the model.

The problems encountered with segment modeling prevented the application of a true transfer function using the Box-Jenkins method. This is a result of the necessity to use ARIMA procedures in the development of this type of model. Subsequently the transfer function method was dropped in favor of the lagged dependent

regressor method which is very similar in form but without the restrictions. Transfer function models would be possible with larger contiguous data series than were available for this analysis.

The segment models must be based on several criteria. The choices of criteria are often limited by the computer software available and the method of calculation for any given method. Minimum AIC and MSE were the primary criteria used for optimal model selection in this thesis. Parsimony was also a key factor as time series models can readily become cumbersome. The AIC provides a measure of the parsimony of the model and the MSE provides a measure of the accuracy and precision of the model.

The application of the Kalman Filter method is relatively straight forward and easy to apply as it is simply a restricted ARMA(1,1) model. It is necessary to stabilize the temperature data and thus the data often require differencing. A single differencing is performed in order to avoid the introduction of false harmonics into the series. The Kalman Filter method thus becomes an ARIMA(1,1,1) model.

An ARIMA model is fit to each data segment and as a result the models may be different in parameterization and have a wide variety of configurations. One potential inconsistency is that individual segments may

have better model fits than others and thus the forecasts for any given missing segment will not be uniform in accuracy throughout the reconstructed data set. The skill of the modeler in determining the optimum model for a given data segment thus becomes a key factor in the applicability of this method.

The relatively poor results of the ARIMA method application in this thesis may be a direct result of data segment model optimization methods and skill. As such, the evaluation of this method compared to others may be slightly biased. This illustrates the complex nature and potential pitfalls of potential application of this method.

The Kalman Filter provides an easy method of ARIMA model development. It is also a constant form of time series model, in that an ARIMA(1,1,1) model is used, which is applied to all data segments in the series and should afford relatively consistent forecasting capabilities over the whole of the data set. This, however, does not necessarily hold true as the ARIMA(1,1,1) model does not always produce the optimal description of any given data segment. It was found that, for longer contiguous data segments, the Kalman Filter method produced better forecasts than for shorter segments and that this held true for all data sets used in this thesis. This suggests that, over a long period

of time, an ARIMA(1,1,1) model may provide an adequate model for estimation purposes. This may provide a relatively simple method for application of ARIMA type procedures to this problem but must be verified through further testing using longer data sets.

Comparison of the results of the different methods between data sets showed that no one method is superior in all situations. The ratio and difference methods tended to perform better with data sets where climatic variation was smaller and the related stations had a more constant relationship such as the cases of the Green Mountain and Driftwood Creek data sets. The performance of these two methods dropped significantly when forced to contend with the inconsistent relationships due to wide climatic variations between stations as demonstrated with the Stony Lake data sets. The regression based methods such as the ordinary least squares regression, polynomial regression, and lagged dependent regressor proved to be better overall estimators for all data sets on a more consistent basis than the other four methods.

While ordinary least squares regression and polynomial regression models may be good solutions for many data types such as the Driftwood Creek and Green Mountain sites where a large number of related stations are relatively close, the lagged dependent regressor

method has been demonstrated to be a superior method in situations where large relative distances may separate related stations, such as the case with the Stony Lake data. This also holds for situations where high variability of topography or land use have a marked influence on the microclimate where the station with the missing data is located since closely related stations may not be readily available. The lagged dependent regressor method is a better method than standard regression methods for situations where related stations may be in different physioclimatic regimes because it incorporates the long term relationship between two stations, the immediate past history of the short series station data and the current data from the related station. The use of the spatial and temporal aspects of this model make it preferable for applications where the selection of nearby stations is limited as was the case of the Stony Lake example.

The selection of the most appropriate method of missing data estimation to use for a given data set must be based on several criteria. These include the availability of appropriate computing capabilities, concurrent data from a nearby station or data from other associated environmental variables from the short series station, and a sufficient amount of contiguous data from the short series station to establish trends. Method

selection will be determined by this information. The ratio and difference methods offer ease of calculation but suffer from inconsistency of good estimation. The ARIMA and Kalman Filter models are complex, relatively difficult to use, and offer little in the way of good estimation except where long segments of contiguous data exist. These two time series methods may prove much better estimators given more data than used in this thesis. Regression and polynomial models provide good consistent results in this application. The polynomial method, however, has been shown to produce unrealistic results in some temperature estimation applications (Thiébaux and Pedder, 1987). Furthermore, the highly correlated spatial and temporal nature of climate (particularly temperature) data do not meet key assumptions made in the application of least squares regression methods.

The simple form lagged dependent regressor model is capable of incorporating many of the advantages of ordinary least squares regression while accounting for the spatial and temporal components. This method produced consistently good estimations of the missing data in the three varied physioclimatic regimes presented here as well as the artificial series generated for test purposes. The lagged dependent regressor method is, perhaps, a little more difficult to

apply in that it is a three step process but all steps are readily achieved using common computer software. As such, the lagged dependent regressor method proved itself to be the preferred method of missing temperature data estimation. The application of this method to other climatic variables would require an evaluation of the nature of the variable to be estimated.

Bibliography

Atmospheric Environment Service. Canadian Climate Normals. Volume 2. Temperature. 1951 - 1980. Environment Canada, Downsview. 1982a.

Atmospheric Environment Service. Canadian Climate Normals. Volume 3. Precipitation. 1951 - 1980. Environment Canada, Downsview. 1982b.

W.F. Baird and Associates Coastal Engineers Ltd., Hydrotek Resource Consultants, and Lawless J.F. Review and Assessment of Procedures for Extreme Value Analysis for Selected Geophysical Data. Report No. 86-7. Unpublished Manuscript. Atmospheric Environment Service. Downsview. 1986.

W.F. Baird and Associates Coastal Engineers Ltd., Hydrotek Resource Consultants, and Lawless J.F. Review and Assessment of Procedures for Extreme Value Analysis for Selected Geophysical Data. Phase II Report. Report No. 87-5. Unpublished Manuscript. Atmospheric Environment Service. Downsview. 1987.

Barham, S. Y. and Dunstan, F. D. J. "Missing Values in Time Series" in Time Series Analysis: Theory and Practice 2. O.D. Anderson (editor). North-Holland Publishing Co., New York, 1982.

Box, G. E. and Jenkins, G. M. Time Series Analysis: forecasting and control. Revised Edition. Holden-Day Inc., Toronto and San Francisco, 1976.

Box, G. E. and Tiao, G. C. "Comparison of Forecast and

Actuality." Applied Statistics, volume 25, number 3, 1976.

Box, M. J. "A Parameter Estimation Criterion for Multiresponse Models Applicable when Some Observations are Missing." Applied Statistics, volume 20, number 1, 1971.

Bradshaw, L. S. and Slazar, L. A. "On Using a Fourier Series Model for Estimating Diurnal Temperatures at Mountainous Locations in the Western United States." Journal of Climatology and Applied Meteorology, volume 24, number 3, 1985.

Brooks, C. E. P., and Carruthers, N. Handbook of Statistical Methods in Meteorology. Her Majesty's Stationery Office, London. 1953.

Brubacher, S. R. and Wilson, G. T. "Interpolating Time Series with Application to the Estimation of Holiday Effects on Electricity Demand." Applied Statistics, volume 25, number 2, 1976.

Burkhardt, T. and Hense, A. "On the Reconstruction of Temperature Records From proxy Data in Mid Europe." Archives For Meteorology, Geophysics, and Bioclimatology, Series B, volume 35, number 4, 1985.

Burt, J. E. "Time Averages, Climatic Change, and Predictability", Geographical Analysis, volume 18, number 4, 1986.

Byrd, G. P. "An Adjustment for the Effects of Observation Time on Mean Temperature and Degree-Day Computations." Journal of Climate and Applied

Meteorology, volume 24, 1985.

Cameron, M. A. and Turner, T. R. "Fitting Models to Spectra using Regression Packages." Applied Statistics, volume 36, number 1, 1987.

Chaghaghi, François S. Time Series Package (TSPACK). Springer-Verlag. New York, 1985.

Chang, T. J., Kavvas, M. L., and Delleur, J. W. "Modelling of Sequences of Wet and Dry Days by Binary Discrete Autoregressive Moving Average Processes", Journal of Climatology and Applied Meteorology, volume 23, number 3, 1984.

Chatfield, C. The Analysis of Time Series: An Introduction. Chapman and Hall. London, New York, 1980.

Chatfield, C. and Pepper, M. "Time Series Analysis: An Example from Geophysical Data." Applied Statistics, volume 20, number 3, 1971.

Chow, G. C. and Lin, A. "Best Linear Unbiased Estimation of Missing Observations in an Economic Time Series." Journal of the American Statistical Association, volume 71, number 355, 1976.

Clarke, B. R. "Algorithm AS 194 - An Algorithm for Testing Goodness of Fit of ARMA(P,Q) Models." Applied Statistics, volume 32, 1983.

Cliff, A.D., Haggett, P., Ord, J. K., Bassett, K. A., and Davies, R. B. Elements of Spatial Structure. Cambridge University Press. New York, 1975.

Cliff, A.D. and Ord, J. K. "Model Building and Analysis of Spatial Pattern in Human Geography", Journal of the Royal Statistical Society, volume 37, number 3, 1975.

Conrad, V. and Pollack, L. W. Methods of Climatology. Harvard University Press, Cambridge. 1950.

Dempster, A. P., Laird, N. M., and Rubin, D. B. "Maximum Likelihood from Incomplete Data via the EM Algorithm." Applied Statistics, volume 26, number 1, 1977.

Doran, H.E. "Prediction of Missing Observations in the Time Series of an Economic Variable." Journal of the American Statistical Association, volume 69, number 346, 1974.

Dunsmuir, W. "Asymptotic Theory for Time Series Containing Missing and Amplitude Modulated Observations." Sankhyã: The Indian Journal of Statistics, Series A, volume 43, number 3, 1981.

Dunsmuir, W. "Estimation for Stationary Time Series When Data are Irregularly Spaced or Missing." in Applied Time Series Analysis II. D. F. Findley (editor). Academic Press, Inc. Toronto, 1981.

Essenwanger, O. Applied Statistics in Atmospheric Science. Part A. Frequency and Curve Fitting. Elsevier Scientific Publishing Company, New York, 1976.

Essenwanger, O. M. World Survey of Climatology Volume 1B. General Climatology, 1B. Elements of Statistical Analysis. Elsevier Science Publishing Company Inc., New York. 1985.

Findlay, B. F. Climatological Interpolation / Extrapolation Techniques. Report No. 85-4. Unpublished Manuscript. Atmospheric Environment Service. Downsview. 1985.

Fleer, H.E. "Rainfall Fluctuations and Sunspot Variability." Archives for Meteorology, Geophysics, and Bioclimatology, Series B, volume 30, 1982.

Flocas, A. A., Giles, B. D., and Angouridakis, V. E. "On the Estimation of Annual and Monthly Mean Values of Air Temperature over Greece Using Stepwise Regression Analysis." Archives for Meteorology, Geophysics, and Bioclimatology, Series B, volume 32, 1983.

Friedman, Milton. The Interpolation of Time Series By Related Series. National Bureau of Economic Research. New York, 1962.

Fuenzalida, H. and Rosenbluth, B. "Distortion Effects of the Anomaly Method of Removing Seasonal or Diurnal Variations from Climatological Time Series." Journal of Climate and Applied Meteorology, volume 25, 1986.

Fuller, W. A. Introduction to Statistical Time Series. John Wiley and Sons, Inc. Toronto, 1976.

Fuller, W. A. Measurement Error Models. John Wiley and Sons, Inc. Toronto, 1987.

Gandin, L. S. Objective Analysis of Meteorological Fields. Translated from Russian. Gidrometeorologicheskoe Izdatel'stvo. Leningrad 1963.

Gottman, John M. Time Series Analysis. A Comprehensive

Introduction for Social Scientists. Cambridge University Press. New York, 1981.

Granger, C. W. J. and Newbold, P. Forecasting Economic Time Series. Academic Press. New York, 1977.

Gullett, D. W. Smoothing Daily Climatic Normals. Report No. 86-7. Unpublished Manuscript. Atmospheric Environment Service. Downsview. 1986.

Guiot, J. "The Extrapolation of Recent Climatological Series with Spectral Canonical Regression." Journal of Climatology, volume 5, number 3, 1985.

Hansen, J. E. and Driscoll, D. M. "The Numerical Simulation of Hourly Temperatures." Fourth Conference on Probability and Statistics in Atmospheric Sciences. American Meteorological Society, Boston. 1975.

Harvey, A. C. The Econometric Analysis of Time Series. Philip Allen. London. 1981.

Hopkins, T. R. "Algorithm AS 193 - A Revised Algorithm for the Spectral Test." Applied Statistics, volume 32, 1983.

Jarrett, R. G. "The Analysis of Designed Experiments with Missing Observations." Applied Statistics, volume 27, number 1, 1978.

Jenkins, G. M. "General Considerations in the Analysis of Spectra" in Time Series Analysis Papers. Emanual Parzen (editor). Holden-Day Inc. San Francisco, 1970.

Jenkins, G. M. Practical Experiences with Modelling and

Forecasting Time Series. Gwilym Jenkins and Partners (Overseas) Ltd. St. Helier, 1979.

Johnstone, Kirk. Climate Network Design: A Gandin Approach with Computer Procedures. Report No. 85-1. Unpublished Manuscript. Atmospheric Environment Service. Downsview. 1985.

Jones, R. H. "Maximum likelihood Fitting of ARMA Models to Time Series With Missing Observations." Technometrics, volume 22, number 3, 1980.

Jones, R. H. "Fitting a Continuous Time Series to Discrete Data." in Applied Time Series Analysis II. D. F. Findley (editor). Academic Press, Inc. Toronto, 1981.

Jones, R. H. "Time Series Analysis - Time Domain." in Probability, Statistics, and Decision Making in the Atmospheric Sciences. Murphy, A. H. and Katz, R. W. (editors). Westview Press, Inc., Boulder and London. 1985.

Kemp, W. P., Burnell, D. G., Everson, D. O., Thompson, A. J. "Estimating Missing Daily Maximum and Minimum Temperatures." Journal of Climate and Applied Meteorology, volume 22, 1983.

Lee, A. C. L. "Smoothing and filtering of meteorological data." The Meteorological Magazine, volume 110, 1981.

Lawson, M. P., and Cerveny, R. S. "Seasonal Temperature Forecasts as Products of Antecedent Linear and Spatial Temperature Arrays." Journal of Climate and Applied

Meteorology, volume 24, 1985.

Laughlin, G.P. "Minimum Temperature and Lapse Rate in Complex Terrain: Influencing Factors and Prediction." Archives for Meteorology, Geophysics, and Bioclimatology, Series B, volume 30, 1982.

Lutkepohl, Helmut. " The Impact of Omitted Variables on the Structure of Multiple Time Series: Quenouille's Data Revisited" in Time Series Analysis: Theory and Practice 2. O.D. Anderson (editor). North-Holland Publishing Company. New York, 1982.

Makridakis, S. and Wheelwright, S. Interactive Forecasting. Holden-Day, Inc. San Francisco, 1978.

McCutchan, M. H. "A Model for Diagnosing and Predicting Surface Temperature." Fourth Conference on Probability and Statistics in Atmospheric Sciences. American Meteorological Society, Boston. 1975.
pp. 25 - 30.

Mélard, G. "Algorithm AS 197 - A Fast Algorithm for the Exact Likelihood of Autoregressive-Moving Average Models." Applied Statistics, volume 33, 1985.

Murphy, A. H. and Katz, R. W. (editors) Probability, Statistics, and Decision Making in the Atmospheric Sciences. Westview Press, Inc., Boulder and London. 1985.

Myers, R. M. Classical and Modern Regression with Applications. Duxbury Press. Boston, 1986.

Neave, H. R. "Spectral Analysis of a Stationary Time

Series Using Initially Scarce Data." Biometrika, volume 57, number 1, 1970.

Nerlove, M., Grether, D. M., and Carvalho, J. L. Analysis of Economic Time Series. A Synthesis. Academic Press. New York, 1979.

Nicholls, N. "The potential for long-range prediction of seasonal mean temperature in Australia." Australian Meteorological Magazine, volume 31, 1983.

Panofsky, H. A. and Brier, G. W. Some Applications of Statistics to Meteorology. The Pennsylvania State University, University Park, Pennsylvania. 1963.

Parzen, Emanuel. "An Approach to Empirical Time Series Analysis" in Time Series Analysis Papers. Emanual Parzen (editor). Holden-Day, Inc. San Francisco, 1970a.

Parzen, Emanuel. "On Spectral Analysis With Missing Observations and Amplitude Modulation" in Time Series Analysis Papers. Emanual Parzen (editor). Holden-Day, Inc. San Francisco, 1970b.

Quenouille, M. H. The Analysis of Multiple Time Series. Hafner Publishing Company. New York, 1968.

Reuter, H. "Zur Frage der Erfassung des Temperaturtagesganges durch eine diskrete Anzahl von Beobachtangen." Archives for Meteorology, Geophysics, and Bioclimatology, Series B, volume 28, 1980.

Roodenburg, J. "Forecasting urban temperatures from rural observations." The Meteorological Magazine,

volume 112, 1983.

SAS Institute Inc., SAS User's Guide: Basics Version 5 Edition. Cary, NC: SAS Institute Inc., 1985.

SAS Institute Inc., SAS/ETS User's Guide, Version 5 Edition. Cary, NC: SAS Institute Inc., 1984.

Saucier, W. J. Principles of Meteorological Analysis. The University of Chicago Press, Chicago and London. 1965.

Scheaffer, R. L., Mendenhall, W., Ott, L. Elementary Survey Sampling. Third Edition. Duxbury Press. Boston, 1986.

Schickedanz, P. T. and Bowen, E. G. "Computation of Climatological Power Spectra Using Variable Record Lengths." Fourth Conference on Probability and Statistics in Atmospheric Sciences. American Meteorological Society, Boston. 1975.

Schönwiese, C. D. "Central England Temperature and Sunspot Variability 1660-1975." Archives for Meteorology, Geophysics, and Bioclimatology, Series B, volume 26, 1977.

Schönwiese, C. D. "On the Problem of Statistical Climate Modelling." Archives for Meteorology, Geophysics, and Bioclimatology, Series B, volume 26, 1978.

Seaman, R. S. "Distance-time autocorrelation functions for winds in the Australian Region." Australian Meteorological Magazine, volume 23, 1975.

Seaman, R. S. "Objective analysis accuracies of statistical interpolation and successive correction schemes." Australian Meteorological Magazine, volume 31, 1983.

Seaman, R. S. "Generation of synthetic meteorological data sets by linear filtering of white noise." Australian Meteorological Magazine, volume 34, 1986.

Seaman, R. S. and Hutchinson, M. F. "Comparative real data tests of some objective analysis methods by withholding observations" Australian Meteorological Magazine, volume 33, 1985.

Shugart, H. H. (editor). Time Series and Ecological Processes. Society for Industrial and Applied Mathematics, Philadelphia. 1978.

Smith, P. "The Use of Analysis of Covariance to Analyze Data from Designed Experiments with Missing or Mixed-up Values." Applied Statistics, volume 30, number 1, 1981.

Smith, W. P. "Reconstruction of Precipitation in North Eastern Nevada Using Tree Rings, 1600 - 1982." Journal of Climate and Applied Meteorology, volume 25, 1986.

Snijders, T. A. B. "Interstation Correlations and Nonstationarity of Burkina Faso Rainfall." Journal of Climate and Applied Meteorology, volume 25, 1986.

Steinberg, L. Arima Forecast Methods. Report No. 84-1. Unpublished Manuscript. Atmospheric Environment Service. Downsview. 1984.

Stol, P. T. Conceptual rainfall interstation

correlation functions for time-integrated rainfall depths. IAHS Publication number 158. Institute for Land and Water Management Research, Wageningen, Netherlands. 1986.

Stringer, E. T. Techniques of Climatology. W. H. Freeman and Company, San Francisco. 1972.

Tabony, R. C. "The Estimation of Missing Climatological Data." Journal of Climatology, volume 3, number 3, 1983.

Tabony, R. C. "A comparison of the principal component and near neighbour methods for the areal quality control of minimum temperature and sunshine duration." The Meteorological Magazine, volume 115, 1986.

Thiébaux, H. J. and Pedder, M. A. Spatial Objective Analysis: with applications in atmospheric science. Academic Press. Toronto, 1987.

Van Loon, H., and Labitzke, K. "When the wind blows", New Scientist, volume 8, 1988.

Wigley, T. M. L., Briffa, K. R., and Jones, P. D. "On the Average Value of Correlated Time Series, with Applications in Dendroclimatology and Hydrometeorology", Journal of Climatology and Applied Meteorology, volume 23, number 1, 1984.

Woodcock, F., and Southern, B. "The use of linear regression to improve official temperature forecasts." Australian Meteorological Magazine, volume 31, 1983.

Young, Peter. Recursive Estimation and Time-Series

Analysis An Introduction. Springer-Verlag, Heidelberg. 1984.

Zerefos, C. S. "Long Term Stratospheric Temperature Fluctuations and Solar Activity." Archives for Meteorology, Geophysics, and Bioclimatology, Series B, volume 32, 1983.

Appendix A.

Method Estimation Distribution Histograms

The following 28 figures are graphical representations of the data found in tables 5, 6, 7, and 8. They are ordered by data series and method such that figures 1 through 7 represent the results from the methods applied to the artificial data series, 8 through 14 represent the results from the methods applied to the Driftwood Creek data series, etc.

Each figure contains the result for a particular method applied to a given data set and shows the results of the application of that method with ten, twenty, and thirty percent of the data missing. The histograms are grouped by percent missing data and coded to correspond to the range classifications used in the tables listed above (Table 10).

Table 10.

Histogram Coding and Ranges Covered

Histogram Code	Range Covered
0.5	±0.5
1.0	±1.0
1.5	±1.5
2.0	±2.0
2.5	±2.5
3.0	±3.0
3.5	±3.5
4.0	>±3.5°C and unrecoverable

Figure 1.
Artificial Data: Ratio Method Prediction Distribution

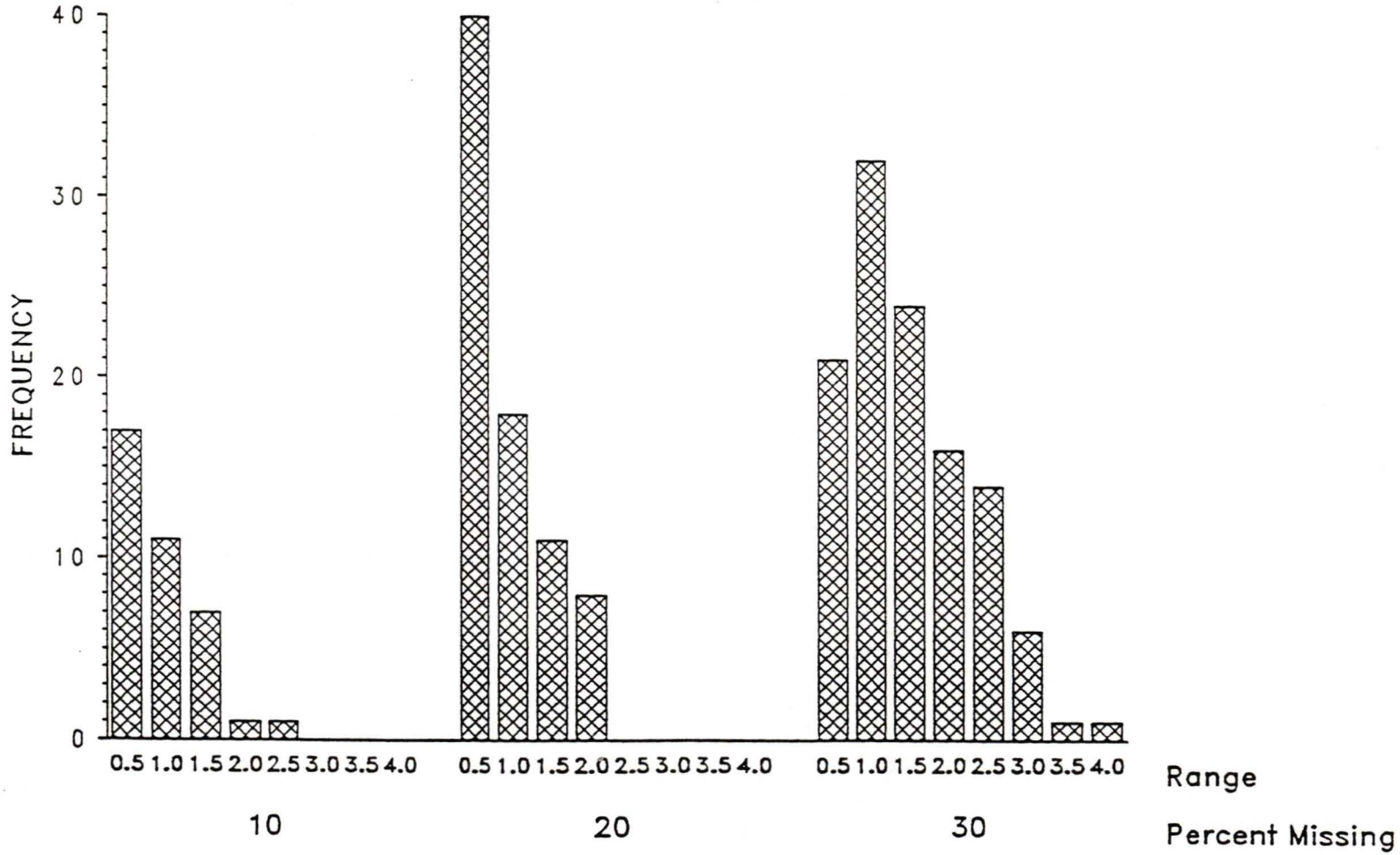


Figure 2.
Artificial Data: Difference Method Prediction Distribution

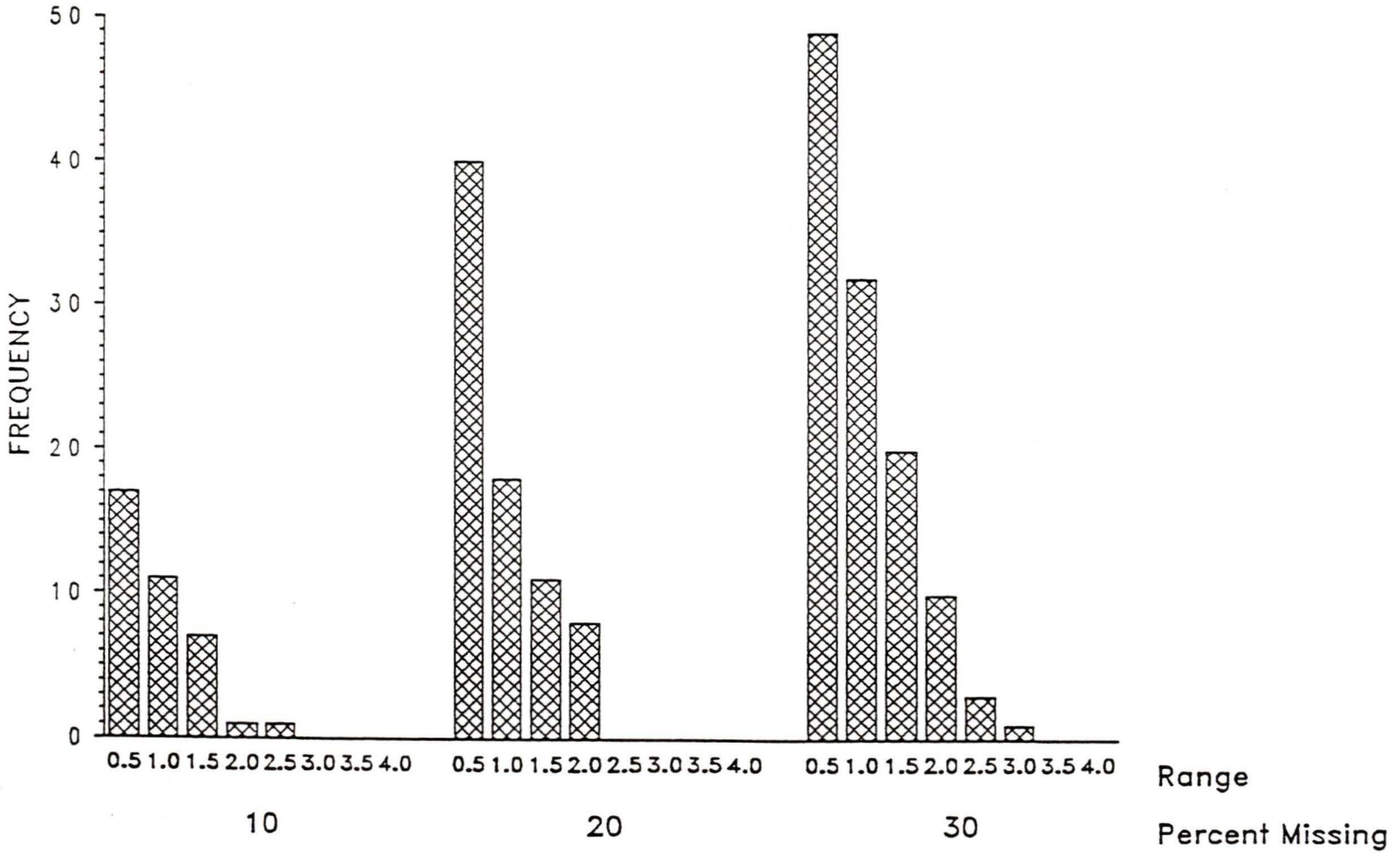


Figure 3.
Artificial Data: Regression Model Prediction Distribution

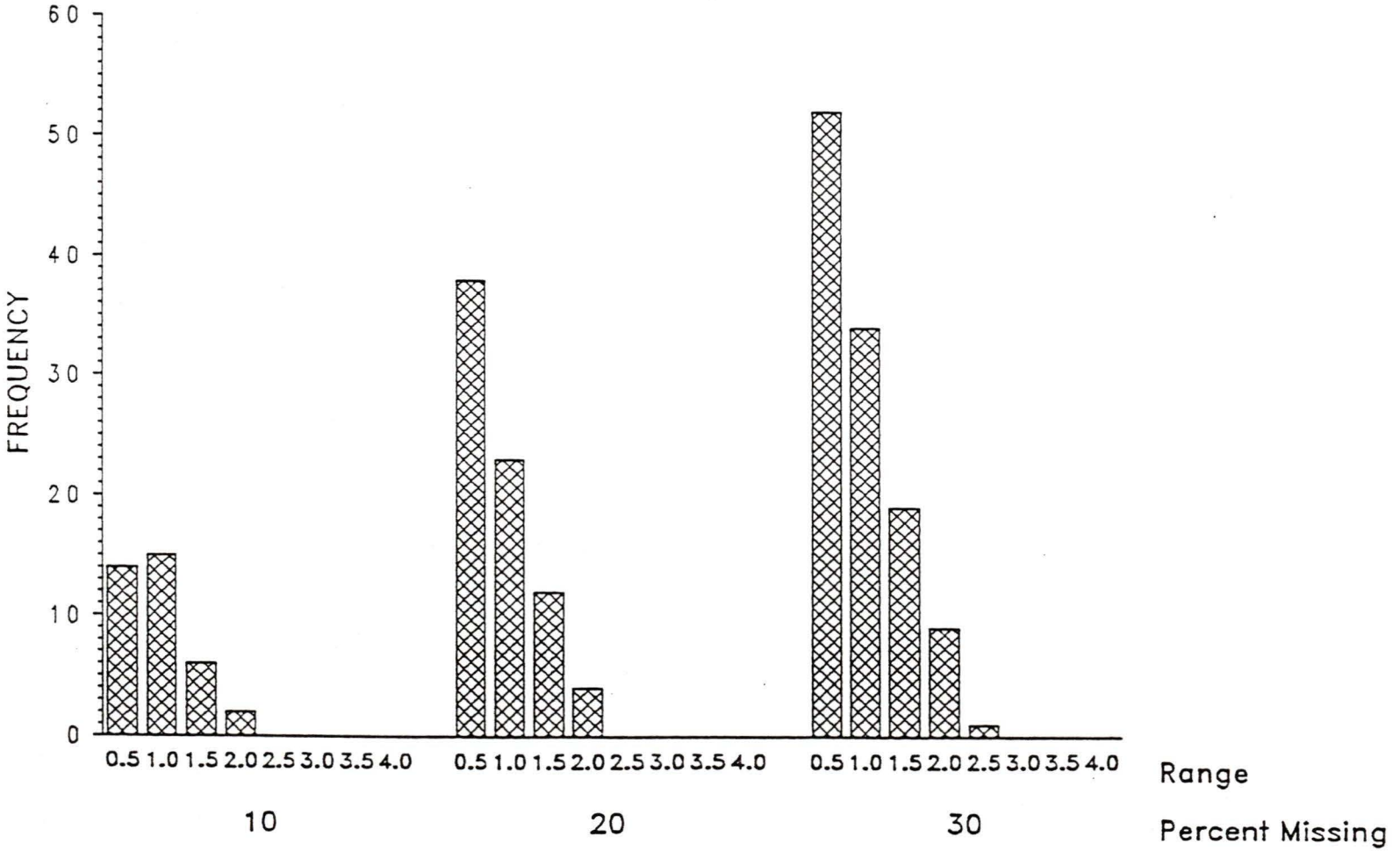


Figure 4.
Artificial Data: Polynomial Model Prediction Distribution

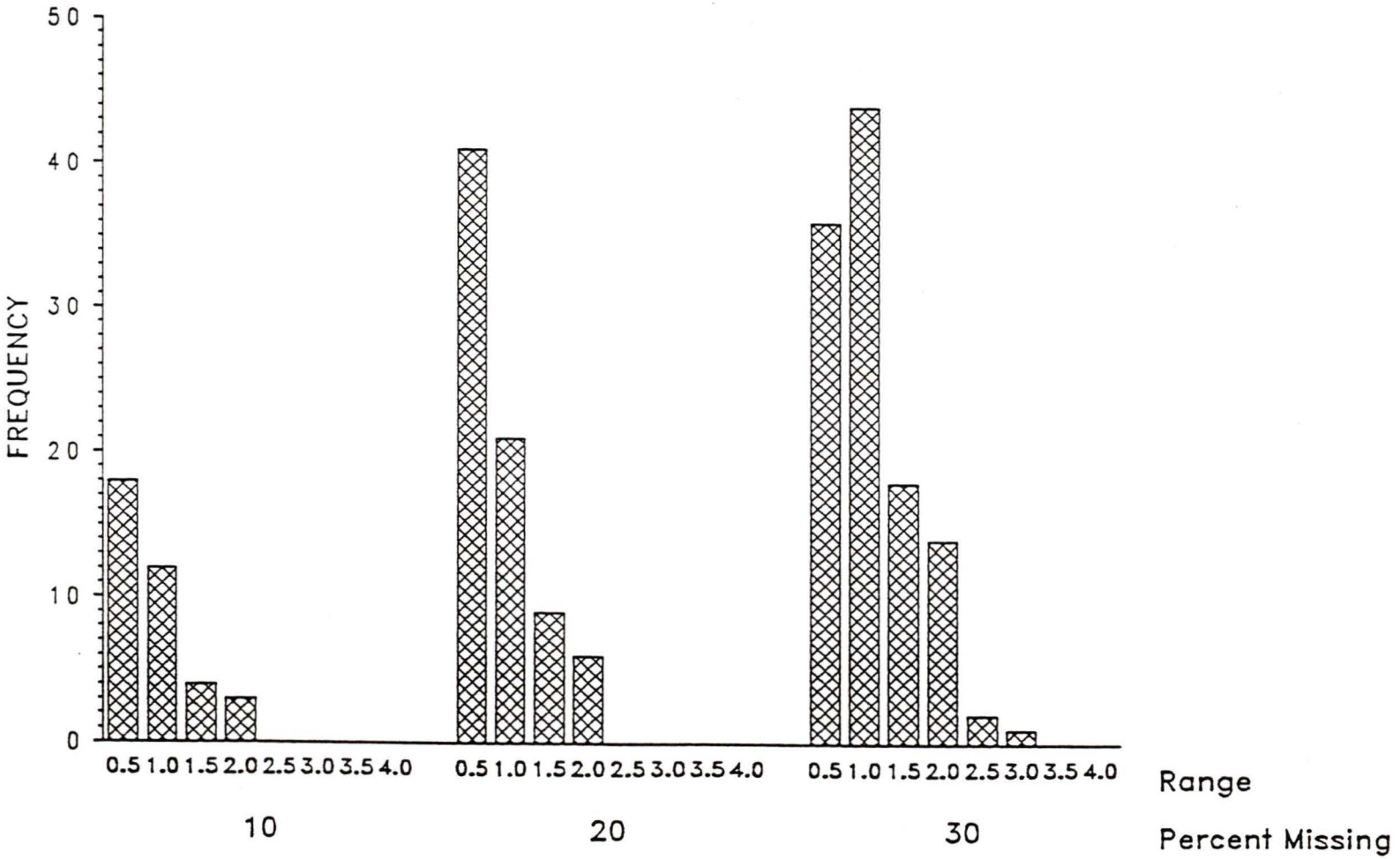


Figure 5.
Artificial Data: ARIMA Model Prediction Distribution

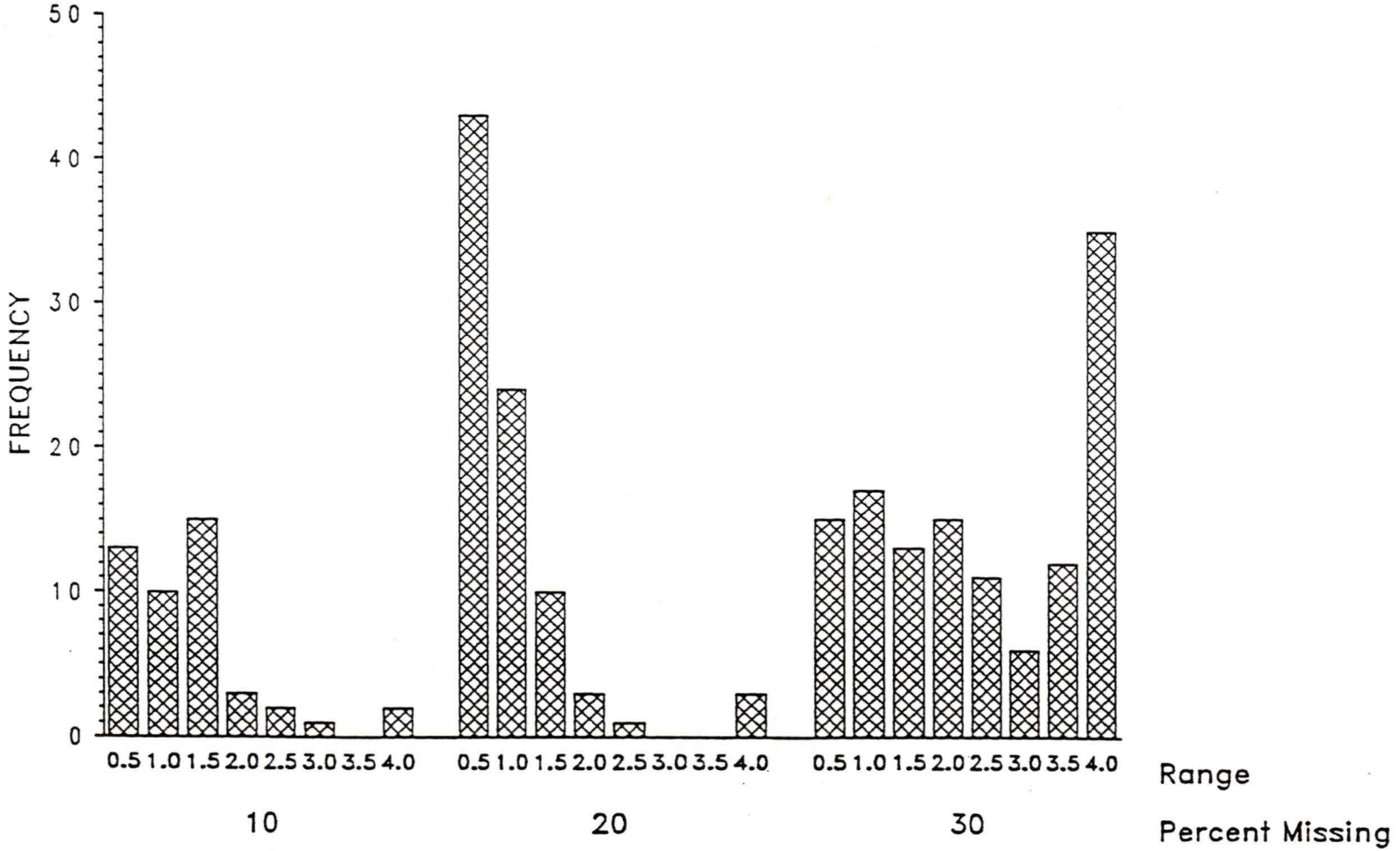


Figure 6.
Artificial Data: Lagged Dependent Regressor Prediction Distribution

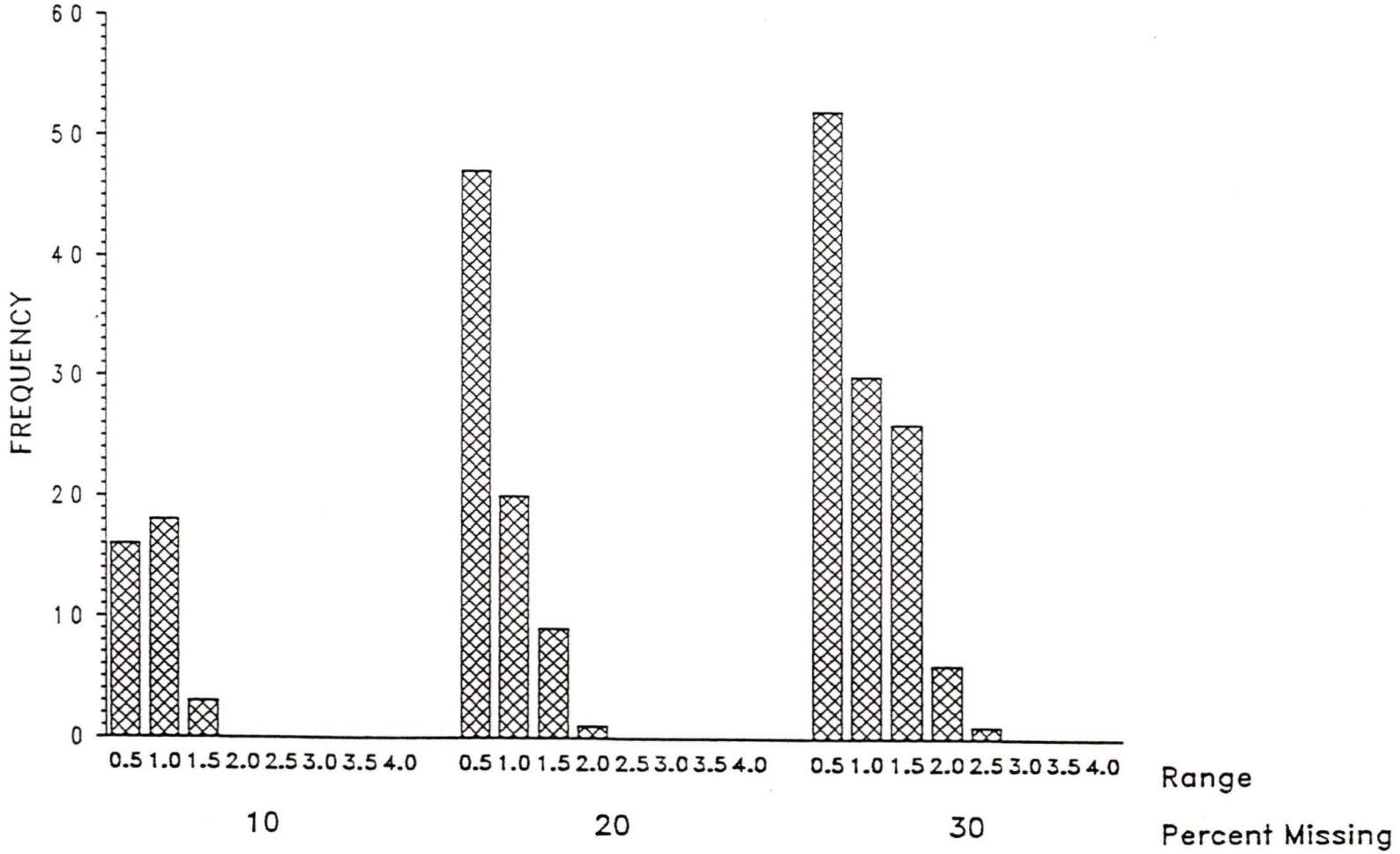


Figure 7.
Artificial Data: Modified Kalman Filter Prediction Distribution

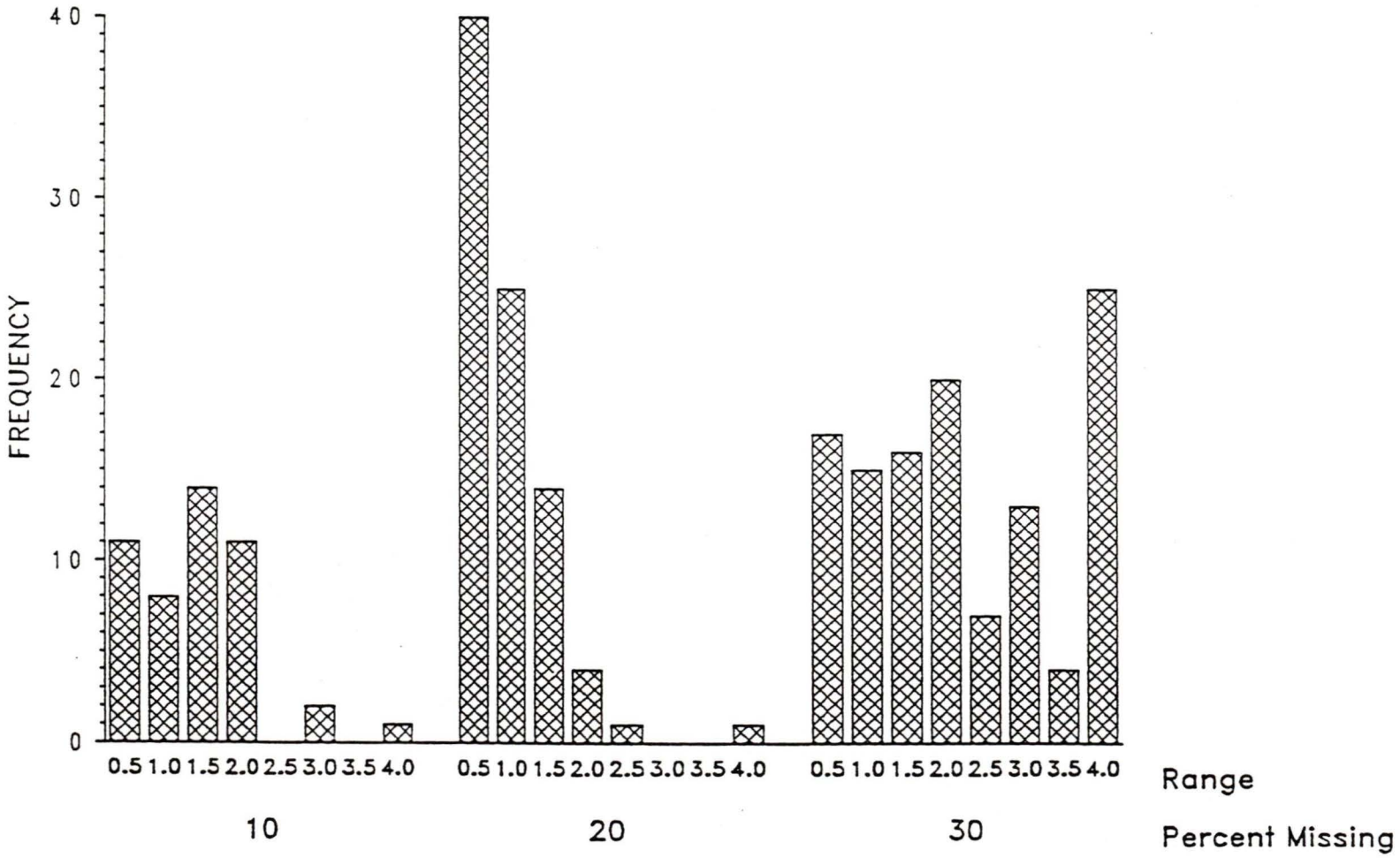


Figure 8.
Driftwood Creek Data: Ratio Method Prediction Distribution

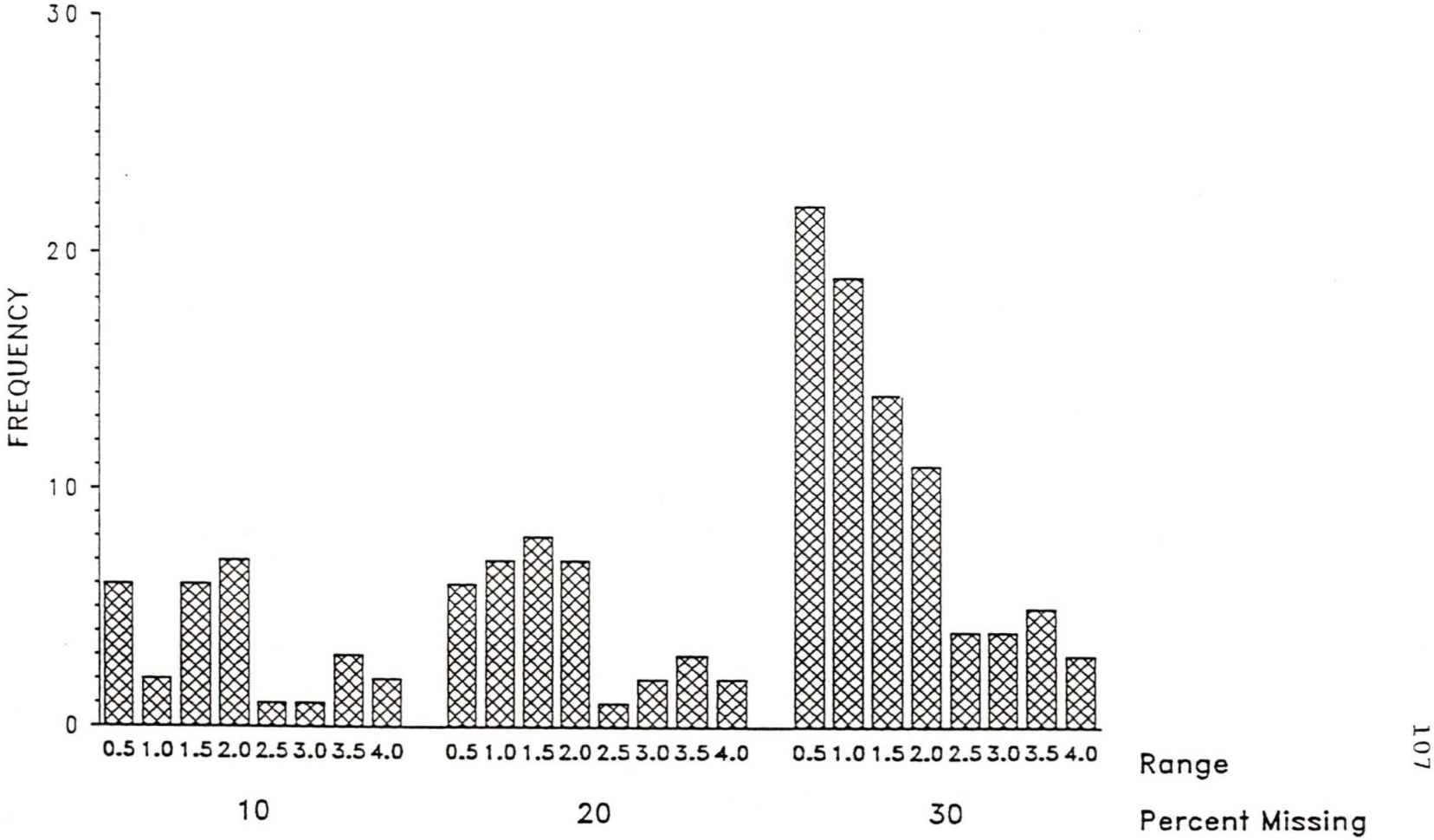


Figure 9.
Driftwood Creek Data: Difference Method Prediction Distribution

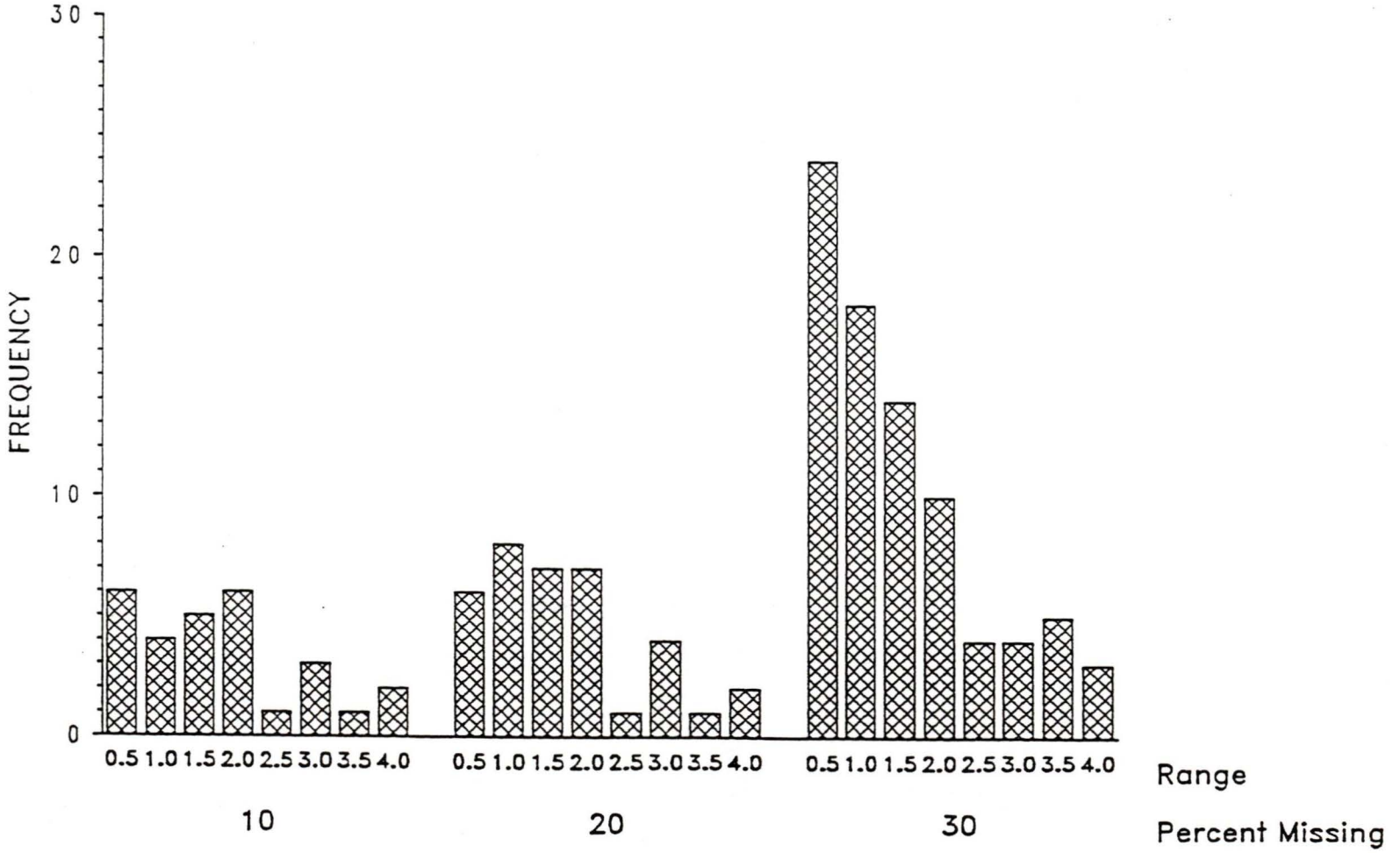


Figure 10.
Driftwood Creek Data: Regression Model Prediction Distribution

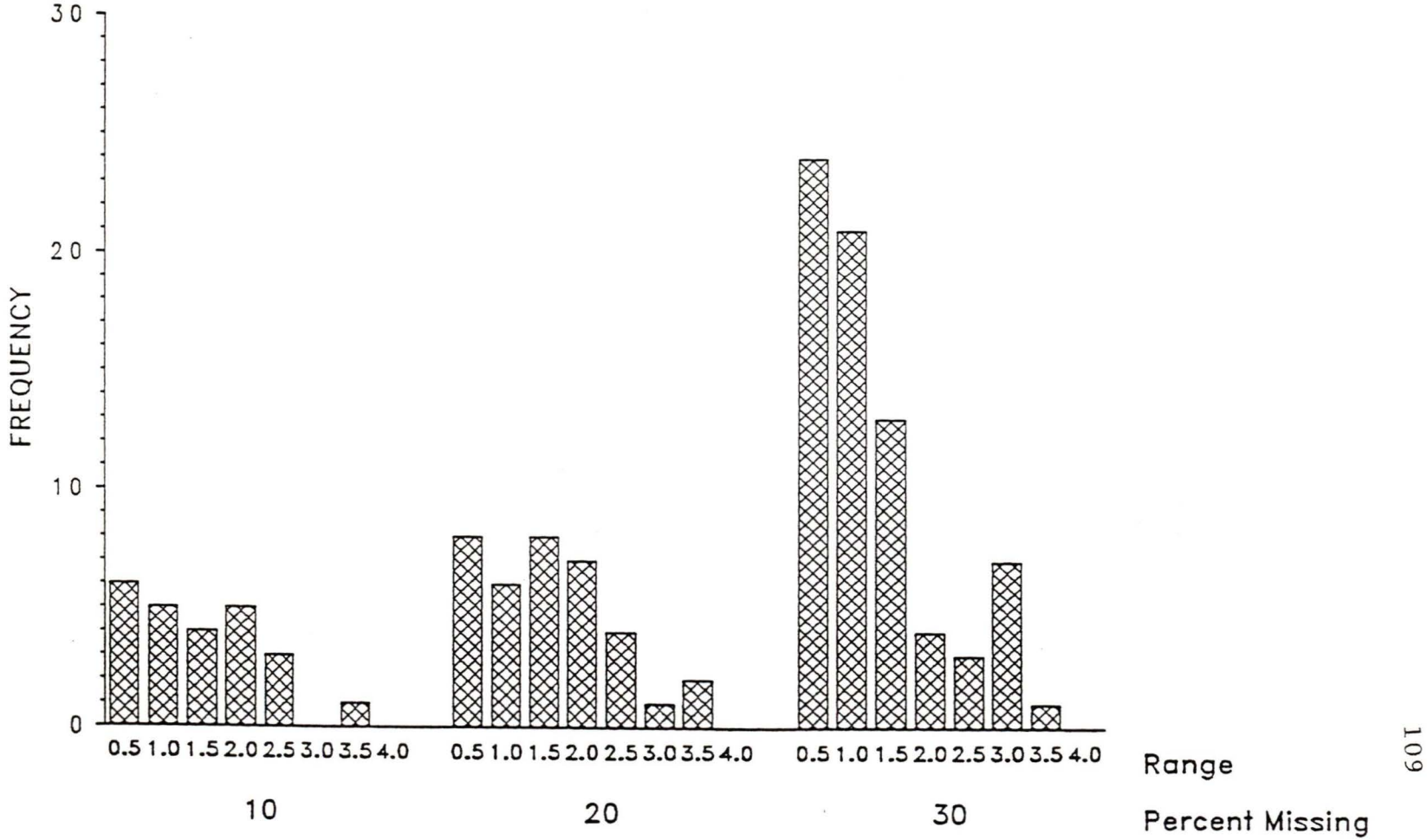


Figure 11.
Driftwood Creek Data: Polynomial Model Prediction Distribution

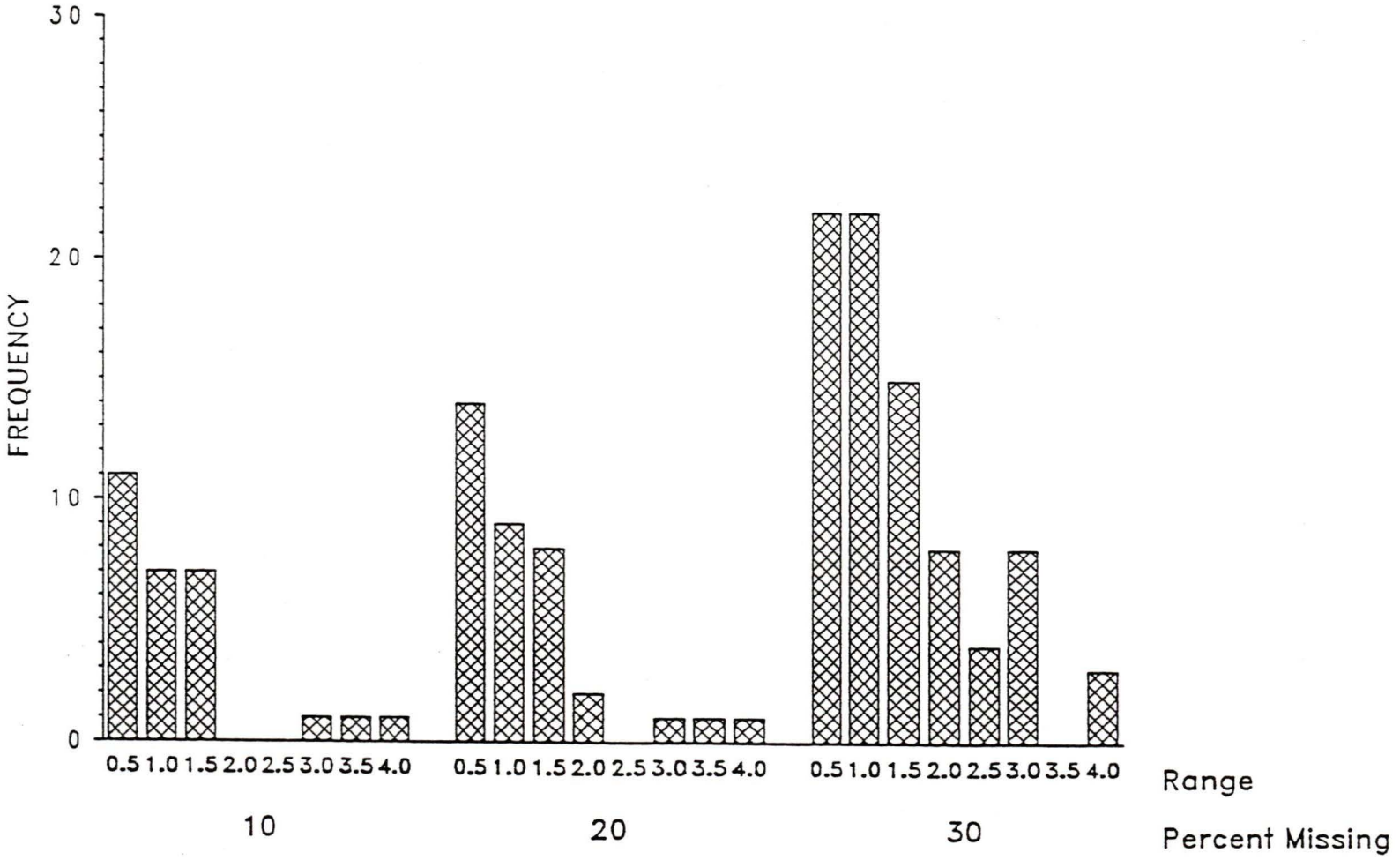


Figure 12.
Driftwood Creek Data: ARIMA Model Prediction Distribution

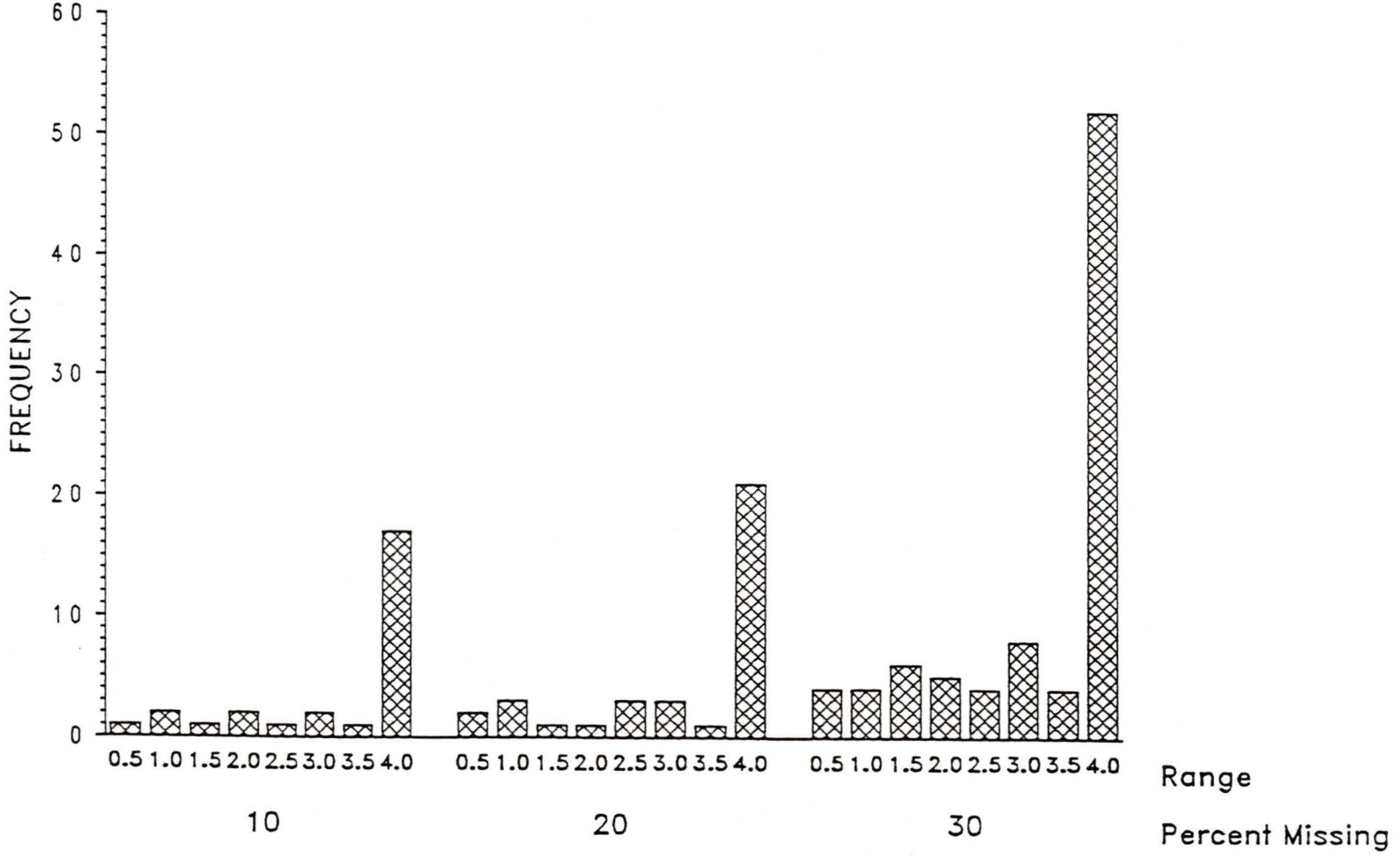


Figure 13.
Driftwood Creek Data: Lagged Dependent Regressor Prediction Distribution

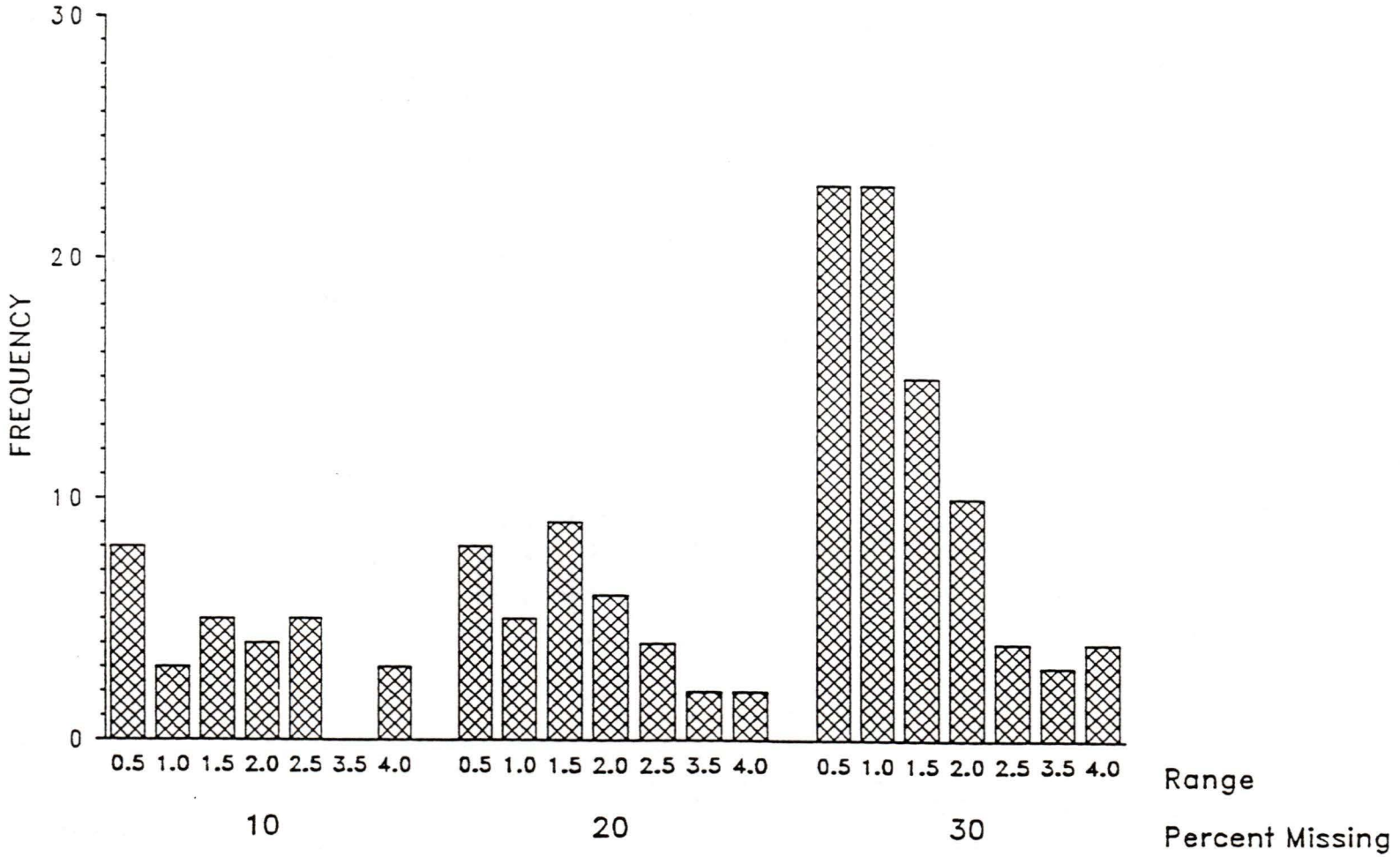


Figure 14.
Driftwood Creek Data: Modified Kalman Filter Prediction Distribution

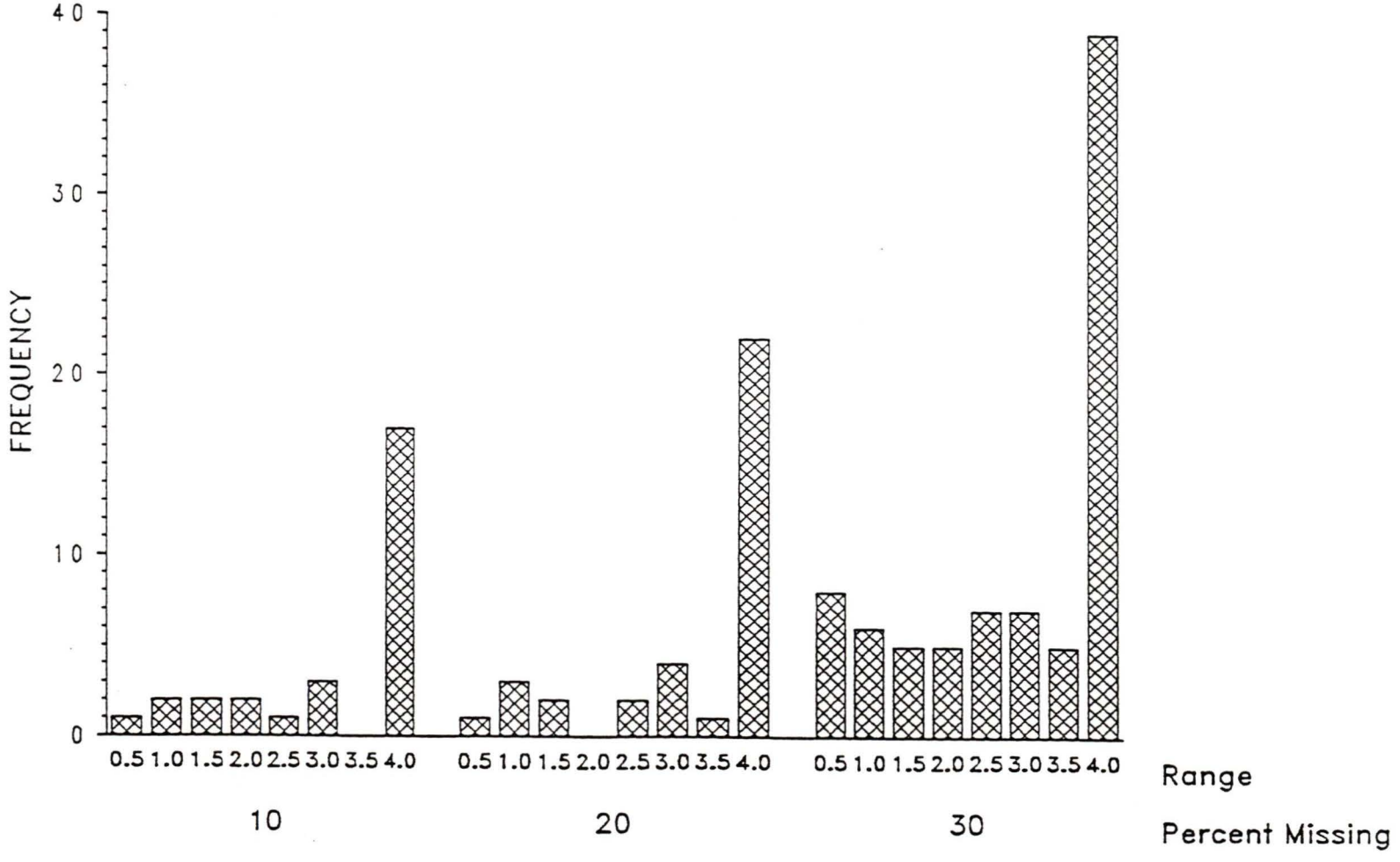


Figure 15.
Green Mountain Data: Ratio Method Prediction Distribution

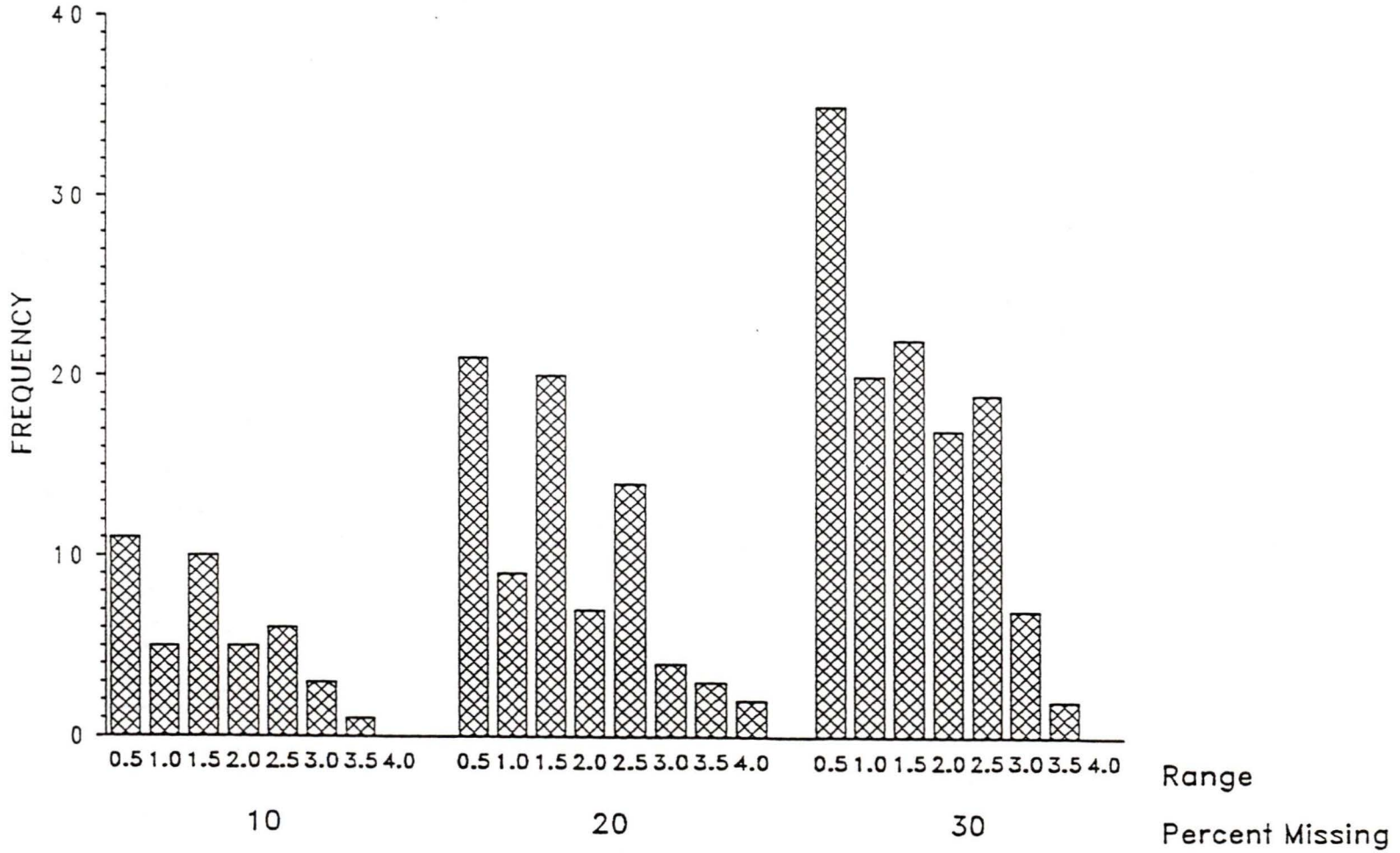


Figure 16.
Green Mountain Data: Difference Method Prediction Distribution

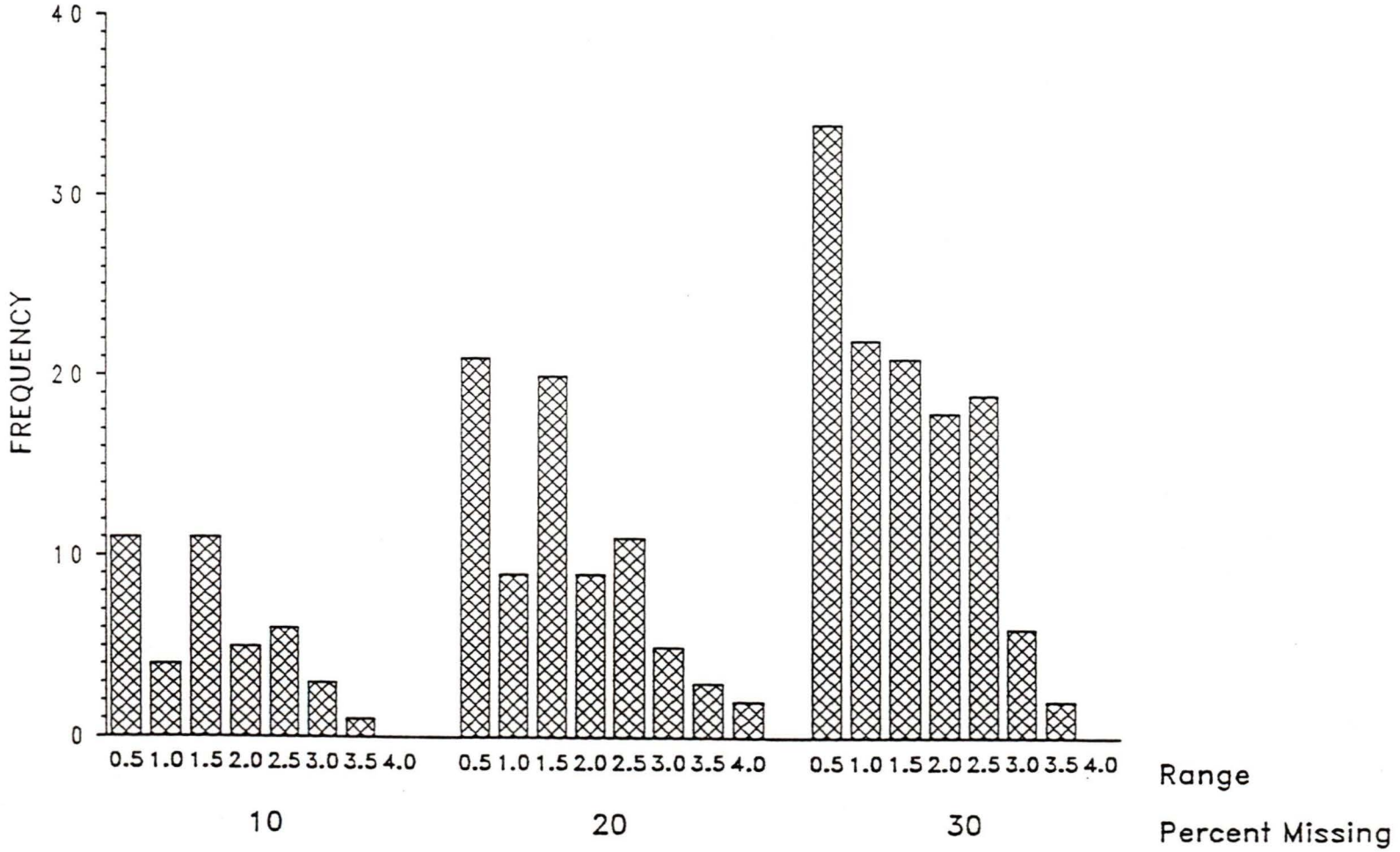


Figure 17.
Green Mountain Data: Regression Model Prediction Distribution

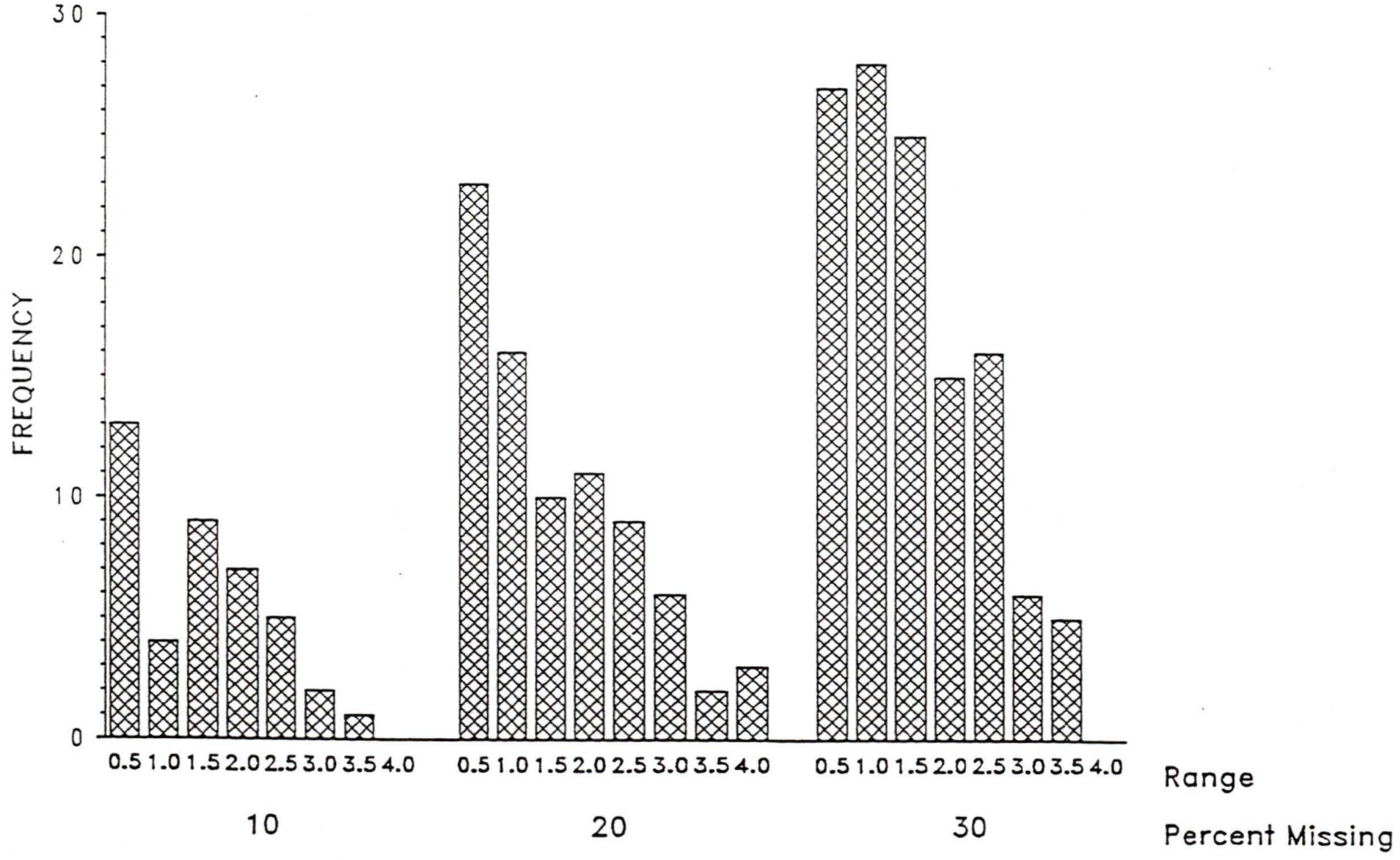


Figure 18.
Green Mountain Data: Polynomial Model Prediction Distribution

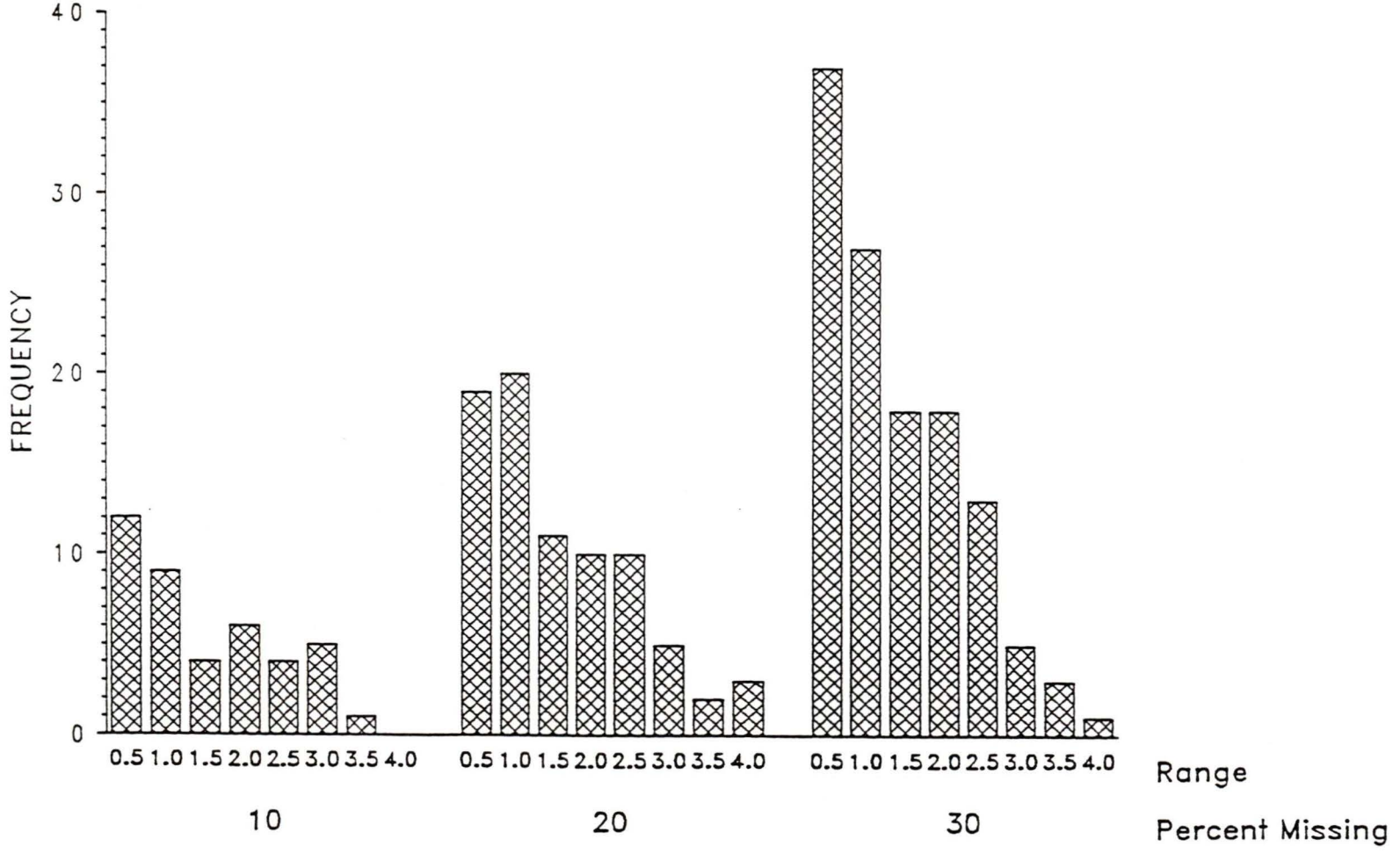


Figure 19.
Green Mountain Data: ARIMA Model Prediction Distribution

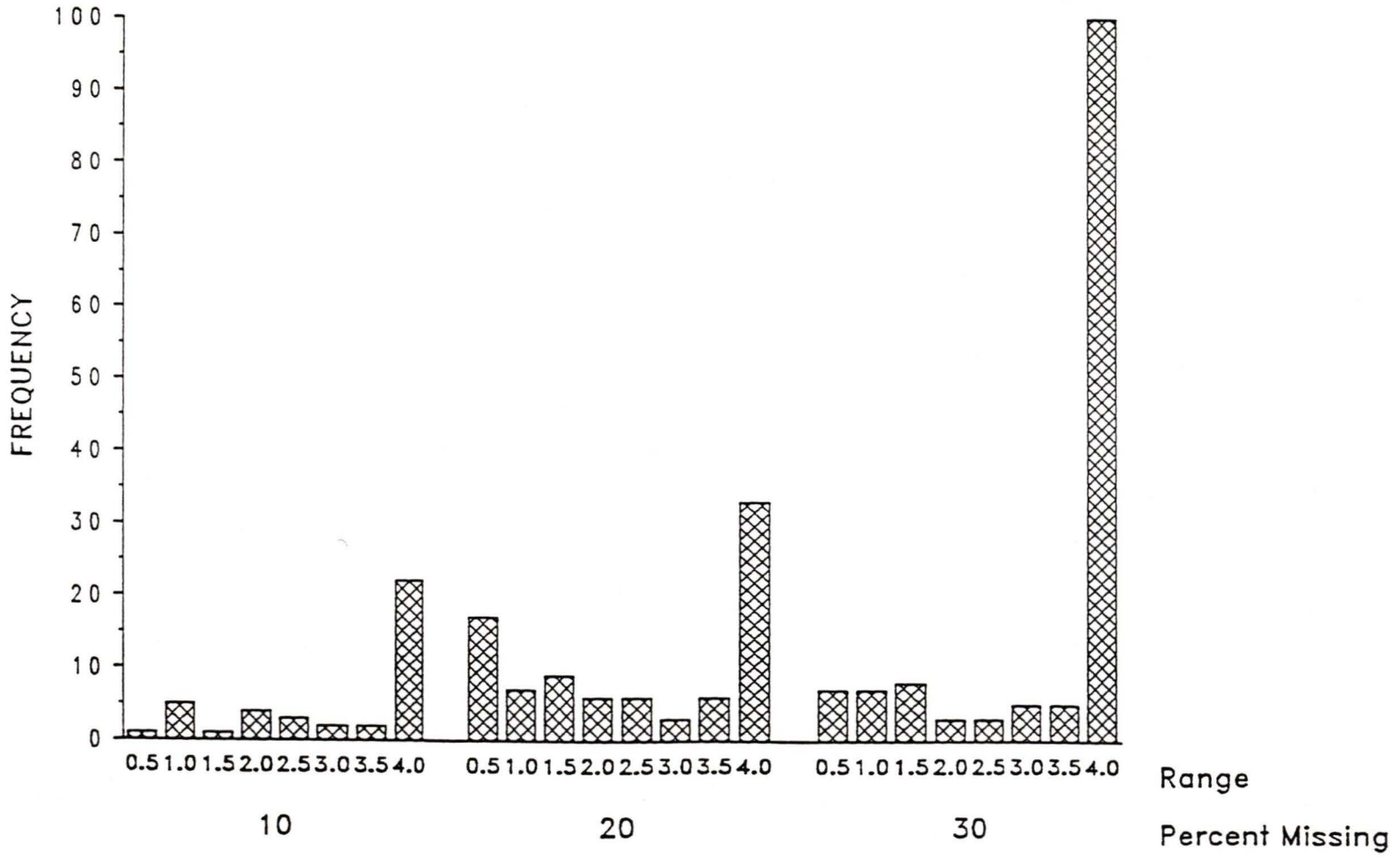


Figure 20.
Green Mountain Data: Lagged Dependent Regressor Prediction Distribution

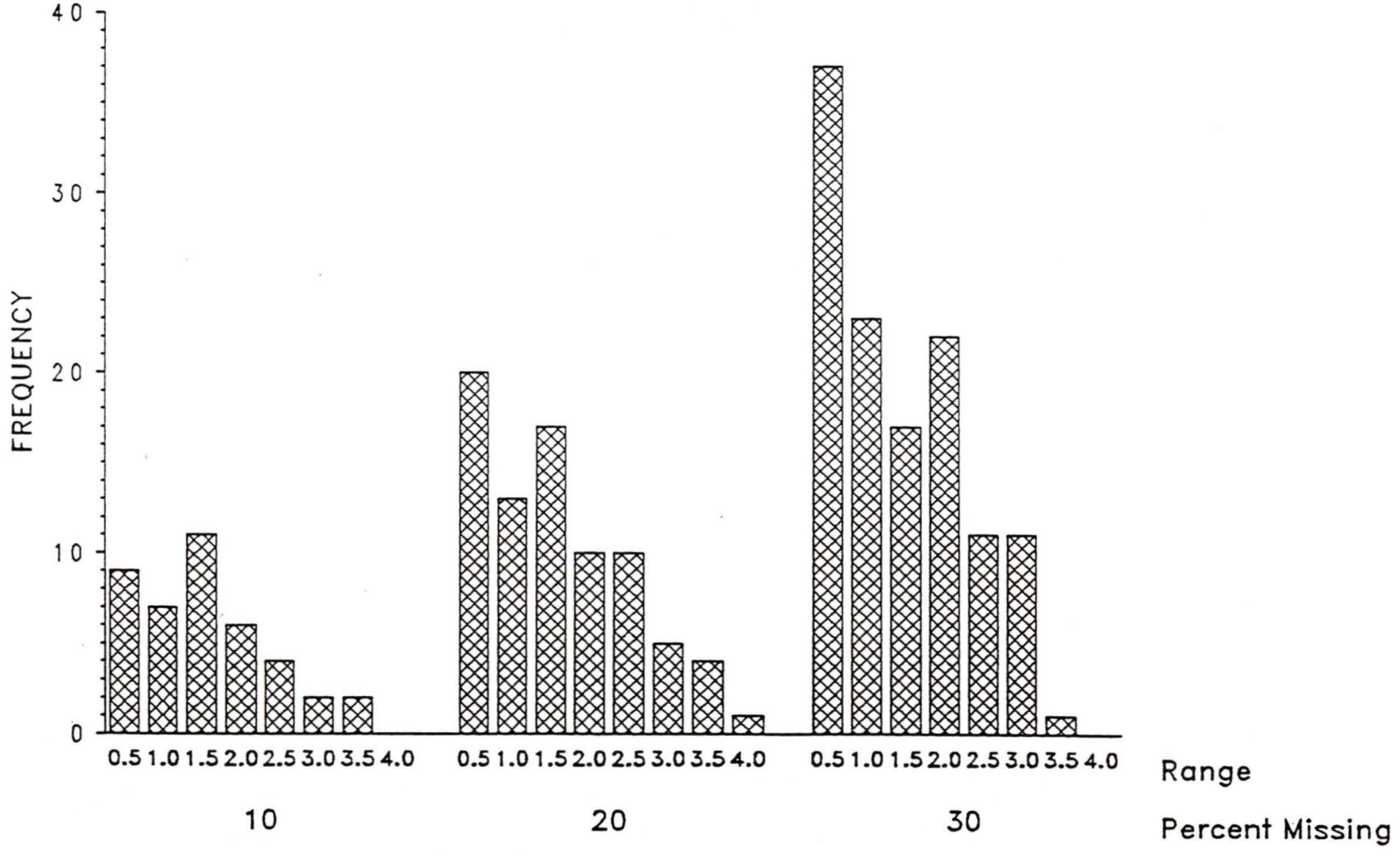


Figure 21.
Green Mountain Data: Modified Kalman Filter Prediction Distribution

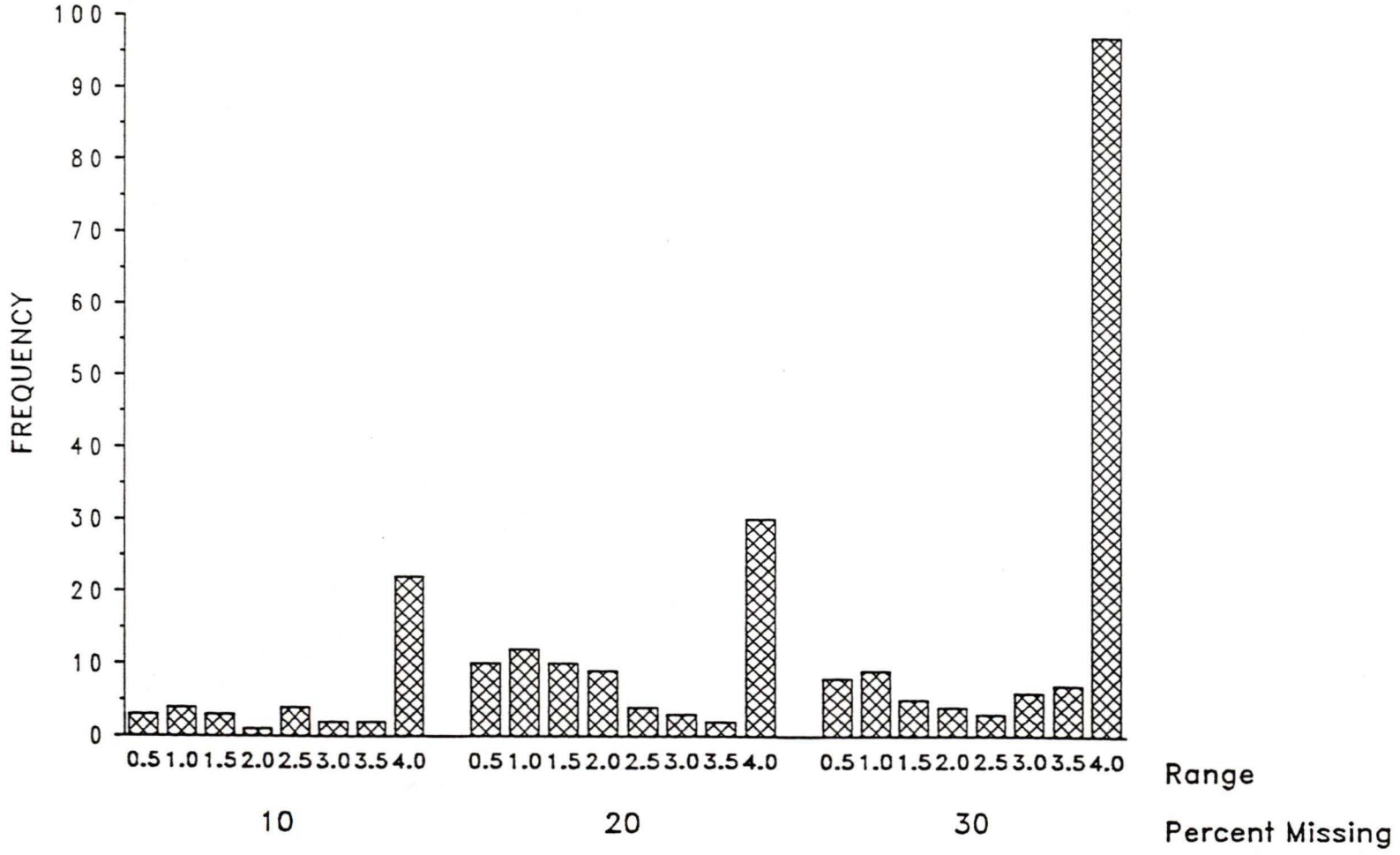


Figure 22.
Stony Lake Data: Ratio Method Prediction Distribution

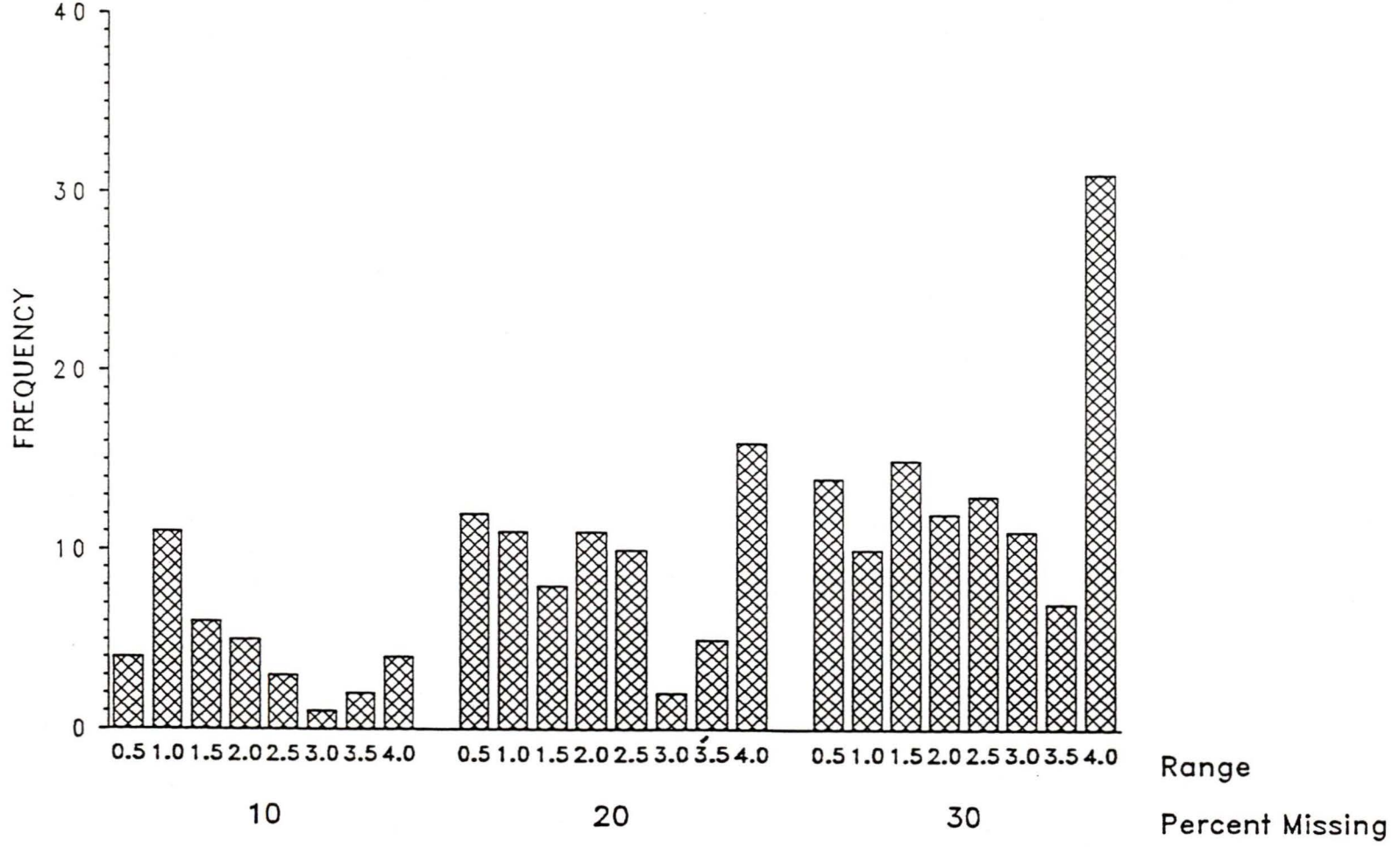


Figure 23.
Stony Lake Data: Difference Method Prediction Distribution

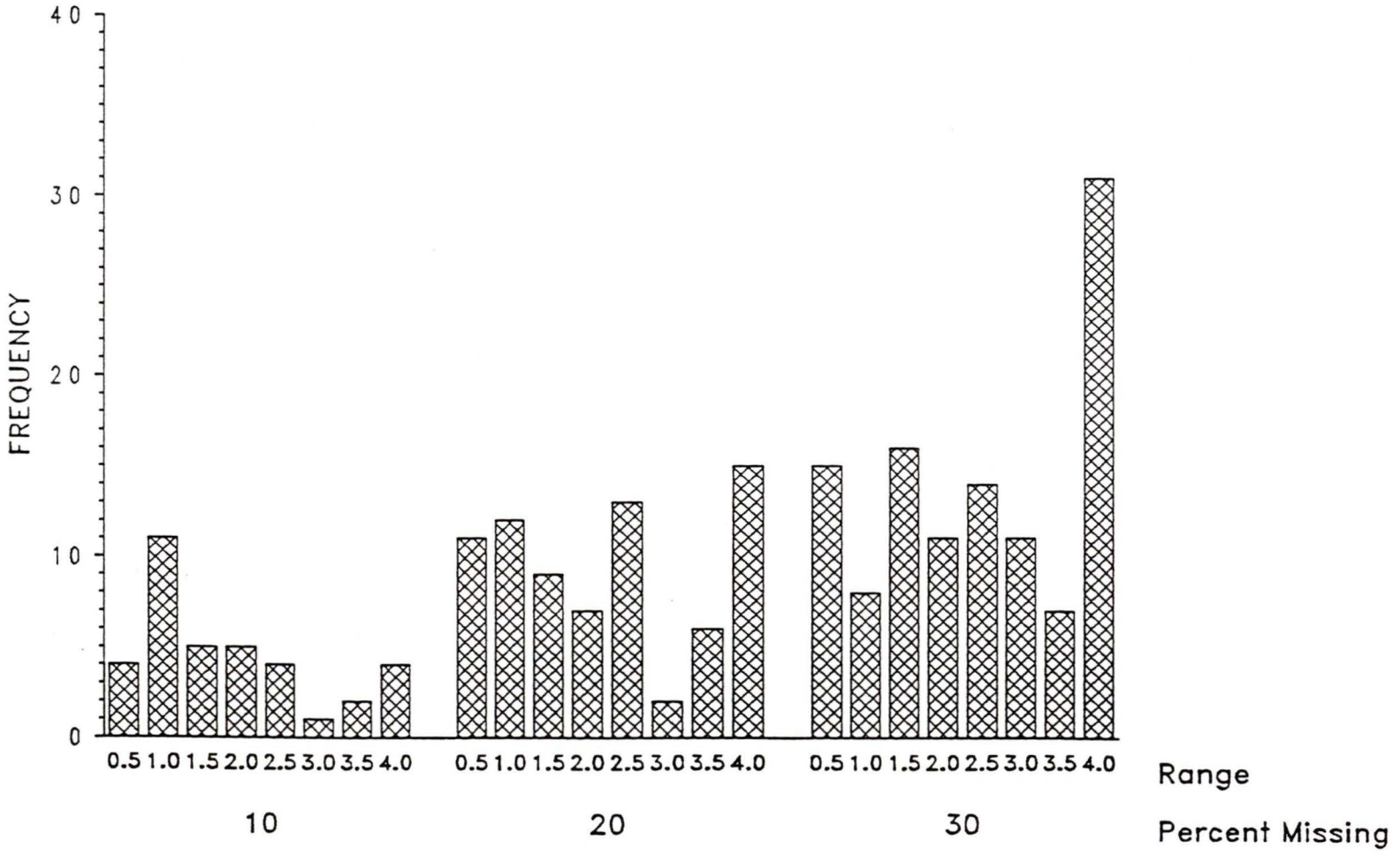


Figure 24.
Stony Lake Data: Regression Model Prediction Distribution

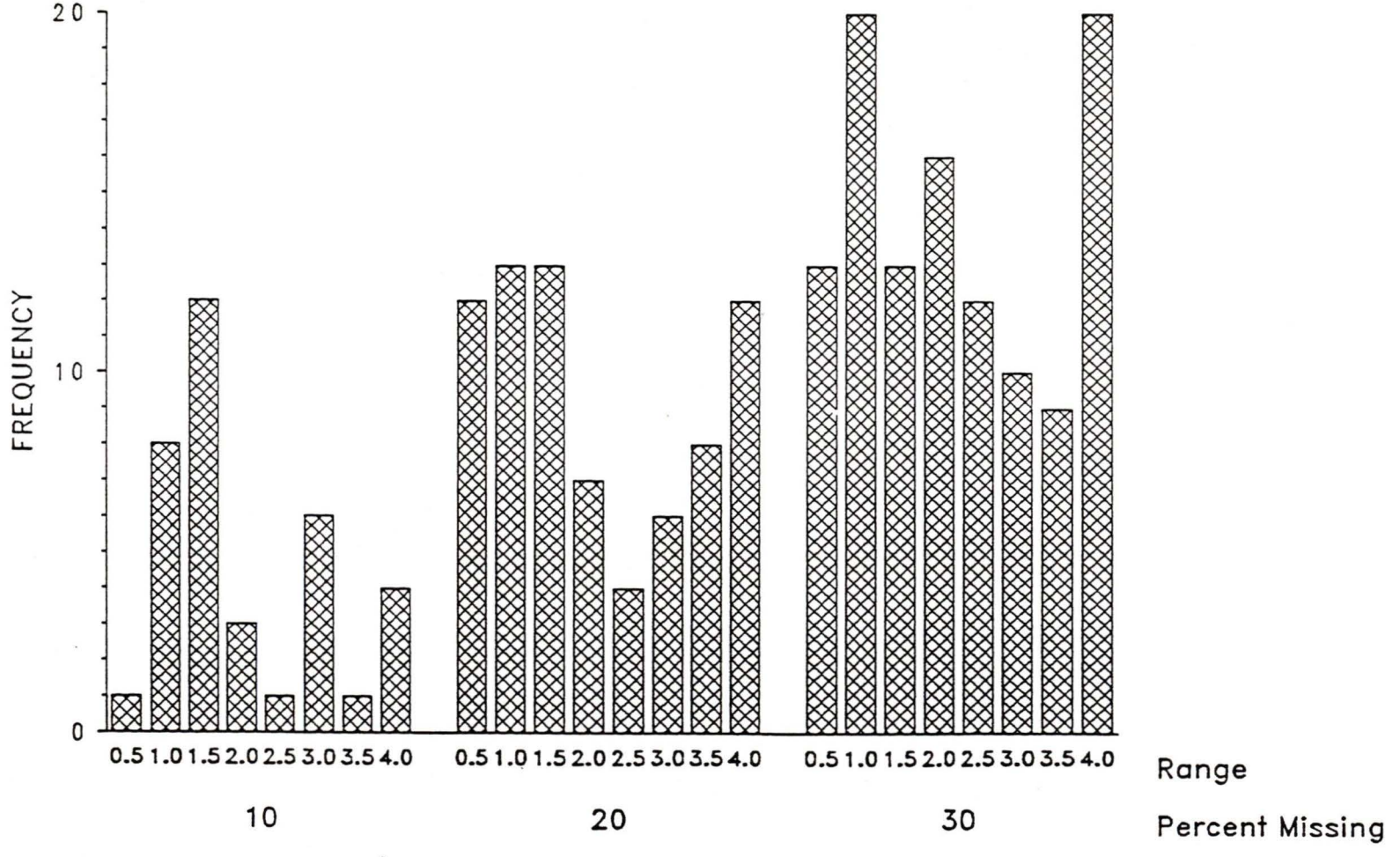


Figure 25.
Stony Lake Data: Polynomial Model Prediction Distribution

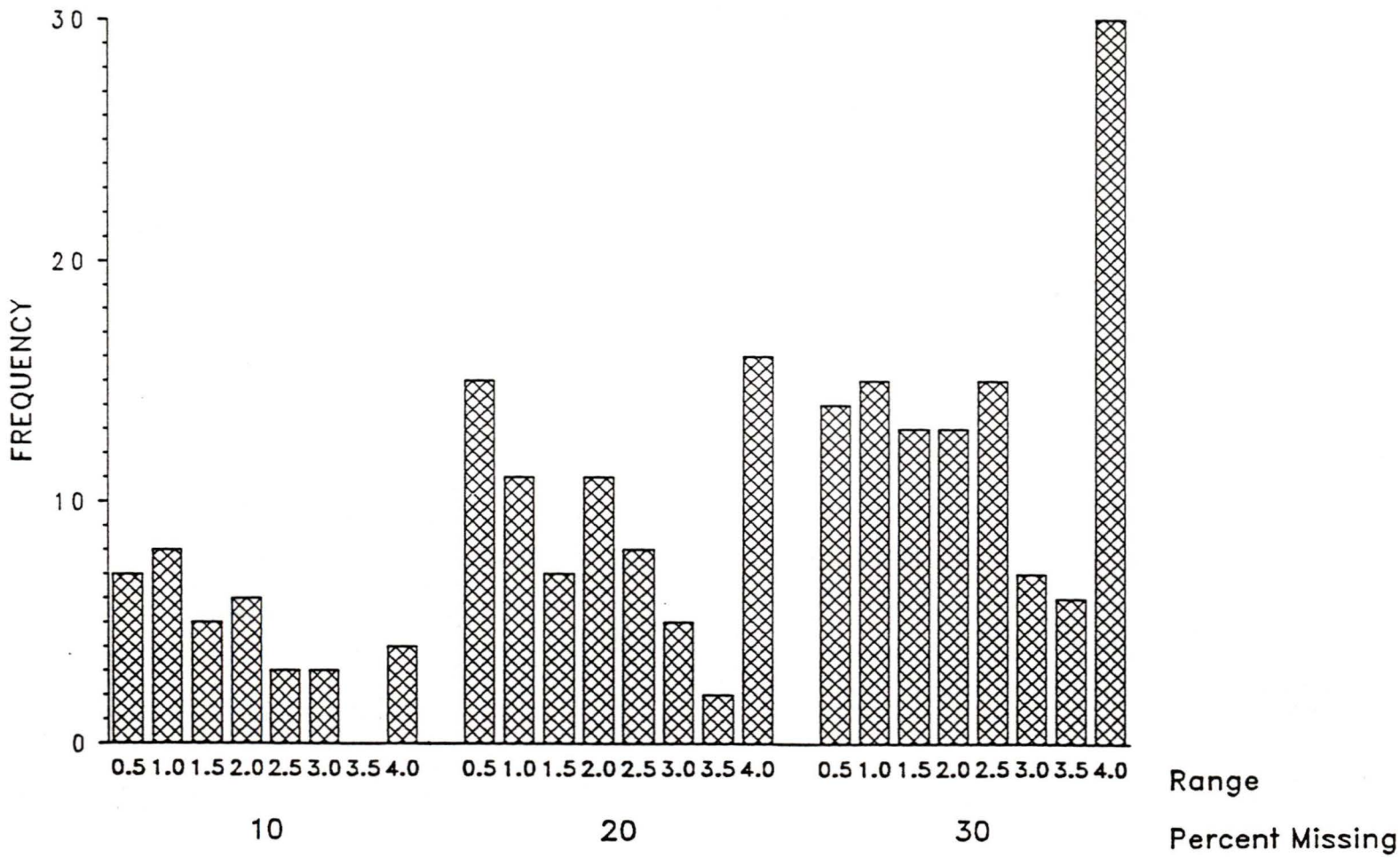


Figure 26.
Stony Lake Data: ARIMA Model Prediction Distribution

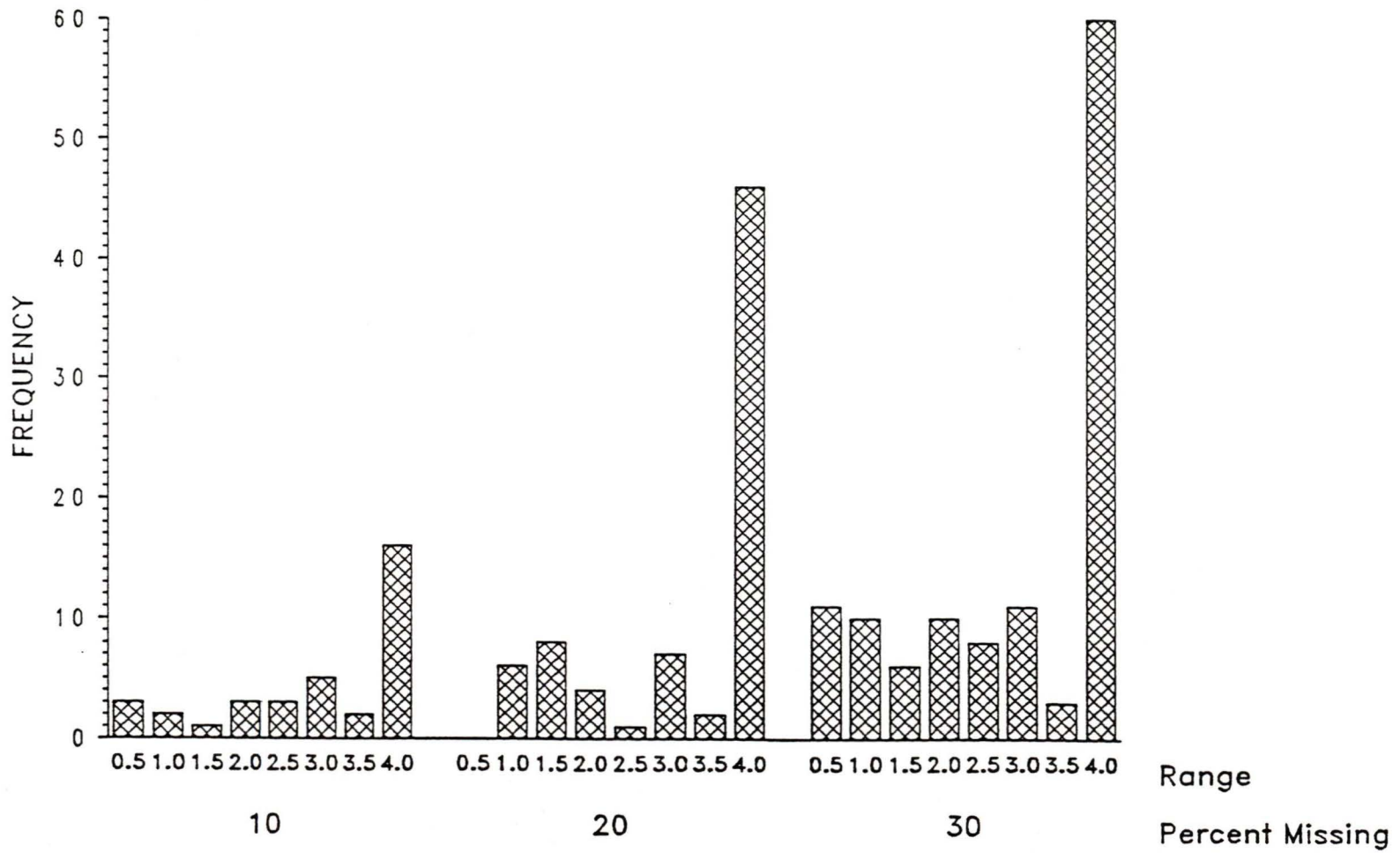


Figure 27.
Stony Lake Data: Lagged Dependent Regressor Prediction Distribution

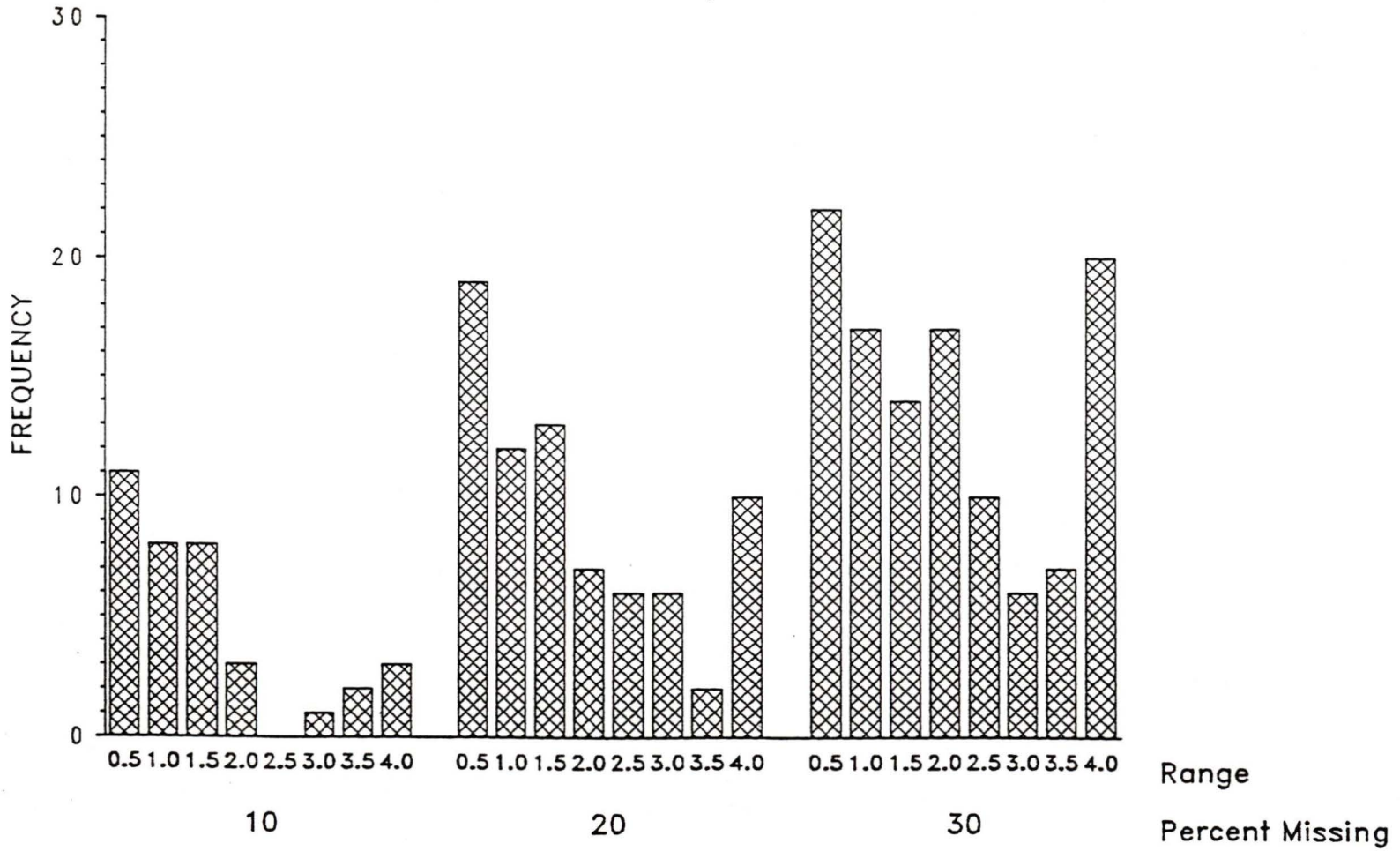
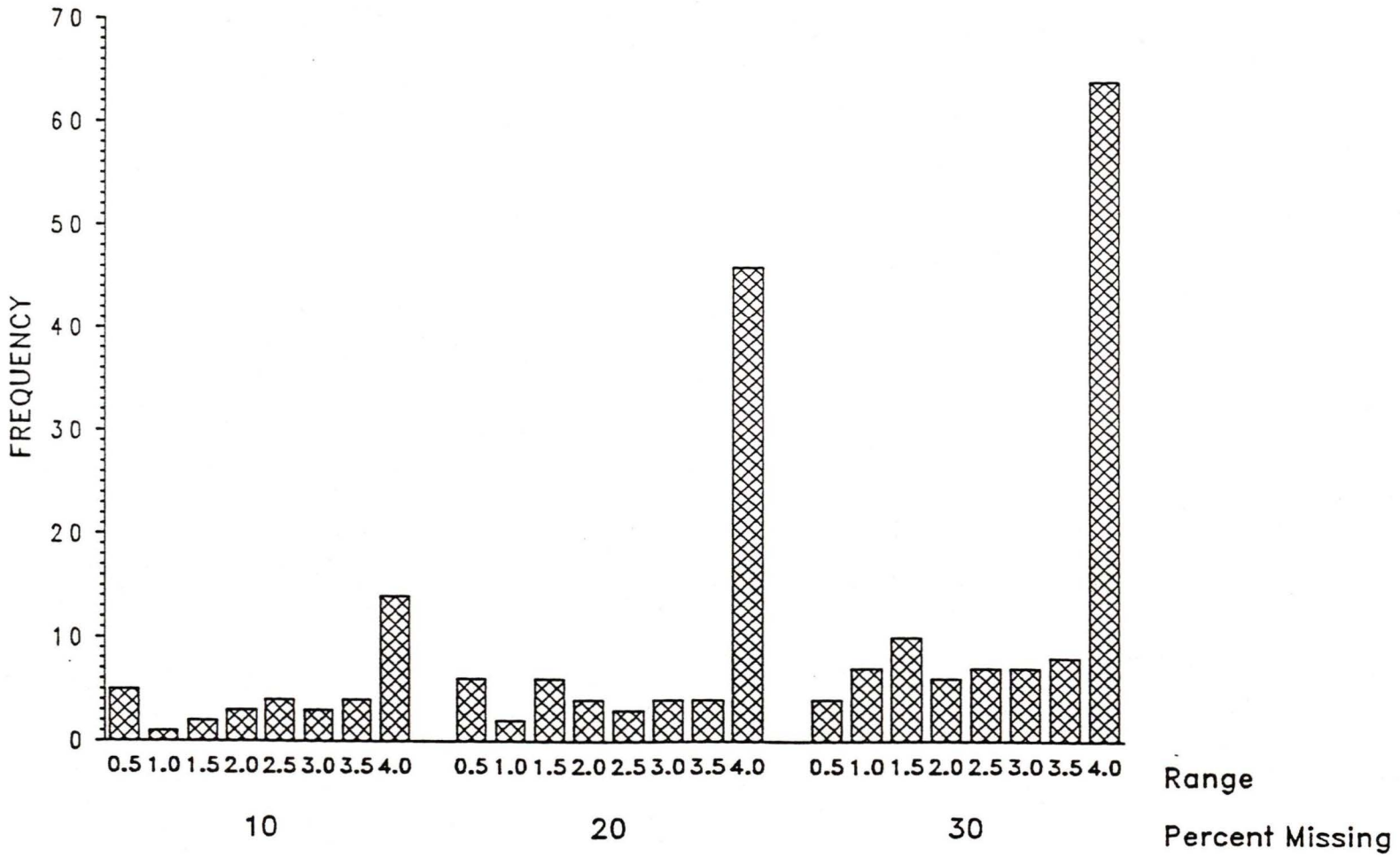


Figure 28.
Stony Lake Data: Modified Kalman Filter Prediction Distribution



Appendix B.

Artificial Data Estimation Models

Difference Method

The difference method is the standard method used by the Atmospheric Environment Service (AES) to correct short series climate normals. The best correlated AES standard station is used for correction purposes in all cases.

Missing 10 percent

Series 1 (Y) Mean 8.83
Series 3 (X) Mean 8.92

Model: $y = -0.09 + x$

Missing 20 percent

Series 1 (Y) Mean 9.32
Series 3 (X) Mean 9.40

Model: $y = -0.08 + x$

Missing 30 percent

Series 1 (Y) Mean 9.24
Series 3 (X) Mean 9.35

Model: $y = -0.09 + x$

Ratio Method Models

The ratio method, while not used as a standard correction method for temperature data, may be applied if the data has been transformed to the Kelvin scale. The models here were developed using the transformed data. They must be applied only to transformed data in order to produce meaningful results.

Missing 10 percent

Series 1 (Y) Mean 281.98
Series 3 (X) Mean 282.07

Model: $y = 0.9997 * x$

Missing 20 percent

Series 1 (Y) Mean 282.47
Series 3 (X) Mean 282.55

Model: $y = 0.9997 * x$

Missing 30 percent

Series 1 (Y) Mean 281.39
Series 3 (X) Mean 282.50

Model: $y = 0.9961 * x$

Ordinary Least Squares Regression

Missing 10 percent

Y	Series 1
X_2	Series 2
X_3	Series 3

Model: $y = -0.64 + 0.51*x_2 + 0.53*x_3$

M.S.E. = 0.76 $R^2 = 0.98$

Missing 20 percent

Y	Series 1
X_2	Series 2
X_3	Series 3

Model: $y = -0.65 + 0.52*x_2 + 0.53*x_3$

M.S.E. = 0.76 $R^2 = 0.98$

Missing 30 percent

Y	Series 1
X_2	Series 2
X_3	Series 3

Model: $y = -0.67 + 0.49*x_2 + 0.55*x_3$

M.S.E. = 0.76 $R^2 = 0.99$

Polynomial Curve Fit

Missing 10 percent

Y Series 1
X Series 3

Model: $y = -.06 + 0.84*x + 0.03*x^2 - 0.001*x^3$

M.S.E. = 0.72 $R^2 = 0.98$

Missing 20 percent

Y Series 1
X Series 3

Model: $y = 0.05 + 0.77*x + 0.04*x^2 - 0.002*x^3$

M.S.E. = 0.74 $R^2 = 0.98$

Missing 30 percent

Y Series 1
X Series 3

Model: $y = 0.46 + 0.50*x + 0.08*x^2 - 0.003*x^3$

M.S.E. = 0.61 $R^2 = 78$

ARIMA Models

The models listed here represent the best available fitted model for each data segment based on a combination of minimum AIC value and minimum estimated variance. The ARIMA(p,d,q) values represent the order of the autoregressive term, differencing, and moving-average term respectively.

Missing 10 Percent

Data Segment A: ARIMA(1,1,3)

σ^2 = 0.36 AIC = 409.10

Data Segment C: ARIMA(0,1,2)

σ^2 = 0.78 AIC = 189.33

Data Segment D: ARIMA(0,1,2)

σ^2 = 0.50 AIC = 56.54

Missing 20 Percent

Data Segment A: ARIMA(1,1,2)

σ^2 = 0.37 AIC = 201.59

Data Segment B: ARIMA(0,1,1)

σ^2 = 0.52 AIC = 77.92

Data Segment C: ARIMA(0,1,1)

σ^2 = 0.46 AIC = 47.01

Data Segment D: ARIMA(1,1,2)

σ^2 = 0.52 AIC = 255.46

Missing 30 Percent

Data Segment A: ARIMA(1,1,2)

σ^2 = 0.34 AIC = 111.28

Data Segment C: ARIMA(1,1,2)

σ^2 = 0.20 AIC = 19.40

Data Segment D: ARIMA(0,1,1)

σ^2 = 0.38 AIC = 218.76

Data Segment F: ARIMA(0,1,1)

σ^2 = 0.59 AIC = 59.19

Data Segment G: ARIMA(0,1,1)

σ^2 = 0.52 AIC = 62.64

Lagged Dependent Regressor Models

Missing 10 Percent

$$AVGTEMP1_t = -0.135 + 0.613*AVGTEMP1_{t-1} + 0.398*AVGTEMP3_t$$

M.S.E. = 0.519 R^2 = 0.982

Missing 20 Percent

$$AVGTEMP1_t = -0.118 + 0.614*AVGTEMP1_{t-1} + 0.395*AVGTEMP3_t$$

M.S.E. = 0.53 R^2 = 0.983

Missing 30 Percent

$$AVGTEMP1_t = -0.170 + 0.648*AVGTEMP1_{t-1} + 0.366*AVGTEMP3_t$$

M.S.E. = 0.43 R^2 = 0.989

Appendix C.

Driftwood Creek Data Estimation Models

Difference Method

The difference method is the standard method used by the Atmospheric Environment Service (AES) to correct short series climate normals. The best correlated AES standard station is used for correction purposes in all cases.

Missing 10 percent

Forestry Canada (Y)	Mean 4.28
Bobbie Burns Lodge (X)	Mean 2.84

Model: $y = 1.44 + x$

Missing 20 percent

Forestry Canada (Y)	Mean 4.45
Bobbie Burns Lodge (X)	Mean 3.03

Model: $y = 1.42 + x$

Missing 30 percent

Forestry Canada (Y)	Mean 4.38
Bobbie Burns Lodge (X)	Mean 2.94

Model: $y = 1.44 + x$

Ratio Method Models

The ratio method, while not used as a standard correction method for temperature data, may be applied if the data has been transformed to the Kelvin scale. The models here were developed using the transformed data. They must be applied only to transformed data in order to produce meaningful results.

Missing 10 percent

Forestry Canada (Y)	Mean	277.44
Bobbie Burns Lodge (X)	Mean	275.99

Model: $y = 1.005 * x$

Missing 20 percent

Forestry Canada (Y)	Mean	277.60
Bobbie Burns Lodge (X)	Mean	276.18

Model: $y = 1.005 * x$

Missing 30 percent

Forestry Canada (Y)	Mean	277.53
Bobbie Burns Lodge (X)	Mean	276.09

Model: $y = 1.005 * x$

Ordinary Least-squares Regression

Missing 10 percent

Y	Forestry Canada Station
X_2	Bobbie Burns Lodge
X_3	Golden
X_4	Rogers Pass
X_5	Mount Fidelity

Model: $y = 1.37 + 0.39*x_2 + 0.22*x_3 + 0.13*x_4 + 0.28*x_5$

M.S.E. = 1.25 $R^2 = 0.99$

Missing 20 percent

Y	Forestry Canada Station
X_2	Bobbie Burns Lodge
X_3	Golden
X_4	Rogers Pass
X_5	Mount Fidelity

Model: $y = 1.35 + 0.36*x_2 + 0.23*x_3 + 0.14*x_4 + 0.31*x_5$

M.S.E. = 1.14 $R^2 = 0.99$

Missing 30 percent

Y	Forestry Canada Station
X_2	Bobbie Burns Lodge
X_3	Golden
X_4	Rogers Pass
X_5	Mount Fidelity

Model: $y = 1.21 + 0.37*x_2 + 0.25*x_3 + 0.16*x_4 + 0.26*x_5$

M.S.E. = 1.17 $R^2 = 0.99$

Polynomial Curve Fit

Missing 10 percent

Y Forestry Canada Station
X Bobbie Burns Lodge

Model: $y = 0.73 + 0.96*x + 0.01*x^2$

M.S.E. = 1.77 $R^2 = 0.98$

Missing 20 percent

Y Forestry Canada Station
X Bobbie Burns Lodge

Model: $y = 0.73 + 0.96*x + 0.01*x^2$

M.S.E. = 1.79 $R^2 = 0.98$

Missing 30 percent

Y Forestry Canada Station
X Bobbie Burns Lodge

Model: $y = 0.56 + 1.04*x + 0.01*x^2$

M.S.E. = 1.53 $R^2 = 0.99$

ARIMA Models

The models listed in this appendix represent the best available fitted model for each data segment based on a combination of minimum AIC value and minimum estimated variance. The ARIMA(p,d,q) values represent the order of the autoregressive term, differencing, and moving-average term respectively.

Missing 10 Percent

Data Segment A: ARIMA(0,1,6)

σ^2 = 4.44 AIC = 483.14

Data Segment B: ARIMA(0,1,1)

σ^2 = 5.84 AIC = 162.11

Data Segment C: ARIMA(0,1,1)

σ^2 = 5.12 AIC = 367.64

Missing 20 Percent

Data Segment A: ARIMA(0,1,5)

σ^2 = 3.71 AIC = 356.54

Data Segment B: ARIMA(0,1,1)

σ^2 = 5.15 AIC = 251.70

Data Segment C: ARIMA(1,1,1)

σ^2 = 4.96 AIC = 366.37

Missing 30 Percent

Data Segment A: ARIMA(1,1,4)

σ^2 = 3.73 AIC = 323.70

Data Segment C: ARIMA(0,1,1)

σ^2 = 6.04 AIC = 339.43

Data Segment D: ARIMA(0,1,1)

σ^2 = 3.43 AIC = 74.31

Lagged Dependent Regressor Models

Missing 10 Percent

$$AVGDCFC_t = 1.260 + 0.176*AVGDCFC_{t-1} + 0.794*AVGBOBB_t$$

M.S.E. = 1.73 R^2 = 0.980

Missing 20 Percent

$$AVGDCFC_t = 1.224 + 0.193*AVGDCFC_{t-1} + 0.777*AVGBOBB_t$$

M.S.E. = 2.14 R^2 = 0.980

Missing 30 Percent

$$AVGDCFC_t = 1.287 + 0.179*AVGDCFC_{t-1} + 0.796*AVGBOBB_t$$

M.S.E. = 1.79 $R^2 = 0.984$

Appendix D.

Green Mountain Data Estimation Models

Difference Method

The difference method is the standard method used by the Atmospheric Environment Service (AES) to correct short series climate normals. The best correlated AES standard station is used for correction purposes in all cases.

Missing 10 percent

Forestry Canada (Y)	Mean 6.09
Cowichan Forestry (X)	Mean 8.91

Model: $y = -2.82 + x$

Missing 20 percent

Forestry Canada (Y)	MEAN 5.75
Cowichan Forestry (X)	Mean 8.58

Model: $y = -2.83 + x$

Missing 30 percent

Forestry Canada (Y)	Mean 4.97
Cowichan Forestry (X)	Mean 7.71

Model: $y = -2.74 + x$

Ratio Method Models

The ratio method, while not used as a standard correction method for temperature data, may be applied if the data has been transformed to the Kelvin scale. The models here were developed using the transformed data. They must be applied only to transformed data in order to produce meaningful results.

Missing 10 percent

Forestry Canada (Y)	Mean	279.24
Cowichan Forestry (X)	Mean	282.06

Model: $y = 0.990 * x$

Missing 20 percent

Forestry Canada (Y)	Mean	278.90
Cowichan Forestry (X)	Mean	281.73

Model: $y = 0.990 * x$

Missing 30 percent

Forestry Canada (Y)	Mean	278.12
Cowichan Forestry (X)	Mean	280.86

Model: $y = 0.990 * x$

Ordinary Least-squares Regression

Missing 10 percent

Y	Forestry Station
X_2	Cowichan Forestry
X_3	Port Alberni Airport

Model: $y = -3.07 + 1.20*x_2 - 0.17*x_3$

M.S.E. = 2.22 $R^2 = 0.94$

Missing 20 percent

Y	Forestry Canada Station
X_2	Cowichan Forestry
X_3	Port Alberni Airport

Model: $y = -3.00 + 1.26*x_2 - 0.24*x_3$

M.S.E. = 2.06 $R^2 = 0.94$

Missing 30 percent

Y	Forestry Canada Station
X_2	Cowichan Forestry
X_3	Port Alberni Airport

Model: $y = -3.03 + 1.29*x_2 - 0.25*x_3$

M.S.E. = 2.23 $R^2 = 0.94$

Polynomial Curve Fit

Missing 10 percent

Y Forestry Canada Station
X Cowichan Forestry

Model: $y = -2.55 + 0.98*x - 0.01*x^2 + 0.001*x^3$

M.S.E. = 2.10 $R^2 = 0.95$

Missing 20 percent

Y Forestry Canada Station
X Cowichan Forestry

Model: $y = -2.44 + 0.97*x - 0.01*x^2 + 0.001*x^3$

M.S.E. = 1.97 $R^2 = 0.95$

Missing 30 percent

Y Forestry Canada Station
X Cowichan Forestry

Model: $y = -2.54 + 0.98*x - 0.01*x^2 + 0.001*x^3$

M.S.E. = 2.18 $R^2 = 0.94$

ARIMA Models

The models listed in this appendix represent the best available fitted model for each data segment based on a combination of minimum AIC value and minimum estimated variance. The ARIMA(p,d,q) values represent the order of the autoregressive term, differencing, and moving-average term respectively.

Missing 10 Percent

Data Segment A: ARIMA(0,1,2)

σ^2 = 4.59 AIC = 403.45

Data Segment B: ARIMA(0,1,3)

σ^2 = 4.93 AIC = 717.60

Data Segment C: ARIMA(0,1,1)

σ^2 = 4.26 AIC = 65.31

Data Segment D: ARIMA(1,1,1)

σ^2 = 3.20 AIC = 202.61

Missing 20 Percent

Data Segment A: ARIMA(0,1,2)

σ^2 = 4.68 AIC = 387.64

Data Segment B: ARIMA(0,1,3)

σ^2 = 4.77 AIC = 457.09

Data Segment C: ARIMA(0,1,3)

σ^2 = 4.68 AIC = 111.72

Data Segment E: ARIMA(0,1,3)

σ^2 = 2.53 AIC = 74.61

Data Segment F: ARIMA(0,1,1)

σ^2 = 1.18 AIC = 54.98

Data Segment G: ARIMA(0,1,1)

σ^2 = 2.06 AIC = 66.51

Missing 30 Percent

Data Segment A: ARIMA(0,1,2)

σ^2 = 4.17 AIC = 684.45

Data Segment D: ARIMA(0,1,4)

σ^2 = 4.94 AIC = 187.63

Data Segment F: ARIMA(0,1,3)

σ^2 = 0.94 AIC = 63.87

Lagged Dependent Regressor Models

Missing 10 Percent

$$AVGGMFC_t = -2.607 + 0.095*AVGGMFC_{t-1} + 0.914*AVGCOFOR_t$$

M.S.E. = 1.87 $R^2 = 0.953$

Missing 20 Percent

$$AVGGMFC_t = -2.599 + 0.098*AVGGMFC_{t-1} + 0.909*AVGCOFOR_t$$

M.S.E. = 1.81 $R^2 = 0.951$

Missing 30 Percent

$$AVGGMFC_t = -2.569 + 0.110*AVGGMFC_{t-1} + 0.905*AVGCOFOR_t$$

M.S.E. = 1.80 $R^2 = 0.954$

Appendix E.

Stony Lake Data Estimation Models

Difference Method

The difference method is the standard method used by the Atmospheric Environment Service (AES) to correct short series climate normals. The best correlated AES standard station is used for correction purposes in all cases.

Missing 10 percent

Forestry Canada (Y)	Mean 5.24
Hixon (X)	Mean 6.57

Model: $y = -1.33 + x$

Missing 20 percent

Forestry Canada (Y)	Mean 5.01
Hixon (X)	Mean 6.12

Model: $y = -1.11 + x$

Missing 30 percent

Forestry Canada (Y)	Mean 4.97
Quesnel Airport (X)	Mean 6.87

Model: $y = -1.90 + x$

Ratio Method Models

The ratio method, while not used as a standard correction method for temperature data, may be applied if the data has been transformed to the Kelvin scale. The models here were developed using the transformed data. They must be applied only to transformed data in order to produce meaningful results.

Missing 10 percent

Forestry Canada (Y)	Mean	278.39
Hixon (X)	Mean	279.72

Model: $y = 0.995 * x$

Missing 20 percent

Forestry Canada (Y)	Mean	278.16
Hixon (X)	Mean	279.27

Model: $y = 0.996 * x$

Missing 30 percent

Forestry Canada (Y)	Mean	278.12
Quesnel Airport (X)	Mean	280.02

Model: $y = 0.993 * x$

Ordinary Least-squares Regression

Missing 10 percent

Y	Forestry Station
X_2	Barkerville
X_3	Prince George Airport
X_4	Quesnel Airport

Model: $y = 0.13 + 0.48*x_2 + 0.26*x_3 + 0.26*x_4$

M.S.E. = 7.50 $R^2 = 0.90$

Missing 20 percent

Y	Forestry Station
X_2	Barkerville
X_3	Prince George Airport
X_4	Quesnel Airport

Model: $y = 0.42 + 0.46*x_2 + 0.26*x_3 + 0.27*x_4$

M.S.E. = 6.99 $R^2 = 0.90$

Missing 30 percent

Y	Forestry Station
X_2	Barkerville
X_3	Prince George Airport
X_4	Quesnel Airport

Model: $y = 0.12 + 0.42*x_2 + 0.24*x_3 + 0.33*x_4$

M.S.E. = 7.21 $R^2 = 0.91$

Polynomial Curve Fit

Missing 10 percent

Y Forestry Canada Station
X Hixon

Model: $y = -1.15 + 0.94*x + 0.01*x^2 - 0.0003*x^3$

M.S.E. = 8.93 $R^2 = 0.88$

Missing 20 percent

Y Forestry Canada Station
X Hixon

Model: $y = -0.90 + 0.96*x + 0.005*x^2 + 0.0003*x^3$

M.S.E. = 8.46 $R^2 = 0.88$

Missing 30 percent

Y Forestry Canada Station
X Quesnel Airport

Model: $y = -1.69 + 0.92*x + 0.005*x^2 - 0.0002*x^3$

M.S.E. = 8.34 $R^2 = 0.89$

ARIMA Models

The models listed in this appendix represent the best available fitted model for each data segment based on a combination of minimum AIC value and minimum estimated variance. The ARIMA(p,d,q) values represent the order of the autoregressive term, differencing, and moving-average term respectively.

Missing 10 Percent

Data Segment A: ARIMA(1,1,2)

σ^2 = 10.70 AIC = 951.25

Data Segment B: ARIMA(1,1,2)

σ^2 = 7.49 AIC = 634.49

Data Segment C: ARIMA(0,1,3)

σ^2 = 3.12 AIC = 182.68

Missing 20 Percent

Data Segment A: ARIMA(1,1,2)

σ^2 = 9.66 AIC = 789.66

Data Segment B: ARIMA(0,1,3)

σ^2 = 7.98 AIC = 367.42

Data Segment C: ARIMA(0,1,2)

σ^2 = 7.56 AIC = 328.47

Data Segment D: ARIMA(0,1,1)

σ^2 = 3.36 AIC = 90.17

Missing 30 Percent

Data Segment A: ARIMA(0,1,1)

σ^2 = 5.18 AIC = 296.98

Data Segment B: ARIMA(1,1,1)

σ^2 = 18.97 AIC = 274.35

Data Segment C: ARIMA(0,1,1)

σ^2 = 14.86 AIC = 111.75

Data Segment D: ARIMA(0,1,2)

σ^2 = 7.62 AIC = 72.23

Data Segment E: ARIMA(0,1,3)

σ^2 = 6.77 AIC = 251.21

Data Segment F: ARIMA(0,1,1)

σ^2 = 6.90 AIC = 77.41

Data Segment H: ARIMA(1,1,2)

σ^2 = 3.23 AIC = 221.97

Lagged Dependent Regressor Models

Missing 10 Percent

$$AVGSLFC_t = -0.502 + 0.503*AVGSLFC_{t-1} + 0.472*AVGHIXON_t$$

M.S.E. = 5.89 $R^2 = 0.917$

Missing 20 Percent

$$AVGSLFC_t = -0.375 + 0.491*AVGSLFC_{t-1} + 0.481*AVGHIXON_t$$

M.S.E. = 5.79 $R^2 = 0.916$

Missing 30 Percent

$$AVGSLFC_t = -0.801 + 0.484*AVGSLFC_{t-1} + 0.487*AVGQUAIR_t$$

M.S.E. = 5.68 $R^2 = 0.927$

VITA

Surname: Benton Given Names: Ross Andrew

Place of Birth: Victoria Date of Birth: May 17, 1958

Educational Institutes Attended:

University of Victoria 1976 to 1981

Degrees Awarded:

B.Sc. (Co-op) University of Victoria 1981

PARTIAL COPYRIGHT LICENSE

I hereby grant the right to lend my thesis to users of the University of Victoria Library, and to make single copies only for such users or in response to a request from the Library of any other university, or similar institution, on behalf of its users. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by me or a member of the University designated by me. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Title of Thesis: A Comparative Analysis of Seven Methods for the Estimation of Values for Observations Missing from Temperature Climate Data Series

Author [redacted]

Ross Benton

Date November 2, 1990