The Subtleties of Distribution and Interpolation

David M. Aadland
Utah State University

Follow this and additional works at: https://digitalcommons.usu.edu/eri

Recommended Citation
https://digitalcommons.usu.edu/eri/156

This Article is brought to you for free and open access by the Economics and Finance at DigitalCommons@USU. It has been accepted for inclusion in Economic Research Institute Study Papers by an authorized administrator of DigitalCommons@USU. For more information, please contact digitalcommons@usu.edu.
Economic Research Institute Study Paper
ERI #99-06

THE SUBTLETIES OF DISTRIBUTION AND INTERPOLATION

by

DAVID M. AADLAND

Department of Economics
Utah State University
3530 Old Main Hill
Logan, UT 84322-3530

February 1999
THE SUBTLETIES OF DISTRIBUTION AND INTERPOLATION

David M. Aadland, Assistant Professor

Department of Economics
Utah State University
3530 Old Main Hill
Logan, UT 84322-3530

The analyses and views reported in this paper are those of the author(s). They are not necessarily endorsed by the Department of Economics or by Utah State University.

Utah State University is committed to the policy that all persons shall have equal access to its programs and employment without regard to race, color, creed, religion, national origin, sex, age, marital status, disability, public assistance status, veteran status, or sexual orientation.

Information on other titles in this series may be obtained from: Department of Economics, Utah State University, 3530 Old Main Hill, Logan, Utah 84322-3530.

Copyright © 1999 by David M. Aadland. All rights reserved. Readers may make verbatim copies of this document for noncommercial purposes by any means, provided that this copyright notice appears on all such copies.
THE SUBTLETIES OF DISTRIBUTION AND INTERPOLATION

David M. Aadland

ABSTRACT

This paper addresses several issues associated with distribution and interpolation of time series, including model selection and various data transformations. Monte Carlo experiments are performed, which suggest that failure to account for these data transformations may lead to serious errors in estimation.

Key words: missing data, temporal aggregation, systematic sampling, Kalman smoother, state-space model
1. INTRODUCTION

Missing data is a widespread problem in the social sciences. It arises in many different forms: irregularly-spaced missing observations, small blocks of missing observations (often at the beginning or end of a series), and missing data associated with temporal aggregation or systematic sampling. These last two cases, which are the focus of this paper, are arguably the most widespread and least transparent. For example, if the capital stock is sampled once per year, whereas it is believed to be generated by quarterly investment decisions, then observations on the capital stock in the first three quarters of the year will be systematically missing. This is referred to as systematic sampling. Likewise, if personal expenditures are generated by weekly decisions of households, but the data are only recorded once a month, we will only observe the sum of the series. This is referred to as temporal aggregation and is a missing-data problem because the sums between observations (i.e., the improper aggregates) can be thought of as missing. Both temporal aggregation and systematic sampling may lead to problems associated with model specification, parameter estimation, inference and prediction; see Sims (1971), Brewer (1973), Geweke (1978), Weiss (1984), Ermini (1989) and Rossanna and Seater (1992) to mention a few.

Several options are available for handling missing data. One option receiving much attention is based on the state-space representation (SSR) and the Kalman filter. These two items provide the foundation for several different smoothing algorithms that, under some fairly general conditions, provide optimal (in the sense of minimum mean-square error) estimates of the missing observations conditional on the observed data set. Other methods include the related-series method of Chow and Lin (1971), the dummy variable method of Sargan and Drettakis...
(1974), exponential smoothing, and a dynamic programming (DP) algorithm implemented in Regression Analysis of Time Series (RATS).

This paper outlines a systematic methodological approach (based on Kalman smoothing) for distributing or interpolating a sampled or aggregate time series which have been subject to data transformations. While the theoretic properties of the various interpolation and distribution procedures have been well documented (Harvey 1989), their performance in specific applications has been relatively unexplored. In response, this paper delves into some of the rather subtle issues that can arise during interpolation and distribution by performing Monte Carlo experiments which attempt to replicate the problem facing econometricians in practice. For example, the DISTRIB and INTERPOL procedures contained within the popular econometrics package RATS, if naively applied to data that has been subject to various transformations, may result in misleading estimates. However, as will be made clear, if these data transformations are explicitly accounted for at the modeling stage, the estimates of the missing data can often be greatly improved.

The paper proceeds as follows. In section 2, I introduce the SSR, the Kalman filter and the fixed-interval smoother. In section 3, I perform several Monte Carlo experiments which contrast benchmark and Kalman-smoothed estimates of several different ARIMA processes and discuss the implications of these results. In section 4, I summarize the paper's most important findings.

2. STATE-SPACE REPRESENTATION, KALMAN FILTER AND KALMAN SMOOTHER

---

1 While I focus exclusively on systematically sampled and temporally aggregated data, the techniques can in many instances be generalized to handle other patterns of missing data; see Brockwell and Davis (1987).
In order to work with the Kalman filter it is necessary to write the dynamic system in its SSR. In the context of missing data, the SSR is a convenient representation because it provides the framework for a number of different smoothing algorithms that, conditional on the entire observed series, generate optimal estimates of the missing observations. Those well-versed in the state-space model, the Kalman filter and smoothing algorithms in the context of missing observations may choose to skim the material in sections 2.1 through 2.4 and focus on the subsequent material.

2.1 The State-Space Representation

Begin by considering the following class of univariate ARIMA(p,d,q) processes:

\[ \phi_p(L) \Delta^d x_{nt-i} = \theta_q(L) \varepsilon_{nt-i} \quad \tau = 1, \ldots, T / n; \quad i = 0, \ldots, n - 1 \]  

(2.1)

where the roots of \( \phi_p(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \) lie outside the unit interval,

\[ \theta_q(L) = 1 + \theta_1 L + \cdots + \theta_q L^q, \quad \Delta^d = (1 - L)^d, \quad L \] is the lag operator, and \( \varepsilon \sim \text{iid } N(0, \sigma^2) \). In practice, the basic variable \( x_{nt-i} \) is not always observed at its natural timing interval. Rather, it is often observed at its sampling interval, which due to institutional constraints may be longer than the timing interval. In particular, if the data-sampling interval for a flow series is longer than the timing interval (i.e., \( n > 1 \)), we observe the temporally aggregated variable \( B_1(L) x_{nt} = X_{nt} \forall \tau \), where \( B_1(L) = 1 + L + \cdots + L^{n-1} \) is the aggregation operator. Alternatively, for a stock series, we observe the systematically sampled variable \( B_0(L) x_{nt} = x_{nt-j} \forall \tau \), where \( j \in \{0, \ldots, n-1\} \) and \( B_0(L) = L^j \) is the sampling operator (hereafter, assume without loss of generality that \( j = 0 \) such that \( B_0(L) = 1 \)). Occasionally for a stock series, we may also observe the temporally averaged variable \( (B_1(L) / n) x_{nt} = X_{nt} / n \forall \tau \). Throughout the remainder of the paper, I will commonly
refer to the disaggregate or basic model as the monthly model and the aggregate model as the quarterly model. This is done only for expositional purposes and can be easily generalized to other data frequencies and degrees of aggregation. Also, as a clarifying note, if by chance the timing and data-sampling intervals coincide (i.e., $n = 1$), then apart from any other data limitations, missing data will not be considered a problem.

The time-invariant SSR is comprised of two equations, the first being the state equation:

$$\xi_{nt-i} = F\xi_{nt-i-1} + Rv_{nt-i}, \quad \tau = 1, \ldots, T / n; \quad i = 0, \ldots, n - 1 \quad (2.2a)$$

which describes the law of motion for the state vector. The second is the measurement equation:

$$X_{nt} = H\xi_{nt}, \quad \tau = 1, \ldots, T / n \quad (2.2b)$$

which relates the $(r \times 1)$ state vector to the observable variable, where $r = \max[\eta, p, 1]$,

$$\eta = \lambda^{\mu}n + (1 - \mu)(d + \lambda - 1)(n - 1), \quad \mu = 0 \text{ when the SSR is stationary, } \mu = 1 \text{ when the SSR is nonstationary, } \lambda = 0 \text{ under systematic sampling, and } \lambda = 1 \text{ under temporal averaging or aggregation.}$$

For the stationary SSR, the various components are defined as follows. First, the state vector is given by

$$\xi_{nt-i} = (\Delta^d x_{nt-i} \Delta^d x_{nt-i-1} \ldots \Delta^d x_{nt-i-(r-1)})^\prime,$$

while the error vector is given by

$$v_{nt-i} = (e_{nt-i} e_{nt-i-1} \ldots e_{nt-i-q})^\prime,$$

with diagonal variance-covariance matrix $\Sigma_v$. Second, the transition matrix is

$$F = \begin{bmatrix}
\phi_1 & \cdots & \phi_p & \cdots & 0 & 0 \\
1 & 0 & 0 & 0 \\
\vdots & & & & & \\
0 & 1 & 0 & 0 \\
\vdots & & & & & \\
0 & \cdots & 0 & 1 & 0
\end{bmatrix}_{(r \times r)}.$$
where the columns with lead zeros are only necessary if \( n > p \) and

\[
R = \begin{bmatrix}
1 & \theta_1 & \cdots & \theta_q \\
0 & 0 & 0 & \vdots \\
0 & 0 & \cdots & 0
\end{bmatrix}_{(r \times q+1)}.
\]

Third and finally, the measurement vector \( H \) is comprised of zeroes and the coefficients on the polynomial \( B_i(L)^{d+\lambda} \) such that the coefficient on the 0\(^{th}\) order term is in the first position, the coefficient on the 1\(^{st}\) order term is in the second position, etc. If \( r \), the length of \( H \), is greater than one plus the order of \( B_i(L)^{d+\lambda} \), then zeroes need to be placed in remaining positions. Table 1 presents various combinations of \( H \) for different orders of integration and degrees of sampling and aggregation.

[Insert table 1]

The non-stationary SSR is useful for directly estimating the non-stationary basic series and can be created with a few simple modifications to the stationary SSR. First, the non-stationary state vector, \( \xi_{n,i}^* \), is formed by concatenating the column vector

\[(x_{n-i-\alpha} \cdots x_{n-i-\alpha-(d-1)})'\]

to the end of \( \xi_{n,i} \), where \( \alpha = n^\lambda \). Second, the transition matrix needs to be augmented as shown below:

\[
F^* = \begin{bmatrix}
F & 0 & 0 \\
1 & \Gamma_{d-1} & \gamma_d \\
0 & 1 & 0
\end{bmatrix}_{(r+d) \times (r+d)}
\]

where 0 and I are the appropriately dimensioned null and identity matrices, \( \gamma_d \) is the coefficient on the \( d \)\(^{th}\) order term in the lag polynomial \(-\Delta^d\) and \( \Gamma_{d-1} = (\gamma_1 \cdots \gamma_{d-1}) \). The final modification to the SSR is to the measurement vector, which becomes \( H^* = (H_\alpha \ 0 \ \pi_\alpha \cdots \pi_{\alpha+d-1}) \), where
Hα indicates the first α elements of H, πj is the coefficient on the jth order term in

\[-(B_{1,n}(L)^{d+λ})^\lambda \Delta^d,\]

and B_{1,n}(L)^{d+λ} indicates the first n elements in the polynomial B_1(L)^{d+λ}.

Table 2 depicts the H* vector for various combinations of n and d.

2.2 The Kalman Filter

In the context of distribution and interpolation, the Kalman filter is a recursive procedure for making optimal forecasts of the state variable using past observations on the sampled or aggregated series. If we assume Gaussian errors, it can be summarized by the equations

\[
\hat{\xi}_{nt-i|nt-n} = E(\xi_{nt-i}|X_{nt-n}, \ldots, X_n) = F^{n-i}\hat{\xi}_{nt-n} + K_{nt-i}(X_{nt} - H\hat{\xi}_{nt-n}) \tag{2.3a}
\]

\[
P_{nt-i|nt-n} = E((\xi_{nt-i} - \hat{\xi}_{nt-i|nt-n})(\xi_{nt-i} - \hat{\xi}_{nt-i|nt-n})') = (F^{n-i}P_{nt|nt-n} - K_{nt-i}HP_{nt|nt-n})F^{n-i} + \sum_{j=0}^{n-i-1}F^jRQR'(F^j)' \tag{2.3b}
\]

where \(K_{nt-i} = F^{n-i}P_{nt|nt-n}H'(HP_{nt|nt-n}H')^{-1}\) is the gain matrix (as Ansley and Kohn (1985) point out, if the SSR is nonstationary, it is not appropriate to interpret \(\hat{\xi}\) as the linear projection of \(\xi\) on past X). Derivations of the Kalman filter in the presence of missing data can be found in Harvey (1989, chap. 6). Together with a set of initial conditions, equations (2.3) completely specify the Kalman filter. The initial conditions for the stationary model are commonly specified as the unconditional mean and variance of the state vector

\[
\hat{\xi}_{i0} = E(\xi_i) = 0 \tag{2.4a}
\]

\[
\text{vec}(P_{i0}) = \text{vec}(E(\xi_{i0} - \hat{\xi}_{i0})(\xi_{i0} - \hat{\xi}_{i0})') = (I - (F \otimes F))^{-1}\text{vec}(R\Sigma_y R') \tag{2.4b}
\]

where \(i = 1, \ldots, n\) and the vec operator stacks the columns of an (m×n) matrix into an (mn×1) column vector.
The initial conditions for the non-stationary case are more problematic because the unconditional mean and variance of the state vector are not well defined. Nevertheless, several methods have been suggested in the literature to deal with initial conditions for the nonstationary SSR. The most straightforward method is to specify a non-informative (partly diffuse) prior distribution for $\xi_n$ using the "large $\kappa$" approximation. Alternatively, if a run of observations are available at the beginning (or end) of the series, an equivalent method is available that forms a proper prior; see Harvey (1989, chap. 3). Koopman (1997) provides a nice summary of the recent advances for handling the initial conditions in nonstationary SSR and presents a new, computationally efficient exact solution as well.

2.3 The Likelihood Function

Once the Kalman filter has been specified, estimates of the parameters

$\Theta = (\phi_1 \ldots \phi_p \theta_1 \ldots \theta_q \sigma^2)'$ 

can be obtained by maximizing the likelihood function. Note that if $\varepsilon_{nt}$ and the initial values are normally distributed, $X_{nt|nt-n}$ will also be normally distributed. As a consequence, the prediction errors can be used to write the exact log likelihood function (apart from a constant and proportionality factor) as

$$L(\Theta) = -\log(\|HP_{nt|nt-n}H'\|) - \sum_{t=1}^{T/n} \nu_{nt} (HP_{nt|nt-n}H')^{-1} \nu_{nt}$$

(2.5)

where $\nu_{nt} = X_{nt} - \hat{H}_{nt|nt-n}$ is the prediction error. The estimate $\hat{\Theta}$ that maximizes (2.5) can be calculated using any of several numerical optimization routines given a set of initial values for $\Theta$. One particularly efficient algorithm in this context is the EM algorithm; see Dempster, Laird and Rubin (1977).
2.4 The Kalman Smoother

While the filtered estimates themselves may at times be of interest, often we are interested in estimates of the basic series based on the entire observable series rather than estimates based only on past data. Estimates based on the entire observable series are called smoothed estimates of which there are several different variations. The most commonly used for economic data is the fixed-interval smoother:

\[
\hat{x}_{nt-i|T} = \hat{x}_{nt-i|nt-n} + J_{nt-i}(\hat{x}_{nt-i|T} - \hat{x}_{nt-i|nt-n})
\]  

(2.6)

where \( J_{nt-i} = P_{nt-i|nt-n}F'P_{nt-i|nt-n}^{-1} \). Notice that the series \( \{P_{nt-i|nt-n}\} \) and \( \{\hat{x}_{nt-i|nt-n}\} \) used in the smoother are produced by the Kalman filter, and thus will need to be stored for use in (2.6). The smoothed series \( \{\hat{x}_{nt-i|T}\} \) is produced by a sequence of backward recursions starting with \( \hat{x}_{T|T} \) which is given by the final iteration of the updating equations for the Kalman filter (the updating equations are formed by setting \( i = n \) in (2.3a)).

2.5 Model Selection

Selecting the correct disaggregate model is a formidable task when only aggregate or sampled data are observed. Harvey (1989, p. 309) states that "...ARIMA model specification is rather difficult in the absence of any available observations at the model timing interval." In particular, simply applying univariate model-selection methods (e.g., Box-Jenkins analysis) to the aggregated or sampled observations can be misleading. As several authors have noted (e.g., Working (1960), Tiao (1972) and Weiss (1984)), this is true because the basic models are not generally invariant to temporal aggregation or systematic sampling. In fact, Weiss provides a
detailed analysis describing the aggregate process that results when a basic \( \text{ARIMA}(p,d,q) \) process is subject to temporal aggregation or systematic sampling.

In response, I propose the following strategy for selecting a monthly model given quarterly data. First the "best" quarterly model is chosen using Box-Jenkins methods. Each chosen quarterly model is then mapped to a monthly model using tables 1 and 2 in Weiss (1984). For our purposes, the relevant parts of tables 1 and 2 can be summarized by noting that upon systematic sampling or temporal aggregation, a monthly \( \text{ARIMA}(p,d,q) \) process becomes a quarterly \( \text{ARIMA}(p,d,\omega) \) process where \( \omega = \lfloor(n - 1)(p + d + \lambda + q) / n \rfloor \) and \( \lfloor \eta \rfloor \) refers to the integer part of \( \eta \). A certain degree of caution needs to be applied when taking this approach because there is not a one-to-one mapping between monthly and quarterly models. As an illustration, both the monthly \( \text{AR}(1) \) and \( \text{ARMA}(1,1) \) flow processes map to a quarterly \( \text{ARMA}(1,1) \) process. Consequently, there is some subjectivity involved with choosing an appropriate disaggregate model (not unlike choosing the aggregate model). Notwithstanding this point, it still seems more theoretically pleasing to try and account for the change in model structure imposed by aggregation or sampling rather than incorrectly assume model invariance.

2.6 Modifications to the SSR: Natural Logarithms and Ratios

It is a widely accepted practice to make preliminary transformations of many economic variables. Two of the more common transformations include taking the natural logarithm of exponentially growing series and the ratio of two series (e.g., per-capita and nominal-real transformations). Both transformations require small and often subtle modifications to the SSR.
Natural Logarithms.

The primary difficulty of working with variables that have been transformed as natural logs is that the sum of the logs is not equal to the log of the sums. For example, if a monthly flow series is measured in logs, simply aggregating the observations will not produce the log of the quarterly series. Harvey and Pierse (1984) deal with this problem by assuming that the observable logged aggregates are normally distributed and then use the extended Kalman filter (Harvey 1989, pp. 160-162) to handle the implied nonlinear measurement equation.

I take a related approach. Begin by writing the sum of the basic logged series as
\[
\log(\tilde{X}_{nt}) = \log(x_{nt}) + \log(x_{nt-1}) + \cdots + \log(x_{nt-(n-1)}),
\]
where \(\tilde{X}_{nt}\) is the n-period geometric sum of the basic series. Now take a first-order Taylor series approximation around the sample mean \((\bar{X}_{nt})\) for each of the \(T/n\) sets of consecutive nonoverlapping basic observations, which results in
\[
\log(\tilde{X}_{nt}) \equiv n \log(\bar{X}_{nt}) + \frac{1}{\bar{X}_{nt}} (x_{nt} - \bar{X}_{nt}) + \cdots + \frac{1}{\bar{X}_{nt}} (x_{nt-(n-1)} - \bar{X}_{nt}).
\]
For a temporally aggregated flow series, the sample mean for each \(n\) consecutive basic observations is observable and given by \(X_{nt}/n\). Since the last \(n\) terms above always sum to zero, the working approximation for the sum of the basic logs becomes
\[
\log(\tilde{X}_{nt}) \equiv n(\log(X_{nt}) - \log(n)), \quad (2.7)
\]
which can be used on the left-hand side of (2.2b) to linearize the measurement equation. If the basic series is instead temporally averaged, the first-order approximation of the time average of the logs is the log of the average such that no adjustments are necessary.

\[\text{Actually, } p \text{ and } \omega \text{ represent the maximum AR and MA orders of the aggregated process as noted in Weiss (1984).}\]
Ratios.

Another potential problem is estimating missing observations of the ratio of two series. The problem is similar to the case of natural logs, in that, the sum of a ratio is not equal to the ratio of the sums. However, provided the individual aggregates are available, one alternative is to estimate the missing values of each of the individual series and then form the ratio of the two estimates. In the case where the aggregates are only available as a ratio, a second alternative is to linearize the unobservable sum of the basic ratios and write it in terms of the observable ratio. There are several cases to be considered depending on whether the variables are flows or stocks and whether the stock(s) is systematically sampled or temporally averaged.

Begin by assuming that all stocks are temporally averaged. Now take a first-order Taylor series approximation (around \( \overline{Y}_{nt} \) and \( \overline{X}_{nt} \)) of the unobserved sum of ratios:

\[
\frac{x_{nt}}{y_{nt}} + \cdots + \frac{x_{n(t-1)}}{y_{n(t-1)}} = \frac{x_{nt}}{y_{nt}} + \sum_{i=0}^{\eta - 1} \frac{1}{y_{nt}} (x_{n(t-i)} - \overline{X}_{nt}) - \sum_{i=0}^{\eta - 1} \frac{\overline{X}_{nt}}{y_{nt}} (y_{n(t-i)} - \overline{Y}_{nt}) = \frac{x_{nt}}{y_{nt}}.
\]  

(2.8)

In case one, assume \( x \) is a flow and \( y \) is a stock (e.g., income per capita). According to (2.8), then at least to a first order, the ratio of the within-period sum of \( x \) to the within-period average of \( y \) can be used to approximate the sum of ratios. The quality of this approximation will vary inversely with the amount of variation within each \( n \)-length interval of the basic series, that is, the greater the within-sum variation, the worse the approximation. In case two, when the roles of \( x \) and \( y \) are reversed, such that \( x \) is instead a stock and \( y \) is a flow (e.g., price-earnings ratio), then multiplying the top and bottom of (2.8) by \( n \) indicates that \( n^2 \) times the ratio of the within-period average of \( x \) to the within-period sum of \( y \) is the appropriate approximation. In case three and four where \( x \) and \( y \) are either both flows (e.g., average labor productivity) or both stocks (e.g., the unemployment rate), by the same logic as above, the appropriate approximation is \( n \) times the

I thank Massimiliano Marcellino for pointing this out to me.
ratio of the within-period sums for $x$ and $y$ when considering flows and $n$ times the ratio of the within-period averages when considering stocks.

Of course, linearizing around the within-period averages is not feasible when the stock is observed as a systematically sampled series. This creates no problems in cases three and four. In case three there are no stocks and in case four where $x$ and $y$ are both stocks, the observed ratio can then be treated as a single stock. However, for cases one and two, if one linearizes around the systematically sampled stock, then the summation terms in (2.8) do not vanish. A natural solution, although not particularly pleasing from a theoretical standpoint, is to linearize around $\bar{Y}_{nt}$ and $\bar{X}_{nt}$ and then replace the average with its systematically sampled counterpart. This can be partially justified by taking a first-order approximation of the left-hand side of (2.8) around $y_{nt}$ and ignoring the term

$$\frac{X_{nt}}{ny_{nt}} \left( \frac{y_{nt-1} + \ldots + y_{nt-(n-1)}}{y_{nt}} - n \right),$$

which should be approximately zero if, again, the within-period variation in the basic series is sufficiently small.

Another possible transformation combines logarithms and ratios. Assume that one wishes to distribute the log of a ratio. The relevant sum is then

$$\log \left( \frac{x_{nt}}{y_{nt}} \right) + \ldots + \log \left( \frac{x_{nt-(n-1)}}{y_{nt-(n-1)}} \right) = \log(\bar{X}_{nt}) - \log(\bar{Y}_{nt}),$$

(2.9)

where $\bar{X}_{nt}$ and $\bar{Y}_{nt}$ are again the geometric within-period sums. Thus, the appropriate approximation for a flow variable in (2.9) is given by (2.7), for a systematically sampled stock it is $n^* \log(y_{nt})$, and for a temporally averaged stock it is $\log(\bar{Y}_{nt})$. 
3. EXPERIMENTAL DESIGN AND SIMULATION RESULTS

In this section I describe both the design and results of several Monte Carlo experiments. These experiments are intended to highlight some of the effects of not appropriately accounting for data transformations when distributing or interpolating sampled or aggregated time series. In general, distribution and interpolation should be applied cautiously to time series and may not always be the optimal strategy for handling missing data. At the same time, these procedures may be an attractive option when one has several series measured at, say, a monthly frequency and a single series measured at a quarterly frequency. Rather than aggregate all the monthly series to a quarterly frequency it may be desirable to distribute or interpolate the quarterly series to the monthly frequency, in essence, exchanging the smaller degrees of freedom associated with the quarterly interval for the additional uncertainty associated with the estimated series (another option is to specify a model that directly incorporates mixed-frequency data; see Zadrozny (1990)). The following sections discuss the intricacies of such a procedure for various artificially generated time series which have been subject to data transformations.

3.1 Experimental Design

The first step in this Monte Carlo experiment is to generate 100 time series from each of the following three processes:

(1) \[ x_{3t-i} = c_1(x_{3t-i-1})^{0.7} \exp(\varepsilon_{3t-i}) \]

(2) \[ x_{3t-i} = c_2(x_{3t-i-1})^{0.6} \exp(\varepsilon_{3t-i} + 0.4\varepsilon_{3t-i-1} + 0.15\varepsilon_{3t-i-2}) \]

(3) \[ x_{3t-i} = c_3(x_{3t-i-1})^{1+0.5}(x_{3t-i-2})^{-0.3-0.5}(x_{3t-i-3})^{0.3} \exp(\varepsilon_{3t-i} + 0.25\varepsilon_{3t-i-1}) \]

where \( c_1 = c_2 = \exp(1), \ c_3 = \exp(0.01), \tau = 1, \ldots, 50 \) and \( i = 0, 1, 2 \). Using standard ARIMA notation, (1), (2) and (3) are AR(1), ARMA(1,2) and ARIMA(2,1,1) processes in logs,
respectively. For each of the three ARIMA processes, \{\varepsilon}\ is drawn from a mean-zero normal distribution with a constant variance such that \text{var}(\ln(x)) = 1 \times 10^{-4}\ for the stationary-in-levels processes (1) and (2), and \text{var}(\Delta_1 \ln(x)) = 1 \times 10^{-4}\ for the difference-stationary process (3). Each series is initialized at \(x = 1\) and \(\varepsilon = 0\). Once each 150-length realization is generated it is then temporally aggregated or systematically sampled. First, I treat \{x\} as a flow process and temporally aggregate each set of three consecutive observations, beginning with the first observation. For the three processes above, this creates 100 aggregate series, each with a sample size equal to 50. The entire procedure is then repeated, that is, the processes are simulated and then treated as stock series, where \{x\} is then systematically sampled by selecting every third observation, beginning with the first observation in the series.

Once the aggregate data are in hand, I then distribute and interpolate the series under various transformations and contrast the estimated series with the actual series. Depending upon the nature of the data, I either interpolate or distribute the logged aggregate data using (i) the benchmark estimator (i.e., DISTRIB or INTERPOL procedures) with logarithmic and ratio corrections where appropriate; (ii) the Kalman smoother without the logarithmic correction; and (iii) the Kalman smoother with the logarithmic correction. In addition, I also create and then distribute the logged flow-stock ratio of the previously generated aggregate data using (i) and (iii); (iv) the Kalman smoother with the logarithm but without the ratio correction; (v) the ratio of the individual estimates from (iii); and (vi) the Kalman smoother with both the logarithm and ratio corrections.\textsuperscript{3}

\textsuperscript{3} Since \(x\) and \(y\) are generated from the same models with the same parameters, the log of \((x/y)\) will have the same structure as the log of \(x\) or the log of \(y\), with twice the error variance.
I begin by estimating benchmark monthly values for the various aggregate series using the procedures DISTRIB and INTERPOL included in the software package RATS. In our notation, the DISTRIB and INTERPOL procedures solve the following problem using a dynamic programming (DP) algorithm:

\[
\minimize \sum_{t=1}^{T} \varepsilon_i^2 \text{ subject to } \xi_{nt-i} = F\xi_{nt-i-1} + \nu_{nt-i} \text{ and } X_{nt} = H\xi_{nt} \forall \tau,i
\]

where \( F \) is given, \( H = (1 \ldots 1) \) for the DISTRIB procedure, and \( H = (1 \ldots 0) \) for the INTERPOL procedure. In light of our previous discussion, it is clear that, in general, these procedures will not properly account for data that is in the form of logs, differences, or ratios.

The problem with using the DISTRIB and INTERPOL procedures when the data has been subject to transformations is twofold. One, the measurement vector \( H \) is fixed at two values, either \( H = (1 \ldots 1) \) or \( H = (1 \ldots 0) \), whereas, tables 1 and 2 indicate that \( H \) varies systematically with \( d, \lambda \) and \( n \). And two, they allow only a limited number of basic ARIMA models. The extent to which these limitations are important is an empirical one and will be partially addressed in this paper. At the same time, despite the clear limitations for handling transformed data, the two RATS procedures are simple to use and should serve as reasonable benchmarks from which to compare the quality of the estimates from the Kalman smoother.

As for model identification, since I assume that the parameters are known with certainty, the fact that the models vary with respect to sampling and aggregation is a moot point. This is due to the dual fact that aggregation and sampling influence only the MA component (i.e., \( R \)) of the model and the variance matrix \( RQR' \) does not have an effect on the estimates of the state vector produced by the Kalman filter algorithms; see Harvey 1989, p.107. Hence, a changing model structure is relevant only to the extent that it influences parameter estimates if the wrong basic ARIMA model is chosen. As a consequence, I assume that the econometrician applies the
correct basic ARIMA model. An investigation of the effects of model structure and parameter estimation is left for future research.

3.2 Contrasting the Results

The various estimates of the monthly series are then contrasted with their actual values using Theil’s U statistic

$$\sqrt{\frac{\sum (x_{nt-i} - \hat{x}_{nt-i|T})^2}{\sum x_{nt-i}^2}},$$

where the summations run over all $\tau$ and $i$. Theil’s U statistic is preferable to other criterion such as root mean-square errors because it removes any scaling issues and allows for comparisons across series and transformations; see Greene (1993). Comparisons of distribution and interpolation performance across data types and transformations will therefore be possible.

Table 3 presents Theil’s U statistics for the benchmark and smoothed estimates. Several salient features of table 3 deserve further attention. First, notice that the RATS and KS estimates which implement logarithmic and ratio corrections are quite similar. The Kalman smoother appears to perform slightly better when distributing flow series and the RATS procedure performs slightly better when distributing flow/stock data. In addition, notice that the Theil U statistic for the Kalman smoother in the ARIMA(2,1,1) case is approximately 15 percent lower than its RATS counterpart. This is not surprising given that table 1 indicates that the measurement vector should be $H = (1 2 3 2 1)$ while the DISTRIB procedure has $H$ fixed at $(1 1 1)$. This implies that while the Kalman smoother correctly estimates the growth rate of $x_{nt}$, $\log(x_{nt} / x_{nt-1})$, the DISTRIB procedure instead estimates the growth rate of $X_{nt}$ in basic time,
\begin{equation}
\log(X_{nt} / X_{nt-1}) = \log(x_{nt} / x_{nt-1}) + \cdots + \log(x_{nt-(n-1)} / x_{nt-n}).
\end{equation}

Figures 1a and 1b, which depict the actual, RATS and KS estimate for the first 48 observations of the first run of the Monte Carlo experiment, confirm this relationship. Figure 1b shows that the RATS and KS estimates of the stock growth rate are quite similar, which is to be expected given that $H = (1\ 1\ 1)$ is the appropriate measurement vector (see table 1). The RATS estimates of the flow growth rate as shown in figure 1a, on the other hand, appear to be shifted forward by one to two periods. This is exactly what one would expect given the approximation above and the oscillatory nature of the data.

[Insert table 3 and figures 1a, 1b]

Second, notice that for all three processes, the Theil U statistics are smaller for flows than stocks. This suggests that aggregation over time, while resulting in a loss of information, is generally preferable to point-in-time sampling. Figures 1a and 1b also confirm this pattern. The flow estimates in figure 1a, clearly come closer to the actual series than do their stock counterparts.

Third, the correction for logarithms greatly reduces the prediction errors. Moreover, since the logarithmic correction is simply a linear transformation of the logged aggregates, it does not influence the shape of the estimated time series but rather affects its position and amplitude. Another interesting result is that failure to impose the logarithmic transformation for the flow variable results in much greater errors for the first two $I(0)$ processes than the $I(1)$ process. This can readily be seen in both table 3 and in figures 2a and 2b. In table 3, notice that the logarithmic correction reduces Theil's U statistic by 99.8% and 99.7% for the two stationary
series, while it is reduced by only 56% for the nonstationary series. Figures 2a and 2b present the actual and Kalman smoothed estimates, with and without the logarithmic correction. Based on these simulations, it is clear that if one were not to impose a logarithmic correction for the AR(1) and ARMA(1,2) level processes, it would be much more evident than when estimating growth rates in the ARIMA(2,1,1) case.

[Insert figures 2a, 2b]

Fourth, and finally, when distributing the (log) ratio of two series, it appears that the first-best method is to use the individually distributed or interpolated series to form the estimate as opposed to directly distributing or interpolating the aggregate ratio. For the cases being considered, there is an approximate 20% reduction in the Theil U statistics when the individual estimates are used. Figures 3a and 3b depict the actual data along with the Kalman smoothed data using both the aggregate ratio and the individual aggregates. The superiority of the estimates based on the individual series is more evident in figure 3a. Close inspection of the ratio and individual estimates shows that the estimates based on the aggregate ratio tend to have more variation than those based on the aggregate ratio and, as a consequence, tend to have larger prediction errors. In addition, table 3 shows that failure to impose a correction when distributing an aggregate ratio, similar to the logarithmic case considered earlier, is likely to produce grossly inferior estimates.

[Insert figures 3a, 3b]

4. CONCLUDING REMARKS

Social scientists are often constrained in their research by the availability of data. When this constraint comes in the form of missing observations, it may at times be advisable to
estimate these missing observations given the data in hand. Several procedures have been suggested in the literature to accomplish this, perhaps the most widely used being the Kalman smoother. While the theoretical properties of these techniques have been well documented, there remain some rather subtle issues that may influence their performance in practice. This paper attempts to address some of the errors that can be made during interpolation and distribution if problems associated with model selection and various data transformations are not appropriately addressed.

While the discussion is in terms of regularly spaced missing observations and applied to only a limited number of ARIMA models and types of data transformations, the results are suggestive of an important practical consideration. That is, when estimating missing data, failure to explicitly account for data transformations such as differences, logs, or ratios, may lead to serious errors in estimation.
Table 1. Measurement Vector (H) for Various Combinations of \( n \) and \( d \) — Stationary SSR

<table>
<thead>
<tr>
<th>Integration Order</th>
<th>Stock - Flow</th>
<th>Degree of Aggregation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( n = 2 )</td>
</tr>
<tr>
<td>( d = 0 )</td>
<td>( \lambda = 0 )</td>
<td>(1)</td>
</tr>
<tr>
<td></td>
<td>( \lambda = 1 )</td>
<td>(1 1)</td>
</tr>
<tr>
<td>( d = 1 )</td>
<td>( \lambda = 0 )</td>
<td>(1 1)</td>
</tr>
<tr>
<td></td>
<td>( \lambda = 1 )</td>
<td>(1 2 1)</td>
</tr>
<tr>
<td>( d = 2 )</td>
<td>( \lambda = 0 )</td>
<td>(1 2 1)</td>
</tr>
<tr>
<td></td>
<td>( \lambda = 1 )</td>
<td>(1 3 3 1)</td>
</tr>
</tbody>
</table>

NOTE: I have assumed that \( \eta \geq p \). If \( \eta < p \), \( p-\eta \) zeros need to be appended to the end of \( H \).
Table 2. Measurement Vector (H*) for Various Combinations of \( n \) and \( d \) — Nonstationary SSR

<table>
<thead>
<tr>
<th>Integration Order</th>
<th>Stock - Flow</th>
<th>Degree of Aggregation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \lambda = 0 )</td>
<td>( n = 2 )</td>
</tr>
<tr>
<td>( d = 1 )</td>
<td>(1 1)</td>
<td>(1 1)</td>
</tr>
<tr>
<td>( \lambda = 1 )</td>
<td>(1 2 2)</td>
<td>(1 2 3 3)</td>
</tr>
<tr>
<td>( d = 2 )</td>
<td>(1 2 -1)</td>
<td>(1 2 -1)</td>
</tr>
<tr>
<td>( \lambda = 1 )</td>
<td>(1 3 5 -3)</td>
<td>(1 3 6 9 -6)</td>
</tr>
</tbody>
</table>

NOTE: I have assumed that \( \eta \geq p \). If \( \eta < p \), \( p - \eta \) zeros need to be inserted immediately before the last \( d \) terms in \( H^* \).
Table 3. Theil’s U Statistics for Benchmark and Smoothed Monthly Estimates

<table>
<thead>
<tr>
<th>Series Type</th>
<th>Estimation</th>
<th>Model Type (in logs)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AR(1)</td>
<td>ARMA(1,2)</td>
<td>ARIMA(2,1,1)</td>
<td></td>
</tr>
<tr>
<td>flow</td>
<td>RATS</td>
<td>1.36 e-3</td>
<td>1.52 e-3</td>
<td>3.70 e-1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>KS(log)</td>
<td>5.63 e-1</td>
<td>5.26 e-1</td>
<td>7.11 e-1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>1.37 e-3</td>
<td>1.54 e-3</td>
<td>3.13 e-1</td>
<td></td>
</tr>
<tr>
<td>stock</td>
<td>RATS</td>
<td>1.62 e-3</td>
<td>1.79 e-3</td>
<td>3.15 e-1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>1.61 e-3</td>
<td>1.86 e-3</td>
<td>3.28 e-1</td>
<td></td>
</tr>
<tr>
<td>flow</td>
<td>RATS</td>
<td>6.09 e-1</td>
<td>5.30 e-1</td>
<td>7.18 e-1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>KS(ratio)</td>
<td>1.62 e+2</td>
<td>1.19 e+2</td>
<td>1.15 e+0</td>
<td></td>
</tr>
<tr>
<td>stock</td>
<td>KS(ind)</td>
<td>5.06 e-1</td>
<td>4.23 e-1</td>
<td>6.45 e-1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>6.11 e-1</td>
<td>5.32 e-1</td>
<td>7.84 e-1</td>
<td></td>
</tr>
</tbody>
</table>

NOTES: The notation 1.0 e-1 indicates 1.0 to the -1 power. The estimation categories are as follows: (i) RATS -- RATS DISTRIB or INTERPOL procedures where appropriate; (ii) KS(log): Kalman smoother using logged arithmetic sum without correction; (iii) KS(ind): estimates formed using individual KS flow and stock estimates; (iv) KS(ratio): Kalman smoother using ratio of aggregate data with log correction but without ratio correction; and (v) KS: Kalman smoother using logarithm and ratio correction where appropriate. The parameters for the three ARIMA models are given in section 3.1.
REFERENCES


Figure 1a. Distribution of a Differenced ARIMA(2,1,2) Flow Process

Figure 1b. Distribution of a Differenced ARIMA(2,1,2) Stock Process
Figure 2a. Log Correction and Distribution -- AR(1) Process

Figure 2b. Log Correction and Distribution -- Differenced ARIMA(2,1,2) Process
Figure 3a. Ratio Estimates -- ARMA(1,2) Process

Figure 3b. Ratio Estimates -- Differenced ARIMA(2,1,2) Process