Standard methods for imputing missing values in financial panel/time series data[1]

Philip Kokic

Working Paper Series
No. 2, June 2001

QANTARIS GmbH

Hostatostraße 25
D-65929 Höchst
Frankfurt, Germany
Tel.: ++49-69/3140 2311
Fax: ++49-69/3140 2323
Email: qantaris@freenet.de

[1]EUREDIT WP 5.7: Part A of deliverables D 5.7.1 and D 5.7.2

1 Introduction

The purpose of this note is to describe the standard, most commonly used approaches for imputing missing values in financial time series. Although more sophisticated approaches may be used by individual financial institutions and data providers, specific information about their methodologies is usually unavailable.

The data examined in the EurEdit project is confined to shares, European style call and put options on certain of these shares, and bonds (non-callable), as well as several indexes. There are four types of standard methods that we consider: last-value carried forward and linear interpolation, which are applied to all types of financial instruments; Black Scholes pricing, which only applies to the options time series, and finally term-structure pricing which only applies to the bond time series. The purpose of this paper is to describe each of these basic methods and to present some results for these using the evaluation criteria described in Insiders (2000).

All of these methods have the feature that they operate on a single time series (except that other instruments may be used to derive volatilities or interest rates). Otherwise they largely ignore the information contained in the other instruments and indexes that is relevant for prediction. For this reason, one would expect that these basic methods can be significantly improved upon. This is one of the main objectives of work package 5.7 in the EurEdit project.

2 Last-value carried forward and linear interpolation

Both the last-value carried forward and linear interpolation methods are extremely simplistic and hardly need definition other than for completeness. Both methods are frequently used by many data providers and financial institutions for imputing missing values in financial time series, particularly the last-value carried forward. In summary, for the last-value carried forward method the price P_{it} for instrument i at the current time point t is

$$\hat{P}_{it} = P_{i,t-1}. \tag{2.1}$$

Of course, if the price is not known at time $t-1$, then it is carried forward from an earlier time point.

For the linear interpolation method we assume that the price is know at two time points $r < t < s$. Then

$$\hat{P}_{it} = \frac{(t-s)P_{r,t-1} + (r-t)P_{s,t-1}}{r-s}. \tag{2.2}$$

If either r or s does not exist, for example at the beginning or end of the price time series, then another method must be used at these extremes.

3 The Black-Scholes put and call pricing formulae

The derivation of this famous formula is standard in many financial texts, see for example Hull (1997), and so there is no need to describe it in detail here. In summary, it is assumed that the underlying asset (in this case a share) follows exponential Brownian motion with volatility σ. Let

- S_{it} be the price of the underlying asset at time t,
- X_i be the excise price of the option at maturity,
- τ_i be the time to maturity of the option (i.e. it matures at $t + \tau_i$), and
- r be the annual interest rate.

The Black-Scholes formula for the price of a call option at time s is:

$$\hat{P}_{it} = S_{it}\Phi(d_1) - X_i e^{-r\tau_i}\Phi(d_2), \tag{3.1}$$

where

$$d_1 = \frac{\ln(S_{it}/X_i) + (r + \sigma^2/2)\tau_i}{\sigma\sqrt{\tau_i}},\ d_2 = d_1 - \sigma\sqrt{\tau_i}, \tag{3.2}$$

and Φ is the standard normal cumulative distribution function. The corresponding formula for a put option is:

$$\hat{P}_{it} = X_i e^{-r\tau_i}\Phi(-d_2) - S_{it}\Phi(-d_1). \tag{3.3}$$

Both (3.1) and (3.3) are used as a standard in financial institutions to impute the price of a call or put option. The value of sigma is typically estimated from other option prices from the current trading day by calibrating either of the above two formulae to observed prices. These values are usually referred to as implied volatilities.

4 Bond pricing using a term structure model

Consider a set of bonds, B_t, where t denotes today. Assume that we know the price histories $(P_{it})_{i \in B_t}$ of all bonds. Our aim in this section is to estimate a pricing function based on the term structure (interest rate function) $r_t(\tau_i)$ of a risk free zero coupon bond with maturity date $t + \tau_i$. In finance this is considered to be the standard method for imputing bond prices. We describe how this is done below.

Consider a bond issued by firm i at time t. Let

- J_i denote the set of cash flows for bond i,
- c_{ij} the j-th cash flow, $j \in J_i$,
- τ_{ij} the time ahead (in years) from time t that cash flow c_{ij} occurs (this can be negative), and
- $r_t(\tau_{ij})$ the discretely compounded annual interest rate at time t associated with cash flow c_{ij}.

The (model) price $\tilde{P}_{it}$ of bond i at time t can then be written as

$$\tilde{P}_{it} = \sum_{j \in J_i} \frac{c_{ij}}{\big(1 + r_t(\tau_{ij})\big)^{\tau_{ij}}}, \tag{4.1}$$

and

$$P_{it} = \tilde{P}_{it} + \varepsilon_{it}, \tag{4.2}$$

where ε_{it} has mean zero. In equation (4.1) all variables are known except the interest rates $r_t(\tau)$. The standard approach is to model this interest rate function as a polynomial in τ:

$$r_t(\tau) = \sum_{k=0}^{K} \alpha_{tk} \tau^k, \tag{4.3}$$

where K is usually set to the value 3 or 4.

Combining (4.1)-(4.3) we see that the term structure model is actually a non-linear statistical model, and that the parameters $\alpha_0, \ldots, \alpha_K$ can be estimated by minimise the least squares criteria:

$$\sum_{i \in B_t} (P_{it} - \tilde{P}_{it})^2. \tag{4.4}$$

Minimisation of this function is achieved using standard techniques and so there is no need to go into details in this paper. The resulting estimates $\hat{\alpha}_0, \ldots, \hat{\alpha}_K$ are plugged back into equation (4.3), which is then applied to (4.1) to produce the pricing function:

$$\hat{\tilde{P}}_{it} = \sum_{j \in J_i} \frac{c_{ij}}{\left(1 + \hat{r}_t(\tau_{ij})\right)^{\tau_{ij}}}, \tag{4.5}$$

where

$$\hat{r}_t(\tau) = \sum_{k=0}^{K} \hat{\alpha}_{tk} \tau^k. \tag{4.6}$$

The function (4.5) can be used to price other bonds (on day t), or to impute missing values.

5 Analysis

Analysis of the data was performed using the financial panel/time series data from the Eur$^{\text{E}}$dit project. For a full description of this data and how missing observations were generated see the documentation associated with the data. A total of 87 daily time series covering the time period from the beginning of 1995 to the end of 1999 were used in the analysis, 36 of which are option time series and 36 of which are bond time series. The first two methods, last value carried forward (LVCF) and linear interpolation (LIP), was applied to all time series, the Black-Scholes pricing formula (BlackScholes) was applied to the options time series only, and the term structure model (TermStruct) was applied to the bond time series only.

Assessment was performed on the basis of two criteria, distributional accuracy and prediction accuracy as defined in Chambers (2000). Note that a fuller set of assessments will be performed in a later stage of the Eur$^{\text{E}}$dit project. For the first assessment criterion the Wald statistic was used, see expression (14) of Chambers (2000). Specifically, this statistic and the corresponding p-value, computed on the basis of a χ^2 approximation, was determined over all imputed observations separately for each time series. The resulting set of p-values were then summarised using box plots as shown in figures 1-3. Note that in these figures small values of p close to zero indicate significant departure from preservation of distribution. For predictive accuracy a relative version of expression (19) in Chambers (2000) with $w_i = 1/(\text{number of imputes})$ was used. This statistic

can be interpreted as the average relative error of imputation. Again it was computed separately for each time series and then the set of resulting statistics were summarised using box plots (figures 4-6).

To briefly describe these results let us begin by looking at distributional accuracy. Examining figures 1 and 3, one immediately sees that the analytical based methods (BlackScholes and TermStruct) perform worse than LVCF and LIP and, not surprisingly, all methods perform worse the greater degree of missingness. The method that seems to hold up best against this downward trend with increasing degree of missingness is LVCF, while the worst seems to be TermStruct. There is little difference in performance between type of instrument, although there is a hint of evidence that distributional accuracy of bond prices can be best achieved using basic imputation methods (figure 2).

In terms of predictive accuracy a similar story holds: both the simplistic methods, LVCF and LIP perform much better that either of the analytical based methods, with the Termstrut method performing worst of all. Some erratic behaviour of the linear interpolation method can be observed. A higher degree of missingness seems to reduce the effectiveness of LVCF, while it has virtually no effect on the other methods. In general, option prices appear to be more difficult to predict accurately than other instruments although, somewhat surprisingly, the simplistic methods seem to work better than the more sophisticated pricing formulae in this case.

6 Remarks

From the analysis presented above it appears to be the case that the standard financial pricing formula (Black-Scholes for options and term structure methods for bonds) are worse than the simplistic last value carried forward and linear interpolation methods.

In order to develop a method that is better, it would seem necessary to incorporate the last value carried forward into any new imputation approach. This can easily be achieved by taking the log-returns of the data before defining the imputation model.

References

Chambers, R. (2000). Evaluation Criteria for Statistical Editing and Imputation. EUREDIT working paper, University of Southampton, Southampton, UK.

Hull, J. C. (1997). *Options, Futures, and Other Derivatives.* Upper Saddle River: Prentice Hall, Inc.

Insiders (2000). Specification of criteria needed for Insiders' financial time series data in the EUREDIT project. Working paper, Insiders Financial Solutions, GmbH, Mainz, Germany.

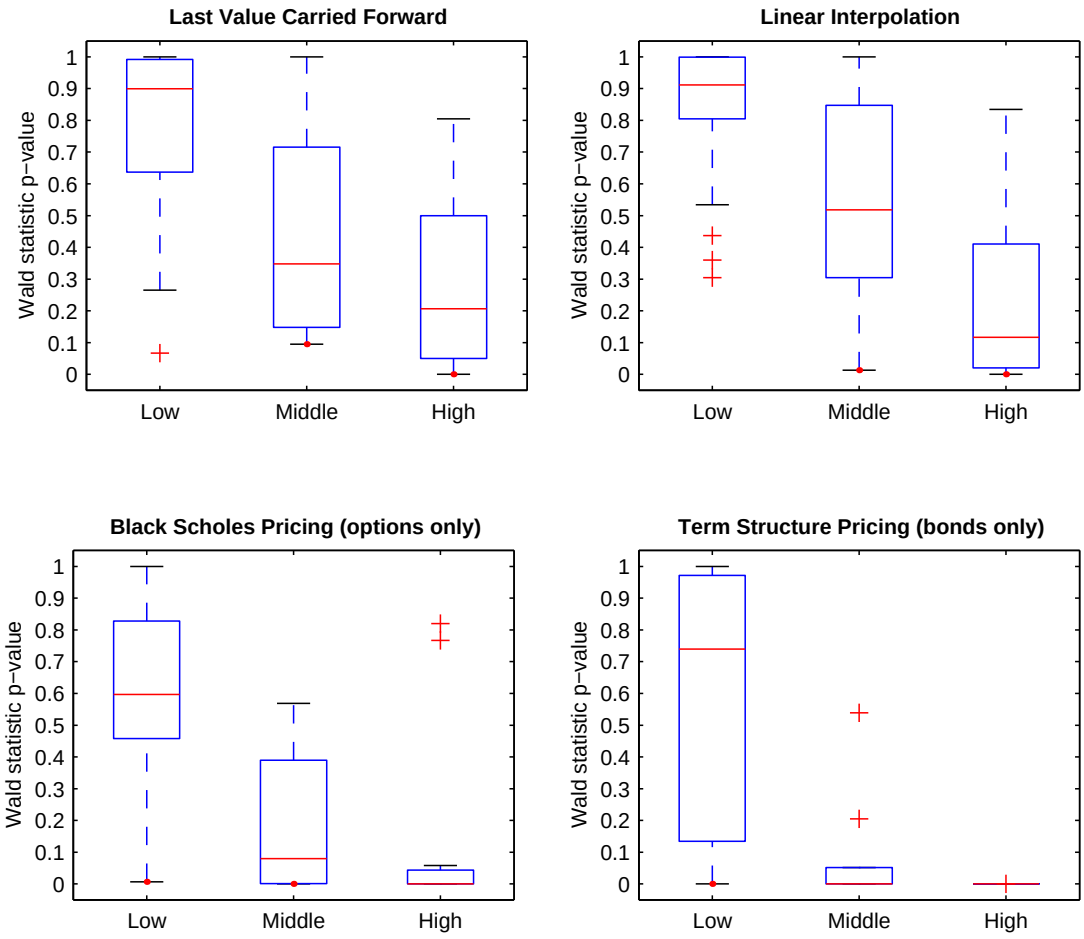

Figure 1: Distributional accuracy by degree of missingness (Wald statistic p-value)

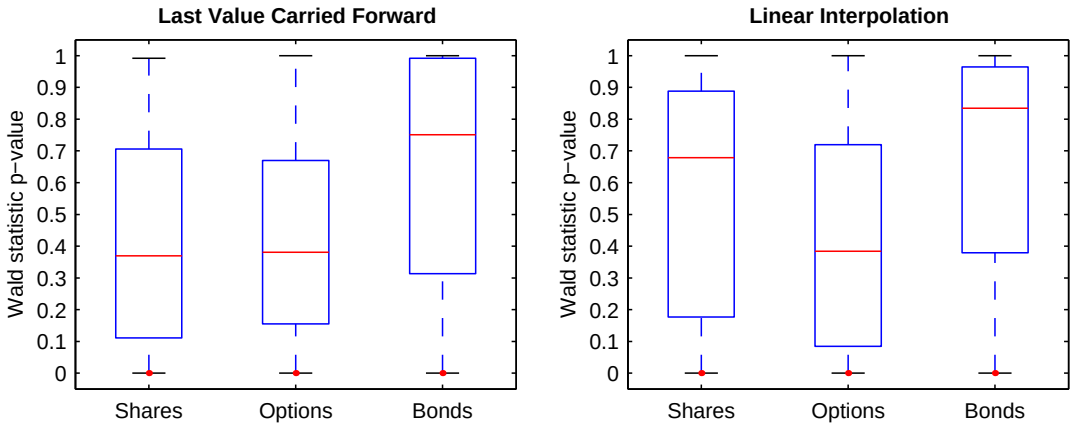

Figure 2: Distributional accuracy by variable type (Wald statistic p-value)

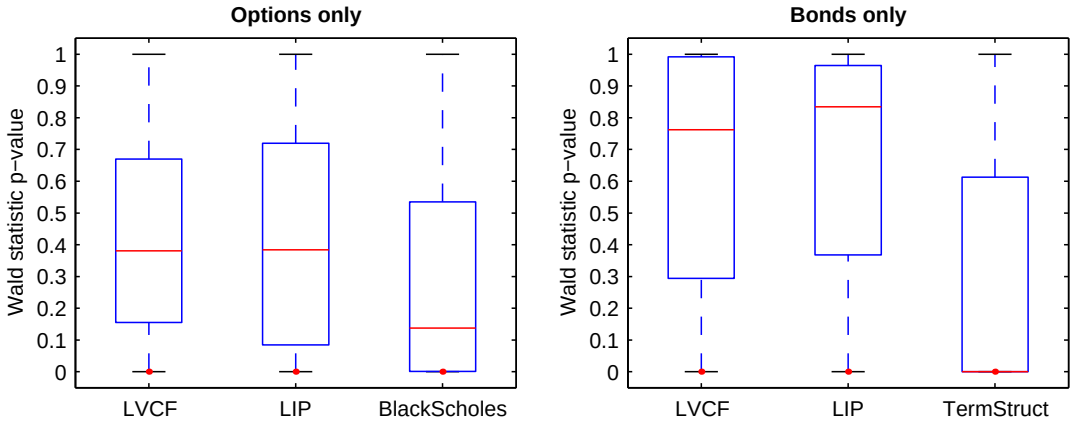

Figure 3: Distributional accuracy by method of imputation (Wald statistic p-value)

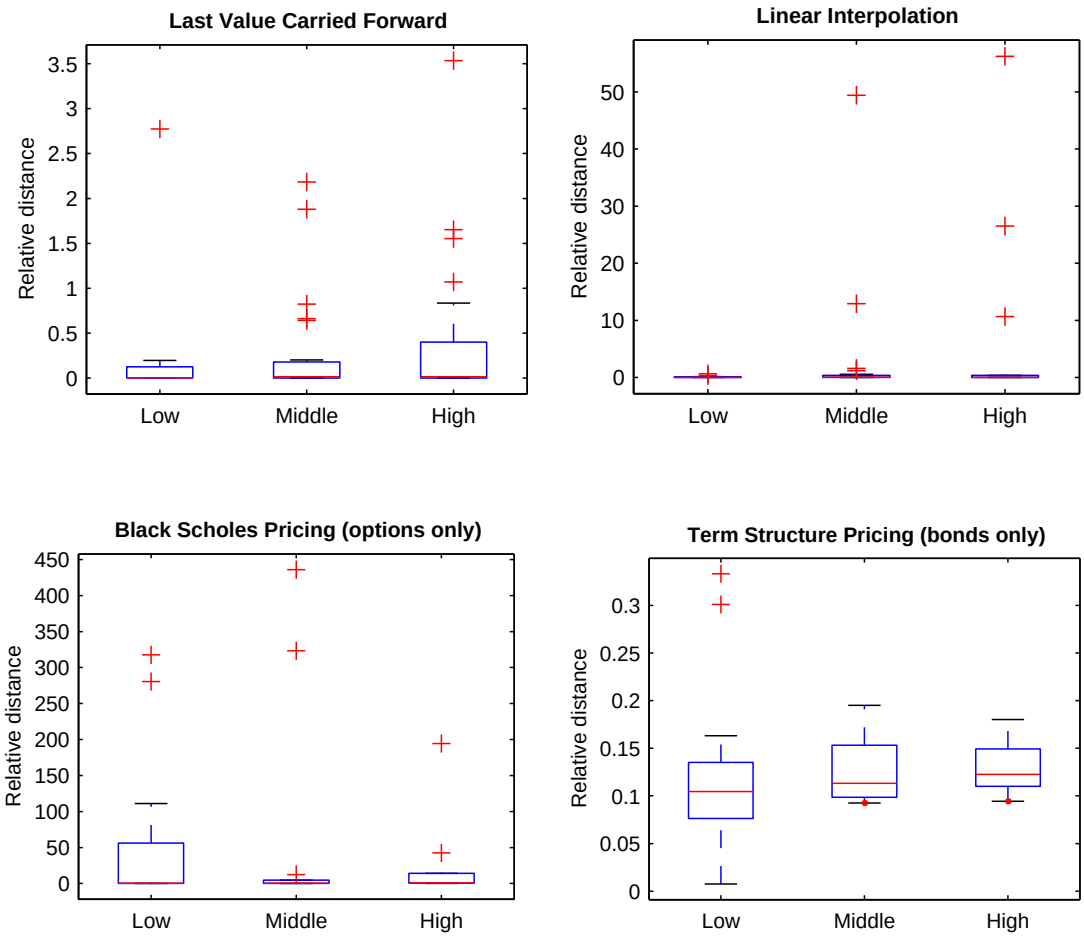

Figure 4: Predictive accuracy by degree of missingness

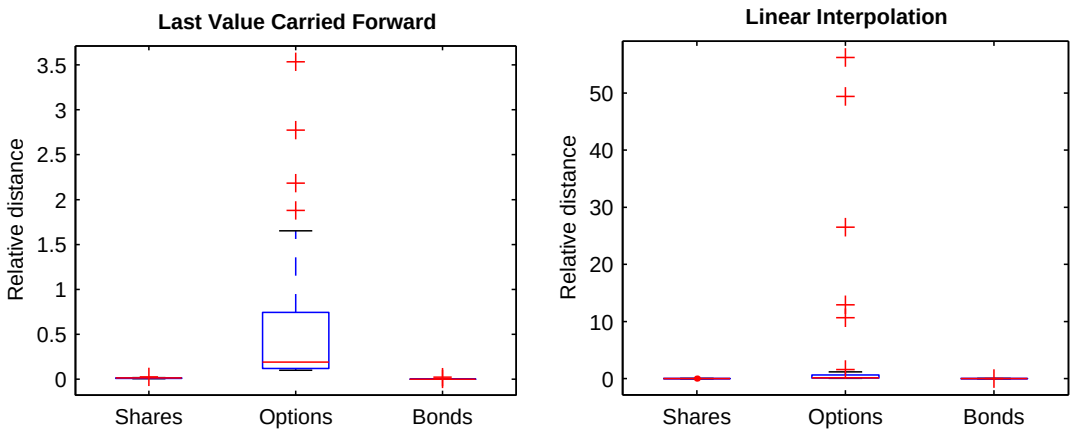

Figure 5: Predictive accuracy by variable type

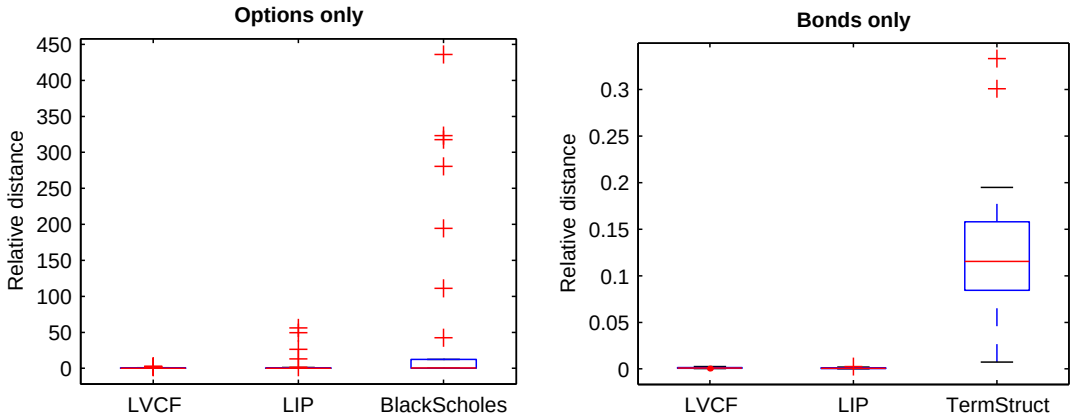

Figure 6: Predictive accuracy by method of imputation