LCERPA
Laurier Centre for Economic Research & Policy Analysis

LCERPA Working paper No. 2024-5

October 2024

Predictable by Construction: Assessing Forecast Directional Accuracy of Temporal Aggregates

Martin McCarthy
Reserve Bank of Australia

Stephen Snudden
Wilfrid Laurier University

Predictable by Construction: Assessing Forecast Directional Accuracy of Temporal Aggregates*

Martin McCarthy
Reserve Bank of Australia

Stephen Snudden†
Wilfrid Laurier University

October 3, 2024

Abstract

In macroeconomics, forecasted data are often constructed as an aggregate over some time interval, such as monthly average prices or quarterly total output. We demonstrate that assessing the null of no directional accuracy requires computing the sign of change relative to the latest disaggregated observation rather than the latest aggregated observation. Changes are predictable *by construction* when assessed relative to the latest aggregated observation, resulting in inflated type I error, and loss of power under the alternative. In contrast, assessing the change in the temporal aggregate relative to the latest disaggregated observation results in expected success ratios of 0.5, hypothesis tests with the intended type I error rate, and high power under the alternative.

JEL classification: C1, C43, C53, F47
Keywords: Directional Accuracy, Temporal Aggregation, Forecasting and Prediction Methods

[*] The views expressed in this paper are those of the authors; no responsibility for them should be attributed to the Reserve Bank of Australia.

[†] Corresponding author; Wilfrid Laurier University, Department of Economics; ORCID iD 0000-0002-4033-3235; Email: ssnudden@wlu.ca

1 Introduction

The directional accuracy of forecasts is arguably the most important forecast evaluation criteria, as it is most closely related to decision-making (Granger and Pesaran, 2000; Öller and Barot, 2000) and profits (Lai, 1990; Pesaran and Timmermann, 1995; Leitch and Tanner, 1991; Anatolyev and Gerko, 2005). However, the study of directional accuracy in macroeconomic forecasting involves unique challenges when dealing with temporally aggregated data, such as monthly averages or quarterly sums. Drawing parallels from finance, where returns are calculated using end-of-period observations to avoid spurious predictability, this paper examines a similar phenomenon when assessing directional accuracy in macroeconomics. Under the usual approach to assessing directional accuracy, the aggregate will appear predictable even if the underlying high-frequency series is a random walk. Since the direction of changes of a random walk are by definition unpredictable, the apparent predictability of the aggregate series arises *by construction.* This creates a risk that macroeconomists or policymakers may rely on models that appear useful even when their apparent good performance is spurious. This could also lead to policymakers attributing movements in the aggregates to the explanatory variables in the apparently useful forecasting model.

The issue is how one computes the sign of changes in the variable of interest. The standard approach uses the change in the aggregate relative to the last aggregate observation. For example, surveys have shown that this approach was applied universally for forecasts of period average primary commodity prices (Ellwanger and Snudden, 2023; Farag et al., 2024) and for period-average bilateral and effective exchange rates (McCarthy and Snudden, 2024). In this paper we demonstrate that only comparisons against the last disaggregated observation provide valid inference against the random walk hypothesis for forecast directional accuracy of temporally aggregated data. There are two main contributions.

The first contribution is to study the random walk null in the assessment of forecast directional accuracy of temporal aggregates. In particular, we focus on mean directional accuracy, commonly referred to as the Success Ratio, (SR). A SR is the proportion of times that a forecast correctly predicts the direction of changes over a sample period. A SR of 0.5 is interpreted as showing that a model is no better than a coin flip at guessing the direction in which a series will move. We prove that if one computes changes relative to the latest aggregate then the expected SR is above 0.5 even when the high-frequency data is a random walk, and hence unpredictable. In contrast, a SR of 0.5 arises only when directional accuracy of the temporal aggregate is computed relative to the latest end-of-period level. Therefore, a forecaster can maintain the goal of forecasting the temporal aggregate while comparing against the random walk forecast only by evaluating directional accuracy relative to the latest high-frequency observation.

Second, we conduct simulation analysis to evaluate the validity of hypothesis testing for forecasts of temporal aggregates. Under the random walk null hypothesis,[1] we quantify that sizable gains in SRs of over 20 percentage points arise spuriously for period-average forecasts of weekly, monthly, and quarterly data. We then conduct power analysis of Pesaran and Timmermann (2009) tests which are commonly used to test the null that these two categorical variables are independent of one another, which is interpreted as no directional accuracy.[2] We find that the null hypothesis is rejected more frequently than the intended level of significance when changes in model-based forecasts are computed relative to the latest aggregate (an example of spurious predictability), but not when changes are relative to the latest end-of-period observation. Under the alternative, when the high-frequency data is persistent but not a random walk, we find the test has reduced power to detect this violation of the null when changes are computed relative to the latest aggregate. We conclude that the test of Pesaran and Timmermann (2009) exhibits substantial power and valid inference for forecasts of temporal aggregates when directional accuracy is assessed relative to the latest high-frequency observation.

An empirical application to United States Treasury yields is then presented, which confirms the importance of the insights herein to the predictability of temporal aggregates in practice.

The findings help bridge the gap between the theoretical and practical assessment of forecast directional accuracy for temporally aggregated macroeconomic variables. The seminal works of Working (1960), Weiss (1984) and Marcellino (1999), demonstrate that the aggregation of (V)ARIMA processes leads to inherent predictability. The current paper furthers the understanding of this effect on forecast accuracy, previously studied in terms of mean-squared precision by Tiao (1972), Wei (1978), and Ellwanger and Snudden (2023), by extending the analysis to the assessment of directional accuracy. Moreover, our analysis extends hypothesis testing of directional forecasts (Pesaran and Timmermann, 1992, 2009) to the context of temporal aggregation. Together, the results indicate that care must be taken to assess the directional accuracy of forecasts of temporal aggregates to avoid spurious predictability.

[1] I.e. when the high-frequency data is a random walk, and hence unpredictable.

[2] One can view the sign of the actual change in the variable as a categorical variable, and the sign of the forecasted change in the variable as another categorical variable.

2 Success Ratios for Directional Accuracy

We analyse forecast directional accuracy when a high-frequency series is temporally aggregated to a lower time frequency (e.g. daily to monthly). Of paramount importance is the case where the underlying high-frequency series is a random walk, so the direction of change is unpredictable by definition. In this case, the SR (mean directional accuracy) for any method of forecasting the high-frequency series will be 0.5 in expectation, though the SR may be higher or lower in a finite sample by chance. In contrast, a SR above 0.5 would be evidence that a forecast is more useful than random guesses in predicting the direction of change.

To make our claims precise, we introduce some notation. For concreteness, we present the theorem in terms of month-averages of a daily series, but it can be applied to any frequency. The daily level on day $t = 1, 2, \ldots, T$ is denoted D_t. The number of days in a month is denoted n, and is assumed equal across all months. The average level in month $m = 1, 2, \ldots, M$, where $T = nM$, is $A_m = \frac{1}{n}\sum_{i=1}^{n} D_{(m-1)n+i}$. The end-of-month level in month m is denoted Z_m and equals D_{mn}. The theorem (and the simulations in the next section) are stated in terms of period-averages, but also apply to period-sums.[3]

Consider a forecaster in month m making a forecast for the level of the average h-months ahead, A_{m+h}. The directional accuracy of a 'candidate' forecast is defined as:

$$s_{m,h}^{\text{candidate vs bench}} \equiv \mathbb{1}\left\{sgn(A_{m+h} - \hat{A}_{m+h|m}^{bench}) = sgn(\hat{A}_{m+h|m}^{candidate} - \hat{A}_{m+h|m}^{bench})\right\}, \tag{1}$$

where $\hat{A}_{m+h|m}^{candidate}$ is the candidate forecast, $\hat{A}_{m+h|m}^{bench}$ is the benchmark forecast, and $\mathbb{1}[\cdot]$ is an indicator function with 1 if true and 0 otherwise. The sign function is:

$$sgn(x) \equiv \begin{cases} 1 & x > 0 \\ -1 & x \leq 0. \end{cases}$$

The benchmark forecast is always a no-change forecast, so directional accuracy can be interpreted as whether the candidate forecast correctly guessed the direction in which the series moved.

[3] Suppose one is forecasting a period-sum $\left(\sum_{i=1}^{n} D_{(m-1)n+i}\right)$ One can equivalently forecast the period-average $\left(\frac{1}{n}\sum_{i=1}^{n} D_{(m-1)n+i}\right)$ and multiply the result by the result by n. Multiplying by n does not change the sign of actual or forecasted changes. This has two implications. Firstly, the theorem in this section can be applied to period-sums, as multiplying by n doesn't affect success ratios. Secondly, the simulations in section 3 apply to period-sums, as the hypothesis test for directional accuracy is a Pesaran and Timmermann (2009) test of the independence of the sign of actual and forecasted changes, which are unaffected by multiplying by n.

When the benchmark is the month-average no-change (avg), directional accuracy of the candidate forecast equals 1 if the change in the month-average series, $(A_{m+h} - A_m)$ is in the same direction expected by the candidate, $\left(\hat{A}^{candidate}_{m+h|m} - A_m\right)$.

$$s^{\text{candidate vs avg}}_{m,h} \equiv \mathbb{1}\left\{sgn(A_{m+h} - A_m) = sgn\left(\hat{A}^{candidate}_{m+h|m} - A_m\right)\right\}, \tag{2}$$

An alternative benchmark is the end-of-month no-change (EOM), which is the last disaggregated observation in the forecasters information set. In this case, the change in the series is measured from the last day of the forecast month m to the month-average of the future month $(m + h)$.

$$s^{\text{candidate vs EOM}}_{m,h} \equiv \mathbb{1}\left\{sgn(A_{m+h} - Z_m) = sgn\left(\hat{A}^{candidate}_{m+h|m} - Z_m\right)\right\}, \tag{3}$$

The SR is the sample mean of the directional accuracy of a sample of forecasts (Pesaran and Timmermann, 1992). Given a full sample of months $m = 1, ..., M$, one selects a subsample of months $m_{min}, (m_{min} + 1), ..., m_{max}$, where $m_{min} \geq 1$ and $m_{max} \leq (M - h)$. For each month in the subsample, a candidate model is used to generate an h-step-ahead forecast using the data available up to the end of that month. The success ratio of this sample of h-step-ahead forecasts is defined as:

$$SR^{\text{candidate vs bench}}_h \equiv \frac{1}{(m_{max} - m_{min} + 1)} \sum_{m=m_{min}}^{m_{max}} s^{\text{candidate vs bench}}_{m,h} \tag{4}$$

Suppose the data generating process of the daily frequency data is a random walk:

$$D_t = D_{t-1} + e_t \quad \forall\ t \tag{5}$$

where the initial level D_0 is a constant, each error e_t is independent of the error in other periods and the past levels of all other variables, and the distribution of each error e_t is continuous and symmetric about 0.

We now show that differences between period-average and end-of-period exchange rates can be expressed in terms of the error in the random walk. This will be used in the proof of the main theorem.

Lemma 1. *The difference between the current end-of-month level and current month-average level,*

$(Z_m - A_m)$, is a weighted sum of errors in month m.

$$Z_m - A_m = \sum_{j=1}^{n} e_{(m-1)n+j} - \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{i} e_{(m-1)n+j}\right) \tag{6}$$

Lemma 2. *The difference between the future month-average and current end-of-month, $(A_{m+h} - Z_m)$, is a weighted sum of errors in months $(m+1)$ to $(m+h)$.*

$$A_{m+h} - Z_m = \sum_{j=n+1}^{hn} e_{(m-1)n+j} + \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{i} e_{(m+h-1)n+j}\right) \tag{7}$$

Proof of lemmas 1 and 2. These are derived from repeated substitution, see appendix A □

These lemmas show that the terms $(Z_m - A_m)$ and $(A_{m+h} - Z_m)$ are linear combinations of the random walk errors of different months. Since the errors are independent of one another, these terms must be independent of each other. Independence implies that the sign of the former term is not useful in predicting the sign of the latter term, which plays a key role in the proof that follows.

We now introduce our main theorem, which shows how the SRs compare to a half in the case when the high-frequency series is a random walk.

Theorem 1. *Suppose the data generating process of the daily frequency data is a random walk, as given by equation (5).*

(a) Suppose the candidate is end-of-month no-change and the benchmark is month-average no-change. Then the expected SR is strictly greater than half.

$$E\left[SR_h^{EOM\ vs\ avg}\right] = E\left[DA_{m,h}^{EOM\ vs\ avg}\right] > \frac{1}{2}$$

(b) Suppose the candidate is any function of the levels of variables up to the time the forecast is made, and the benchmark is end-of-month no-change. Then the expected SR is a half.

$$E\left[SR_h^{candidate\ vs\ EOM}\right] = E\left[DA_{m,h}^{candidate\ vs\ EOM}\right] = \frac{1}{2}$$

Proof of Theorem 1(a). For convenience, define terms u and v by:

$$u \equiv \mathbb{P}[(A_{m+h} - Z_m) > 0 \cap (Z_m - A_m) > 0]$$

$$v \equiv \mathbb{P}[(A_{m+h} - A_m) > 0 \cap (Z_m - A_m) > 0]$$

To prove theorem 1(a), we will show that:

$$E[DA_{m,h}^{\text{EOM vs avg}}] = 2v > 2u = 2 \times \frac{1}{4} = \frac{1}{2} \tag{8}$$

First, we prove the first equality in equation (8), which is that directional accuracy equals $2v$. When the candidate is end-of-month no-change and the benchmark is month-average no-change, directional accuracy can be written:

$$\begin{aligned}
E\left[DA_{m,h}^{\text{EOM vs avg}}\right] &= E\left[\mathbb{1}\left\{sgn(A_{m+h} - A_m) = sgn(Z_m - A_m)\right\}\right] \\
&= \mathbb{P}[sgn(A_{m+h} - A_m) = sgn(Z_m - A_m)] \\
&= \mathbb{P}[(A_{m+h} - A_m) > 0 \cap (Z_m - A_m) > 0] + \mathbb{P}[(A_{m+h} - A_m) \leq 0 \cap (Z_m - A_m) \leq 0] \\
&= \mathbb{P}[(A_{m+h} - A_m) > 0 \cap (Z_m - A_m) > 0] + \mathbb{P}[(A_{m+h} - A_m) < 0 \cap (Z_m - A_m) < 0] \qquad (9) \\
&= 2 \times \mathbb{P}[(A_{m+h} - A_m) > 0 \cap (Z_m - A_m) > 0] = 2v \qquad (10)
\end{aligned}$$

Equality (9) holds because the terms $(A_{m+h} - A_m)$ and $(Z_m - A_m)$ are continuous random variables, as they are the sum of errors (by lemmas 1 and 2), and the errors are continuous. Equality (10) holds because the terms $(A_{m+h} - Z_m)$ and $(Z_m - A_m)$ are symmetric about zero, since they are sums of errors that are symmetric about zero.

Second, we prove the inequality in equation (8), which is $2v > 2u$, or equivalently, $v > u$.

$$\begin{aligned}
v &\equiv \mathbb{P}[(A_{m+h} - A_m) > 0 \cap (Z_m - A_m) > 0] \\
&= \mathbb{P}[((A_{m+h} - Z_m) + (Z_m - A_m)) > 0 \cap (Z_m - A_m) > 0] \\
&= \mathbb{P}[((A_{m+h} - Z_m) + (Z_m - A_m)) > 0 | (Z_m - A_m) > 0] \times \mathbb{P}[(Z_m - A_m) > 0] \\
&> \mathbb{P}[(A_{m+h} - Z_m) > 0 | (Z_m - A_m) > 0] \times \mathbb{P}[(Z_m - A_m) > 0] \qquad (11) \\
&= \mathbb{P}[(A_{m+h} - Z_m) > 0] \times \mathbb{P}[(Z_m - A_m) > 0] \\
&= u
\end{aligned}$$

Inequality (11) holds because $\mathbb{P}\left[x + y > 0 | y > 0\right] > \mathbb{P}\left[x > 0 | y > 0\right]$ for any numbers x and y.

Third, we show that u equals $\frac{1}{4}$.

$$u = \mathbb{P}[(A_{m+h} - Z_m) > 0] \times \mathbb{P}[(Z_m - A_m) > 0] \tag{12}$$

$$= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \tag{13}$$

Equality (12) holds because the terms $(A_m - Z_m)$ and $(A_{m+h} - Z_m)$ are independent of each other. We know the terms are independent because the former term is the sum of errors in m (see lemma 1), the latter is the sum of errors in $(m+1)$ to $(m+h)$, and $\{e\}$ is an independent sequence. Equality (13) holds because the terms $(A_m - Z_m)$ and $(A_{m+h} - Z_m)$ are symmetric about zero, since they are linear combinations of errors that are symmetric about zero. □

Proof of Theorem 1(b). The benchmark is the end-of-month no-change. Hence, expected directional accuracy can be written:

$$E\left[DA^{\text{candidate vs EOM}}_{m,h}\right] \equiv E\left[\mathbb{1}\left\{sgn(A_{m+h} - Z_m) = sgn\left(\hat{A}^{candidate}_{m+h|m} - Z_m\right)\right\}\right]$$

$$= \mathbb{P}\left[sgn(A_{m+h} - Z_m) = sgn\left(\hat{A}^{candidate}_{m+h|m} - Z_m\right)\right]$$

$$= \mathbb{P}\left[(A_{m+h} - Z_m) > 0 \cup \left(\hat{A}^{candidate}_{m+h|m} - Z_m\right) > 0\right] + \mathbb{P}\left[(A_{m+h} - Z_m) \leq 0 \cup \left(\hat{A}^{candidate}_{m+h|m} - Z_m\right) \leq 0\right]$$

$$= \mathbb{P}[(A_{m+h} - Z_m) > 0] \times \mathbb{P}\left[\left(\hat{A}^{candidate}_{m+h|m} - Z_m\right) > 0\right] + \mathbb{P}[(A_{m+h} - Z_m) \leq 0] \times \mathbb{P}\left[\left(\hat{A}^{candidate}_{m+h|m} - Z_m\right) \leq 0\right] \tag{14}$$

$$= \frac{1}{2} \times \mathbb{P}\left[\left(\hat{A}^{candidate}_{m+h|m} - Z_m\right) > 0\right] + \frac{1}{2} \times \mathbb{P}\left[\left(\hat{A}^{candidate}_{m+h|m} - Z_m\right) \leq 0\right] \tag{15}$$

$$= \frac{1}{2}$$

Equality (14) holds because $(A_{m+h} - Z_m)$ and $(\hat{A}^{candidate}_{m+h|m} - Z_m)$ are independent. This is true because: i) the former term depends solely on errors from $(m+1)$ to $(m+h)$, as shown in lemma 2; ii) the latter term depends solely on the daily exchange rate's initial level and errors up to month m, and the levels of other variables up to month m;[4] iii) the errors from $(m+1)$ to $(m+h)$ are independent of the errors up to month m and the past levels of all other variables, by the random walk assumption (5).

Equality (15) holds because the term $(A_{m+h} - Z_m)$ is the sum of the errors of a random walk (see lemma 2). The term is continuous and symmetric about zero because the errors are continuous

[4] A candidate forecast $\hat{A}^{candidate}_{m+h|m}$ only depends on variables up to the time it is made, so it only depends on the daily data's initial level and errors up to m, and on the levels of other variables up to m. The end-of-period level, $Z_m = D_{(m-1)n} + \sum_{j=1}^{n} e_{(m-1)n+j}$, depends only on the initial level and errors up to month m.

and symmetric about zero. Hence, the probability that the term is strictly positive is half, and the probability it is negative or zero is also a half. □

Theorem 1(a) shows that if the SR is computed relative to the latest period-average level, it is trivial to find candidates whose expected SR is above 0.5, with the end-of-period no-change being one such example. This is undesirable, as it implies that the SR will often suggest that a candidate model can guess the direction of changes, but this arises by construction, and hence does not indicate that the model is useful. Theorem 1(b) shows that an SR computed relative to the end-of-period level will equal 0.5 in expectation for any candidate model, which is appropriate given the underlying series is unpredictable.

3 Simulation Evidence

Simulation experiments now quantify the assessment of directional accuracy for forecasts of temporally aggregated data. Let the daily data, D_t, be represented by an autoregressive model with one lag, AR(1), $D_t = \rho D_{t-1} + e_t$, such that $\rho = 1$ is the random walk model, equation 5, and $e_t \sim N(0, 1)$.[5] The simulated data is aggregated, A_m to weekly, monthly, or quarterly frequency, with $n = 5,\ 21,$ or 62, respectively.[6] The baseline simulations burn the first 500 daily observations and use 40 years worth of daily data, consistent with applications where daily data has been available since the early-1980s. The simulations are described in terms of period-averages, but also apply to period-sums (see footnote 3).

3.1 Comparison of No-Change Forecasts

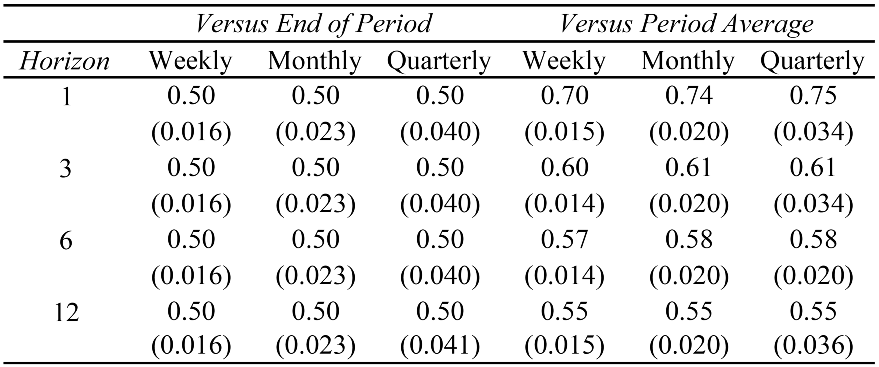
Out-of-sample period-average no-change forecasts, A_m, are compared to the end-of-period no-change forecasts, Z_m, and vice versa, see Table 1. In this table, the daily series is a random walk, $\rho = 1$, and hence unpredictable. The simulations help us quantify theorem 1. Columns 2 to 4 show that the expected SR of a period-average no-change forecast against the end-of-period no-change benchmark is 0.5, consistent with theorem 1(b).

In contrast, the expected SR of an end-of-month forecast against the month-average forecast is greater than 0.5, as shown by theorem 1(a). However, the theorem cannot say how much above 0.5.

[5] Success ratios and tests of Pesaran and Timmermann (2009) do not depend on the initial level or variance of the daily series.

[6] Temporally aggregated economic data are typically built using daily data such as simple averages of closing values on business days (e.g. exchange rates, commodity prices, and interest rates), which is why we assume there are $n = 5$ days in a week, $n = 21$ in a month and 62 in a quarter.

Table 1. Directional Accuracy is Expected when Compared to the Period Average No-Change



	Versus End of Period			*Versus Period Average*		
Horizon	Weekly	Monthly	Quarterly	Weekly	Monthly	Quarterly
1	0.50	0.50	0.50	0.70	0.74	0.75
	(0.016)	(0.023)	(0.040)	(0.015)	(0.020)	(0.034)
3	0.50	0.50	0.50	0.60	0.61	0.61
	(0.016)	(0.023)	(0.040)	(0.014)	(0.020)	(0.034)
6	0.50	0.50	0.50	0.57	0.58	0.58
	(0.016)	(0.023)	(0.040)	(0.014)	(0.020)	(0.020)
12	0.50	0.50	0.50	0.55	0.55	0.55
	(0.016)	(0.023)	(0.041)	(0.015)	(0.020)	(0.036)

Notes: Success ratios from forecasts of end-of-period and period average no-change forecasts. 5000 simulations, using 40 years of data. Standard deviation of the success ratios in brackets.

Table 1, columns 5 to 7, quantifies that the success ratios for the end-of-period no-change forecast relative to the period-average no-change forecast in the case where the random walk error, e_t, is normally distributed. The SR is 70, 74, and 75 percent at the one-step-ahead for weekly, monthly, and quarterly forecasts, respectively. Thus, sizable gains in directional accuracy are spuriously expected when forecasts are compared to the period-average no-change, even though the daily data is inherently unpredictable.

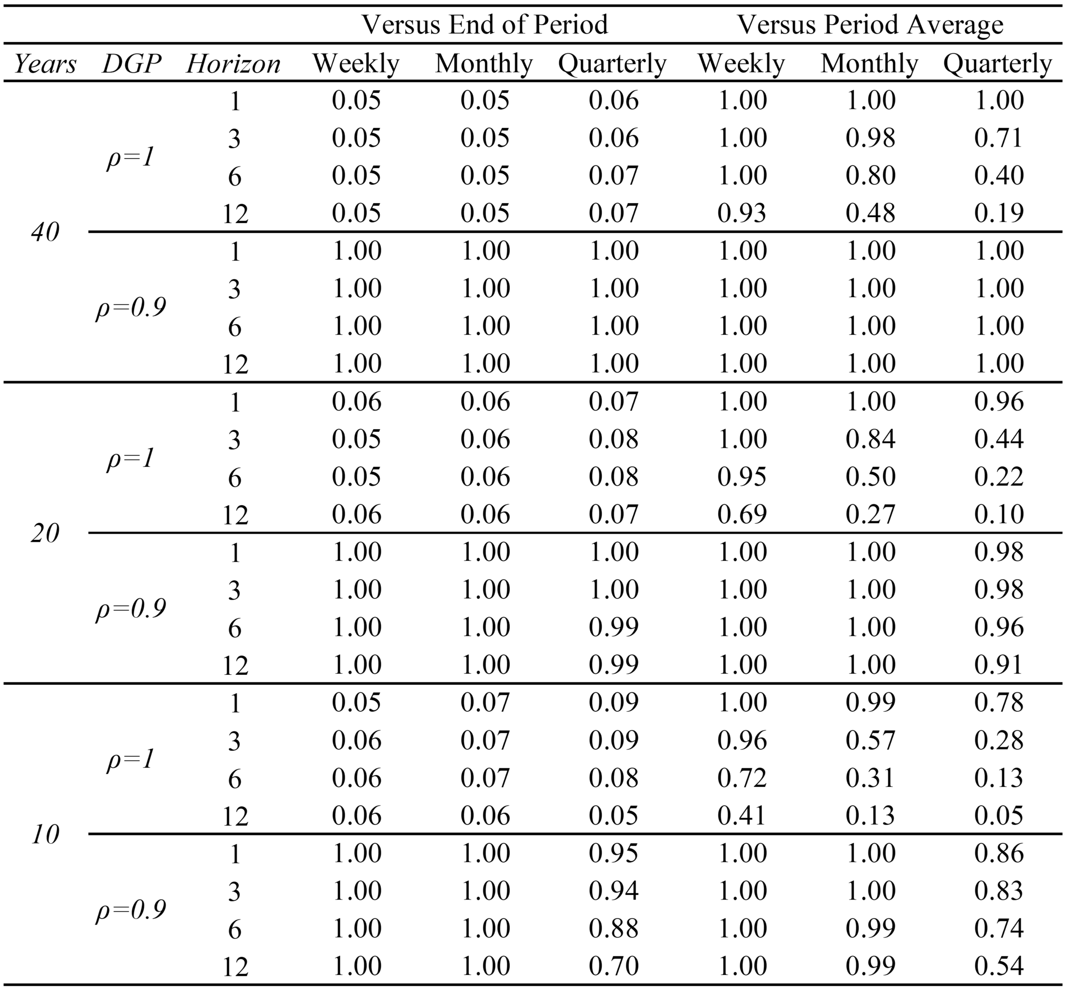
3.2 Tests of Directional Accuracy for Bottom-up Forecasts

Suppose one wanted to test the null hypothesis that the expected success ratio is 0.5. That is, the candidate forecast is no better than a coin flip at predicting the direction in which the actual outcome will differ from the benchmark. The sign of the actual change in the period-average $sgn(A_{m+h} - \hat{A}^{bench}_{m+h|m})$ and the sign of the forecasted change $sgn(\hat{A}^{candidate}_{m+h|m} - \hat{A}^{bench}_{m+h|m})$ are both categorical variables. Given a time series of these categorical variables, we test the null hypothesis that they are independent of each other using the Pesaran and Timmermann (2009) test.[7] This test is ideal for our situation, as it is valid even if each of the categorical random variables is serially correlated. Under the null of independence, the sign of the forecasted change contains no information about the sign of the actual change, so the forecasting method has no directional accuracy. However, if we reject the null of independence, then the forecasting method either has directional accuracy (if the SR is above 0.5) or could be reversed to obtain a method with directional accuracy (if the SR is below 0.5).

[7] The test uses Newey and West (1987) standard errors with $4(N/100)^{2/9}$ lags, where N is the observations in the forecast evaluation sample. The test statistics are evaluated using the standard normal.

To see how the benchmark itself affects power, we need to keep the candidate forecast the same and vary only the benchmark. A bottom up forecast is a good candidate forecast of the period average forecast of the AR(1) model. Specifically, an AR(1) model is estimated with ordinary least squares on the daily data. The estimate of ρ is used it to construct dynamic forecasts of the daily series, and then these forecasts are ex-post averaged to the desired lower frequency. We reserve the second half of the sample to evaluate out-of-sample forecasts, and reestimate the model with an expanding window in every period.

Table 2. Comparison Relative to the Period Average Results in Type I Error and Power Loss



			Versus End of Period			Versus Period Average		
Years	*DGP*	*Horizon*	Weekly	Monthly	Quarterly	Weekly	Monthly	Quarterly
40	*ρ=1*	1	0.05	0.05	0.06	1.00	1.00	1.00
		3	0.05	0.05	0.06	1.00	0.98	0.71
		6	0.05	0.05	0.07	1.00	0.80	0.40
		12	0.05	0.05	0.07	0.93	0.48	0.19
	ρ=0.9	1	1.00	1.00	1.00	1.00	1.00	1.00
		3	1.00	1.00	1.00	1.00	1.00	1.00
		6	1.00	1.00	1.00	1.00	1.00	1.00
		12	1.00	1.00	1.00	1.00	1.00	1.00
20	*ρ=1*	1	0.06	0.06	0.07	1.00	1.00	0.96
		3	0.05	0.06	0.08	1.00	0.84	0.44
		6	0.05	0.06	0.08	0.95	0.50	0.22
		12	0.06	0.06	0.07	0.69	0.27	0.10
	ρ=0.9	1	1.00	1.00	1.00	1.00	1.00	0.98
		3	1.00	1.00	1.00	1.00	1.00	0.98
		6	1.00	1.00	0.99	1.00	1.00	0.96
		12	1.00	1.00	0.99	1.00	1.00	0.91
10	*ρ=1*	1	0.05	0.07	0.09	1.00	0.99	0.78
		3	0.06	0.07	0.09	0.96	0.57	0.28
		6	0.06	0.07	0.08	0.72	0.31	0.13
		12	0.06	0.06	0.05	0.41	0.13	0.05
	ρ=0.9	1	1.00	1.00	0.95	1.00	1.00	0.86
		3	1.00	1.00	0.94	1.00	1.00	0.83
		6	1.00	1.00	0.88	1.00	0.99	0.74
		12	1.00	1.00	0.70	1.00	0.99	0.54

Notes: Rejection rate of the bottom-up forecasts using the Pesaran and Timmermann (2009) test relative to alternative no-change benchmarks. The daily random walk is given by $\rho = 1$; whereas $\rho = 0.9$ implies a predictable data generating process. 5000 simulations, using 40, 20, and 10 years worth of data. Tested at the 5 percent level.

The simulations in Table 2 provide evidence on the power of hypothesis tests of directional accuracy using 40, 20 and 10 years of data. When the daily data is unpredictable, $\rho = 1$, testing relative to the end-of-period no-change closely reflects the 5 percent significance level. Moreover, the corresponding SRs take on a value of 0.5, see Table A1. This is desirable as when ρ is 1, then the daily series is a random walk and by definition, no candidate model should exhibit improvements in directional accuracy.

Another way to implement the test of Pesaran and Timmermann (2009) is to compare relative to the period-average no-change benchmark. When forecasting a month-average at a 1-month horizon, the null is rejected 100 percent of the time when $\rho = 1$. This may be a valid test of whether $sgn(A_{m+h} - \hat{A}^{bench}_{m+h|m})$ is independent of $sgn(\hat{A}^{candidate}_{m+h|m} - \hat{A}^{bench}_{m+h|m})$. However, it is an invalid test of our underlying question, which is whether there are candidates with directional accuracy, since it rejects more than 5 percent of the time when the daily series is a random walk.

Now consider when the daily data is a predictable stationary AR(1) with $\rho = 0.9$. When comparing relative to the end of period no-change forecast, the test exhibits substantial power at all horizons, even with only tens of years of data (i.e. 5 years of forecasts). In contrast, when comparing relative to the period average no-change, the last three columns, the null is often rejected, but not quite as often. The testing of directional accuracy relative to the monthly average no-change exhibits loss of predictive power relative to comparisons against the end-of-period no-change.

Together, the findings show that care needs to be taken to assess directional accuracy of temporal aggregates. Assessing directional accuracy of the temporal aggregate relative to the period average no-change results in inflated type 1 error under the null and loss of power under the alternative. To assess directional accuracy, one should instead use the test of Pesaran and Timmermann (2009) relative to the end-of-period no-change.

4 Application

Our application focuses on forecasting the monthly average of the nominal yield on U.S. 10-Year Treasury securities. Previous studies have compared the directional accuracy of interest rate forecasts against either the period-average no-change benchmark or alternative models (see, e.g. Johannsen and Mertens, 2021). In this analysis, we test directional accuracy against both the monthly average and the end-of-month no-change benchmarks.

The daily closing data was obtained from the FRED database of the Federal Reserve Bank of St. Louis (ticker DGS10). End-of-month observations are the closing prices on the last trading day of each month, and the monthly average is calculated as the simple average of daily closing prices. This data is available in real-time and is not subject to historical revisions. The time series begins in January 1990, with the forecast evaluation period running from January 2000 to June 2023.

To construct a model-based forecast, we estimate an autoregressive moving average (ARIMA) model, with parameterization selected using information criterion (Akaike, 1974), consistent with Rossana and Seater (1995). The model is intentionally kept simple to serve as an illustrative example rather than to advocate for a specific forecasting method or to conclusively assess the predictability of the series. We employ a bottom-up forecasting approach: the ARIMA model is estimated using daily frequency data in differences using all available data at the end of each month using expanding window estimation. The estimated model then forecasts the daily changes in the rate, which are converted into levels using the model-implied cumulative sum. These daily level forecasts are subsequently averaged to produce monthly forecasts.

Table 3. Success Ratios for Monthly Average Forecasts of the Nominal 10-Year Treasury Bonds Against Alternative No-change Benchmarks

Horizon	**1**	**3**	**6**	**9**	**12**
No-change benchmark					
Monthly average	0.69 (0.000)	0.60 (0.000)	0.57 (0.020)	0.57 (0.026)	0.60 (0.061)
End-of-month	0.47 (0.659)	0.49 (1.000)	0.51 (1.000)	0.54 (1.000)	0.58 (1.000)

Note: Out-of-sample forecasts of the nominal monthly average 10-Year Treasury Securities yield in levels, 2000M1–2023M6. Forecast criteria reported relative to the end-of-month and monthly average no-change forecast, with p-values of the Pesaran and Timmermann (2009) tests reported in brackets.

Table 3 presents the success ratio of the (exact same) monthly average forecasts of the nominal 10-Year Treasury yields when evaluated against both the monthly average and end-of-month no-change benchmarks. The p-values for the null hypothesis of no directional accuracy are calculated following Pesaran and Timmermann (2009) and are reported in brackets.

When the forecasts are evaluated against the monthly average no-change, the forecast gains are substantial and statistically significant. However, when the forecasts are evaluated against the end-of-month no-change benchmark, which aligns with the random walk hypothesis, the gains do not exceed random chance (0.5) and are not statistically significant at any forecast horizon. This serves as a clear example of how spurious predictability can arise in practice due to temporal aggregation.

5 Conclusion

We have demonstrated that only comparisons against the end-of-period no-change benchmark provide a valid test against the random walk hypothesis for forecast directional accuracy of temporally aggregated data. Our findings challenge conventional methods of assessing forecast performance of temporally aggregated data by demonstrating that comparisons against the period-average no-change benchmark can lead to spurious results. Specifically, such comparisons inflate type I errors and reduce the power of hypothesis tests, which can mislead researchers and practitioners into believing that a forecasting model has predictive power when, in fact, any perceived gains in accuracy are merely artifacts of the aggregation process.

This insight is particularly relevant for macroeconomic forecasting, where data is often aggregated from daily observations, such as in the study of interest rates, exchange rates, as well as primary and non-primary commodity prices. As the use of high-frequency data continues to grow (for instance, through scanner and online sales data for consumer prices) adhering to the principles outlined in this paper will become increasingly applicable. Fortunately, a forecaster can maintain the goal of forecasting the temporal aggregate and ensure reliable evaluations of forecast directional accuracy.

Importantly, our theorem applies to series for which disaggregated data is not observed. For such series, our results imply two critical points. First, improvements over the aggregate no-change benchmark should be expected and do not constitute sufficient evidence against the random walk hypothesis. Second, such forecasts should be benchmarked against alternative forecasts, particularly those that leverage higher-frequency data. The broad applicability of the insight bridges a gap between the theoretical underpinnings of macroeconomic forecasting and the practical challenges faced when dealing with real-world data.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Anatolyev, S. and Gerko, A. (2005). A trading approach to testing for predictability. *Journal of Business & Economic Statistics*, 23(4):455–461.

Ellwanger, R. and Snudden, S. (2023). Forecasts of the real price of oil revisited: Do they beat the random walk? *Journal of Banking and Finance*, 154(106962):1–8.

Farag, M., Snudden, S., and Upton, G. (2024). Can futures prices predict the real price of primary commodities? *LCERPA Working Paper No. 2024-3*.

Granger, C. W. and Pesaran, M. H. (2000). Economic and statistical measures of forecast accuracy. *Journal of Forecasting*, 19(7):537–560.

Johannsen, B. K. and Mertens, E. (2021). A time-series model of interest rates with the effective lower bound. *Journal of Money, Credit and Banking*, 53(5):1005–1046.

Lai, K. S. (1990). An evaluation of survey exchange rate forecasts. *Economics Letters*, 32(1):61–65.

Leitch, G. and Tanner, J. E. (1991). Economic forecast evaluation: profits versus the conventional error measures. *The American Economic Review*, 81(3):580–590.

Marcellino, M. (1999). Some consequences of temporal aggregation in empirical analysis. *Journal of Business & Economic Statistics*, 17(1):129–136.

McCarthy, M. and Snudden, S. (2024). Forecasts of period-average exchange rates: New insights from real-time daily data. *LCERPA Working Paper No. 2024-6*.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Öller, L.-E. and Barot, B. (2000). The accuracy of European growth and inflation forecasts. *International Journal of Forecasting*, 16(3):293–315.

Pesaran, M. H. and Timmermann, A. (1992). A simple nonparametric test of predictive performance. *Journal of Business & Economic Statistics*, 10(4):461–465.

Pesaran, M. H. and Timmermann, A. (1995). Predictability of stock returns: Robustness and economic significance. *The Journal of Finance*, 50(4):1201–1228.

Pesaran, M. H. and Timmermann, A. (2009). Testing dependence among serially correlated multicategory variables. *Journal of the American Statistical Association*, 104(485):325–337.

Rossana, R. J. and Seater, J. J. (1995). Temporal aggregation and economic time series. *Journal of Business & Economic Statistics*, 13(4):441–451.

Tiao, G. C. (1972). Asymptotic behaviour of temporal aggregates of time series. *Biometrika*, 59(3):525–531.

Wei, W. W. (1978). Some consequences of temporal aggregation in seasonal time series models. In *Seasonal analysis of economic time series*, pages 433–448. NBER.

Weiss, A. A. (1984). Systematic sampling and temporal aggregation in time series models. *Journal of Econometrics*, 26(3):271–281.

Working, H. (1960). Note on the correlation of first differences of averages in a random chain. *Econometrica*, 28(4):916–918.

Online Appendix

A Proof of Lemmas

The daily data is a random walk (equation (5)), so we can write it as:

$$\underbrace{D_{(m-1)n+i}}_{\text{Level on day } i \text{ of month } m} = \underbrace{D_{(m-1)n}}_{\text{Level on last day of month } (m-1)} + \underbrace{\sum_{j=1}^{i} e_{(m-1)n+j}}_{\text{Errors in month } m} \tag{16}$$

This allows us to rewrite end-of-month and month-average observations in terms of the errors.

- The end-of-month level in month m is:

$$Z_m = D_{(m-1)n} + \sum_{j=1}^{n} e_{(m-1)n+j} \tag{17}$$

- The month-average level in month m is:

$$A_m = \frac{1}{n}\sum_{i=1}^{n}\left(D_{(m-1)n} + \sum_{j=1}^{i} e_{(m-1)n+j}\right) = D_{(m-1)n} + \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{i} e_{(m-1)n+j}\right) \tag{18}$$

- The month-average level in month $(m+h)$ is:

$$\begin{aligned} A_{m+h} &= \frac{1}{n}\sum_{i=1}^{n}\left(D_{(m+h-1)n} + \sum_{j=1}^{i} e_{(m+h-1)n+j}\right) & (19)\\ &= D_{(m+h-1)n} + \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{i} e_{(m+h-1)n+j}\right) & (20)\\ &= D_{(m-1)n} + \sum_{j=1}^{n} e_{(m-1)n+j} + \sum_{j=n+1}^{hn} e_{(m-1)n+j} + \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{i} e_{(m+h-1)n+j}\right) & (21) \end{aligned}$$

The difference between the current end-of-month level and current month-average level, $(Z_m -$

A_m), is a weighted sum of errors in month m. Equations (18) and (17) imply:

$$\begin{aligned} &Z_m - A_m \\ &= \left[D_{(m-1)n} + \sum_{j=1}^{n} e_{(m-1)n+j} \right] - \left[D_{(m-1)n} + \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{i} e_{(m-1)n+j} \right) \right] \\ &= \underbrace{\sum_{j=1}^{n} e_{(m-1)n+j} - \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{i} e_{(m-1)n+j} \right)}_{\text{Errors in } m} \end{aligned} \tag{22}$$

Similarly, the difference between the future month-average and current end-of-month, ($A_{m+h} - Z_m$), is a weighted sum of errors in months $(m+1)$ to $(m+h)$. Equations (21) and (17) imply:

$$\begin{aligned} &A_{m+h} - Z_m \\ &= \left[D_{(m-1)n} + \sum_{j=1}^{n} e_{(m-1)n+j} + \sum_{j=n+1}^{hn} e_{(m-1)n+j} + \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{i} e_{(m+h-1)n+j} \right) \right] \\ &- \left[D_{(m-1)n} + \sum_{j=1}^{n} e_{(m-1)n+j} \right] \\ &= \underbrace{\sum_{j=n+1}^{hn} e_{(m-1)n+j}}_{\text{Errors in } (m+1) \text{ to } (m+h-1)} + \underbrace{\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{i} e_{(m+h-1)n+j} \right)}_{\text{Errors in } m+h} \end{aligned}$$

B Success Ratios for Bottom Up Forecasts

Table A1. Corresponding Success Ratios for Bottom Up Forecasts in Table 2

			Versus End of Period			Versus Period Average		
Years	*DGP*	*Horizon*	Weekly	Monthly	Quarterly	Weekly	Monthly	Quarterly
40	*ρ=1*	1	0.50	0.50	0.50	0.70	0.74	0.75
		3	0.50	0.50	0.50	0.60	0.61	0.61
		6	0.50	0.50	0.50	0.57	0.58	0.58
		12	0.50	0.50	0.50	0.55	0.55	0.55
	ρ=0.9	1	0.64	0.75	0.84	0.71	0.74	0.75
		3	0.72	0.80	0.85	0.71	0.75	0.75
		6	0.75	0.80	0.85	0.74	0.75	0.75
		12	0.76	0.80	0.85	0.75	0.75	0.75
20	*ρ=1*	1	0.50	0.50	0.50	0.70	0.74	0.75
		3	0.50	0.50	0.50	0.60	0.61	0.61
		6	0.50	0.50	0.50	0.57	0.58	0.58
		12	0.50	0.50	0.50	0.55	0.55	0.55
	ρ=0.9	1	0.64	0.75	0.84	0.71	0.74	0.75
		3	0.72	0.80	0.85	0.71	0.75	0.75
		6	0.75	0.80	0.85	0.74	0.75	0.75
		12	0.76	0.80	0.85	0.75	0.75	0.75
10	*ρ=1*	1	0.50	0.50	0.50	0.70	0.74	0.75
		3	0.50	0.50	0.50	0.60	0.61	0.61
		6	0.50	0.50	0.50	0.57	0.58	0.58
		12	0.50	0.50	0.50	0.55	0.55	0.55
	ρ=0.9	1	0.64	0.75	0.84	0.71	0.74	0.75
		3	0.72	0.80	0.85	0.71	0.75	0.75
		6	0.75	0.80	0.85	0.74	0.75	0.75
		12	0.76	0.80	0.85	0.75	0.75	0.75

Notes: Success ratios for the bottom-up forecasts from Table 2. The daily random walk is given by $\rho = 1$; whereas $\rho = 0.9$ implies a predictable data generating process. 5000 simulations, using 40, 20, and 10 years worth of data.