SMU Data Science Review

Volume 7 | Number 1 Article 1

# Extending the M3-Competition: Category and Interval-Specific Time Series Forecasting

Will Sherman
*Southern Methodist University*, wsherman@mail.smu.edu

Kati Schuerger
*Southern Methodist University*, kschuerger@mail.smu.edu

Randy Kim
*Southern Methodist University*, randyk@mail.smu.edu

Bivin Sadler
*Southern Methodist University*, bsadler@mail.smu.edu

## Recommended Citation

Sherman, Will; Schuerger, Kati; Kim, Randy; and Sadler, Bivin () "Extending the M3-Competition: Category and Interval-Specific Time Series Forecasting," *SMU Data Science Review*: Vol. 7: No. 1, Article 1.
Available at: https://scholar.smu.edu/datasciencereview/vol7/iss1/1

# Extending the M3-Competition: Category and Interval-Specific Time Series Forecasting

Will Sherman[1], Kati Schuerger [1], Randy Kim [1], Dr. Bivin Sadler [1]
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
{wsherman, kschuerger, randyk, bsadler}@smu.edu

**Abstract.** The M3-Competition found that simple models outperform more complex ones for time series forecasting. As part of these competitions, several claims were made that statistical models exceeded machine learning (ML) techniques, such as recurrent neural networks (RNN), in prediction performance. These findings may over-generalize the capabilities of statistical models since the analysis measured the total forecasting accuracy across a wide range of industries and fields and with different interval lengths. This investigation aimed to assess how statistical and ML methods compared when individuating series by category and time interval. Utilizing the M3 data and building individual models using Facebook© Prophet and R packages: *tswge*, *forecast*, and *nnfor,* there were significant differences in model performance. The statistical models performed better for monthly – industry, macro, and micro combinations (Wilcoxon signed-rank adjusted p-value < 0.0001) for short-term forecast horizons (h=5). However, the multilayer perceptron (MLP) surpassed the statistical models in quarterly – industry data (p-value < 0.001) for the same forecast length. The statistical models also outperformed ML methods for long-term forecasts in the same category by interval combinations (p-value < 0.01). Thus, identifying which model may have increased performance in specific category, interval and horizon combinations provides direct value for time series analysis.

## 1 Introduction

As society continues to become more technologically interconnected, more data are being generated than ever before. Much of this data (possibly most of it, or some might argue almost all) could be considered through a time series lens.

Time series data appear in many real-world situations—from heart rate monitoring to retail sales numbers to temperature data. Time series forecasting accuracy holds great importance as poor performing or inaccurate forecasts could have substantial operational and/or financial consequences. The capabilities of different modeling approaches present a unique opportunity to optimize time series forecasting. Anything from stock market re-analysis after significant fluctuation (*e.g.* 2008 financial crisis) to modeling disease transmission among community systems (*e.g.* dengue fever or COVID-19) may benefit from the knowledge of which type of time series modeling to employ.

The Makridakis Competitions, referred to hereafter as the M-Competitions, are a series of competitions aimed at producing the "best" forecasts for a variety of time series data from a variety of time series models and methods. The M-Competitions primarily relate to business and economic time series; despite this, their conclusions may still be applicable to other fields. The findings of these competitions serve as a foundation for understanding traditional statistical methods and kernel methods, as well as the capabilities of these approaches to achieve accurate forecast predictions. The M-Competitions are highly referenced by researchers and featured heavily in the International Journal of Forecasting.

Among the conclusions drawn from the M-Competitions, the most interesting was that the simpler models outperformed the more complicated (more complex, some may call "more superior or advanced") ones.

This research aimed to do a deep dive into time series methods, specifically as they relate to the M3-Competitions (the third iteration of this series of forecasting competitions). A key focus was model performance concerning the category of the time series and the frequency of the observations composing the series. The M3 data had 6 categories; Demographic, Finance, Industry, Macro, Micro, and Other; it also had 3 noted time intervals: Monthly, Quarterly, and Yearly. This research sought to evaluate modeling methods to determine the most useful type of model given a particular category of time series data with observations at a particular interval. What analytic methodologies are the most appropriate when forecasting time series data? Can consensus be drawn on these analytics when considering forecast horizon, category, and interval?

Research since the M3-Competition has identified a host of additional investigative questions: which metrics represent the most accurate prediction, how should cross-validation be approached, *etc*. One commonality of the body of work surrounding the M3-Competition was that the researchers evaluated model performance on all the series—with different categories and time intervals—in aggregate. Researchers have noted that this may not account for tendencies toward bias of some methods used. This is relevant, especially in forecasting and prediction interval evaluation. This research used different time series methods and statistical testing to evaluate which models perform better given a specific category of time series: querying the benefits to using different models given different categories and measurement intervals of time series data.

## 2 Literature Review

### 2.1 M-Competitions history

In 1979, Makridakis and Hibon made one of the first efforts to compare several time series methods across multiple data series. They selected 111-time series out of a large cross-section of available data, covering a broad range of subjects: business firms, industry, and macro data.

Among the conclusions drawn from this analysis, the most interesting/relevant is that the simpler models outperformed, the more complicated (more complex, some may call "more superior or advanced") ones. This conclusion challenged the accepted paradigm of the time—that more advanced models are inherently "better"—and sparked a lot of responses from the scientific community, some of which are quite critical (Hyndman, 2020).

The M-Competitions are a series of competitions in which researchers have conducted univariate time series analysis on different types of time series data

(Makridakis et al., 2000). Each category has multiple series with various measurement intervals (e.g., quarterly, yearly). The goal was to assess model performance using different time series methodologies (e.g., ARMA, ARIMA, neural nets) across different time series types with different measurement intervals.

In response to these criticisms and to incorporate some of the suggestions for improvement from commentators, the first M-Competition was created in 1982. This competition increased the number of time series to 1,001, and the number of methods to 15.

A key innovation of the M-Competition was that different experts were tasked to analyze and model data related to their specific domain of expertise (*i.e.*, financial data). This meant that, rather than having only two researchers—Makridakis and Hibon—assess the results, additional participants were engaged to contribute to the research effort specific to their expertise.

Each expert provided their forecasts, and then those forecasts were compared with actual values to determine forecast errors and compute various measures of accuracy. The results of the M-Competition were quite like those of Makridakis and Hibon's original study in 1979.

### 2.2 Key findings of the M-Competitions

One key finding that has been supported heavily throughout the M-Competitions is that sophisticated or complex methods do not necessarily provide more accurate forecasts than simpler ones (Koning et al., 2005). Additional findings that appear consistent throughout later competitions include:

(a) The relative ranking of the performance of the various methods varies according to the accuracy measure being used.

(b) The accuracy when various methods are combined outperforms, on average, the individual methods being combined and does very well compared to other methods.

(c) The accuracy of the various methods depends upon the length of the forecasting horizon involved.

The original findings of the competition (1982) had a large impact on forecasting research, influencing researchers to focus their attention on what models produce good (more useful) forecasts rather than on the mathematical properties of those models. It also called attention to the dangers of over-fitting and suggested a framework for treating forecasting as a different problem from time series analysis.

The M2- (1993) and M3- (1998) Competitions further extended the original competition, including more methods (specifically extending to include neural networks and expert systems), more researchers, and more series.

This research focused mainly on the results and methodologies of the M3-Competition while keeping in mind the M4-Competition. The M4 increased the number of series being forecasted from 3,003 (M3-Competition) to 100,000 (M4-Competition), and researchers placed a slightly heightened emphasis on accuracy scoring and testing the statistical significance of results.

The first competition after the M3-Competition, the M4 introduced the improvement of complex hybrid models which rely on combined statistical and kernel methods. Twelve of the 17 most accurate prediction methods were combination approaches to forecasting (Makridakis, 2018). However, the length of the forecasting window played heavily into outcomes; standardization of the forecasting horizon is essential during competitions but may vary drastically when designing studies (Darin et al., 2020).

The M5, which focused on forecasting data made available by Walmart, included explanatory variables such as price, promotion, and special events. This was a departure from the M3 and M4, which used univariate analysis, only using past values of that series to make future predictions.

The M6 is currently taking place, and the competition has been extended to take place over several months. The focus of the M6 is to forecast financial (stock) prices and explore the connection between forecasts' accuracy and returns on investments made from those forecasts. At the time of this publication, the expected release of results from the M6 is 2024.

### 2.3 M-Competition modeling methods

Several different methods were used by participants in the M3-Competition, from the simplest AR (autoregressive) model to the more complex MLP (multilayered perceptron) and hybrid methods such as exponential smoothing combined with recurrent neural networks. Some analyses used pre-processing methods (such as data transformation) to achieve stationarity, log transformation, de-seasonalizing, scaling the data, and differencing the data to remove unit roots. Each of these were relevant to this research. Model methods employed were autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), seasonal ARIMA, MLP, and the Facebook© Prophet model.

A further consideration is the terminology ascribed to the model generation techniques. The term "machine learning" when describing models may have an inherent flaw since machine learning models are "maximum likelihood estimators[,] meaning they are statistical in nature" (Barker, 2020). Therefore, it was important to consider the models more deeply to address a key assertion from the M3-Competition: that statistical models tend to outperform machine learning unstructured models (Makridakis, 2018).

Finally, as was found in the M4, models which benefited from a hybrid approach tended to improve forecasting accuracy (Pawlikowski et al., 2020). A stretch goal may be to evaluate weighted ensemble statistical models. The idea behind this approach is a mirror of NNs and the "forecasting intervals" that a multi-layer perceptron may generate. Because these models select the median or mode of the generated series for prediction purposes, it presents a possible advantage over traditional statistical forecasting windows through a "law of large numbers" approach to generating forecasting models which may then serve in much the same as the multiple models from NN.

### 2.4 M3 results: Specificity of Data

The goal of forecasters in M3-Competition was to focus on a model's overall performance only. They approached the problem of creating a generalized model and did not consider model performance based on the category (subject area) of time series or observation interval.

The M3-Competition took twenty-four models and applied them to build forecasts for 3,003 different time series, from five different categories, with varying observation intervals. Model performance was assessed based on the performance of each model for the given time series category and interval. (The initial M3 results aggregated results over all 3,003 time series when comparing model performance.) This last piece is important to understand and a driving theme for this research.

Using different time series models, researchers built models to forecast each time series type and each observation interval. These forecasts were then compared based on how well each model type performed (across all the time series).

This led to the conclusion that simpler models perform better in general. While this may be useful to understand, the analysis does not address whether a particular method/model type may be better suited to forecast a specific type of time series. For example, it may be true that the ARMA models tend to perform better across the board, but it may be possible that an MLP model produces much better forecasts on financial data.

This research would like to understand if the type (*i.e.*, Demographic, Financial, Micro, Macro, Industry, and Other), or observation interval (*i.e.*, Monthly, Quarterly, Yearly) of the data being analyzed is meaningful in selecting the model type to use.

To make an analogy, imagine a toolbox containing only one tool: a hammer. This is fine if you only need to pound in or remove nails; but what if the task is to repair a hole in a favorite sweater or tighten the lug nuts on a hubcap after replacing the tire? In these scenarios, there may be a different tool that will reach the desired outcome more effectively. Different tools are more helpful for different situations, and in the same way, there are different models that can help with different types of time series data. The goal is to see if there is evidence that models may provide more useful forecasts depending on the subject area or time interval of the time series.

This research tested the hypothesis that different models may be better suited to forecast (and perform well), depending on the time series category (subject matter) and observation interval. This would suggest that category, observations interval, or both, is helpful as an influencing factor when selecting the time series method to build a model and generate forecasts.

# 3 Methods

## 3.1 Data

The data used in the M3-Competition involves 3,003-time series taken from several fields (business, demographics, finance, and economics). The number of observations for each series ranges in length between 14 and 126 observations. The data are either annual, quarterly, or monthly. All values are positive.

The data come from the M3-Competition and can be found at the provided website: https://forecasters.org/resources/time-series-data/m3-competition/ (Chart provided in Appendix).

The initial approach for this research was to pick two different subject matter areas and several time series methods, create long-term and short-term forecasts and prediction intervals and compare the performance of these methods. Specifically, this research began with Microeconomic and Finance data as put forth by the M3-Competition and contrast exponential smoothing, ARMA, ARIMA, and MLP.

This analysis was univariate, using only the data of each series to fit the model and predict future values. Multivariate methods (for example, leveraging the cross-correlation between multiple explanatory variables) were not considered in the M3-Competition and was not included in this research either.

Model performance was evaluated using rolling window root mean squared error (RMSE) in addition to symmetric mean absolute percent error (sMAPE)—a modified metric to the mean absolute percentage error (MAPE) which was the metric used in the M3-Competition. This research took a dual approach to the forecasting window: mirroring the short-term horizon of the M3-Competition, using five-time points as the forecasting window, and a long-term horizon with variable length determined by sampling 20% of the end of each series.

The final output was the error metrics for each model based on the forecasting window. Model specifications, the series each result is derived from, and the model parameters specific to the method—statistical or machine learning (ML)—were also included.[1]

### 3.2 Assumptions for stationary models

Assumptions are essential to any statistical investigation, and time series modeling is no different. As part of data exploration and pre-processing, this research considered the assumptions relevant to the modeling method used.

A weak stationary process can be considered in a state of "equilibrium." There are three conditions for stationarity in a time series.

Condition 1: The mean does not depend on time. Subpopulations of $X_t$ have the same mean for each value *t*.

$$E[X_t] = \mu$$

Condition 2: The variance does not depend on time. Subpopulations of $X_t$ for a given time have a finite and constant variance for all *t*.

$$Var[X_t] = \sigma^2 < \infty$$

Condition 3: The correlation between data points does not depend on where the points are in time. It only depends on how far apart the data points are in time.

$$Cor(X_t, X_{t+h}) = \rho_h$$

It should be noted that this research does not assume stationarity for all of the time series.

### 3.3 Independence assumption

One of the key assumptions for other statistical methods (non-time series) is that the data are independent. Time series methods are useful when serial correlation exists in

[1] There is not much available information regarding what each of the time series used in the M3-Competition represent. Some assumptions can be made based on domain knowledge, but this may impact the extent to which these results may be generalized. The original organizers of the competitions have been contacted; if appropriate, the response will be included.

the data (meaning that the relative order of the observations is important for identifying patterns in the data). The goal of time series methods is to capture, model, and capitalize on this serial correlation as a means for creating more accurate predictions of future values.

Given what is known about the series, namely that they may be fully or partly coming from stock market data, there is reason to believe the data may be correlated. Example: If a main company has a ‘good’ day (*e.g.*, Alphabet), subsidiaries or related companies may also have a ‘good’ day (*e.g.*, Fitbit). There may be correlation between companies in each category—concept of all ships rise together—maybe tech companies have trends that extend beyond the company to the category. Ultimately, there is not much information available around where this data came from or what it exactly represents. This research proceeded with caution when applying statistical tests.

### 3.4 Data pre-processing: Normalization

The time series provided in the M3-Competition varied in terms of value ranges. This research utilized min/max scaling to preprocess each time series. Normalized data allows intuitive comparisons to be drawn when investigating residuals and scoring metrics. This approach was used by some of the models from the original paper; preprocessing approaches which changed the shape of the data (*e.g.*, logarithmic transformation) were not pursued.

### 3.5 Hypothesis testing for statistical significance (Wilcoxon Signed-Rank)

In addition to evaluating forecasting, determining the statistical significance of the difference in performance must be completed. Suppose the difference in results (*e.g.*, RMSEs on the same series for different models) is not determined to be of statistical significance. In that case, the difference between forecasts is negligible (the results of one model cannot be interpreted as "better" or different from the other).

The method for testing the statistical differences between model performance was a Wilcoxon signed-rank test for paired data. A non-parametric test provided better understanding between models, as the research focused on the distribution of the median of the paired differences. Additionally, an assumption of the paired t-test was questionable: the distribution of the dependent variable may be heavily right skewed. And, finally, some category-by-interval combinations had sample sizes smaller than reliably used in parametric analysis (*e.g.*, Yearly ~ Other, n=11).

Because the models are being generated to forecast against the same time series, they operate as two different treatments of the same population. Therefore, the paired approach evaluating the difference in the absolute value of the sMAPE scores can be interpreted as a test for whether the pseudomedian of the differences is zero.

### 3.6 Cross-validation

Cross-validation is a foundational tool for training machine learning models and statistical models. However, there are significant concerns when utilizing cross-validation (CV) in time series data. Out-of-sample (OOS) evaluation is the preferred training method in many schools of thought. To benefit from CV in time series, several novel approaches have been considered, such as K-Fold CV and leave-one-out cross-validation. While these methods tend to have lower mean absolute predictive absolute error (MAPAE) than traditionally built OOS models—for both statistical and NN approaches (Bergmeir et al., 2018)—this study applied the traditional OOS method to compare to the M3 results directly.

### 3.7 Models

#### ARMA

Autoregressive Moving Average (ARMA) is a time series model that consists of two main components which are autoregressive (AR) and moving average (MA). An AR(p) is similar to a multiple regression model of order *p* which predicts current or future values of a series based on the linear combination of the past values with the set of parameters called autoregressive coefficients. The MA(q) component of order *q* predicts future or current value of a series based on the similarities or errors between past and present with the set of parameters called moving average coefficients.

#### ARIMA

Autoregressive Integrated Moving Average (ARIMA) is an extension of the ARMA model with an additional component that accounts for non-stationary behavior in a series. The AR and MA components function the same as in ARMA model, and the Integrated component accounts for non-stationary by differencing the series until the series becomes stationary (yielding an ARIMA(p,d,q) model where *d* is the degree of differencing).

#### Seasonal ARIMA

A seasonal ARIMA model (denoted as ARUMA for the remainder of the paper) is formed by including seasonal differencing components to ARIMA models. The ARUMA(p,d,q)(P,D,Q) model includes autoregressive, integrated, and moving average seasonal differencing (*P*, *D*, and *Q*, respectively).

Where $X_t$ represents the distribution of the time series at a particular timepoint, *t*, and $a_t$ represents the distribution of the random noise at *t*, the ARUMA model can be expressed in the form below.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $(1-\varphi_1 B)$ | $(1-\Phi_1 B^S)$ | $(1-B)^d$ | $(1-B^S)$ | $X_t$ | = | $(1-\theta_1 B)$ | $(1-\Theta_1 B^S)$ | $a_t$ |
| *p* | *P* | *d* | *D* | | | *q* | *Q* | |

#### Prophet

Facebook© Prophet is an open-source library for time series forecasting. It was designed as a decomposable time series model that produces future predictions based on historical data. Prophet uses a Bayesian time series method that has several components such as trend, seasonality, and holidays. Because Prophet is using a Bayesian approach to estimate the model parameters, it handles missing data, outliers, and other features which makes the model flexible, scalable, and robust. The Prophet model determines the flexibility of the trend and depends on the size of the trend.

### MLP

The multilayered perceptron (MLP) creates an ensemble of neural networks, each of which is trained using a different set of random initial weights. There are several hyperparameters which may be optimized: number of repetitions, number of hidden nodes, number of hidden layers, and lag and seasonality detection. Forecasts from the MLP model are derived from many repetitions of which each has a different set of initial weights. For this research, the number of repetitions was set to 200; the strategy for hidden nodes, hidden layers, lag detection, and seasonality detection was to allow automatic detection in a per-series fashion. The strategy for hidden layer optimization employed 5-fold cross-validation.

In most cases of forecasting analysis for MLP models, the median or mode of the forecasts (one forecast for each repetition) is used to provide the best forecasts—this research utilized the median. To protect against overfitting, the cross-validation strategy was used to help limit ballooning of neural network hidden layers, bringing down model complexity. Additionally, while any single network may overfit the data, the ensemble method with median forecasts utilizes the law of large numbers wherein the median of the forecasts obtained from many networks should be close to the expected value of the distribution from which the time series derives.

### 3.8 Model evaluation: scoring metrics

Model evaluation for the M3 primarily looked at MAPE. This research includes other performance measures such as RMSE, ASE (average squared error), RMSSE (root mean squared scaled error), and sMAPE. The final two metrics are novel in the M-Competitions to M5. They present advantages over percentage errors when the time points have values equal to zero or the relative benchmark errors are zero (Makridakis et al., 2022).

This research also examined fixed origin versus rolling window analyses. There is some heated discussion on the optimal scoring metric and method for forecasting (Tashman, 2000). One of the primary considerations when evaluating RMSE or ASE or mean absolute deviation is that altering numeric values with scaling or normalization substantially changes the error magnitudes; Tashman argues that using percent error measures is superior since they are scale independent. One consideration that should be taken, regardless, is the distribution of the errors. Badly skewed errors may require additional approaches that must be handled circumstantially.

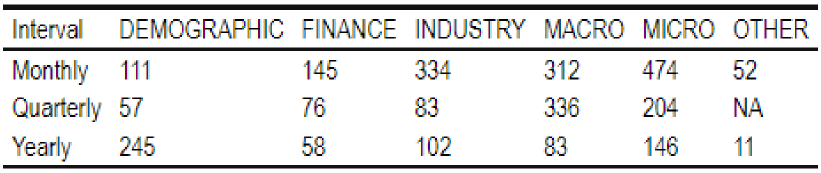
Also, according to Tashman, there are significant drawbacks to fixed-origin evaluation for individual, univariate time series; this technique misses the benefit of distribution-level analysis. Leveraging rolling origins essentially updates the time series and allows forecasting against each new origin producing a more robust analysis using the breadth of available data. This concern, of course, may be offset by using multiple time series. This facet was not investigated within the body of this research as the length of forecasting window was given priority.

# 4 Results

This research aims to identify how the categorical source of the data deriving a time series or the interval of the data inform the choice of method for modeling: statistical or ML technique. The goal was to identify a more useful or common starting point that can be established for forecasting time series (to be used by future time series analysts).

The series investigated from the original M3-Competition are those which had defined time intervals (Table 1). After removing series without provided intervals, the data represented 94.2% of the original M3 data that was investigated.

**Table 1.** Delineation of time series by both interval and data-type



| Interval | DEMOGRAPHIC | FINANCE | INDUSTRY | MACRO | MICRO | OTHER |
|---|---|---|---|---|---|---|
| Monthly | 111 | 145 | 334 | 312 | 474 | 52 |
| Quarterly | 57 | 76 | 83 | 336 | 204 | NA |
| Yearly | 245 | 58 | 102 | 83 | 146 | 11 |

As part of the data normalization process, this research utilized a min-max scaling algorithm for standardizing the amount of error between series—since series ranges were vastly unequal across and within categories and time intervals. The primary evaluation metric selected for analysis, also utilized in the M3-Competition, was the sMAPE. This metric is determined as the absolute value of the difference in the forecast ($F_t$) and the actual value ($A_t$) divided by the average of the two, for each timepoint (t), n being the number of timepoints in the forecast horizon. This metric was chosen over the MAPE; while MAPE is scale-independent, it becomes undefined when the actual values approach zero. The sMAPE has both lower (0%) and the upper (200%) bounds and is less restricted when actuals are close to zero.

$$\text{sMAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|\mathrm{F}_t - \mathrm{A}_t|}{(\mathrm{F}_t + \mathrm{A}_t)/2}$$

The distribution of best-performing models for sMAPE scores in both short-term and long-term forecasts appear to differ drastically depending on both the datatype and the periodic time interval (Figure 1). Taking Short Term sMAPE for Monthly ~ Macro time series as an example, the model with the lowest scores for the most time series forecasts was the ARMA (orange: 52.2%); the model with the fewest top performing scores was the ARIMA (red: 9.3%). On the other hand, the contrast for Quarterly ~ Macro time series suggests that both ARIMA (red: 25.6%) and MLP models (teal: 24.1%) are more frequently useful with respect to ARMA (orange; 16.4%). Notably, it appears that there is a lack of consistency for model performance. Additionally, the forecast horizon (h) for the short-term predictions was set to a constant value. The long-term forecast horizon represents 20% of the total series data; this still provided consistency when comparing models since the length of horizon is consistent per series.

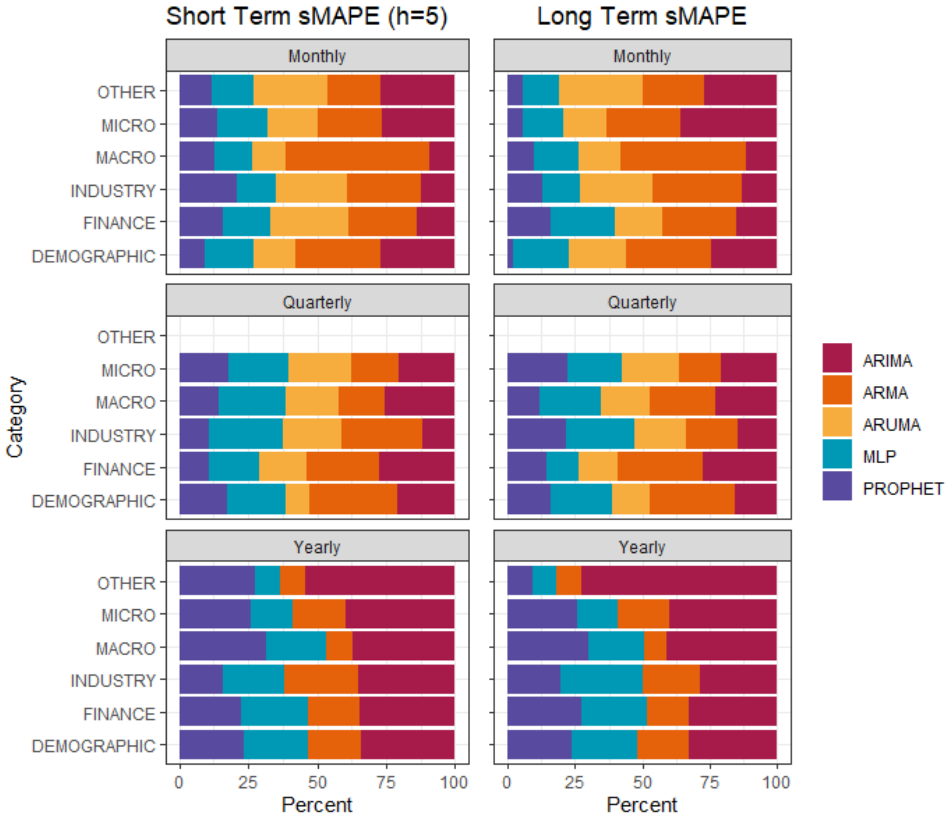


**Figure** 1. (a) Left. Distribution of the model with the lowest sMAPE scores for short-term forecasts (h=5) across all time series with a given Category and Time Interval. (b) Right. Distribution of model with the lowest sMAPE scores for long-term forecasts.

However, this representation of the number of top performing models doesn't evaluate differences in distribution or whether these differences are statistically significant. Therefore, these distributions can be visualized on a per-model basis to validate the scale at which these differences occur for short-term (Figure 2) and long-term forecasts. The distribution of each model shows low scores across all series with a general right-skew. Figure 2.B also shows that some difference in performance exists across models with the most optimal scores—particularly, the Prophet model has fewer samples with sMAPE scores below 5% (*i.e.*, bin 0).

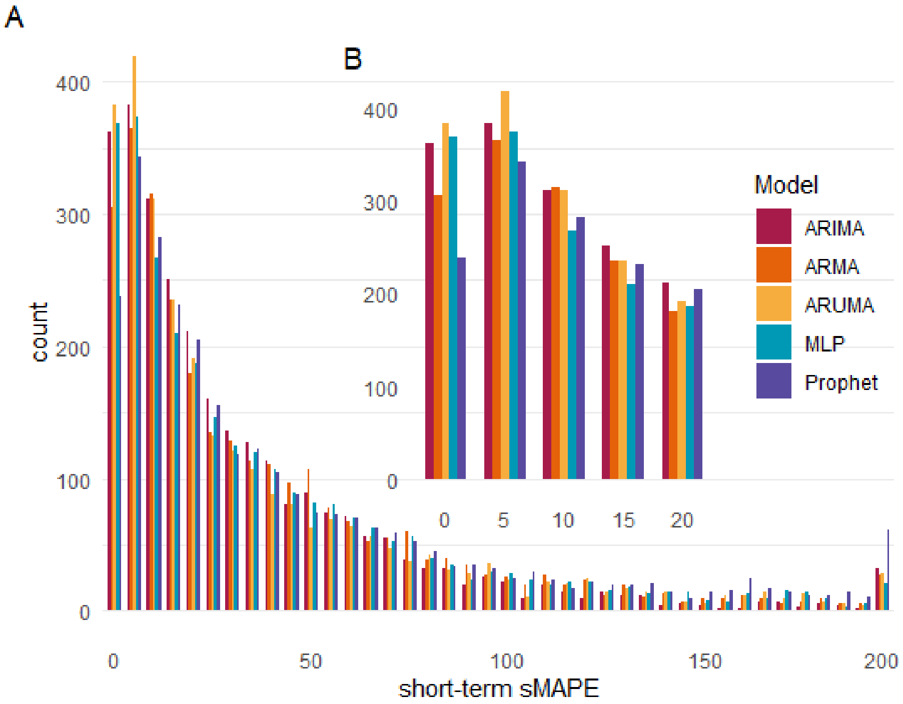


**Figure 2.** (A) The overall distribution of short-term (n=5) sMAPE scores for all models binned in intervals of 5. (B) Condensed distribution for binned sMAPE scores under 20%.

The trend for all series shown in Figure 2 becomes markedly different when evaluating the distributional differences of models by Interval and Category (Figure 3). For Figure 3.A and 3.B, the Industrial category for time series data was selected for having roughly similar distribution. As shown, there are drastic changes between the top performing models depending on the period of the time series. Figures 3.B and 3.C represent two categories, Industrial and Financial, which had similar distributions with the same period. Again, major variations in model performance appear to occur in a category-dependent fashion.

One additional consideration for these comparisons in model performance is that the model may be performing optimally on its own set of series. That is, the series most easily modeled by the ARUMA for Monthly~Finance data may not be those most easily modeled by the MLP, though their counts are equal for bin 0. Therefore, a Wilcoxon Signed-Rank test was performed to compare performance, this included a *post hoc* multiple test correction based on the total number of comparisons performed (n=170). The method used for multiple tests was the Bonferroni correction; the number of tests was computed as the product of the number of model comparison combinations (*i.e.*, $5^C2$) and the number of category-interval groups (*i.e.*, 17).

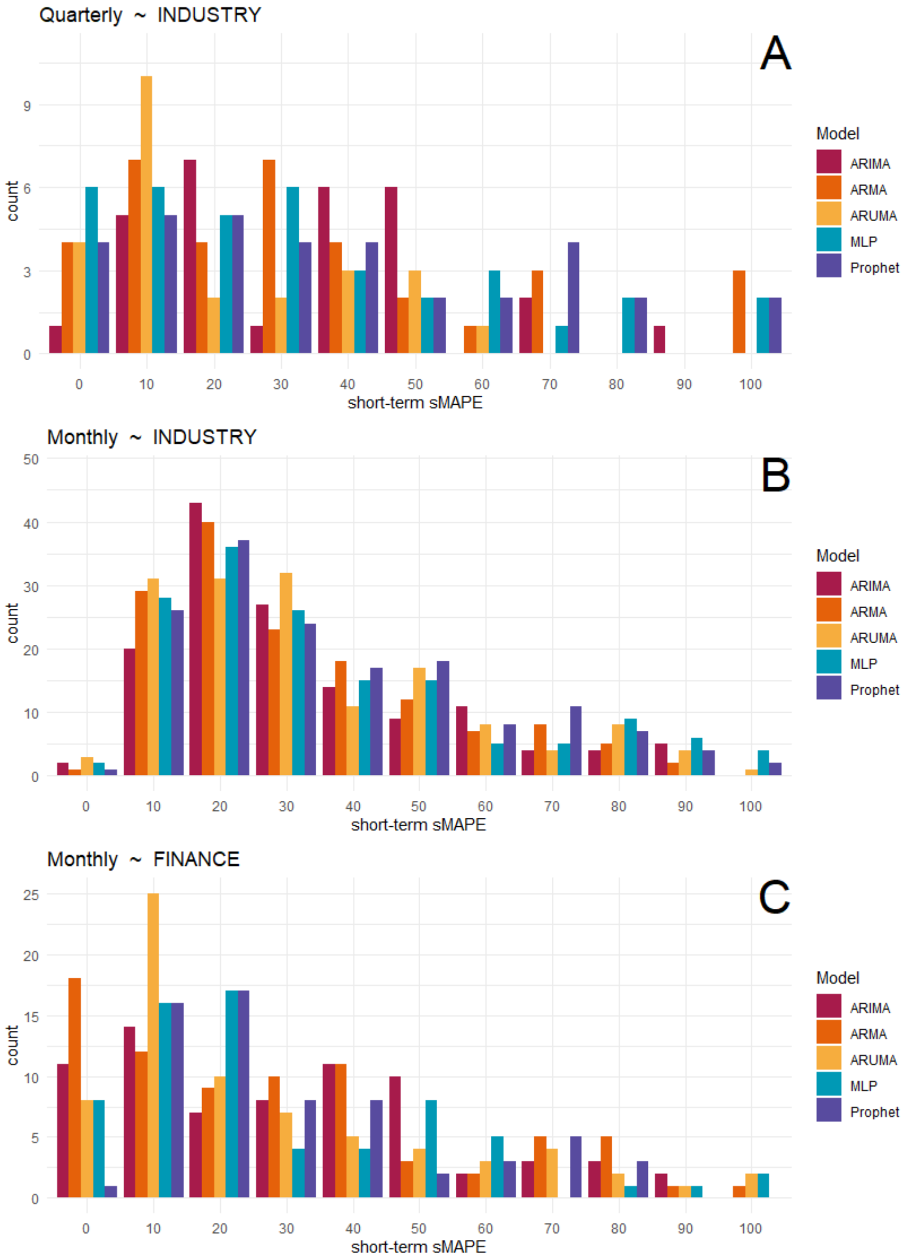
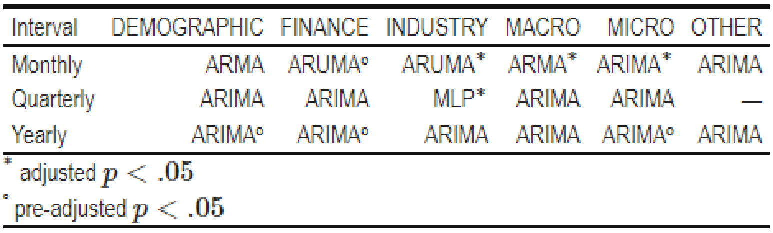


**Figure 3.** (A) Distribution of short-term (n=5) sMAPE scores with values less than 100% for time series with Quarterly frequency and from the Industry category. (B) Short-term sMAPE scores for time series with Monthly frequency and from the Industry category. (C) Short-term sMAPE scores for time series with Monthly frequency and from the Finance category.

The distribution of differences of model residuals can be assessed as $X_i - Y_i$, where $X_i$ and $Y_i$ are the distributions of the two models being paired by time series. In this context, the Wilcoxon signed-rank test is used to evaluate if the distribution of the differences in the residuals is significantly different from zero (*i.e.*, $H_0$: observations $X_i - Y_i$ are symmetric about $\mu = 0$). After comparing the models via Wilcoxon signed-rank, the top performing model was identified; from each comparison to that model, the maximum p-value was taken, and the Bonferroni correction was applied to assess significance after accounting for multiple comparisons (n=170). Table 2 represents the top performing models in Category and Interval dependent fashion for the sMAPE metric over the short-term prediction interval (h=5).

**Table 2.** Best Model from Wilcoxon Signed-Rank Test (short-term sMAPE)



| Interval | DEMOGRAPHIC | FINANCE | INDUSTRY | MACRO | MICRO | OTHER |
|---|---|---|---|---|---|---|
| Monthly | ARMA | ARUMA° | ARUMA* | ARMA* | ARIMA* | ARIMA |
| Quarterly | ARIMA | ARIMA | MLP* | ARIMA | ARIMA | — |
| Yearly | ARIMA° | ARIMA° | ARIMA | ARIMA | ARIMA° | ARIMA |

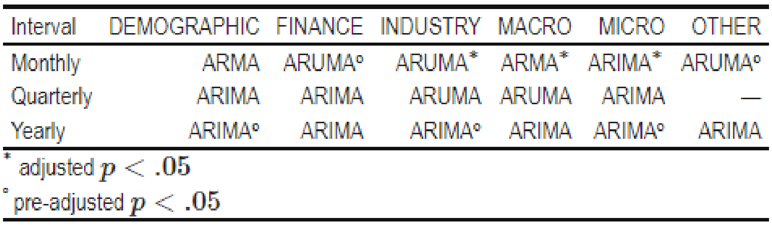
* adjusted $p < .05$

° pre-adjusted $p < .05$

In this case, the statistical models tend to dominate—though not in a strictly significant fashion. Of particular interest are the Monthly results, as three different statistical models show strong evidence of statistical significance (Bonferroni adjusted p-value < 1e-4). That three different linear models show distinguishable improvements over their counterparts is of note. However, the final model to show strong evidence of significance was the MLP for Quarterly~Industry data (Bonferroni adjusted p-value < 1e-3). This diverges from the original findings in M3-Competition—though specifically for short-term sMAPE forecasts. One final note for the comparisons is that all ARIMA and ARUMA with Yearly frequency were evaluated to be identical models via the auto.arima function from the *forecast* package. Therefore, model comparisons were done against ARMA, MLP, and Prophet models only. However, the Bonferroni correction was applied at the same level (n=170).

Of additional note, the metric used for analysis influenced the output for Quarterly~Industry data (Table 3). When using RMSE for short-term forecasting, the MLP model lost significance and was superseded by the ARUMA model as the optimal model for these time series (though not to a statistically significant degree). The other models which yielded significant results after multiple-tests correction were unchanged (Bonferroni adjusted p-values < 1e-3).

**Table 3.** Best Model from Wilcoxon Signed-Rank Test (short-term RMSE)



| Interval | DEMOGRAPHIC | FINANCE | INDUSTRY | MACRO | MICRO | OTHER |
|---|---|---|---|---|---|---|
| Monthly | ARMA | ARUMA$^{\circ}$ | ARUMA$^{*}$ | ARMA$^{*}$ | ARIMA$^{*}$ | ARUMA$^{\circ}$ |
| Quarterly | ARIMA | ARIMA | ARUMA | ARUMA | ARIMA | — |
| Yearly | ARIMA$^{\circ}$ | ARIMA | ARIMA$^{\circ}$ | ARIMA | ARIMA$^{\circ}$ | ARIMA |

$^{*}$ adjusted $p < .05$

$^{\circ}$ pre-adjusted $p < .05$

# 5 Discussion

From an aggregate perspective, the findings of the M3-Competition bear out: simpler statistical models outperformed the more ML-based models. Even when evaluating the cross-section of categories and time Intervals, the statistical models were more frequently the models with the most accurate forecasts. However, the fact that there are counterexamples at all speaks to the importance of evaluating time series with respect to both the category of data and frequency at which the observations occur.

Another facet worth noting is that model performance may vary dramatically depending on the metric used for evaluation. The change to RMSE from the sMAPE saw the loss of a statistically significant model in the MLP for Quarterly~Industry data. Additionally, the length of forecasting window may heavily impact model performance. Long-term sMAPE and RMSE may see variations in model performance based on the length of the forecast window (Appendix: Tables 4 and 5), which is another key consideration that researchers must consider when choosing and then optimizing forecast models. Notably, the five models evaluated here—four seen before in the M-Competitions and one novel (*i.e.*, Prophet model)—are but a small subset of the models investigated within the M3-Competition.

Since these results provide some evidence to suggest that different model types may be more useful given the observation interval and/or category of the time series being evaluated, this means that there may be utility in starting with a given model/method when presented with different scenarios. For example, the ARIMA model was most successful in building forecasts for yearly data; therefore, individuals looking for a base model for comparison may choose the ARIMA—considering their data may fall into one of the categories as defined here (*i.e.*, Demographic, Industry, Finance, Macro, Micro, or Other) and they have data points with consistent time steps with interval matching those defined here (*i.e.*, Monthly, Quarterly, and Yearly).

One of the main concerns with secondary analysis such as this is that only associations should be drawn rather than being able to draw causal inference. It is prudent to acknowledge that the results of this study may be related only to the categories and intervals herein or to only the specific series from which the M3-Competition was derived. While this research should be understood to apply primarily to the domains whose data were analyzed (not extrapolated to other domains), these results may still have practical significance for approaching model building with a focus on identifying either baseline models to compete against or a method by which comparisons can more readily be drawn and the steps required to do such comparisons.

As the volume and variety of data continue to accumulate, consideration must also be given to infrastructure requirements, cost, and time-to-train. When dealing with data such as microtransactions, the data volume and velocity may require significant investment; this may become an additional consideration beyond those given in this research. This research was able to leverage the computing power of a high-performance cluster—this infrastructure may not be available for smaller firms, hospitals, or individual practitioners, as the size and scope of their data grows. An important factor to note is that ML methods are more computationally complex than statistical methods. This means that ML methods require more processing power or time to train the model (or both), rendering them potentially less workable options for practitioners without access to appropriate computing resources or in need of delivering results in an expedited timeframe.

This research would like to note that data used in this study did not contain any Personally Identifiable Information (PII) or business Intellectual Property (IP). Data used in this research did not contain any confidential or proprietary information, and all data is publicly available. This is ideal data for field research; however, consideration must still be given to how this research will be received, as firms who are aware that their data were included in the M3-Competition may have strong interests in these outcomes. It should be noted that these methods do not purport to be the single best model for a given category-interval combination. It also may not be appropriate to apply to PII or IP as there are additional ramifications that should be considered when employing linear models and ML methods.

Future researchers have a veritable cornucopia of possible avenues. One suggestion is further univariate analysis of the models employed throughout the M-Competitions or hybrid model ensembling—a focus area for competitions following the M3-Competition. Hybrid model ensembling is a mechanism for combining models in order to generate future predictions. It may be interesting to investigate how forecasts perform when curated by multiple methods, particularly, looking at how hybrid machine learning models perform compared to singular ML methods (such as those in this research).

Additional research into what underlying factors enable statistical methods to beat ML methods when it comes to time series analysis could be very helpful. A diagnostic approach may be able to identify where ML methods "break" so that changes may be deployed to assist ML in closing the gap (in terms of prediction metrics) with traditional statistical methods. Also, research into performance enhancements that reduce the required processing time and power for ML methods may also be a worthwhile enterprise, with the added benefit of potentially making ML methods more accessible to a broader range of practitioners.

Lastly, evaluation of series length and prediction performance would help to gain more understanding of whether ML methods perform differently (better accuracy) when provided more data on which to train their models. It is possible that one of the reasons ML metrics were less favorable may be that there was not enough historical data provided for the model to properly train, as ML methods are often improved when dealing with large datasets and high dimensionality.

## 6 Conclusion

This research finds compelling evidence that time series model comparisons, both for competitions and in real application, should be done with respect to the interval of the series and the category from which the data are derived. Collapsing all categories and observation intervals into the same bucket for evaluation may lack the nuance underpinning the actual results. Metrics and forecasting horizons may also impact model performance and should be assessed, when possible, as this research found some evidence that model dominance may be lost depending on these design choices.

**Acknowledgments**

First and foremost, we thank our research advisor, Dr. Sadler (Southern Methodist University), for encouraging us to pursue this project. Your mentorship, expertise, and guidance have been invaluable. We thank the International Institute of Forecasters for providing the M3-Competition data.

We thank our faculty advisor, Dr. Jacquelyn Cheun-Jensen (Southern Methodist University), for diligent guidance and manuscript review. Finally, a special thank you to our families and all who supported us; we are infinitely grateful for your patience and constancy.

## References

1. Adya, M., Armstrong, J. S., Collopy, F., & Kennedy, M. (2000). An application of rule-based forecasting to a situation lacking domain knowledge. *International Journal of Forecasting*, *16* (4), 477–484. https://doi.org/10.1016/s0169-2070(00)00074-1
2. Barker, J. (2020). Machine learning in M4: What makes a good unstructured model? *International Journal of Forecasting*, *36* (1), 150–155. https://doi.org/10.1016/j.ijforecast.2019.06.001
3. Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, *120*, 70–83. https://doi.org/10.1016/j.csda.2017.11.003
4. Flores, B. E., & Pearce, S. L. (2000). The use of an expert system in the M3-Competition. *International Journal of Forecasting*, *16* (4), 485–496. https://doi.org/10.1016/s0169-2070(00)00068-6
5. Goodwin, P. (2020). Performance measurement in the M4-Competition: Possible future research. *International Journal of Forecasting*, *36* (1), 189–190. https://doi.org/10.1016/j.ijforecast.2019.02.015
6. Hill, C., Li, J., Schneider, M. J., & Wells, M. T. (2020). The Tensor Auto-regressive model. *Journal of Forecasting*, *40* (4), 636–652. https://doi.org/10.1002/for.2735
7. Hyndman, R. J. (2020). A brief history of forecasting competitions. *International Journal of Forecasting*, *36* (1), 7–14. https://doi.org/10.1016/j.ijforecast.2019.03.015

8. Koning, A. J., Franses, P. H., Hibon, M., & Stekler, H. O. (2005). The M3-Competition: Statistical tests of the results. *International Journal of Forecasting*, *21* (3), 397–409. https://doi.org/10.1016/j.ijforecast.2004.10.003
9. Lv, S.-X., Peng, L., Hu, H., & Wang, L. (2022). Effective machine learning model combination based on Selective Ensemble strategy for time series forecasting. *Information Sciences*, *612*, 994–1023. https://doi.org/10.1016/j.ins.2022.09.002
10. Makridakis, S., & Hibon, M. (2000). The M3-competition: Results, conclusions and implications. *International Journal of Forecasting*, *16* (4), 451–476. https://doi.org/10.1016/s0169-2070(00)00057-1
11. Makridakis, S., & Petropoulos, F. (2020). The M4-Competition: Conclusions. *International Journal of Forecasting*, *36* (1), 224–227. https://doi.org/10.1016/j.ijforecast.2019.05.006
12. Makridakis, S., Assimakopoulos, V., & Spiliotis, E. (2018). Objectivity, reproducibility and replicability in forecasting research. *International Journal of Forecasting*, *34* (4), 835–838. https://doi.org/10.1016/j.ijforecast.2018.05.001
13. Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and Ways Forward. *PLOS ONE*, *13* (3). https://doi.org/10.1371/journal.pone.0194889
14. Makridakis, S., Spiliotis, E., &amp; Assimakopoulos, V. (2020). Predicting/hypothesizing the findings of the M4-Competition. *International Journal of Forecasting*, 36 (1), 29–36. https://doi.org/10.1016/j.ijforecast.2019.02.012
15. Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 Accuracy Competition: Results, findings, and conclusions. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2021.11.013
16. Rubio, G., Pomares, H., Rojas, I., & Herrera, L. J. (2011). A heuristic method for parameter selection in LS-SVM: Application to time series prediction. *International Journal of Forecasting*, *27* (3), 725–739. https://doi.org/10.1016/j.ijforecast.2010.02.007
17. Sbrana, G., & Silvestrini, A. (2022). Random coefficient state-space model: Estimation and performance in M3–M4-Competitions. *International Journal of Forecasting*, *38* (1), 352–366. https://doi.org/10.1016/j.ijforecast.2021.06.003
18. Seong, B. (2020). Smoothing and forecasting mixed-frequency time series with vector exponential smoothing models. *Economic Modelling*, *91*, 463–468. https://doi.org/10.1016/j.econmod.2020.06.020
19. Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, *36* (1), 75–85. https://doi.org/10.1016/j.ijforecast.2019.03.017
20. Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, *16* (4), 437–450. https://doi.org/10.1016/s0169-2070(00)00065-0

# Appendix

Table 1 from *The M3-Competition: results, conclusions and implications* describing the origin of the datasets used.

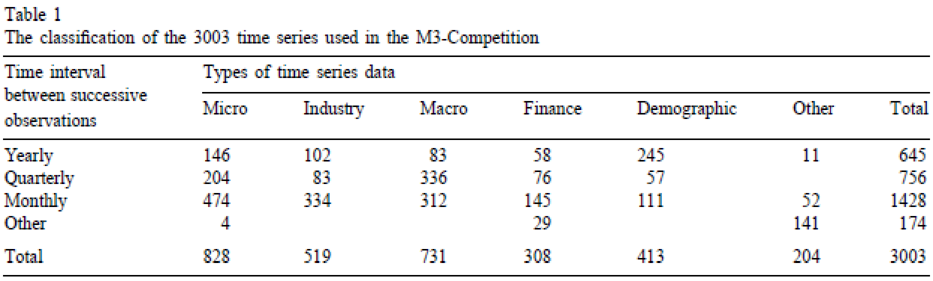


Table 1
The classification of the 3003 time series used in the M3-Competition

| Time interval between successive observations | Types of time series data | | | | | | |
|---|---|---|---|---|---|---|---|
| | Micro | Industry | Macro | Finance | Demographic | Other | Total |
| Yearly | 146 | 102 | 83 | 58 | 245 | 11 | 645 |
| Quarterly | 204 | 83 | 336 | 76 | 57 | | 756 |
| Monthly | 474 | 334 | 312 | 145 | 111 | 52 | 1428 |
| Other | 4 | | | 29 | | 141 | 174 |
| Total | 828 | 519 | 731 | 308 | 413 | 204 | 3003 |

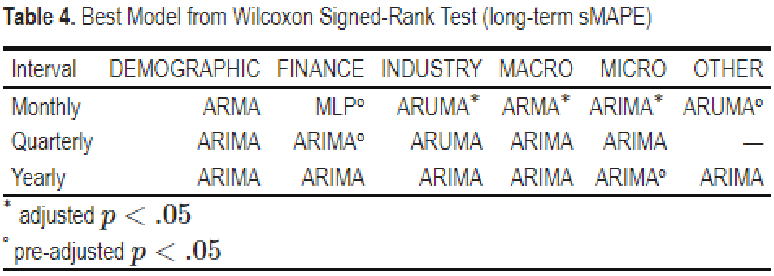


**Table 4.** Best Model from Wilcoxon Signed-Rank Test (long-term sMAPE)

| Interval | DEMOGRAPHIC | FINANCE | INDUSTRY | MACRO | MICRO | OTHER |
|---|---|---|---|---|---|---|
| Monthly | ARMA | MLP° | ARUMA* | ARMA* | ARIMA* | ARUMA° |
| Quarterly | ARIMA | ARIMA° | ARUMA | ARIMA | ARIMA | — |
| Yearly | ARIMA | ARIMA | ARIMA | ARIMA | ARIMA° | ARIMA |

* adjusted $p < .05$

° pre-adjusted $p < .05$

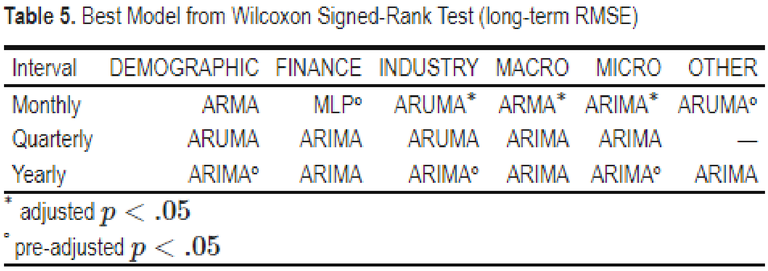


**Table 5.** Best Model from Wilcoxon Signed-Rank Test (long-term RMSE)

| Interval | DEMOGRAPHIC | FINANCE | INDUSTRY | MACRO | MICRO | OTHER |
|---|---|---|---|---|---|---|
| Monthly | ARMA | MLP° | ARUMA* | ARMA* | ARIMA* | ARUMA° |
| Quarterly | ARUMA | ARIMA | ARUMA | ARIMA | ARIMA | — |
| Yearly | ARIMA° | ARIMA | ARIMA° | ARIMA | ARIMA° | ARIMA |

* adjusted $p < .05$

° pre-adjusted $p < .05$