Forecasting Microbiology Laboratory Test Volumes using Time Series Models

LOUIS DANIËL LIZARAZO FUENTES, University of Twente, The Netherlands

Hospitals find themselves in an increasingly uncertain financial situation. A key reason for this being the increase in patient related costs. Specifically laboratory costs stand out after having increased at a compound annual growth rate of 8.54% in the last decade.

This study addresses the predictive capabilities and comparative performance of various time series forecasting models (SARIMAX and Prophet), within a major Dutch hospital in order to aid in the forecasting ability of microbiology laboratory volumes. The findings of the study reveal that SARIMAX and Prophet both exhibit comparable predictive efficacy in forecasting microbiology laboratory test volumes. both models demonstrate a cluster with satisfactory performance, as the error metrics fall below the designated thresholds. Satisfactory being phrased as a Symmetric Mean Absolute Percentage Error (SMAPE) under 30%, and a Mean Absolute Scaled Error (MASE) below 1.00.

Nevertheless, the presence of several outliers suggests that SARIMAX and Prophet may not be optimal fits for certain datasets.

Key Words: Microbiology, Laboratory Volumes, Time Series Forecasting, SARIMAX, Prophet

1 INTRODUCTION

According to a recent report [2] published by the Dutch Assocation of Hosptials (NVZ) using data from 56 Dutch hospitals, which account for 93% of the country's Zorgverzekeringswet (Zvw) budget, 11% of hospitals are suffering from contract violations or liquidity issues. By the end of 2023, this number is expected to climb to 40% . There were a number of issues raised, including rising labor costs, energy costs, and patient-related costs such as laboratory tests. Within the last decade, Dutch national costs for medical laboratories (such as microbiology, pathology or toxicology) have grown at a compound annual growth rate of 8.54% according to the Central Statistics Bureau [1]. Another reason for concern listed was a potential change in contract terms such as tariff compensation [2]. When discussing contract terms and pricing structures hospitals use a variety of data analysis and data management tools to support their understanding of anticipated expenses. However, these could still be improved upon. While addressing contract terms and pricing structures, hospitals employ various methods to enhance their comprehension of expected expenses, yet there is room for further improvement with the increasing accessibility of newer and more powerful methods.

TScIT 40, February 2, 2024, Enschede, The Netherlands

Various studies [3–6; 15; 21; 38; 40; 41] have explored predictive methodologies in relation to forecasting healthcare related metrics, such as medical waste [5; 21], emergency department visits [3; 6; 15; 20; 40] or COVID related metrics [38]. There is an absence of research, using patient data, for the use case of forecasting of laboratory volumes.

Thus this study aims to address this gap by specifically focusing on forecasting microbiology laboratory test volumes using real-world laboratory data from a prominent Dutch hospital. The primary objective of this study is to assess the predictive efficacy of SARIMAX and Prophet in forecasting microbiology laboratory test volumes. Additionally, the study aims to evaluate the comparative performance of SARIMAX and Prophet in this context. Finally the financial impact of the forecasts and their implications for management will also be discussed.

2 RELATED WORK

The practice of making forecasts in healthcare related settings has a longstanding history. Noteworthy are the recent studies examining machine learning models and time series models with the aim of forecasting a diverse range of healthcare related metrics. Even now a distinct knowledge gap exists in the forecasting of laboratory volumes. This section provides an overview of literature focusing on forecasting and comparing accuracy of machine learning and time series models in healthcare settings.

Numerous studies, as detailed in Table 1, have extensively examined the performance of diverse models in forecasting a broad spectrum of medical metrics. The effectiveness of these models appears to be highly dependent on the specific use case, with performance exhibiting variations across the examined time series. Notably the Seasonal Autoregressive Integrated Moving Average Exogenous model (SARIMAX) emerged as a consistently robust performer in multiple studies. SARIMAX however is described to be computationally intensive [38] and requires proper parameter selection. In a notable work by Ghysels et al. [23] focusing on forecasting seasonal time series, the authors highlighted that linear models and some of their variants require fewer observations compared to their more complex peers. This attribute can prove advantageous in scenarios with limited data availability. Additionally another time series model, Prophet, is recognized for its flexibility and built in support for covariates such as Dutch holidays. Moreover, Prophet demonstrates resilience against outliers and trend shifts, exhibiting flexibility even when dealing with non-stationary data. This adaptability is particularly advantageous, considering that real-world time series data often deviates from perfect stationarity even after undergoing mathematical transformations.

Models	Metrics	Accuracy Metric	Validation	Authors
SARIMA, Prophet, LR, RF, XGBoost, LSTM	Pharmacy purchase orders	MAE	Growing Window Prequential	Almentero et al. (2021) [4]
Rolling Average, Holt-winters, VAR, Schweigler et al, Whitt et al, SARIMAX	Emergency department hourly occupancy	MAE, MAPE, MSE	Holdout OOS (73/27)	Cheng et al. (2021) [15]
AR, Holt-Winters, SARIMA, Prophet, LR, ElasticNet, XGBoost, GLM, Ensamble	Daily emergency department arrivals	MAE, MAPE, R-squared, Pearson correlation	Holdout OOS (86/14), Growing Window Prequential	Álvarez-Chaves et al. (2023) [6]
Cubist tree, SVM	Hospital Length of Stay	PMAE, R-squared, precision, Recall, Accuracy, AUC	Holdout OOS (75/25)	Turgeman et al. (2017) [41]
RF, AdaBoost, GBMs, Ensamble	Medical waste quantities	MAE, RMSE, R-squared, MAPE	K-fold CV (k=5)	Erdebilli et al. (2022) [21]
Kernel-based SVM, Maxout activation Deep Learning	Medical waste quantities	MAE, RMSE, R-Squared	Holdout OOS (70/30)	Altin et al. (2023) [5]
LR, SVR, ESM, SARIMAX, BSTS, Prophet, RF, XGBoost, LSTM	Number of COVID fatalities	RSME	Growing Window Prequential	Simmons et al. (2023)[38]
SARIMA, SARIMAX, GLM, Prophet	Total daily arrivals, Daily peak occupancy	MAPE, Accuracy, Sensitivity, Specificity, AUC	Growing Window Prequential, Holdout OOS (68/32)	Tuominen et al. (2021)[40]
XGBoost, AdaBoost, MLP	Admission status of emergency patients	AUC, sensitivity, specificity, F1, accuracy	Holdout OOS (80/20)	Ahmed et al. (2022) [3]

Table 1. **LR** Linear Regression, **AR** AutoRegressive, **SARIMA** Seasonal AutoRegressive Integrated Moving Average, **VAR** Vector AutoRegressive, **ESM** Exponential Smoothing Model, **GLM** Generalized Linear Model, **RF** Random Forest, **SVM** Support Vector Machine, **GBM** Gradient Boosting Machine, **SVR** Support Vector Regression, **BSTS** Bayesian Structural Time Series, **LSTM** Long Short-Term Memory, **MLP** Multi-Layer Perceptron, **MAE** Mean Absolute Error, **MAPE** Mean Absolute Percentage Error, **PMAE** Precision Mean Absolute Error, **RMSE** Root Mean Squared Error, **AUC** Area Under the Curve, **OOS** Out-of-Sample, **CV** Cross-Validation

3 METHODOLOGY

As previously mentioned two models stand out, SARIMAX and Prophet, which will be implemented, tuned and evaluated in this study.

3.1 SARIMAX

Seasonal Autoregressive Integrated Moving-Average with Exogenous Regressors (SARIMAX) is an extension of the ARIMA model which was initially proposed by Box and Jenkins [11]. SARIMAX incorporates the capability to account for both seasonality and exogenous variables, making it a versatile time series forecasting model. SARIMAX among other time series models have been extensivly lined out by Hyndman and Athanasopoulos in their book *Forecasting: Principles and Practice* [27].

The SARIMAX model consists of several components:

(1) **Autoregressive order AR(p):**
In an autoregression model, we forecast the variable of interest using a linear combination of past values of the variable. The term autoregression indicates that it is a regression of the variable against itself.

Thus, an autoregressive model of order p can be written as

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

Where ε_t is white noise. This resembles a multiple regression but with lagged values of y_t as predictors. We refer to this as an AR(p) model, an autoregressive model of order p.

(2) **Moving Average order, MA(q):**
Rather than using past values of the forecast variable in a regression, a moving average model uses past forecast errors in a regression-like model.

$$y_t = c + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q},$$

Where ε_t is white noise. We refer to this as an MA (q) model, a moving average model of order q.

(3) **Differencing order I(d) :** An ARIMA model is an ARMA model yet with a preprocessing step included in the model that we represent using I(d). I(d) is the difference order, which is the number of transformations needed to make the data stationary.

$$\underbrace{(1-B)^d y_t}_{\substack{\uparrow \\ d \text{ differences}}}$$

(4) **Seasonality**
A seasonal ARIMA model is formed by including additional seasonal terms in the ARIMA model. It is written as follows:

$$\text{ARIMA} \quad \underbrace{(p,d,q)}_{\substack{\uparrow \\ \text{Non-seasonal part} \\ \text{of the model}}} \quad \underbrace{(P,D,Q)_m}_{\substack{\uparrow \\ \text{Seasonal part of} \\ \text{of the model}}}$$

Where m = number of observations per year (sometimes also denoted as s). Uppercase notations are used for the seasonal parts of the model, and lowercase notations are used for the non-seasonal parts of the model.

The seasonal part of the model consists of terms that are similar to the non-seasonal components of the model, but involve backshifts of the seasonal period. The additional seasonal terms are simply multiplied by the non-seasonal terms.

$$y_t = c + \sum_{n=1}^{p} \alpha_n y_{t-n} + \sum_{n=1}^{q} \theta_n \epsilon_{t-n} + \sum_{n=1}^{P} \phi_n y_{t-sn} + \sum_{n=1}^{Q} \eta_n \epsilon_{t-sn} + \epsilon_t$$

(5) **Exogenous (X):** This component includes the impact of external factors or predictors on the time series. Exogenous variables are additional features that are not part of the time series but can influence its behavior.

$$\sum_{n=1}^{r} \beta_n x_{n_t}$$

This results in the following overall model for SARIMAX:

$$y_t = c + \sum_{n=1}^{p} \alpha_n y_{t-n} + \sum_{n=1}^{q} \theta_n \epsilon_{t-n} + \sum_{n=1}^{r} \beta_n x_{n_t} + \sum_{n=1}^{P} \phi_n y_{t-sn} + \sum_{n=1}^{Q} \eta_n \epsilon_{t-sn} + \epsilon_t$$

Where:

- y_t is the observed time series at time t.
- α_n, ϕ_p are autoregressive parameters.
- θ_n, η_n are moving average parameters.
- β_n is the coefficient for the exogenous variable X_t.
- ϵ_t is the white noise error term.

Various manners of calculating or estimating the aforementioned parameters exist. A widely used approach is to employ automated model selection techniques, and one popular method is the AutoARIMA algorithm.

AutoARIMA is an automated time series forecasting algorithm that systematically searches through different combinations of parameters to identify the optimal model for a given time series. The algorithm uses criteria such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) to evaluate the goodness of fit for each model candidate [27].

The steps involved in the AutoARIMA algorithm are as follows:

Step 1: Initial Model Fitting
Start with an initial model, often a simple (S)ARIMA model, and fit it to the time series data.

Step 2: Iterative Model Search
Iteratively explore different combinations of parameters, including autoregressive order (p), differencing order (d), moving average order (q), seasonal autoregressive order (P), seasonal differencing order (D), and seasonal moving average order (Q). The search is guided by minimizing AIC or BIC.

Step 3: Select Optimal Model
Choose the model with the lowest AIC or BIC as the optimal AutoARIMA model.

Step 4: Refit and Forecast
Refit the selected model to the entire time series data and make future forecasts.

The AutoARIMA algorithm provides a convenient way to automatically identify the most suitable model for a given time series, saving time and effort in manual model selection and parameter tuning. However this process is time consuming and puts a lot of weight into a single evaluator. Another method is to determine the parameters using various mathematical methods. This can save time due to not having to explore the entire solution space. Due to the number of time series being evaluated in this study this was chosen as the preferred method.

- **Auto-regression order (p)**
The estimation of p is done through the use of the auto correlation function (ACF) [11; 27; 30; 37]. The ACF defines how data points in a time series are related, on average, to the preceding data points.
- **Differencing order (d)**
Various methods of calculating the required order of differncing are available. Very commonly used ones are the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) [28], the augmented Dicky-Fuller test (ADF) [19] and the Phillips-Perron test (PP) [32].

(1) **ADF**:
ADF is commonly used to test for the presence of a unit root in the time series, indicating non-stationarity. It is suitable when you want to determine the order of differencing (d) needed to achieve stationarity.

(2) **KPSS**:
KPSS complements the ADF test by testing the null hypothesis of stationarity around a deterministic trend. It is useful for checking whether a series is trend-stationary.

(3) **PP:**
PP is an alternative to the ADF test and is used to test for the presence of a unit root in a time series.

Some of these methods complement each other thus generally multiple tests are examined [18]. In this study only ADF is used to determine the required differencing order, consequently additional tests remain a point for future work.

- **Moving average order (q)**
The estimation of q is done through the use of the partial auto correlation function (PACF) [27; 30; 37]. The PACF provides the partial correlation of a stationary time series with its own lagged values, adjusting for the values of the time series at all shorter lags. This is in contrast to the ACF, which does not account for the influence of other lags.

- **Seasonal auto-regressive order (P)**
The estimation of P is done through the use of the ACF, similarly to the calculation of the auto-regression order p[27; 30; 37].

- **Seasonal differencing order (D)**
Literature mentiones different preferences for seasonal unit root tests in order to estimate the appropriate seasonal differencing order. Tests also complement each other [26].

(1) **Dickey, Hasza, Fuller (DHF) [17]**
This test is employed for assessing the presence of unit roots in a time series. It is particularly useful for testing the stationarity of a series.

(2) **Phillips-Perron (PP) [32]**
The Phillips-Perron test is another method for testing the null hypothesis of a unit root in a time series. It is often used as an alternative to the Dickey-Fuller test.

(3) **HylleBerg, Engle, Granger, Yoo (HEGY) [25]**
The HylleBerg, Engle, Granger, Yoo test is a test for unit roots that considers various lag structures. It offers flexibility in capturing different patterns in the data.

(4) **Osborn, Chui, Smith and Birchenhall (OCSB) [31]**
The Osborn, Chui, Smith, and Birchenhall test is designed to assess the null hypothesis that a seasonal unit root is present in a time series. It specifically focuses on detecting seasonality-related non-stationarity.

(5) **Canova and Hansen (CH) [12]**
The Canova and Hansen test examines the absence of a seasonal unit root in a time series. It is particularly relevant for analyzing time series data with seasonal patterns, providing insights into the stationarity of the seasonal component.

Various studies, as illustrated by Lopes, Rodrigues and Osborn [16; 34], showcase preferences for different methods. This diversion highlights the absence of a consensus on a single method suitable for all cases. In this paper, the OCSB method will be employed based on the findings of Rodrigues and Osborn, who identified it as a performing and effective approach [34].

- **Seasonal moving average order (Q)**
Similarly to its non-seasonal counterpart, the seasonal moving average order (Q) is calculated using the PACF[27; 30; 37]. It helps determine the number of lagged observations of the seasonal moving average component to include in the model.

- **Length of the seasonal cycle (s/m)**
This parameter (s or m) denotes the number of observations between successive occurrences of a seasonal pattern. In this case, as there are 12 months in a year, the length of the seasonal cycle (s/m) is set to 12 for monthly data. In the case of quarterly data this parameter would be set to 4 [27].

3.1.1 Invertible models. Ensuring invertibility is essential for both Moving Average (MA) and AutoRegressive (AR) models, and subsequently expanded models such as SARIMAX. An invertible MA model ensures that lagged forecast errors' coefficients remain stable, contributing to the model's interpretability and forecast accuracy. Similarly, invertibility in AR models guarantees well-behaved autoregressive coefficients, maintaining model stability [27].

3.2 Prophet

The Prophet model was proposed by Taylor and Letham [39]. It is designed for time series forecasting and decomposes a time series into three main components: trend ($g(t)$), seasonality ($s(t)$), and holidays ($h(t)$), along with an error term (ϵ_t). The model can be expressed as:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

3.2.1 Trend Function ($g(t)$). The trend function captures nonperiodic changes in the time series. For forecasting problems without saturating growth, the trend is modeled as a piece-wise constant rate of growth:

$$g(t) = \left(k + \mathbf{a}(t)^\top \boldsymbol{\delta}\right) t + \left(m + \mathbf{a}(t)^\top \boldsymbol{\gamma}\right)$$

Where k and m are constants, $\mathbf{a}(t)$ is a vector of binary indicators for holidays, and $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ are vectors of coefficients.

3.2.2 Seasonality ($s(t)$). Seasonality captures periodic changes in the time series, such as weekly and yearly patterns. The seasonality component is modeled as a sum of harmonic functions:

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi n t}{P}\right) + b_n \sin\left(\frac{2\pi n t}{P}\right) \right)$$

Where P is the period of the seasonality, and a_n and b_n are coefficients.

3.2.3 Holiday Effects ($h(t)$). The holiday component represents the effects of holidays, which may occur irregularly. Holiday effects are incorporated by assuming independence and assigning a parameter k_i for each holiday i. A matrix of regressors $Z(t)$ is generated, representing indicator functions for each holiday.

$$Z(t) = [\mathbf{1}(t \in D_1), \ldots, \mathbf{1}(t \in D_L)]$$

The holiday component is then defined as:

$$h(t) = Z(t)\kappa$$

Where κ is a vector of parameters for each holiday. To capture effects for a window of days around a holiday, additional parameters are included, treating each day in the window as a holiday itself.

In summary, Prophet provides a flexible framework for time series forecasting by considering trends, seasonality, and holiday effects in a decomposed manner such as initially described by Harvey and Peters. [24]. The model parameters are estimated using a Bayesian approach, and the forecast is generated by combining these components.

3.3 Validation methods

As shown in the related works section (Table 1), a plethora of accuracy metrics are used for evaluating forecasts. Botchkarev [10] categorized numerous performance metrics and provided a clear overview of available metrics based on the method of determining point distance, method of normalization, and method of aggregation of point distances over a dataset.

Several were considered for application in this study, Botchkarev [10] noted that no consensus exists on the ever increasing list of metrics. In this context, the metrics chosen are Mean Absolute Error (MAE), Symmetric Mean Absolute Percentage Error (SMAPE) and Mean Absolute Scaled Error (MASE), each with distinct advantages such as laid out in Table 2.

Various methods for estimating model performance are used to showcase accuracy metrics. However in the literature, a consensus regarding the optimal approach has not been found, as evidenced by several studies on the subject [7–9; 13; 14; 29; 35]. In the context of time series forecasting, Cerqueira et al. [13] systematically examined various techniques. The general categories and commonly used methods examined are laid out in Table 3. As laid out by Cerqueira et al. [13] a repeated OOS Holdout approach showed competitive performance in real world data sets. Consequently, the same method is used in this study.

In order to compare the performance of both models, a paired t-test is employed. The paired t-test is a widely used statistical method for comparing the means of paired samples. The paired t-test is deemed suitable under certain assumptions. These include a paired design, normality of differences, independence of observations, and interval data [42].

3.4 Anomalies

The presence of anomalous historical data points can exert a notable impact on forecasting accuracy. In addressing this challenge, Freeman et al. [22] provided a comprehensive framework for selecting anomaly detection methods. Ultimately, inspired by the work of Almentero et al. [4], a binary dummy variable approach was adopted to accommodate prolonged anomalous periods, specifically during the COVID-19 pandemic.

4 STUDY DESIGN

A systematic approach for designing and executing data-related projects is the Cross-Industry Standard Process for Data Mining (CRISP-DM) [36], comprising six main components:

(1) **Business Understanding:** To comprehend project objectives and requirements from a business perspective.

(2) **Data Understanding:** Collection and analysis of data to gain a preliminary understanding of information availability and quality.

(3) **Data Preparation:** Cleaning, transforming, and preparing data for analysis.

(4) **Modeling:** Selection and application of various modeling techniques to prepared data.

(5) **Evaluation:** Assessing the efficacy of the model in meeting business objectives.

(6) **Deployment:** Ensuring the project delivers value to stakeholders.

4.1 Business Understanding

The hospital sought deeper insights into future costs and worrying trends, making volume forecasting a crucial step towards achieving this objective.

4.2 Data Understanding

The dataset utilized in this paper comprises confidential historical microbiology order records obtained from a prominent Dutch hospital. The data spans from January 2016 to August 2023. The key attributes included in the dataset which were considered relevant for the purposes of this study were the inquiry code, count, and amount.

4.3 Data preparation

The first step was to remove redundant columns, then split by code and sum the counts on a monthly basis. Certain codes only came into use more recently and thus did not have a full history from January 2016 to August 2023.

Series with less than 48 months of non-zero counts or an overall count of tests under 700 in their respective time periods were omitted. This was done in order to ensure the relevance of the predictions for the stakeholders.

Metric	Description
Mean Absolute Error (MAE)	MAE measures the average absolute differences between predicted and actual values. It is particularly useful for comparing different models on the same dataset, providing a clear indication of the magnitude of errors in the predictions. • Less sensitive to outliers then squared metrics. • Not scale-independent
Symmetric Mean Absolute Percentage Error (SMAPE)	SMAPE is a percentage-based metric that considers both the magnitude and direction of errors. It is chosen over the traditional MAPE due to its symmetry, treating overestimation and underestimation equally. SMAPE is scale-independent, making it suitable for comparing forecasting accuracy across different datasets. It expresses errors as a percentage of the actual values, providing a relative measure of accuracy. • Symmetric, treating overestimation and underestimation equally. • Scale-independent, suitable for diverse datasets. • Percentage-based interpretation. • Sensitive to small actual values.
Mean Absolute Scaled Error (MASE)	MASE is a scale-independent metric that evaluates the accuracy of predictions by considering the mean absolute error relative to the mean absolute error of a naïve baseline model (calculated from the training set mean). It provides a normalized measure of accuracy, allowing for comparisons between models and datasets. • Scale-independent. • Provides a normalized measure of accuracy. • Assumes the baseline model is always available and effective.

Table 2. Accuracy Metrics [10]

Method	Description
Out-of-Sample (OOS)	
Holdout	A method where a portion of the dataset is set aside for validation.
Rep-Holdout	Repeated application of the holdout method to different subsets of the data.
Cross-Validation (CV)	
Standard, Randomized K-Fold Cross-Validation	Randomly partitions the data into K folds for training and testing.
Blocked K-Fold Cross-Validation	Divides data into blocks before applying K-fold cross-validation.
Prequential	
Prequential Evaluation in Blocks in a Growing Fashion	Incrementally evaluates model performance as new data is added.
Prequential Evaluation in Blocks in a Sliding Fashion	Continuously assesses model performance in a sliding window fashion.

Table 3. Evaluation Methods [7–9; 13; 14; 29; 35]

4.4 Modeling

SARIMAX and Prophet were implemented using Jupyter Notebooks in Visual Studio Code. SARIMAX was chosen as it performed quite well in several studies and lower data requirements while Prophet was chosen due to it's flexibility in adding covariates such as the Dutch holidays, handling outliers, handling trend shifts and handling non stationary data.

Both models were further enhanced by incorporating a binary COVID dummy variable similarly to Alementero et al. [4]. This variable served as a reinforcing factor, introducing a binary distinction related to the presence or absence of COVID-related influences. The period from February 2020 to February 2022 was classified as the COVID period.

Forecasts were clipped at a value of zero as volumes can only be zero or positive.

4.4.1 SARIMAX. SARIMAX parameters were calculated in the following manner:

(1) **Autoregressive Order (p):** Significant lags are selected based on the common 95% confidence interval leading to a significance value of 0.5 to be applied on the ACF [11; 27; 30; 37].

(2) **Differencing Order (d):** The input is differenced until the null hypothesis can be rejected from a ADF test [19], with a maximum value of 2.

(3) **Moving Average Order (q):** Significant lags are selected based on the common 95% confidence interval leading to a significance value of 0.5 to be applied on the PACF [27; 30; 37].

(4) **Seasonal Autoregressive Order (P):** Significant seasonal lags are selected based on the common 95% confidence interval, leading to a significance value of 0.5 to be applied on the seasonal ACF [11; 27; 30; 37].

(5) **Seasonal Integrated Order (D):** The input is seasonally differenced until the null hypothesis can be rejected using the OCSB test [31], with a maximum value of 2.

(6) **Seasonal Moving Average Order (Q):** Significant seasonal lags are selected based on the common 95% confidence interval, leading to a significance value of 0.5 to be applied on the seasonal PACF[27; 30; 37].

(7) **Seasonal cycle length (s/m):** The seasonal cycle length (s/m) is predetermined as 12, given the monthly nature of the data.

4.4.2 Prophet. Certain configurations for the Prophet model were adjusted from the default settings to align with the unique characteristics of the data. Given that seasonality could not be presumed to have a constant additive factor, Prophet was configured to adopt multiplicative seasonality. This deliberate choice ensures that seasonality, holiday effects, and any additional regressors are all modeled in a multiplicative manner. Subsequently Dutch holidays and a binary COVID dummy variable were incorporated into the model to account for their potential impact. The implementation adhered to established conventions outlined in [33].

4.5 Evaluation

This study uses a Repeated Out-of-Sample (REP-OOS) approach based on the findings of Cerqueira [13] in order to estimate model performance.

The REP-OOS testing approach had 10 repetitions per code. In each iteration, a random starting point was selected within the time series where at least 60% of the dataset was available for training. The subsequent 10% was allocated for testing.

The training and testing sets were defined by slicing the time series accordingly, allowing for a comprehensive evaluation of the models' predictive capabilities. The evaluation metrics, including Mean Absolute Error (MAE), Symmetric Mean Absolute Percentage Error (SMAPE), and Mean Absolute Scaled Error (MASE), were calculated for each repetition.

Statistical significance tests, more specifically standard T-tests, are applied to determine differences in overall averages between SMAPE and MASE between SARIMAX and Prophet.

4.6 Deployment

The primary objective of this study is to assess the predictive efficacy of SARIMAX and Prophet on real-world microbiology laboratory data. The results are communicated and provided through various channels to ensure clarity and usability.

(1) **Confidential Results:** Post-sample forecast plots, extending up to December 2025, are exclusively shared with hospital stakeholders. These visualizations incorporate confidence intervals, providing a nuanced understanding of future forecasts, an illustrative plot 3. The results are further aggregated into an annual table for each model. Due to the nature of the data a mathematical transformation has been applied on the absolute values, additionally historical volumes are multiplied by average prices (post sample volumes are multiplied by the same average price). The values will be indexed on the last year of full data (2022).

(2) **Performance Metrics:** Performance metrics for each model and their respective results are systematically recorded into a local database. These metrics serve as key indicators of the models' predictive capabilities. The aggregated results are presented through various figures, allowing for both individual and comparative performance assessments. To enhance interpretability, an interactive 3D illustration is offered, see Figure 4 for an illustrative example, enabling stakeholders to explore the performance of each model on specific codes across the three metrics: MAE, SMAPE, and MASE.

Adopting this deployment strategy ensures that stakeholders receive not only a thorough evaluation of the models' predictions but also an interface for exploring and understanding the performance metrics at a granular level.

5 RESULTS

The primary focus of this study is the evaluation of the predictive efficacy of SARIMAX and Prophet in forecasting microbiology laboratory test volumes. Overall results are assessed using SMAPE and MASE metrics. Per code values are recorded for SMAPE and MASE and split per model.

For SARIMAX the average SMAPE was 30.41%. For Prophet the average SMAPE was 35.62%. Among the forecasts with SMAPE < 30%, Prophet achieved this threshold in 20 out of 32 cases, while SARIMAX achieved it in 23 out of 32 cases. Prophet outperformed SARIMAX on this metric in 13/32 cases while in the remaining 19/32 cases SARIMAX outperformed Prophet. Refer to Figure 1 and 5 respectively for an overview and the full detailed results. Overall in total 43 out of 64 instances had a SMAPE under 30%. Prophet demonstrates superior performance on SMAPE in 13 out of 32 codes, while SARIMAX surpasses Prophet in 19 cases (refer to Figure 1).

In terms of MASE SARIMAX averaged 1.13 while Prophet averaged 1.31. In terms of MASE < 1, indicating a better than naïve forecast, Prophet achieved this threshold in 10 out of 32 cases, while SARIMAX achieved it in 12 out of 32 cases. Prophet outperformed SARIMAX on this metric in 13/32 cases while in the remaining 19/32 cases SARIMAX outperformed Prophet. Refer to Figure 2) and 4 respectively for an overview and the full detailed results. Overall in total 22 out of 64 instances had a MASE under 1.00. Prophet outperforms SARIMAX on MASE in 13 cases, whereas SARIMAX outperforms Prophet in 19 cases for this particular metric (refer to Figure 2).

As illustrated in Figures 1 and 2 results tend to cluster with a moderate amount of outliers. The clustered codes illustrate a decent fit for SARIMAX and Prophet.

However the models may have encountered difficulties in adequately fitting to specific outlier datasets. Notably, *Code_19* in Table 4 exhibits a remarkably high SMAPE of 159.95%, also the highest in the whole study, coupled with a competative MASE of 0.35. This code is characterized by a pattern of plateauing and sudden steep drops, potentially attributed to management decisions specific to this inquiry type. On the other hand, *Code_2*, also presented in Table 4, records the highest MASE in the study at 4.28. However, it is essential to acknowledge that this code corresponds to a unique dataset featuring an exceptionally large COVID peak, approximately 500% greater than regular seasonal peaks, with subsequent seemingly long term irregularities in seasonality.

While the average SMAPE for SARIMAX is 5.21 points lower than that of Prophet, and the average MASE is 0.18 lower, these differences were not found to be statistically significant.

Moreover, it is crucial to consider the practical implications of these metrics in the context of this study's stakeholders. While SARIMAX exhibits a slight advantage in average performance, the absence of statistical significance suggests that both models may offer comparable forecasting capabilities in this specific setting. Stakeholders indicated the intention to evaluate the forecasts on not yet seen data (September 2023 to December 2023). To aid in this, a table comprising the distinct combined top 5 for both SMAPE and MASE per model. For the purposes of this study the codes are obfuscated and the data indexed. See Table 7 for the SARIMAX post sample forecast and Table 6 for the Prophet post sample forecast.

The challenges posed by the absence of benchmark measures in similar real-world datasets underscore the need for rigorous statistical approaches to draw meaningful and reliable conclusions regarding model performance. This acknowledgment emphasizes the complexity of evaluating forecasting models in healthcare and the necessity for nuanced interpretations of the results.

Additionally the run time for both SARIMAX and Prophet was recorded. For this was SARIMAX between 10 and 20 minutes for 32 codes * (10 Validation repetitions + 1 post sample forecast) forecasts, for Prophet the overall run time was between 2 and 3 minutes for the forecasts indicating a similar discrepancy as noted by [38].

5.1 Limitations

An important note to make is that the naïve forecast in used in calculating MASE was the mean of the train set. One potential bias is when the time series data has a strong seasonality or periodic pattern. If the mean method is used as the naïve forecast, and the training set contains multiple seasons or cycles of the time series, the mean would capture the average value across all seasons. In such cases, the mean might align well with the central tendency of the data, resulting in a low MASE compared to more sophisticated forecasting methods. The mean is also heavily impacted by outliers.

Note that variations in the validation data can result in different accuracy metrics for the same average MAE. This discrepancy arises from the usage of different test sets for the same codes.

6 CONCLUSION

This study adds novel insights to the field of time series forecasting by offering a comparative evaluation of SARIMAX and Prophet in forecasting microbiology laboratory test volumes. By utilizing real-world data from a major Dutch hospital, this research contributes to understanding the applicability and performance of these models in a this setting.

The findings of the study reveal that SARIMAX and Prophet both exhibit comparable predictive efficacy in forecasting microbiology laboratory test volumes. Specifically, SARIMAX achieved SMAPE values below 30% in 23 out of 32 datasets and MASE values below 1.00 in 12 datasets. In contrast, Prophet, with 20 datasets below 30% in SMAPE and 10 datasets below 1.00 in MASE, displays slightly less favorable performance.

The averaged performance metrics provide additional insights into the models' reliability. On average, Prophet exhibits an SMAPE of 35.62% and a MASE of 1.31. In comparison, SARIMAX maintains an average SMAPE of 30.41% and an average MASE of 1.13. Importantly, the comparative analysis reveals no statistically significant differences in predictive accuracy between SARIMAX and Prophet.

Clustering appeared in both SMAPE and MASE around adequate (SMAPE < 30.0%: 43/64, MASE < 1.00: 22/64) performance values,

indicating that SARIMAX and Prophet displayed a good fit. Nevertheless, the presence of several outliers suggests that SARIMAX and Prophet may not be optimal fits for certain datasets.

Stakeholders at the hospital indicated great interest in comparing the results with not yet modeled data and continuing research in this direction.

7 FUTURE WORK

Integration of Change Points in Prophet Model: The potential enhancement of forecasting precision through the integration of change points in the Prophet model is a noteworthy suggestion. Adjusting the default change points allows the model to adapt to abrupt shifts or variations in the time series data more accurately. This adjustment could lead to improved accuracy, especially in the presence of sudden changes such as seen throughout the COVID-19 pandemic.

Consideration of External Factors: Recognizing the impact of external factors on forecasting accuracy, the exploration of additional variables such as weather conditions, regional population dynamics, and other relevant factors is recommended. Including these external variables in the models might provide a more comprehensive understanding of the influencing factors, potentially leading to more accurate predictions.

Investigation of Alternative Models: Evaluating additional models beyond SARIMAX and Prophet could uncover alternatives that might better capture underlying trends better. Combining the strengths of multiple models could lead to more reliable predictions.

Clipping at 0 Instead of Log Transformation: The choice of clipping forecasts and bounds at 0, as opposed to transforming, raises an methodological consideration. The impact of this choice on the bias of the forecasts should be thoroughly examined, as different transformations can influence the handling of outliers and skewed distributions. As Hyndman and Athanasopoulos noted a Box-Cox transformations could be an option given that the backtransformation is adjusted for bias [27].

Additional evaluation procedures: Including additional metrics beyond SMAPE and MASE could provide a more comprehensive evaluation of the models' performance. Additional validation methods such as variants of cross validation and prequential approaches could be beneficial.

Generalizability and Applicability: While the models were applied across a diverse set of financial codes, a deeper understanding of their applicability and generalizability is deemed essential. Investigating how the models perform in relation to underlying test volumes and cross-relations can contribute to refining their practical utility.

8 ACKNOWLEDGMENTS

The author extends his heartfelt gratitude to Marcos Machado and Daniela Guericke for their roles as supervisors, providing invaluable guidance and support throughout this study.

REFERENCES

[1] Medische laboratoria en trombosediensten; financiën en personeel, 12 2022.
[2] Impactanalyse kosteninflatie ziekenhuizen 2023, 7 2023.
[3] Abdulaziz Ahmed, Omar Ashour, Haneen Ali, and Mohammad Firouz. An integrated optimization and machine learning approach to predict the admission status of emergency patients. *Expert Systems with Applications*, 202:117314, 9 2022.
[4] Bruno Kinder Almentero, Jiye Li, and Camille Besse. Forecasting pharmacy purchases orders. pages 1–8. IEEE, 11 2021.
[5] Fatma Gül Altin, İbrahim Budak, and Fatma Özcan. Predicting the amount of medical waste using kernel-based svm and deep learning methods for a private hospital in turkey. *Sustainable Chemistry and Pharmacy*, 33:101060, 6 2023.
[6] Hugo Alvarez-Chaves, Pablo Muñoz, and María D. R-Moreno. Machine learning methods for predicting the admissions and hospitalisations in the emergency department of a civil and military hospital. *Journal of Intelligent Information Systems*, 7 2023.
[7] Christoph Bergmeir and Jose M. Benitez. Forecaster performance evaluation with cross-validation and variants. pages 849–854. IEEE, 11 2011.
[8] Christoph Bergmeir and José M. Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 5 2012.
[9] Christoph Bergmeir, Rob J. Hyndman, and Bonsoo Koo. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83, 4 2018.
[10] Alexei Botchkarev. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. 9 2018.
[11] George E. P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Wiley, 1970.
[12] Fabio Canova and Bruce E. Hansen. Are seasonal patterns constant over time? a test for seasonal stability. *Journal of Business & Economic Statistics*, 13:237–252, 7 1995.
[13] Vitor Cerqueira, Luis Torgo, and Igor Mozetič. Evaluating time series forecasting models: an empirical study on performance estimation methods. *Machine Learning*, 109:1997–2028, 11 2020.
[14] Vitor Cerqueira, Luis Torgo, Jasmina Smailovic, and Igor Mozetic. A comparative study of performance estimation methods for time series forecasting. pages 529–538. IEEE, 10 2017.
[15] Qian Cheng, Nilay Tanik Argon, Christopher Scott Evans, Yufeng Liu, Timothy F. Platts-Mills, and Serhan Ziya. Forecasting emergency department hourly occupancy using time series analysis. *The American Journal of Emergency Medicine*, 48:177–182, 10 2021.
[16] Artur C.B. da Silva Lopes. The robustness of tests for seasonal differencing to structural breaks. *Economics Letters*, 71:173–179, 5 2001.
[17] D. A. Dickey, D. P. Hasza, and W. A. Fuller. Testing for unit roots in seasonal time series. *Journal of the American Statistical Association*, 79:355–367, 6 1984.
[18] David A Dickey. Stationarity issues in time series models. *SAS Users Group International*, 30, 2015.
[19] David A. Dickey and Wayne A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74:427–431, 6 1979.
[20] Diego Duarte and Julio Faerman. Comparison of time series prediction of healthcare emergency department indicators with arima and prophet. pages 123–133. Aircc publishing Corporation, 12 2019.
[21] Babek Erdebilli and Burcu Devrim-İçtenbaş. Ensemble voting regression based on machine learning for predicting medical waste: A case from turkey. *Mathematics*, 10:2466, 7 2022.
[22] Cynthia Freeman, Jonathan Merriman, Ian Beaver, and Abdullah Mueen. Experimental comparison and survey of twelve time series anomaly detection algorithms. *Journal of Artificial Intelligence Research*, 72:849–899, 11 2021.
[23] Eric Ghysels, Denise R. Osborn, and Paulo M.M. Rodrigues. *Chapter 13 Forecasting Seasonal Time Series*, pages 659–711. 2006.
[24] A. C. Harvey and S. Peters. Estimation procedures for structural time series models. *Journal of Forecasting*, 9:89–108, 3 1990.
[25] S. Hylleberg, R.F. Engle, C.W.J. Granger, and B.S. Yoo. Seasonal integration and cointegration. *Journal of Econometrics*, 44:215–238, 4 1990.
[26] Svend Hylleberg. Tests for seasonal unit roots general to specific or specific to general? *Journal of Econometrics*, 69:5–25, 9 1995.
[27] Rob J. Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.
[28] Denis Kwiatkowski, Peter C.B. Phillips, Peter Schmidt, and Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54:159–178, 10 1992.
[29] Zhe Liu and Xiangfeng Yang. Cross validation for uncertain autoregressive model. *Communications in Statistics - Simulation and Computation*, 51:4715–4726, 8 2022.
[30] Robert Nau. Identifying the numbers of ar or ma terms in an arima model.
[31] Denise R. Osborn, A. P. L. Chui, Jeremy P. Smith, and C. R. Birchenhall. Seasonality and the order of integration for consumption. *Oxford Bulletin of Economics and*

Statistics, 50:361–377, 11 1988.
[32] P.C.B. Phillips and P. Perron. Testing for a unit root in time series regression. *Biometrika*, 75:335–346, 1988.
[33] Greg Rafferty. *Forecasting Time Series Data with Facebook Prophet*. Packt Publishing, 3 2021.
[34] Paulo M. M. Rodrigues and Denise R. Osborn. Performance of seasonal unit root tests for monthly data. *Journal of Applied Statistics*, 26:985–1004, 12 1999.
[35] Matthias Schnaubelt. A comparison of machine learning model validation schemes for non-stationary time series data, 2019.
[36] Colin Shearer. The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, 5:13–22, 2000.
[37] Robert H. Shumway and David S. Stoffer. *ARIMA Models*, pages 75–163. 2017.
[38] Susan Simmons, Kornelia Bastin, Aric LaBarr, and Christopher Healey. A comparison of the prediction capabilities of large scale time series algorithms. *Medical Research Archives*, 11, 2023.
[39] Sean J. Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72:37–45, 1 2018.
[40] Jalmari Tuominen, Antti Roine, Taavi Saviauk, Anssi Seppo, Miika Pihlaja, Jani Ovaska, Satu-Liisa Pauniaho, and Niku Oksala. Forecasting daily arrivals and peak occupancy in a combined emergency department, 1 2021.
[41] Lior Turgeman, Jerrold H. May, and Roberta Sciulli. Insights from a machine learning model for predicting the hospital length of stay (los) at the time of admission. *Expert Systems with Applications*, 78:376–385, 7 2017.
[42] Robert S Witte and John S Witte. *Statistics*. John Wiley & Sons, 2017.

9 APPENDIX

Code	MAE	SMAPE	MASE
Code_1	29.3	20.06	1.14
Code_2	577.01	127.56	4.28
Code_3	21.43	97.5	1.51
Code_4	8.4	38.6	1.21
Code_5	20.77	65.28	2.32
Code_6	3.6	50.96	1.01
Code_7	74.25	12.43	0.91
Code_8	67.47	8.19	0.86
Code_9	13.64	28.0	1.63
Code_10	45.36	11.31	0.77
Code_11	51.1	13.95	1.23
Code_12	427.45	32.71	0.86
Code_13	151.57	32.06	1.93
Code_14	133.73	12.96	1.71
Code_15	12.85	26.9	1.48
Code_16	14.66	93.41	1.2
Code_17	4.49	45.14	0.91
Code_18	28.06	16.35	1.4
Code_19	21.2	159.95	0.35
Code_20	30.39	27.53	1.11
Code_21	81.46	14.58	1.03
Code_22	396.61	18.7	1.17
Code_23	130.68	17.91	0.95
Code_24	14.19	16.35	1.19
Code_25	7.54	12.86	0.77
Code_26	98.4	9.05	1.05
Code_27	396.47	8.06	1.09
Code_28	5.99	25.75	1.34
Code_29	9.29	38.47	1.77
Code_30	123.02	13.39	2.02
Code_31	84.52	6.82	0.78
Code_32	10.15	37.13	0.82

Table 4. Average results for each code over 10 repetitions Prophet

Code	MAE	SMAPE	MASE
Code_1	19.68	13.46	0.91
Code_2	396.81	159.76	3.67
Code_3	19.44	75.55	1.09
Code_4	7.94	35.93	1.14
Code_5	13.93	38.96	0.9
Code_6	3.44	50.82	1.0
Code_7	114.74	18.5	1.16
Code_8	75.84	9.31	0.95
Code_9	10.26	22.73	1.33
Code_10	85.28	19.54	1.3
Code_11	50.09	14.07	1.11
Code_12	355.54	22.73	0.71
Code_13	116.61	27.82	1.32
Code_14	231.1	21.46	2.07
Code_15	11.9	25.91	1.41
Code_16	16.5	107.87	1.76
Code_17	5.11	54.52	1.03
Code_18	26.69	14.88	1.37
Code_19	18.35	41.46	0.3
Code_20	31.18	27.36	1.28
Code_21	94.2	16.62	1.02
Code_22	370.91	16.26	1.13
Code_23	129.39	17.12	1.01
Code_24	14.17	15.27	1.01
Code_25	6.64	11.24	0.69
Code_26	113.37	10.22	1.1
Code_27	664.77	13.12	1.64
Code_28	5.51	22.8	1.09
Code_29	10.93	40.55	1.62
Code_30	89.21	9.62	1.38
Code_31	100.11	8.15	0.89
Code_32	11.36	42.59	1.01

Table 5. Average results for each code over 10 repetitions SARIMAX

Code	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
Code_8	101.09	98.03	97.34	96.2	98.54	100.58	100.0	104.03	105.4	106.04
Code_10	104.3	153.25	130.57	113.92	104.28	103.01	100.0	105.47	100.08	97.77
Code_19	1071.26	720.73	502.34	310.86	104.61	95.76	100.0	59.68	0.0	0.0
Code_25	86.48	82.46	86.48	88.36	96.12	99.42	100.0	108.85	112.44	113.4
Code_26	109.0	97.76	97.23	91.89	90.19	94.48	100.0	105.83	109.4	111.97
Code_27	124.44	105.43	104.25	98.27	96.16	102.15	100.0	104.73	105.02	105.64
Code_31	93.37	91.86	97.28	95.04	96.03	101.92	100.0	104.17	105.76	107.09
Code_32	136.09	137.13	131.11	120.21	80.51	73.54	100.0	133.68	119.41	118.04

Table 6. Prophet: Adjusted historical data and forecasts for distinct top 5 of combined SMAPE/MASE, index year 2022

Code	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
Code_5	41.59	26.55	35.39	45.13	143.79	252.21	100.0	63.8	32.4	39.91
Code_8	101.09	98.03	97.34	96.2	98.54	100.58	100.0	104.83	104.35	104.95
Code_12	47.17	43.4	42.13	41.28	51.86	54.67	100.0	99.91	96.12	95.76
Code_19	1071.26	720.73	502.34	310.86	104.61	95.76	100.0	74.29	2.2	0.0
Code_25	86.48	82.46	86.48	88.36	96.12	99.42	100.0	108.14	108.82	111.32
Code_26	109.0	97.76	97.23	91.89	90.19	94.48	100.0	104.38	105.0	105.02
Code_27	124.44	105.43	104.25	98.27	96.16	102.15	100.0	104.48	105.76	105.69
Code_30	112.01	107.04	103.03	98.66	94.48	101.72	100.0	106.24	105.96	105.91
Code_31	93.37	91.86	97.28	95.04	96.03	101.92	100.0	104.59	104.81	106.11

Table 7. SARIMAX: Adjusted historical data and forecasts for distinct top 5 of combined SMAPE/MASE, index year 2022

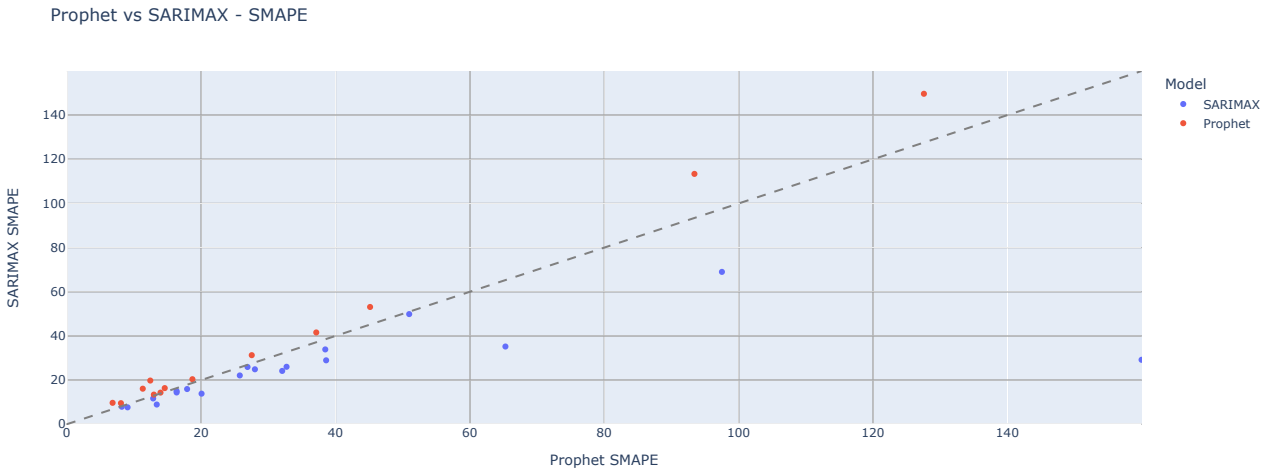

Fig. 1. SMAPE scatterplot for SARIMAX and Prophet on each code.

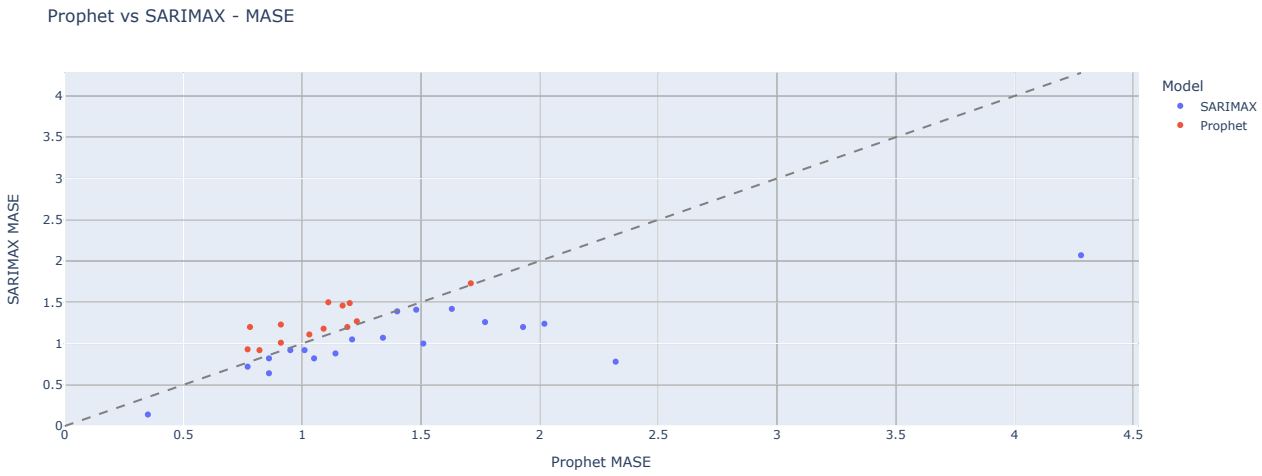

Fig. 2. MASE scatterplot for SARIMAX and Prophet on each code.

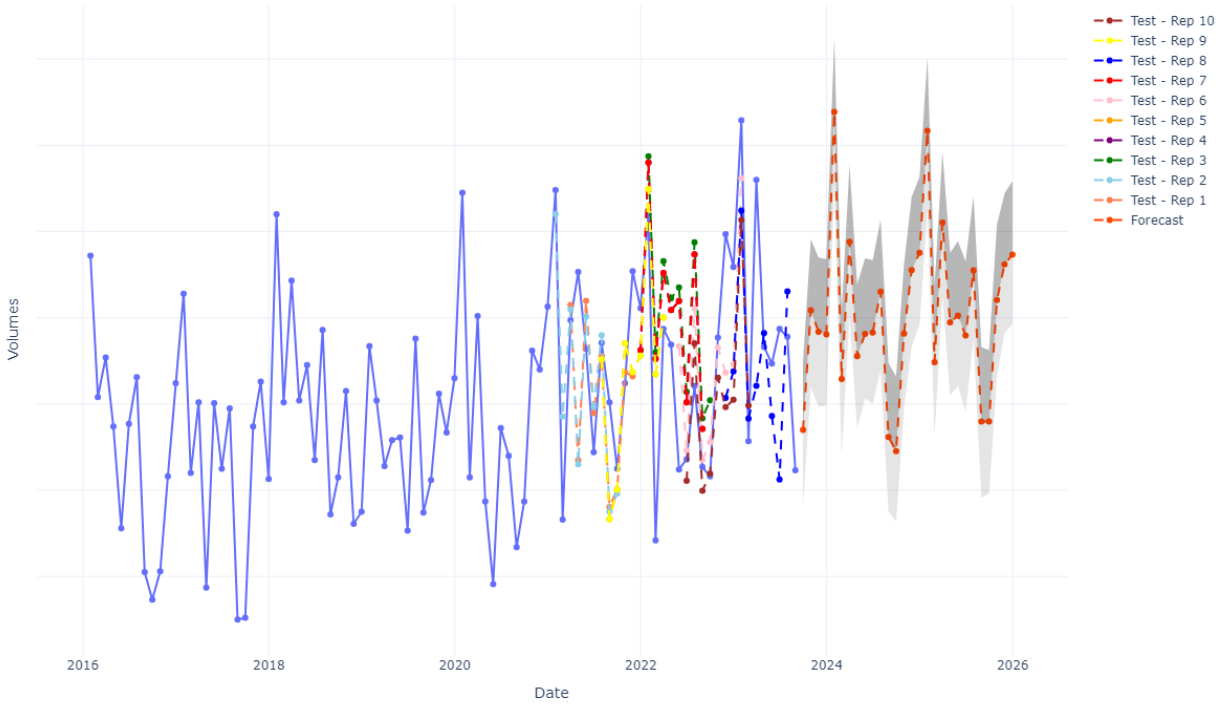

Fig. 3. Illustrative plot of Code_31 including the REP-OOS test forecasts and a post sample forecast with confidence intervals.

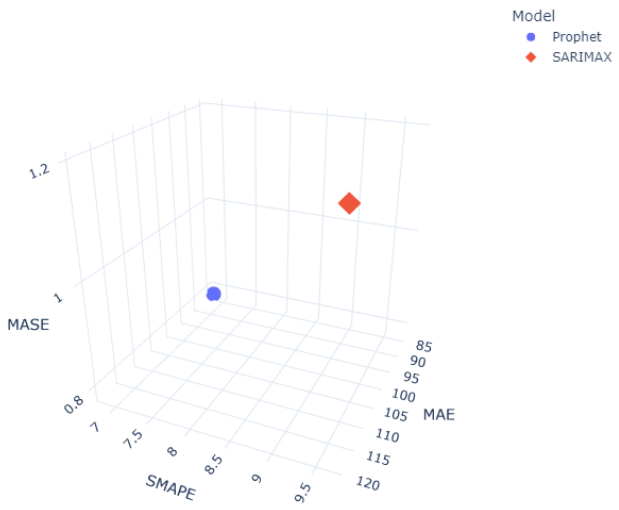

Fig. 4. 3D plot of Code_31 illustrating the model differences on MAE, SMAPE and MASE.