Benchmarking Pre-Trained Time Series Models for Electricity Price Forecasting

Timothée Hornek*, Amir Sartipi*, Igor Tchappi*, and Gilbert Fridgen*
*SnT - Interdisciplinary Center for Security, Reliability and Trust
University of Luxembourg
Kirchberg, Luxembourg
Email: {timothee.hornek, amir.sartipi, igor.tchappi, gilbert.fridgen}@uni.lu

Abstract

Accurate electricity price forecasting (EPF) is crucial for effective decision-making in power trading on the spot market. While recent advances in generative artificial intelligence (GenAI) and pre-trained large language models (LLMs) have inspired the development of numerous time series foundation models (TSFMs) for time series forecasting, their effectiveness in EPF remains uncertain.

To address this gap, we benchmark several state-of-the-art pretrained models—Chronos-Bolt, Chronos-T5, TimesFM, Moirai, Time-MoE, and TimeGPT—against established statistical and machine learning (ML) methods for EPF. Using 2024 day-ahead auction (DAA) electricity prices from Germany, France, the Netherlands, Austria, and Belgium, we generate daily forecasts with a one-day horizon.

Chronos-Bolt and Time-MoE emerge as the strongest among the TSFMs, performing on par with traditional models. However, the biseasonal MSTL model, which captures daily and weekly seasonality, stands out for its consistent performance across countries and evaluation metrics, with no TSFM statistically outperforming it.

Index Terms

electricity price forecasting, generative artificial intelligence, benchmark, pre-trained time-series models

I. Introduction

Recent advances in generative artificial intelligence (GenAI), particularly the widespread adoption of pre-trained large language models (LLMs), have catalyzed the development of analogous models tailored specifically for time series data, commonly known as time series foundation models (TSFMs). Developers describe these models as "universal forecasters", capable of providing robust predictions in various domains [1]. Their training on extensive datasets enables their application in various contexts without the need for additional training, relying solely on their inference capabilities. A specific field of application is electricity price forecasting (EPF), a well-established time series forecasting task focused on wholesale electricity prices [2]. In Europe, for example, the primary electricity auction, namely the day-ahead auction (DAA), takes place daily at noon, determining a single clearing price for each hour of the following day [3].

Researchers are currently discussing several TSFMs in the literature. For example, Chronos models are built on preexisting LLM architectures, using the T5 model architecture, and trained on a combination of existing and synthetic time series data [4]. Similarly, TimesFM is trained on a large corpus of time series data, employing a decoder-only architecture [5]. Time-MoE uses a sparse mixture-of-experts (MoE) architecture, which activates only a subset of its parameters during inference, preserving computational efficiency despite the high total number of parameters. The model is trained on an extensive dataset that contains 300 billion time points [6]. Morai was trained on the LOTSA dataset, containing more than 27 billion observations, to train its masked encoder architecture [1]. The weights of the models mentioned so far are publicly available. In contrast, TimeGPT is a closed source pre-trained model designed to emphasize simplicity and ease of use by enabling predictions through application programming interface (API) calls [7]. All these models provide zero-shot forecasting (i.e., forecasting without fine-tuning or additional training) capabilities, enabling them to use their inference abilities to generate forecasts without requiring any additional prior training.

Forecasting with TSFMs remains an active area of research across various fields, including finance and the energy domain. For instance, the authors in [8] fine-tuned TimesFM [5] for financial time series forecasting, outperforming several statistical and deep learning (DL) benchmark models. In the energy domain, the authors in [9] trained transformer-based models from scratch

This research was funded in part by the Luxembourg National Research Fund (FNR) and PayPal, PEARL grant reference 13342933/Gilbert Fridgen. For the purpose of open access, and in fulfillment of the obligations arising from the grant agreement, the author has applied a Creative Commons Attribution 4.0 International (CC BY 4.0) license to any Author Accepted Manuscript version arising from this submission. The research was carried out as part of a partnership with the energy retailer Enovos Luxembourg S.A.

and compared their performance with TSFMs for household load forecasting. They found that TimesFM [5] outperformed the models trained from scratch, particularly when increasing the input size. In another study focusing on household power consumption data [10], the authors benchmarked statistical and machine learning (ML) models against TSFMs for data imputation tasks. Their findings suggest that TSFMs can improve imputation performance. Focusing on EPF, the authors in [11] investigated the integration of market sentiment and bidding behavior into price forecasting models. By fine-tuning LLMs and using conditional generative adversarial networks (CTSGAN), they achieved improved performance in forecasting price spikes. Similarly, focusing on EPF, the authors in [12] categorized time series forecasting approaches into four groups (econometric, deep learning, transformer-based, and LLMs) and conducted a benchmarking study. Their benchmark emphasized DL models using transformer-based architectures, testing the models on electricity price data from several European countries. The results indicated that transformer-based models achieved superior performance in forecasting electricity prices, with the TSFM TimesFM [5] only slightly less performant.

Despite promising initial applications of TSFMs for EPF, their performance in the EPF domain remains understudied. This unclarity also arises from the wide variety of TSFMs discussed in the literature, leading to questions about their comparative performance. Furthermore, it remains unclear how TSFMs perform relative to traditional statistical and ML models in the context of EPF.

How effective are these so-called universal forecasters [1] in the domain of EPF? This paper seeks to answer this question. We propose a comprehensive benchmark of the state of the art TSFMs, including variations in model sizes, as some models come in versions with different numbers of parameters. Additionally, we include statistical and ML models in the benchmark to allow a comparison not only between TSFMs but also with more traditional forecasting modeling approaches. Our evaluation uses commonly used metrics, such as root mean square error (RMSE), mean absolute Error (MAE), and symmetric mean absolute percentage error (SMAPE), ensuring a robust assessment of forecast performance. Furthermore, we evaluate whether the forecast errors differ statistically using the widely applied Diebold-Mariano (DM) test [13]. To provide timely and geographically diverse results, our test interval covers the year 2024, and contains DAA prices from the largest European power markets by traded volume: Germany, France, the Netherlands, Austria, and Belgium.

We structure the remainder of this paper as follows. Section II outlines our research approach, detailing the models we test and the evaluation metrics we use. Section III describes the time series data, focusing on electricity prices, and specifies the parameterizations of the model. Section IV presents and discusses our results, including our benchmark. Finally, Section V concludes the paper.

II. Research Approach

This section outlines our experiment setup for benchmarking. We test four categories of models: baseline, statistical, and ML models, introduced in Subsection II-A, as well as a diverse set of TSFMs, detailed in Subsection II-B. We test all models in the same manner, adhering to the market schedule of the DAA, whose gate closure time is every day at noon CET [3] for the next day. To align with this market schedule, we conduct experiments daily, with each model forecasting prices for the following day, meaning that the models generate 24 price forecasts. Baseline models and TSFMs forecast prices directly using historical data without requiring training. In contrast, statistical and ML models require training prior to forecasting. To ensure these models incorporate the latest data, we retrain them daily using updated historical price data. Finally, we outline the error metrics and statistical tests used to evaluate and compare model performances in Subsection II-C.

A. Baseline, statistical, and machine learning models

We select a diverse set of time series forecasting models as reference, including three baseline models, three statistical models, and three ML models. Table I summarizes the selected models. All models, except the baseline models, require fitting to the underlying data.

TABLE I
Overview of baseline, statistical, and ML models.

Type	Name	Library	Ref.
Baseline	Naive SeasonalNaiveDay SeasonalNaiveWeek	StatsForecast 2.0.0 [14]	-
Statistical	MSTL TBATS MFLES	StatsForecast 2.0.0 [14]	[15] [16] [17]
ML	ElasticNet KNNRegressor SVR	scikit-learn 1.6.0 [18]	[19] [19] [20]

For the baseline and statistical models, we use the Python library StatsForecast [14], while we use scikit-learn [18] for the ML models. We apply each model in its default settings. The baseline models include Naive, SeasonalNaiveDay, and SeasonalNaiveWeek, which forecast using the last known value, the value from the same time on the previous day, and the value from the same time one week earlier, respectively.

We select statistical models based on their ability to capture multiple seasonalities. Specifically, we choose MSTL, TBATS, and MFLES. To simplify parameterization, we leverage the "Auto" functionality for TBATS and MFLES, referred to as "AutoTBATS" and "AutoMFLES" [14].

For the ML models, we test ElasticNet, KNNRegressor, and SVR, representing a regularized linear model, a nonparametric model, and a nonlinear model, respectively, to ensure a diverse comparison. We fit the ElasticNet model using a cross-validation procedure, denoted "ElasticNetCV" [18].

B. Time series foundation models

We select a set of popular state-of-the-art TSFMs (see Table II) for our benchmark, focusing on models capable of zero-shot forecasting. These models can perform forecasts without additional training, with pre-trained weights typically available for download. The selected models vary in complexity, measured by the number of parameters. Note that the number of activated parameters in Time-MoE is significantly smaller than the total number of parameters, due to the sparse mixture of experts (MoE) architecture [6]. To ensure reproducibility, we provide the release dates of the model weights. A notable exception is the closed-source TimeGPT model, accessible only via API, with its parameter count undisclosed [7]. For TimeGPT, we document the date of access to the API instead of the weight release dates.

TABLE II
OVERVIEW OF TSFMS.

Model	Version(s)	Params [M]	Release	Ref.
Chronos Bolt	Tiny, Mini, Small, Base	9, 21, 48, 205	Nov 26, 2024	[4]
Chronos T5	Tiny, Small, Base, Large	8, 20, 46, 200	Mar 13, 2024	[4]
Morai	Small, Base, Large	14, 91, 311	Jun 17, 2024	[1]
TimesFM	200M, 500M	200, 500	Dec 24, 2024	[5]
TimeGPT	timegpt-1	unknown	Jan 9, 2025†	[7]
TimeMoE	50M, 200M	50††, 200††	Sep 21, 2024	[6]

† Date of API calls, model weights are not public.
†† Number of activated parameters, the total number of parameters is 113M and 453M respectively.

C. Forecasting performance evaluation

We evaluate the forecasting performance of the models using widely adopted error metrics and perform DM tests to statistically compare model performances.

Regarding error metrics, we use three widely adopted metrics in time series forecasting: RMSE, see Eq. (1); MAE, see Eq. (2); and SMAPE, see Eq. (3). Here, y_i denotes the true value, and $\hat{y}_i$ represents its forecast.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

$$\text{SMAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|} \tag{3}$$

The RMSE emphasizes the mean of the errors, making it more sensitive to outliers compared to MAE, which focuses on the median of the errors. However, both metrics are scale-dependent and do not account for the magnitude of the true values. The SMAPE addresses this limitation by normalizing each absolute error relative to the average of the actual and predicted values. This normalization improves robustness against zero values in the true value y_i, which could otherwise cause instability.

The error metrics MAE, RMSE, and SMAPE are used to rank models. However, to assess whether one model statistically outperforms another, we apply the one-sided DM test [13], using the multivariate modification proposed by [21]. Traditionally,

24 independent tests are performed—one for each hour of the day [2], with separate daily error series for each hour of the day. This modification, however, aggregates daily errors into a single daily errors, reducing the number of required tests from 24 to 1. Specifically, the method aggregates the 24 hourly errors using the 1-norm (sum of absolute values) to compute a single daily error metric, which serves as input for the DM test. We use a maximal p-value of 0.1 and reject the null hypothesis if the test statistic falls below the critical value. The null hypothesis states that one model's forecasts are not significantly better than the other's. The alternative hypothesis specifies that one model has significantly superior predictive performance [13].

III. Data and parametrization

This section presents the data we use for our benchmark and the data transformations we apply in Subsection III-A. We then outline the model parameterizations in Subsection III-B.

A. Data

We use publicly available DAA market data from multiple European countries, including Germany, France, the Netherlands, Austria, and Belgium, sourced from the ENTSO-E Transparency Platform[1] for 2024 and parts of 2023. In particular, the German data set also includes Luxembourg, as both belong to the same bidding zone [22]. Although the testing period for our benchmark is limited to 2024, we incorporate data from the end of 2023 for model training and as input for forecasts at the start of 2024. European daylight saving time (DST) causes the DAA to clear for 23 hours (at the start of DST) or 25 hours (at its end) instead of the standard 24 hours. To simplify processing, we transform these days into 24 hour days: at the start of DST, we interpolate to insert an additional hour, while at the end of DST, we average to remove an hour. This preprocessing ensures that all days in the data set have a consistent 24-hour format, facilitating compatibility with statistical and ML models.

Although baselines, statistical methods, and TSFMs use raw data, we preprocess the data for ML models to enhance performance [23]. Specifically, we apply a quantile transformation, a robust technique that normalizes data by mapping them to a normal distribution based on quantiles derived from the training dataset.

B. Parametrization

In this section, we describe the model parameters used in our experiments. The baseline models do not require parameter configuration. For statistical and ML models, we use 12 weeks of historical data for training, with an input size of 1 week. This means that once trained, these models generate forecasts based on prices from the week immediately preceding the forecast period.

We configure all statistical models to account for daily and weekly seasonality. For the MFLES and ElasticNet models, we also set seven one-day windows for cross-validation (CV), corresponding to a one-week CV period necessary for model fitting.

The TSFMs use an input size of one week, consistent with the statistical and ML models.

IV. Results and Discussion

We present our benchmarking results as follows: Subsection IV-A details the error metrics, Subsection IV-B reports the DM test results, and Subsection IV-C concludes with a discussion of our results.

A. Forecasting error metric results

Table III presents the benchmarking results, evaluating the forecasts of all models. We assess each model for every country: Austria (AT), Belgium (BE), Germany (DE), France (FR), and the Netherlands (NL). For each combination, we report three error metrics: MAE, RMSE, and SMAPE. In each row, the **bold underlined** value represents the smallest error, the **bold** value denotes the second smallest, and the underlined value indicates the third smallest.

Within each model type group, we make the following observations: For the baseline models, the SeasonalNaiveDay outperforms across countries for MAE and RMSE, while the Naive model performs best for SMAPE.

Among statistical models, the MSTL model consistently achieves the best results in all countries and metrics, except for the SMAPE in France, where TBATS yields a smaller error.

Among the ML models, the ElasticNet model demonstrates superior performance across all countries and metrics.

Among the TSFMs, two models stand out: Chronos Bolt and TimeMoE. Chronos Bolt achieves strong results across countries for MAE and RMSE, while TimeMoE achieves lower SMAPE. Additionally, the Chronos Bolt models significantly outperform the Chronos T5 models. When comparing models with different numbers of parameters, no clear pattern emerges, except for the smaller models generally underperforming relative to their larger counterparts within the same type (e.g., Chronos Bolt (Tiny), Chronos T5 (Tiny), Moirai (Small)). A notable exception is TimesFM, where performance decreases from the 200M model to the 500M model.

In general, for MAE and RMSE, the best performing model for each country is the MSTL or a Chronos Bolt model (Mini, Small, or Base). For SMAPE, the top models are MSTL or TimeMoE.

[1] https://transparency.entsoe.eu/

TABLE III
PERFORMANCE EVALUATION OF FORECASTING MODELS USING MAE, RMSE, AND SMAPE METRICS ACROSS MODELS AND COUNTRIES.

Metric	Country	Naive	SeasonalNaiveDay	SeasonalNaiveWeek	MSTL	TBATS	MFLES	ElasticNet	KNNRegressor	SVR	Chronos Bolt (Tiny)	Chronos Bolt (Mini)	Chronos Bolt (Small)	Chronos Bolt (Base)	Chronos T5 (Tiny)	Chronos T5 (Mini)	Chronos T5 (Small)	Chronos T5 (Base)	Chronos T5 (Large)	Moirai (Small)	Moirai (Base)	Moirai (Large)	TimesFM (200M)	TimesFM (500M)	TimeGPT	TimeMoE (50M)	TimeMoE (200M)
MAE	AT	26.35	22.92	26.02	**17.52**	19.15	19.99	18.29	21.18	20.58	18.44	**17.48**	17.56	17.55	26.12	23.33	23.97	23.35	21.64	20.46	18.82	19.38	19.06	26.57	20.92	17.96	18.04
	BE	27.15	23.20	28.22	**17.43**	18.99	20.56	18.11	22.84	21.12	18.94	18.05	18.10	18.27	26.57	23.16	23.82	22.69	23.50	20.12	19.01	19.77	19.09	25.00	20.11	**17.80**	18.40
	DE	29.16	27.82	32.81	**20.69**	22.95	23.86	**21.23**	26.52	24.80	22.19	21.61	21.52	21.65	31.34	27.02	29.08	26.74	27.72	23.84	22.67	23.20	22.36	30.88	23.62	22.32	21.61
	FR	26.89	20.36	27.80	16.50	17.19	17.79	17.13	21.65	20.36	16.73	16.15	**16.02**	**16.03**	22.70	21.78	20.54	21.38	21.55	19.07	17.59	18.25	17.25	21.41	18.05	16.69	17.14
	NL	29.31	26.19	29.95	**19.85**	22.27	23.37	20.16	24.20	22.79	20.94	**20.14**	20.39	20.35	27.56	25.29	26.80	27.48	25.83	22.53	21.01	21.88	21.16	28.28	22.46	20.58	20.75
RMSE	AT	43.17	38.40	45.02	31.08	32.57	35.42	32.12	36.07	34.83	31.48	**30.73**	31.00	**29.89**	42.30	38.09	39.98	39.65	36.25	34.77	32.52	33.34	32.55	39.86	34.90	32.90	30.96
	BE	38.74	33.24	40.99	**25.34**	27.07	29.38	26.56	32.99	30.75	27.23	**25.99**	26.38	26.57	38.72	33.56	35.28	32.97	34.27	28.98	27.79	28.72	27.68	35.00	29.16	26.02	27.28
	DE	48.07	44.31	55.25	**35.89**	38.33	40.97	36.82	44.86	41.72	36.50	36.10	36.40	**35.85**	50.00	44.75	48.54	44.15	45.12	38.82	38.49	39.02	36.43	46.00	39.75	39.45	36.86
	FR	35.44	28.34	37.09	22.09	22.58	23.33	23.37	29.62	27.11	22.64	21.86	**21.66**	**21.75**	31.54	30.67	28.54	30.19	30.11	25.51	23.66	24.62	23.13	29.41	24.38	22.46	23.10
	NL	46.33	40.84	48.33	**32.18**	35.04	37.41	33.43	40.53	36.76	33.11	**32.29**	33.08	32.54	42.93	40.63	40.74	43.62	42.30	35.41	33.73	35.52	33.50	41.91	36.01	33.56	33.80
SMAPE [%]	AT	19.81	20.97	22.81	**15.95**	16.52	16.79	16.93	18.91	18.87	17.21	16.51	16.54	16.78	23.46	21.32	21.77	21.33	19.72	17.56	16.66	16.63	17.33	24.38	18.08	**16.09**	16.36
	BE	23.87	25.61	28.79	**19.60**	20.03	20.53	20.50	23.99	22.92	21.48	20.67	20.78	21.09	27.91	25.35	26.08	24.78	25.52	21.24	20.68	20.57	20.72	27.54	21.08	**19.57**	20.12
	DE	23.81	26.32	29.49	**19.92**	20.80	21.14	21.06	24.21	24.66	21.46	21.44	21.34	21.43	30.18	26.23	27.44	25.43	25.87	22.02	21.26	20.77	20.99	29.19	21.29	20.39	**20.23**
	FR	29.98	30.56	36.42	25.61	**25.20**	25.66	26.56	30.91	29.43	26.26	25.68	25.73	25.76	32.86	31.72	30.58	33.02	33.09	26.74	26.03	26.16	26.01	31.73	26.10	**25.17**	25.99
	NL	23.41	25.34	27.84	**19.68**	20.61	20.84	20.82	22.67	23.01	21.15	20.72	21.02	20.78	27.01	23.83	26.35	26.60	24.55	21.43	20.47	20.14	20.50	27.51	21.09	**19.77**	20.03

B. Diebold-Mariano test results

Fig. 1 presents the DM test results for Germany, the largest European power market by traded volume. Due to space constraints, we provide the results for other countries in the appendix. We present the test results as heat maps that illustrate the p-values of the DM test. In these heat maps, p-values above 0.1 are depicted in black, indicating acceptance of the null hypothesis—that neither model is significantly better than the other. Lower p-values indicate cases where the forecast on the x-axis is significantly more performant than that on the y-axis. For example, in Germany, the MSTL model statistically outperforms all Chronos T5 models, but is not significantly more performant than the ElasticNet model. For Germany, we make several observations. The results of the DM test indicate that most of the models outperform the baseline models (Naive, SeasonalNaiveDay, and SeasonalNaiveWeek), with notable exceptions being Chronos Bolt (Tiny) and TimesFM (500M). Furthermore, all Chronos T5 models fail to surpass the SeasonalNaiveDay baseline. Comparing TSFMs with statistical and ML models, none of the TSFMs is significantly superior to the MSTL and ElasticNet models. Among all models, MSTL surpasses the most other models in performance, with only two exceptions: ElasticNet and Chronos Bolt (Small).

The results in other markets follow a similar pattern. The Chronos T5 models generally perform poorly, with only the Large variant surpassing all naive models in Austria but failing to outperform SeasonalNaiveDay in other countries. MSTL performs well across markets, except in France, where Chronos Bolt (Small) achieves significantly higher performance. In both France and Austria, multiple Chronos Bolt models (Mini, Small, and Base) outperform ElasticNet. Additionally, these Chronos Bolt models consistently demonstrate higher performance than other TSFMs, with the exception of TimeMoE models, which they surpass only in specific countries and parameter configurations. The Chronos Bolt models (Mini, Small, and Base) have no statistically significant differences in performance between themselves in all countries tested.

C. Discussion

The superior performance of SeasonalNaiveDay over Naive in terms of MAE and RMSE emphasizes the importance of daily seasonality in EPF. In contrast, Naive's better performance in SMAPE highlights the influence of scale sensitivity, suggesting that error metrics must be carefully chosen based on forecasting objectives.

Extending the analysis to statistical and ML models, our results indicate that the MSTL model is versatile, performing well in both error metrics and statistical tests in all countries tested. This observation suggests that well-calibrated statistical approaches can remain competitive with, and in some cases rival, more complex ML models.

Focusing on TSFMs, the superior performance of Chronos Bolt over Chronos T5 in both error metrics and statistically significant performance validates the improvements introduced in the more recent Chronos Bolt model. These updates demonstrate tangible gains in forecasting performance.

Chronos Bolt models excel in minimizing MAE and RMSE while outperforming most competitors in statistical tests, although TimeMoE achieves better performance for SMAPE. Despite minor variations in performance across metrics, the different Chronos Bolt variants (Mini, Small, and Base) show no statistically significant differences in performance across the countries tested.

Model selection ultimately depends on the error metric most relevant to the application. MSTL stands out as a reliable all-rounder, consistently achieving strong performance across all metrics and statistical tests. Among TSFMs, the Chronos Bolt models (Mini, Small, and Base) consistently outperform most competitors, particularly excelling in minimizing MAE and RMSE. Similarly, TimeMoE (50M or 200M) is well-suited for otimizing SMAPE. For applications requiring parameter

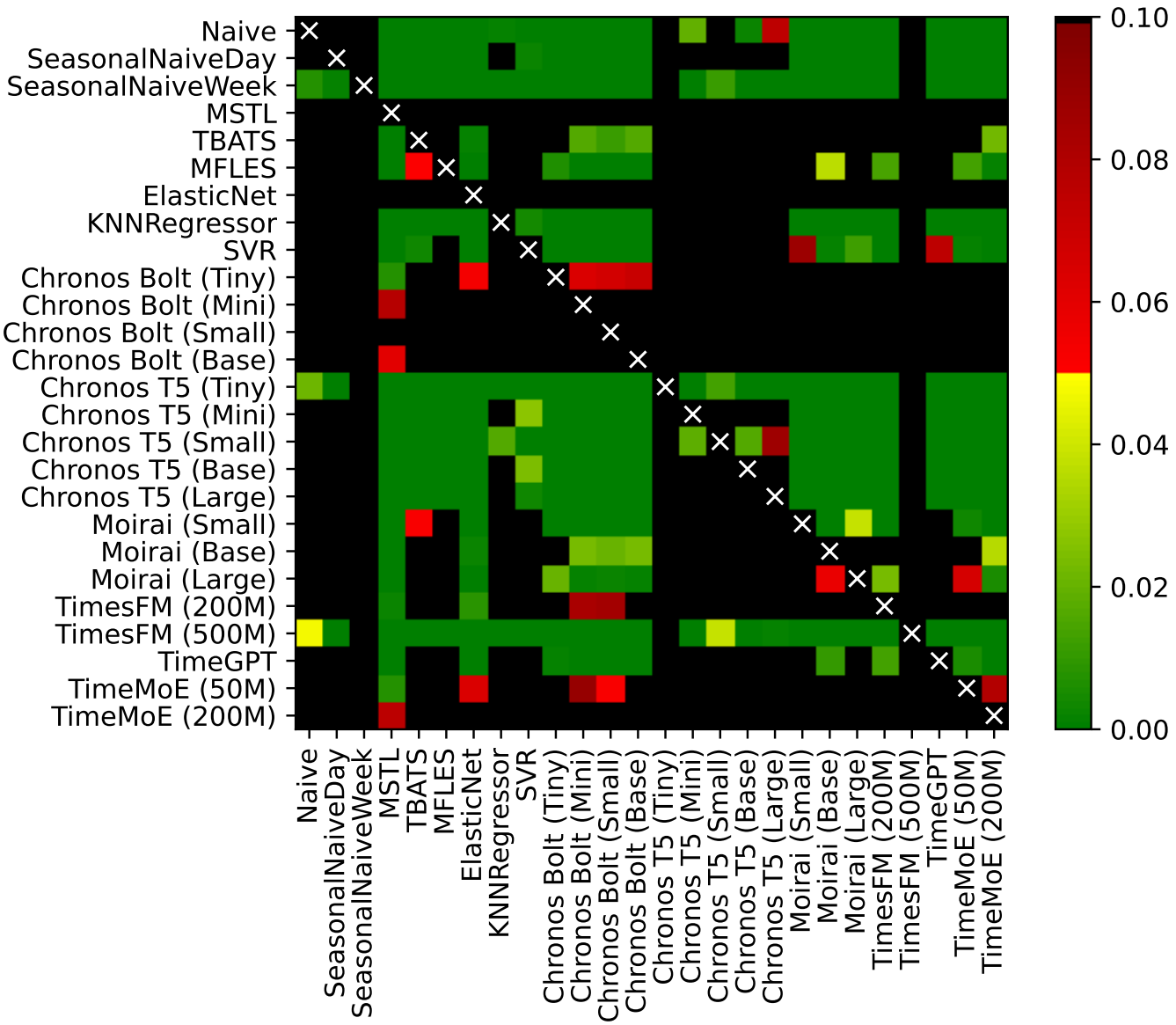

Fig. 1. DM test results for Germany.

efficiency (that is, TSFMs with fewer parameters for faster inference), Chronos Bolt (Mini) and TimeMoE (50M) offer practical choices. Depending on the relevant metric, Chronos Bolt excels in MAE and RMSE, while TimeMoE performs best for SMAPE.

Performance comparison alone does not provide a definitive recommendation, as factors like model interpretability and ease of use also influence the model selection process. A key limitation of TSFMs is their lack of interpretability, which may hinder their adoption by practitioners who require transparency for real-world decision making. However, TSFMs offer a significant advantage in ease of use, enabling inference without the need for prior training. Despite this benefit, the extensive variety of TSFMs may pose a challenge for adoption, as selecting the most suitable model remains a nontrivial task.

V. Conclusion

This paper provides a comprehensive benchmark of TSFMs, evaluating their performance in the largest European power markets using multiple error metrics. Our findings reveal that some TSFMs can compete with statistical and ML forecasting methods, with MSTL consistently demonstrating robust performance across metrics and countries. However, identifying a single optimal TSFM remains challenging, as different models excel for different error metrics (e.g., Chronos Bolt for MAE and RMSE, and TimeMoE for SMAPE). Additionally, increasing the number of parameters does not guarantee improved

performance, complicating the choice of both the type and size of the model. The lack of interpretability in TSFMs further limits their applicability in use cases where explainability is crucial.

For noncritical applications, where ease of use and minimal training requirements outweigh performance and interpretability concerns, TSFMs offer a practical solution. However, for scenarios that require high performance and interpretability, statistical methods remain the most reliable choice.

Our study has several limitations, which present opportunities for future research. We did not incorporate exogenous variables into our forecasting models, although their inclusion could significantly enhance performance. Additionally, the study's geographical scope is restricted to five European countries, leaving room for analysis in other markets. Moreover, our evaluation relies on a single model configuration. Future research could investigate the effects of varying the input sizes for TSFMs and extending the forecast horizon.

ACKNOWLEDGMENT

During the preparation of this work the authors used ChatGPT and Writefull in order to improve readability and language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

[1] G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, and D. Sahoo, "Unified Training of Universal Time Series Forecasting Transformers," May 2024, arXiv:2402.02592 [cs]. [Online]. Available: http://arxiv.org/abs/2402.02592

[2] R. Weron, "Electricity price forecasting: A review of the state-of-the-art with a look into the future," *International Journal of Forecasting*, vol. 30, no. 4, pp. 1030–1081, Oct. 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207014001083

[3] European Commission, "Consolidated text: Commission Regulation (EU) 2015/1222 of 24 July 2015 establishing a guideline on capacity allocation and congestion management (Text with EEA relevance)," 2021. [Online]. Available: http://data.europa.eu/eli/reg/2015/1222/2021-03-15

[4] A. F. Ansari, L. Stella, C. Turkmen, X. Zhang, P. Mercado, H. Shen, O. Shchur, S. S. Rangapuram, S. P. Arango, S. Kapoor, J. Zschiegner, D. C. Maddix, H. Wang, M. W. Mahoney, K. Torkkola, A. G. Wilson, M. Bohlke-Schneider, and Y. Wang, "Chronos: Learning the Language of Time Series," Nov. 2024, arXiv:2403.07815 [cs]. [Online]. Available: http://arxiv.org/abs/2403.07815

[5] A. Das, W. Kong, R. Sen, and Y. Zhou, "A decoder-only foundation model for time-series forecasting," Apr. 2024, arXiv:2310.10688 [cs]. [Online]. Available: http://arxiv.org/abs/2310.10688

[6] X. Shi, S. Wang, Y. Nie, D. Li, Z. Ye, Q. Wen, and M. Jin, "Time-MoE: Billion-Scale Time Series Foundation Models with Mixture of Experts," Oct. 2024, arXiv:2409.16040 [cs]. [Online]. Available: http://arxiv.org/abs/2409.16040

[7] A. Garza, C. Challu, and M. Mergenthaler-Canseco, "TimeGPT-1," May 2024, arXiv:2310.03589 [cs]. [Online]. Available: http://arxiv.org/abs/2310.03589

[8] X. Fu, M. Hirano, and K. Imajo, "Financial Fine-tuning a Large Time Series Model," Dec. 2024, arXiv:2412.09880 [q-fin]. [Online]. Available: http://arxiv.org/abs/2412.09880

[9] M. Meyer, D. Zapata, S. Kaltenpoth, and O. Müller, "Benchmarking Time Series Foundation Models for Short-Term Household Electricity Load Forecasting," Oct. 2024, arXiv:2410.09487 [cs]. [Online]. Available: http://arxiv.org/abs/2410.09487

[10] A. Sartipi, J. D. Fernandez, S. P. Menci, and A. Magitteri, "Bridging Smart Meter Gaps: A Benchmark of Statistical, Machine Learning and Time Series Foundation Models for Data Imputation," Jan. 2025, arXiv:2501.07276 [cs]. [Online]. Available: http://arxiv.org/abs/2501.07276

[11] X. Lu, J. Qiu, Y. Yang, C. Zhang, J. Lin, and S. An, "Large Language Model-based Bidding Behavior Agent and Market Sentiment Agent-Assisted Electricity Price Prediction," *Policy and Regulation IEEE Transactions on Energy Markets*, pp. 1–13, 2024, conference Name: Policy and Regulation IEEE Transactions on Energy Markets. [Online]. Available: https://ieeexplore.ieee.org/document/10804210/?arnumber=10804210

[12] A.-V. Andrei, G. Velev, F.-M. Toma, D. T. Pele, and S. Lessmann, "Energy Price Modelling: A Comparative Evaluation of four Generations of Forecasting Methods," Nov. 2024, arXiv:2411.03372 [cs]. [Online]. Available: http://arxiv.org/abs/2411.03372

[13] F. Diebold and R. Mariano, "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, vol. 13, no. 3, pp. 253–63, 1995, publisher: American Statistical Association. [Online]. Available: https://econpapers.repec.org/article/besjnlbes/v_3a13_3ay_3a1995_3ai_3a3_3ap_3a253-63.htm

[14] F. Garza, M. M. Canseco, C. Challú, and K. G. Olivares, "StatsForecast: Lightning fast forecasting with statistical and econometric models," *PyCon: Salt Lake City, UT, USA*, 2022.

[15] K. Bandara, R. J. Hyndman, and C. Bergmeir, "MSTL: A Seasonal-Trend Decomposition Algorithm for Time Series with Multiple Seasonal Patterns," Jul. 2021. [Online]. Available: https://arxiv.org/abs/2107.13462v1

[16] A. M. De Livera, R. J. Hyndman, and R. D. Snyder, "Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing," *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1513–1527, Dec. 2011, publisher: ASA Website _eprint: https://doi.org/10.1198/jasa.2011.tm09771. [Online]. Available: https://doi.org/10.1198/jasa.2011.tm09771

[17] B. Tyler, "MFLES," 2024. [Online]. Available: https://github.com/tblume1992/MFLES

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html

[19] G. James, D. Witten, T. Hastie, R. Tibshirani, and others, *An introduction to statistical learning*. Springer, 2013, vol. 112.

[20] V. Vapnik, "Statistical learning theory," *John Wiley & Sons*, vol. 2, pp. 831–842, 1998.

[21] F. Ziel and R. Weron, "Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks," *Energy Economics*, vol. 70, pp. 396–420, Feb. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S014098831730436X

[22] ENTSO-E, "Bidding Zone Technical Report 2021," European Network of Transmission System Operators for Electricity, Tech. Rep., 2021. [Online]. Available: https://www.entsoe.eu/news/2021/11/18/entso-e-publishes-its-2021-bidding-zone-technical-report-providing-a-transparent-and-factual-information-on-grid-congestions-in-the-eu/

[23] B. Uniejewski, R. Weron, and F. Ziel, "Variance Stabilizing Transformations for Electricity Spot Price Forecasting," *IEEE Transactions on Power Systems*, vol. 33, no. 2, pp. 2219–2229, Mar. 2018, conference Name: IEEE Transactions on Power Systems.

APPENDIX

The following figures present additional results, specifically the DM test outcomes for the studied countries, excluding Germany, which we reported in the main part of the paper.

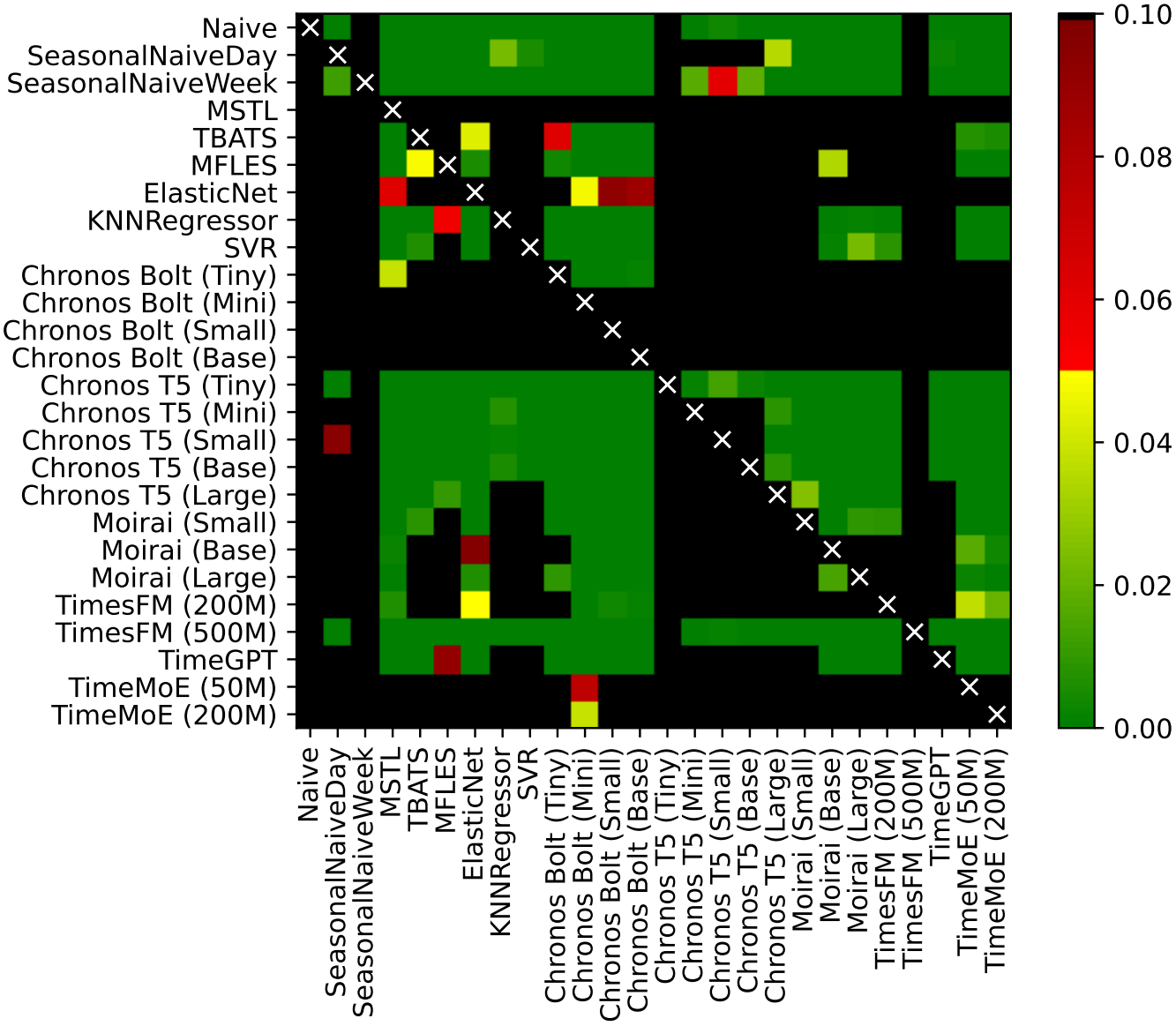

Fig. 2. DM test results for Austria.

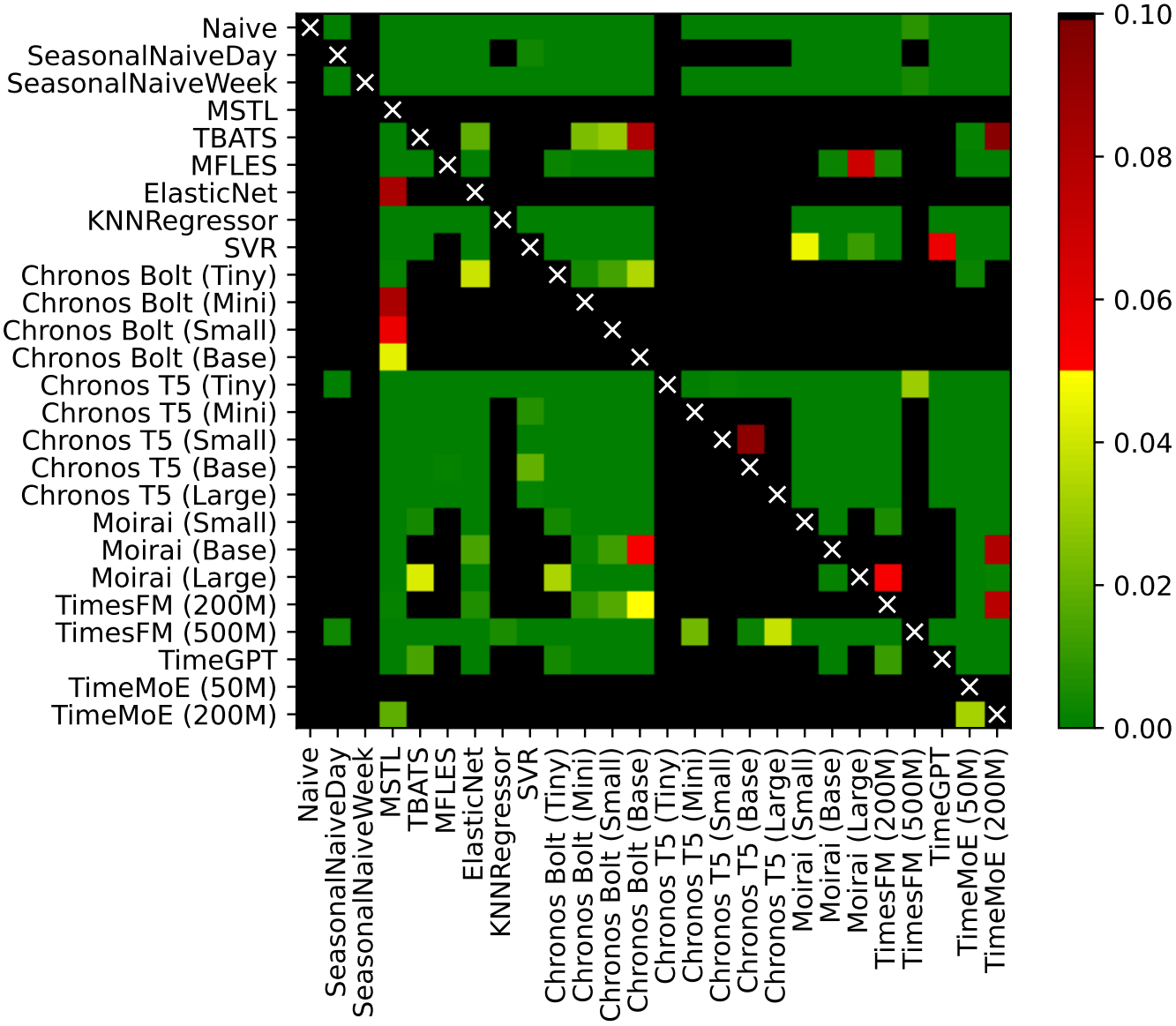

Fig. 3. DM test results for Belgium.

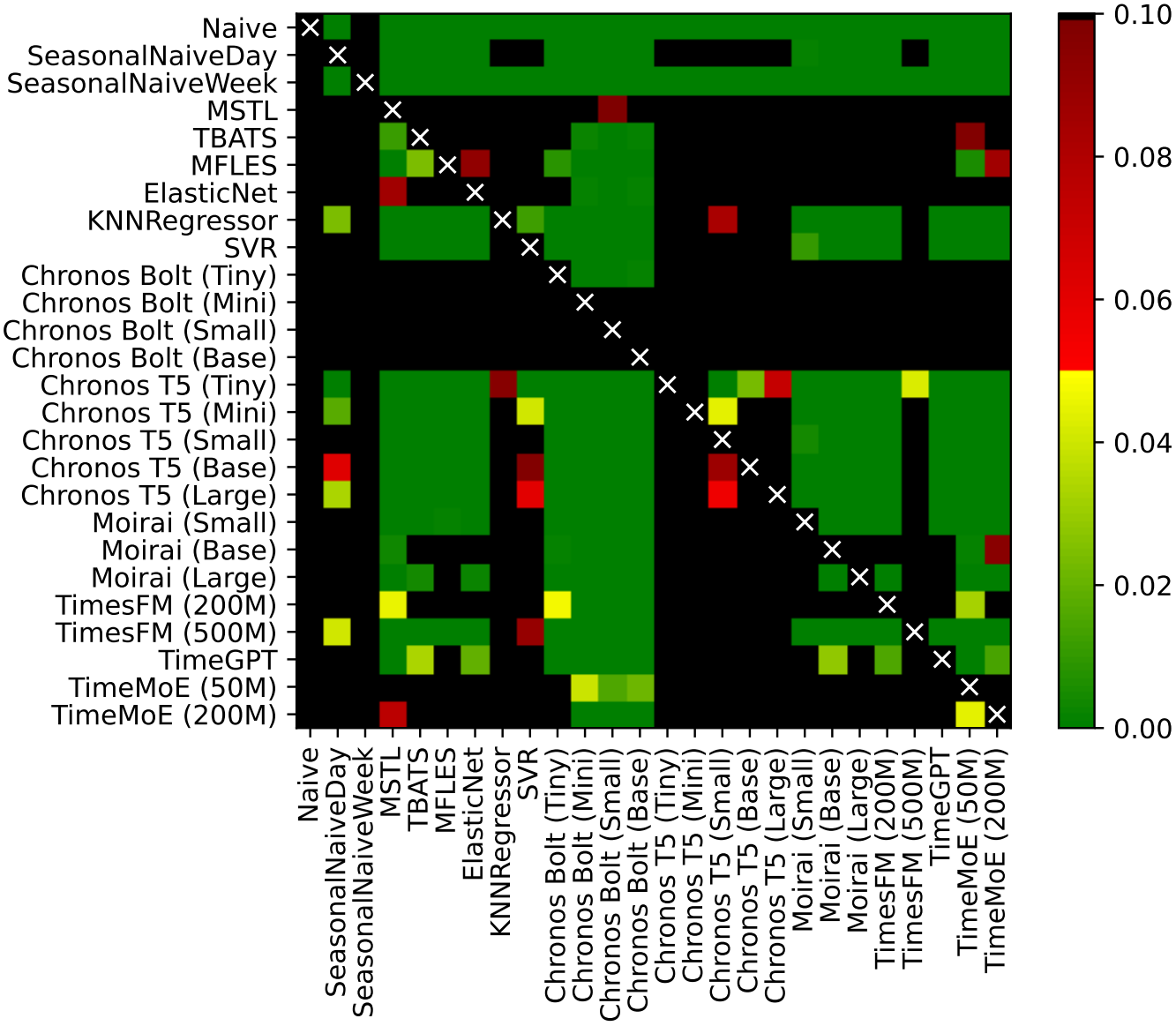

Fig. 4. DM test results for France.

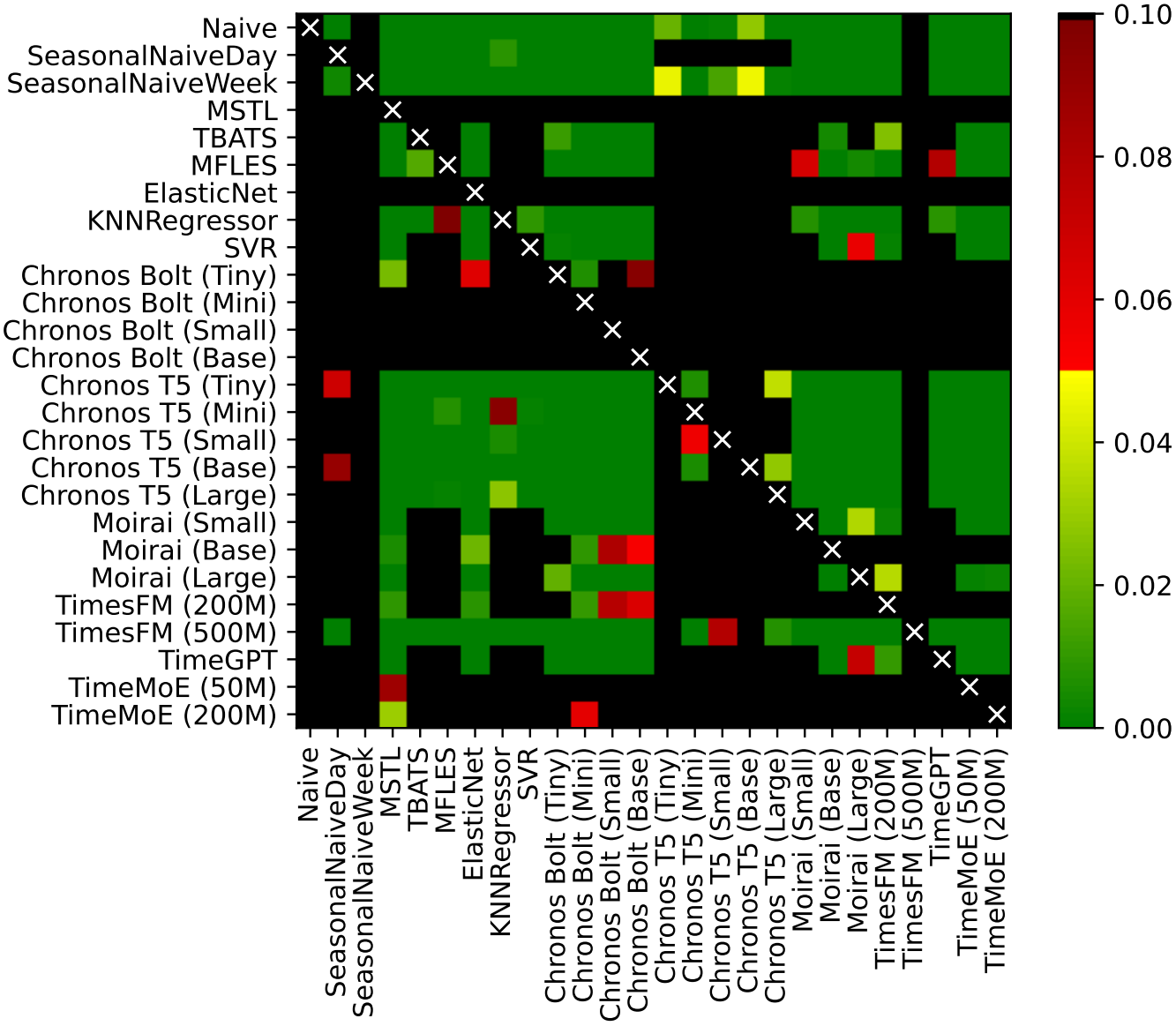

Fig. 5. DM test results for the Netherlands.