Mixed Integer Optimization for SARIMA Specification and Estimation

Luiz Fernando Cunha Duarte
André Ramos
Matheus Alves
Davi Michel Valladão
LAMPS PUC-Rio
Departamento de Engenharia Industrial PUC-Rio

RESUMO

Os modelos SARIMA têm surgido como uma escolha popular para previsão de séries temporais. No entanto, métodos existentes para especificação de seus hiperparâmetros baseiam-se na interpretação subjetiva de testes estatísticos ou abordagens heurísticas devido ao excesso desses. Este artigo propõe uma abordagem de otimização inteira mista para especificação e estimação de um subconjunto específico de modelos SARIMA, denominados modelos autorregressivos integrados (ARI). Nesta abordagem, há a garantia otimalidade global na estimação dos parâmetros e especificação da ordem de integração e da parte autorregressiva. Para validar sua eficácia, foi realizada uma análise comparativa abrangente com o referencial Auto SARIMA utilizando dados de séries mensais da renomada competição M3. A abordagem proposta supera o referencial em um número significativo de cenários.

PALAVRAS CHAVE. Séries temporais, Otimização inteira mista, ARIMA.

EST&MP – Estatística e Modelos Probabilísticos, BDA – Big Data e Analytics, PM – Programação Matemática

ABSTRACT

SARIMA models have emerged as a popular choice for time series forecasting. However, existing methods for specifying their hyperparameters rely on subjective interpretation of statistical tests or heuristic approaches due to their abundance. This article proposes a mixed-integer optimization approach for the specification and estimation of a specific subset of SARIMA models, called autoregressive integrated models (ARI). In this approach, there is a guarantee of global optimality in parameter estimation and specification of the integration order and autoregressive part. To validate its effectiveness, a comprehensive comparative analysis was conducted against the benchmark Auto SARIMA using monthly series data from the renowned M3 competition. The proposed approach outperforms the benchmark in a significant number of scenarios.

KEYWORDS. Time Series, Mixed integer optimization, ARIMA.

EST&MP – Estatística e Modelos Probabilísticos, BDA – Big Data e Analytics, PM – Programação Matemática

DOI: 10.59254/sbpo-2023-174882
https://proceedings.science/p/174882?lang=pt-br

1. Introduction

The rapid and continuous advancement in computer processing capacity has paved the way for significant progress in the field of optimization. Problems that were once deemed intractable can now be efficiently solved using personal computers. This transformative shift has inspired researchers in this domain to explore uncharted territories beyond classical problem domains, seeking to reformulate techniques from diverse disciplines as optimization problems. This synergistic amalgamation combines domain-specific knowledge with the inherent properties, adaptability, and guarantees provided by optimization models [Bertsimas e Dunn, 2019]. One particular topic of interest is time series statistical modeling, in particular ARIMA models.

ARIMA models are widely used for time series prediction. The traditional approach for hyperparameter specification in these models relies on the modeler's utilization of statistical tests and other analytical tools to assess the stationarity of the time series, identify potential seasonal patterns, and determine the orders of the autoregressive and moving-average components. However, due to the possibility of misleading or inconclusive results from certain tests, such as autocorrelation functions, the same approach applied by two different experts may result in different models.

To tackle this challenge, a technique was proposed to automatically fit a SARIMA model, irrespective of seasonality[Hyndman e Khandakar, 2008]. Nevertheless, this method is based on heuristics, thus lacking the assurance of obtaining the optimal model and also excluding a wide range of models. Traditional SARIMA implementations, such as in R [Ripley, 2002], seeks to maximize the log likelihood given six hyperparameters (p, d, q, P, D, Q). The (p, P) hyperparameters are related to the order of the autoregressives components, being the latter related to the seasonal part. While, the (q, Q) are related to the moving-average of the residuals, and the latter is alson related to the seasonal part. The difference orders from both non-seasonal and seasonal, (d, D), are chosen by KPSS [Kwiatkowski et al., 1992] and Canova-Hansen [Canova e Hansen, 1995] tests (as explained in [Hyndman e Khandakar, 2008]), and the seasonal component terms (P, Q) are similar to the non-seasonal components, we present a simplified formulation of the traditional problem

It should be noted that for fixed values of p and q (as well as for P and Q), the model will estimate non-zero coefficients. Thus, this formulation does not produce sparse estimations, which can potentially lead to overfitting. It is important to highlight that this is precisely why it becomes imperative to handle the seasonal and non-seasonal components separately. If only one autoregressive component were employed, the number of lags required to capture seasonality would need to exceed 12, complicating the estimation process.

Recognizing the inherent uncertainty associated with automatic SARIMA fitting, researchers are pursuing alternative approaches that provide stronger guarantees. In response to this imperative, this study introduces a novel methodology that reformulates the specification and estimation process as an optimization problem within the context of autoregressive integrated (ARI) models, a specific subclass of SARIMA models. Shrinking the number of hyperparameters to just one and leveraging the advantageous properties inherent in this framework, this approach enhances both the robustness and reliability of SARIMA fitting. This research represents a significant contribution to the advancement of time series forecasting techniques, offering a promising avenue for addressing uncertainties and improving model selection accuracy.

2. Proposed estimation and specification procedure

The incorporation of sparsity constraints has not been a common practice in the past years, primarily due to the associated complexity of solving an integer optimization problem. However, thanks to advancements in computational power and modern optimization techniques, tackling such problems has become relatively easier and faster. Let us consider the autoregressive formulation of

DOI: 10.59254/sbpo-2023-174882
https://proceedings.science/p/174882?lang=pt-br

type

$$\mathbf{\Delta} y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p-1} \phi_i \mathbf{\Delta} y_{t-i} + \epsilon_t, \tag{1}$$

as in the Augmented Dickey-Fuller (ADF) test [Dickey e Fuller, 1979]. Note that this is a regression model encompassing various components, including an intercept term α, a trend component βt, a differencing component γy_{t-1}, an autoregressive component $\sum_{i=1}^{p-1} \phi_i \mathbf{\Delta} y_{t-i}$, and an error term ϵ_t. Notably, the differencing component plays a crucial role in determining whether the original series or the differenced series is being modeled. This behavior is governed by the coefficient γ. When γ is zero, it indicates that the differenced series is being modelled. If γ is different than zero, it is possible to show that (1) can be reformulated into a traditional autoregressive model

$$y_t = \alpha + \beta t + \sum_{j=1}^{p} \theta_j y_{t-j} + \varepsilon_t.$$

Hence, by utilizing this formulation, it becomes feasible to estimate a regression model that incorporates differencing of the series and incorporates an autoregressive component akin to the SARIMA approach. However, the inclusion of the moving-average component poses a challenge. The residuals can be regarded as variables within the optimization model, representing the discrepancy between the predicted value $\hat{y}_t$ and the observed value y. Additionally, the coefficients that would multiply these residuals in the moving-average component also become variables of the same model. Consequently, the introduction of non-linearity affects certain properties of the model, such as the global optimality of the solution. To retain these desirable properties, the moving-average component will be addressed in future works.

In this research paper, a formulation equivalent to an ARI (Autoregressive Integrated) model is introduced. Notably, the formulation presented below is derived from the Augmented Dickey-Fuller (ADF) test, with a limitation of allowing only one automatic differencing order.

$$\underset{\alpha,\beta,\gamma,\phi}{\text{minimize}} \quad \sum_{t=1}^{T} \epsilon_t^2 \tag{2}$$

$$\text{subject to} \quad \Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \phi_i \cdot \Delta y_{t-i} + \epsilon_t, \quad \forall t \in T \tag{3}$$

$$||\Psi||_0 \leq K, \quad \text{where } \mathbf{\Psi} = \{\alpha, \beta, \gamma, \Phi\} \tag{4}$$

The proposed formulation is designed with the objective of minimizing the sum of squared residuals, while adhering to the dynamic nature of the model described in equation 3. An additional constraint is introduced to promote desirable sparsity in the model, achieved by imposing a limit on the 0-norm of the parameter vector Ψ, using the hyperparameter K.

Given the objective of estimating a fixed number of non-zero coefficients (denoted as K) from a larger set of possibilities, the formulation proposed above can be written using an integer optimization approach. This approach allows the control of the number of non-zero coefficients through the hyperparameter named K. Consequently, the optimization model that focuses on capturing and estimating the K most relevant coefficients can be described as

DOI: 10.59254/sbpo-2023-174882
https://proceedings.science/p/174882?lang=pt-br

$$\underset{\alpha,\beta,\gamma,\phi,I^{\phi},I^{\gamma}}{\text{minimize}} \quad \sum_{t=1}^{T} \epsilon_t^2 \tag{5}$$

$$\text{subject to} \quad \Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \phi_i \cdot \Delta y_{t-i} + \epsilon_t, \quad \forall t \in T \tag{6}$$

$$-MI^{\phi_i} \leq \phi_i \leq MI^{\phi_i}, \quad \forall i \in \{1, \dots, p\} \tag{7}$$

$$-MI^{\gamma} \leq \gamma \leq MI^{\gamma} \tag{8}$$

$$-MI^{\alpha} \leq \alpha \leq MI^{\alpha} \tag{9}$$

$$-MI^{\beta} \leq \beta \leq MI^{\beta} \tag{10}$$

$$I^{\alpha} + I^{\beta} + I^{\gamma} + \sum_{i=1}^{p} I_i^{\phi} \leq K, \tag{11}$$

where $\boldsymbol{I}^{\phi} = \{I^{\phi_t} \mid t = i \in \{1, \dots, p\}\}$ and $\boldsymbol{M}$ is a large number.

The vector I^{ϕ} and the variables $I^{\gamma}, I^{\alpha}, I^{\beta}$ in (7)–(11) are binary variables. I^{ϕ} is a binary variable vector that takes the value 0 when a coefficient is not considered (assumes zero value) and takes the value 1 when it is considered for non-zero estimation. Since this constraint(7) allows the model to disconsider some lags, the interpretation of the parameter p has changed from the number of previous lags used, to the range in which the lags can be selected. Thus, it is easy to notice that the family of autoregressive models considered by this formulation is significantly bigger than the one used by the traditional methods, which is a subset of it.

The values of the binary variables $I^{\gamma}, I^{\alpha}, I^{\beta}$ are also controlled by the model constraint in (11).

It is worth noting that when γ is equal to 0, the model is representing the first-order differenced series (Δy). As such, this methodology does not rely on statistical tests to determine whether differencing is necessary for the time series. Instead, the model estimates it as a parameter.

Given that the model selects only K parameters, an extension of this approach allows for the consideration of seasonality by using a value of p greater than the seasonal period. Let p' be the number of autoregressive lags considered in this approach, and let's assume a yearly seasonality. It can be observed that if $p' \geq 12$, the model could select lags 1 and 12, which is equivalent to choosing $p = 1$ and $P = 1$ in the standard SARIMA approach. However, this extension does not include seasonal differentiation. This flexibility demonstrates that the traditional approach of modeling the autoregressive part is a subcase of the proposed approach.

One another imported aspect of the proposed formulation is the use of the K as the only hyperparameter, which is one of its biggest strengths. Since it reduces the amount of hyperparameters of the traditional ones while enabling a wider range of models to specify.

Thus, a fundamental challenge in the proposed formulation is determining the optimal number of non-zero coefficients. To address this, a straightforward methodology has been developed based on the AICc (Akaike information criterion corrected). In this study, the assumption is made that the residuals of the estimated models are independent and normally distributed. This assumption allows for the approximation of the likelihood function using the estimated variance of the residuals.

To determine the appropriate value of K, the value is incrementally increased and the AICc values of the new models are compared with the previous ones. If the AICc value of the new

DOI: 10.59254/sbpo-2023-174882
https://proceedings.science/p/174882?lang=pt-br

model is smaller, indicating a better fit, the value of K is further increased. However, if the AICc value increases, the process is stopped and the previous model is considered as the final result.

It is worth noting that the presence of outliers in time series data is not uncommon. To ensure robustness against such observations, one may consider reformulating the problem (5)—(11) with an objective function that minimizes the sum of the absolute values of the residuals, rather than the sum of the squared residuals. This formulation offers robustness to outliers. Additionally, this problem can be formulated as a linear integer optimization model, which is computationally more efficient than non-linear integer problems.

One of the great advantages of the proposed approach is its flexibility. By formulating the problem using the optimization framework, it can be easily adapted and customized as needed using any mathematical programming library.

Furthermore, within this framework, it is possible to incorporate additional constraints that enforce specific properties in the estimated model. For instance, constraints can be introduced to ensure stationarity, invertibility, or any other desired characteristic. These constraints provide a means to incorporate domain knowledge and further enhance the quality and interpretability of the estimated model.

3. Results Analysis

Trying to evaluate the accuracy of the proposed model, it was compared against R language Auto ARIMA[Hyndman e Khandakar, 2008]. In the conducted empirical study, 1428 monthly time series from the $M3$ competition were used. This dataset comprises diverse time series that are categorized into six different groups, as summarized in Table 1. The proposed model was implemented using JuMP.jl [Lubin et al., 2023], a mathematical programming framework available in Julia programming language.

For each series in the dataset, the data was divided into a training set and a test set. The last 24 observations were designated as the test set, while the remaining data were utilized to fit three models: SARIMA and two versions of the proposed model, referred to as Optimal SARI. Even though the proposed model does not deal with the seasonal differentiation, it was adopted this name to indicate the extension presented in the section 2

Subsequently, each model was employed to generate forecasts for a 24-step ahead horizon. The accuracy of the forecasts was evaluated using four commonly used metrics: Mean Absolute Percentage Error (MAPE), Mean Absolute Scaled Error (MASE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). These metrics provided comprehensive measures of the forecast error for each combination of model and series.

It is noteworthy that the MASE metric relies on a naive model as a reference. To classify each series based on the presence or absence of the seasonal component, a combination of two seasonal tests was employed. Specifically, the p-value of the Kolmogorov-Smirnov (KS) test [Kruskal e Wallis, 1952] and the p-value of the qs test, a variant of the Ljung-Box test [Ljung e Box, 1978], were utilized. If the p-value of the KS test is below 0.002 or the p-value of the qs test is below 0.01, the series was considered to have a seasonal component. For seasonal series, the seasonal naive model was utilized to compute the MASE metric, whereas for non-seasonal series, the simple naive model was used.

Table 2 presents the percentage of series in which each model outperformed the others. It can be observed that the SARIMA model exhibited superior performance in at least 52.80% and 55.53% of the series, considering the quadratic and absolute error metrics, respectively. Additionally, Table 2 provides a comparison between the two versions of the proposed model. It is evident

DOI: 10.59254/sbpo-2023-174882
https://proceedings.science/p/174882?lang=pt-br

Tabela 1: Number of time series in each category.

Category	Number of series	Percentage
Demographic	111	7.77%
Finance	145	10.15%
Industry	334	23.40%
Macro	312	21.85%
Micro	474	33.19%
Other	52	3.64%
Total	1428	100%

that, across all metrics, the model employing the quadratic error criterion achieved better forecast results in approximately 53% of the series.

Tabela 2: Percentage of times series where a model showed a better forecast result in each metric.

Models	MAPE	MASE	MAE	RMSE
SARIMA	52.80%	58.47%	54.69%	55.11%
Auto ARIMA (quad. error)	47.20%	41.53%	45.31%	44.89%
SARIMA	55.53%	59.17%	55.88%	56.86%
Auto ARIMA (abs. error)	44.47%	40.83%	44.12%	43.14%
Auto ARIMA (quad. error)	53.50%	53.01%	53.01%	53.08%
Auto ARIMA (abs. error)	46.50%	46.99%	46.99%	46.92%

While the analysis presented in Table 2 suggests that the proposed approach did not yield results comparable to SARIMA, it is important to note that the table does not consider the magnitude of the differences between the metrics associated with each model. Therefore, further analysis is necessary to understand the significance of the performance differences observed.

In order to assess the magnitude of these differences, it is crucial to compare the complete distributions of the error metrics. This comparison is illustrated in Figure 1 using box plots. The box plots reveal a similar behavior of the three methods for each of the four metrics, both in terms of median values and variability. However, in terms of the prevalence of outliers, it is evident that SARIMA tended to produce fewer upper outliers. It is worth noting that a logarithmic scale was employed in the figure for the purpose of facilitating the graphical analysis.

In addition to the insights gained from the graphical analysis, it is crucial to employ statistical tests to assess the significance of the observed differences. Due to the nature of the data, particularly the error metrics, the assumption of normality cannot be justified. Consequently, a Wilcoxon test [Wilcoxon, 1945], a non-parametric test that compares the location of two distributions, was conducted. This test is commonly used as a median comparison test and serves as a non-parametric alternative to the traditional t-test.

The bilateral version of the test was employed to compare the distributions of the three model combinations. Table 3 presents the results of each test in terms of p-values. The analysis reveals no compelling evidence to reject the null hypothesis of equality between the location of the distributions when comparing the two versions of the proposed model. Conversely, when comparing the proposed model to the SARIMA model using the MASE metric, there is clear evidence of a significant difference at conventional levels of significance. However, to reach the same conclusion

DOI: 10.59254/sbpo-2023-174882
https://proceedings.science/p/174882?lang=pt-br

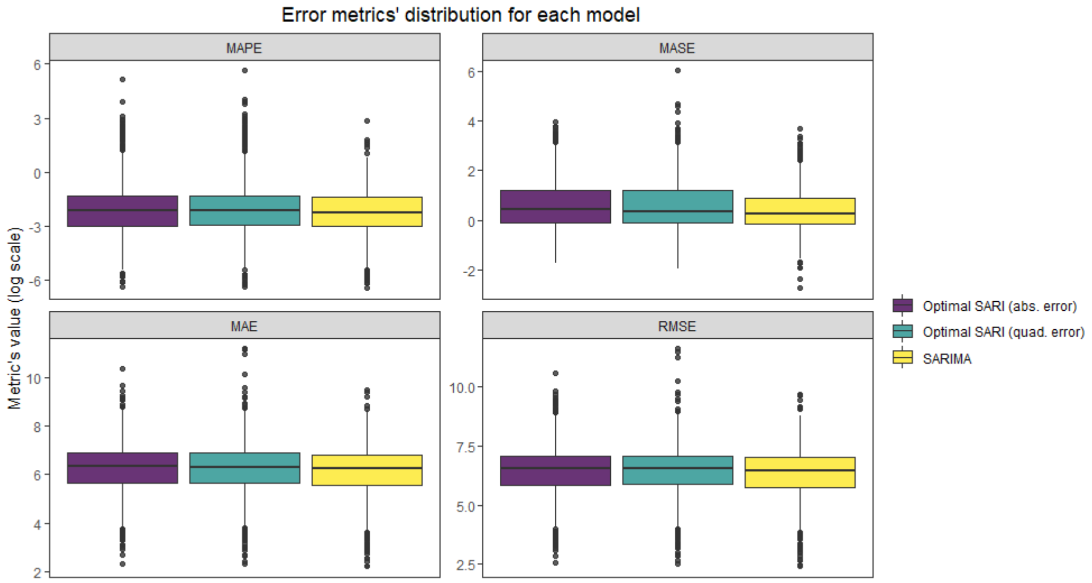

Figura 1: Boxplots comparing the distribution of each error metric (in log scale) for each model.

for the other metrics, a significance level of at least 10% is required for the quadratic error model, and at least 5% for the absolute error model.

Tabela 3: P-value of the bilateral Wilcoxon test, for all three models combinations.

Models Compared	MAPE	MASE	MAE	RMSE
SARIMA Auto ARIMA (quad. error)	0.0585	< 0.01	0.0514	0.0607
SARIMA Auto ARIMA (abs. error)	0.0248	< 0.01	0.0286	0.0258
Auto ARIMA (quad. error) Auto ARIMA (abs. error)	0.7290	0.4195	0.7996	0.7220

Considering the dataset's division into six distinct categories of time series, it is crucial to evaluate the accuracy of the proposed models within each category, as time series within a category may possess unique characteristics. To examine the potential impact of time series category on model accuracy, separate analyses were conducted for each category, following the procedures described earlier.

Table 5 presents the percentage of time series within each category where a specific model yielded superior forecast results, indicated by a lower error metric value. It is worth noting that since these percentages are complementary between the two models considered, the results for the first model in each block are displayed. Upon examining Table 5, it becomes evident that the proposed model with a quadratic error objective function outperformed SARIMA in the Industry and Other categories across all metrics. Similarly, the comparison between SARIMA and the proposed model with an absolute error objective function revealed better accuracy for the latter in these two categories, except for the MASE metric in the Industry category. Notably, the results indicate that the two versions of the proposed model exhibited similar performance, with percentages hovering around 50%, except for the industrial category, where the model with a quadratic error objective function achieved better results in approximately 58% of the time series.

Tabela 4: Percentage of times series, in each category, wherein a model showed a better forecast result in each metric. The percentage showed is referent to the first model in each "Models Compared"block.

Models Compared	Category	MAPE	MASE	MAE	RMSE
SARIMA Auto ARIMA (quad. error)	Demographic	60.36%	72.07%	62.16%	61.26%
	Finance	53.10%	68.97%	53.34%	50.34%
	Industry	46.40%	49.40%	45.81%	44.31%
	Macro	58.33%	60.58%	56.09%	54.49%
	Micro	53.38%	58.65%	59.92%	64.14%
	Other	38.46%	44.23%	44.23%	46.15%
SARIMA Auto ARIMA (abs. error)	Demographic	64.86%	72.07%	65.77%	65.77%
	Finance	54.48%	66.21%	51.03%	52.41%
	Industry	49.40%	52.69%	49.10%	47.90%
	Macro	58.33%	61.22%	56.09%	56.41%
	Micro	57.81%	58.44%	60.5%	63.92%
	Other	40.38%	48.07%	48.07%	46.15%
Auto ARIMA (quad. error) Auto ARIMA (abs. error)	Demographic	51.35%	49.55%	49.55%	51.35%
	Finance	49.66%	48.27%	48.27%	50.34%
	Industry	58.38%	58.68%	58.68%	58.98%
	Macro	53.53%	52.56%	52.56%	50.96%
	Micro	52.11%	51.48%	51.48%	51.69%
	Other	50.00%	53.85%	53.85%	51.92%

Continuing with the previous analysis, Figure 2 presents the distribution of error metrics across different time series categories. Notably, each metric exhibited distinct characteristics across the series categories. For instance, the demographic time series category displayed greater variability in its results compared to the Micro category. This observation highlights the importance of considering the specific characteristics of each time series category when assessing forecast accuracy.

Analyzing the MAPE results, similar behavior was observed, particularly in the Industry and Macro categories. For demographic time series, the SARIMA model exhibited the smallest median MAPE. In contrast, for the Finance and Other categories, the key distinction between the models lay in their variability. SARIMA showed a seemingly smaller variance in the Finance category, while both versions of the proposed model demonstrated lower variability in the MAPE metric than the benchmark methodology in the Other category. In the Micro category, the presence of outliers was more frequent in the proposed models.

Regarding the MASE metric, SARIMA displayed a notably smaller median MASE than both proposed models in the Demographic and Finance categories. For the remaining categories, similar results were observed in terms of median, with SARIMA tending to exhibit less variability. The MAE and RMSE metrics showed comparable patterns, with all three models yielding similar results, except in the Other category where the proposed models showed potentially lower variability than SARIMA.

Furthermore, Table 5 presents the p-values obtained from the bilateral Wilcoxon test used to compare the performance of the three models within each metric and time series category. Speci-

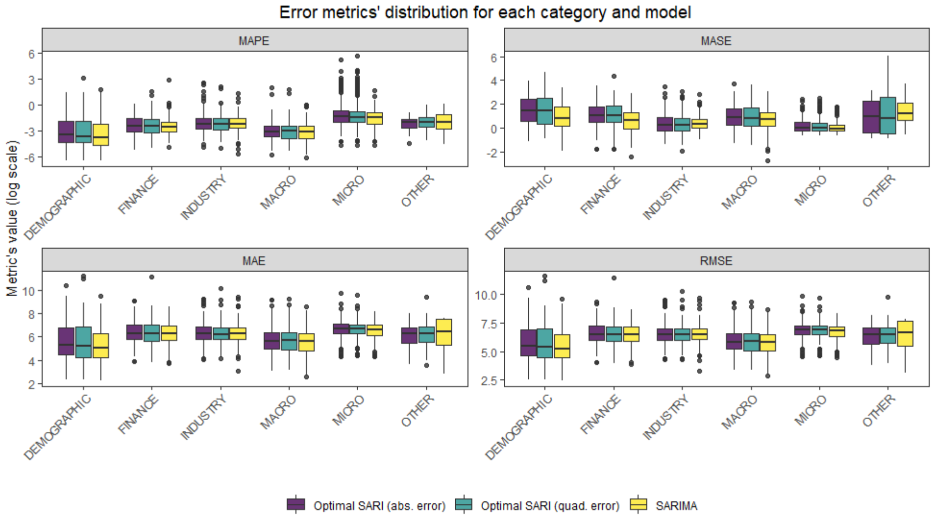

Figura 2: Boxplots comparing the distribution of each error metric (in log scale) for each model.

fically, comparing the proposed model with a quadratic error objective function to SARIMA, the test indicated a significant difference only in the Macro and Micro categories, with significance levels of 10% and at least 2%, respectively. The MASE metric revealed the most pronounced differences between the methodologies, except in the Industry and Other categories where there was insufficient evidence to reject the equality hypothesis. Similarly, the MAE and RMSE metrics exhibited similar patterns, indicating divergent performances between the models primarily in the Micro category.

It is possible to notice that the tests indicate similar conclusions in the comparison of the proposed model with absolute error objective function and SARIMA. Significant differences were detected in time series of the Micro category, in terms of all considered metrics. Using the MAPE metric, the test also indicates significant differences in the Macro category, considering a significance level of at least 10%. Just like before, the MASE metric indicates the major differences, with the industry and other categories being the only ones in which the test did not indicate a significant difference in the forecast results. It is also important to highlight that the test did not find evidence to reject the equality hypothesis for the two versions of the proposed model, considering all metrics and categories.

So, after all this analysis, it is possible to conclude that the proposed methodology was not able to show a better forecast performance than SARIMA. On the other hand, it is necessary to acknowledge that this analysis aimed to identify how distant this first approach is from the SARIMA model. Despite considering only the AR components, this new methodology was able to match the forecast performance of the benchmark in many time series across different categories.

However, in an attempt to understand how the model would perform in a more fair scenario compared to the SARIMA model, the same error metrics were computed, but now only for the series for which SARIMA did not choose any MA component. This enabled the comparison of the forecast performance of the model for series that only have the AR component. Table 6 shows the number of remaining series in each category after applying this filtering.

When considering only the selected set of time series with AR components, Figure 3 demonstrates that the proposed model consistently outperforms SARIMA in terms of each error metric. Moreover, as previously observed, the performance of the two versions of the proposed

Tabela 5: P-value of the bilateral Wilcoxon test, for all three models combinations, categories and metrics

Models Compared	Category	MAPE	MASE	MAE	RMSE
SARIMA Auto ARIMA (quad. error)	Demographic	0.3470	< 0.01	0.3343	0.3250
	Finance	0.5611	< 0.01	0.7516	0.7773
	Industry	0.5145	0.1639	0.4096	0.3728
	Macro	0.0918	0.0155	0.1616	0.1801
	Micro	0.0171	< 0.01	< 0.01	< 0.01
	Other	0.6514	0.3089	0.3343	0.4685
SARIMA Auto ARIMA (abs. error)	Demographic	0.1464	< 0.01	0.1672	0.1441
	Finance	0.5763	< 0.01	0.8886	0.7751
	Industry	0.9685	0.4374	0.8838	0.8730
	Macro	0.0828	< 0.01	0.1453	0.1785
	Micro	< 0.01	< 0.01	< 0.01	< 0.01
	Other	0.2908	0.2084	0.3343	0.3120
Auto ARIMA (quad. error) Auto ARIMA (abs. error)	Demographic	0.6367	0.7571	0.7827	0.7239
	Finance	0.9687	0.8709	0.8599	0.9542
	Industry	0.5208	0.5317	0.4882	0.4427
	Macro	0.8819	0.8248	0.9471	0.9552
	Micro	0.8166	0.6689	0.8609	0.8262
	Other	0.5783	0.8479	0.7923	0.7231

model appears to be highly comparable. The box plots in Figure 4 further support these findings, illustrating consistent superiority across different categories.

In summary, although additional refinements are required for the proposed approach to consistently outperform SARIMA across all cases, recent findings suggest that the model demonstrates superior performance when applied to series with exclusively autoregressive (AR) components, surpassing the benchmark model.

DOI: 10.59254/sbpo-2023-174882
https://proceedings.science/p/174882?lang=pt-br

Tabela 6: Number of time series that SARIMA did not choose any MA component in each category.

Category	Number of series	Percentage
Demographic	21	5.54%
Finance	35	9.23%
Industry	63	16.62%
Macro	78	20.58%
Micro	155	40.90%
Other	27	7.13%
Total	379	100%

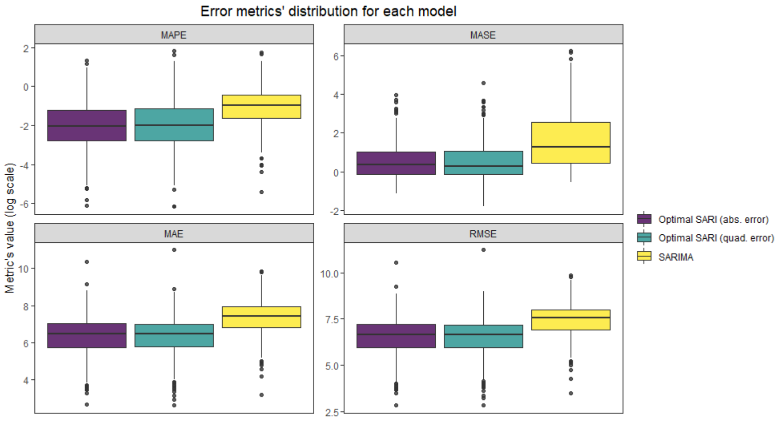

Figura 3: Boxplots comparing the distribution of each error metric (in log scale) for each model, considering only the series without MA component.

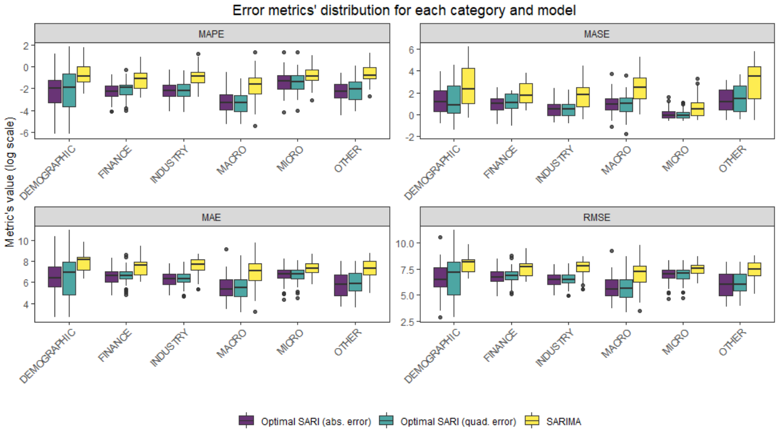

Figura 4: Boxplots comparing the distribution of each error metric (in log scale) for each model and each category, considering only the series without MA component.

DOI: 10.59254/sbpo-2023-174882
https://proceedings.science/p/174882?lang=pt-br

4. Conclusions

In this article, a new methodology was presented for specifying and estimating ARI models by using the mixed-integer optimization framework. As a result, several important properties were obtained, such as global optimality in specification of the order of integration and autoregressive part, as well as model sparsity.

Despite being a subclass of SARIMA, the proposed methodology was able to compete with R's auto SARIMA benchmark and outperform it in more than 40% of the analyzed time series. The performance superiority became evident in cases where Auto Arima estimated models without the MA component. This achievement can be attributed to the intrinsic sparsity achieved by the proposed formulation, which effectively reduces overfitting.

To further improve the work, it is recommended to extend the methodology by incorporating MA terms, automatic seasonal differentiation, stability and invertibility constraints. However, introducing these elements may introduce non-linearity to the problem, potentially leading to local and suboptimal solutions. Additionally, the formulation can be extended to include exogenous variables, resulting in a SARIMAX model.

Referências

Bertsimas, D. e Dunn, J. (2019). *Machine Learning Under a Modern Optimization Lens*. Dynamic Ideas LLC. ISBN 9781733788502. URL https://books.google.com.br/books?id=g3ZWygEACAAJ.

Canova, F. e Hansen, B. (1995). Are seasonal patterns constant over time? a test for seasonal stability. *Journal of Business Economic Statistics*, 13:237–52.

Dickey, D. A. e Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a):427–431. URL https://doi.org/10.1080/01621459.1979.10482531.

Hyndman, R. e Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software*, 26.

Kruskal, W. H. e Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483441.

Kwiatkowski, D., Phillips, P., Schmidt, P., e Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. how sure are we that economic time series have unit root? *Journal of Econometrics*, 54:159–178.

Ljung, G. M. e Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303. ISSN 00063444. URL http://www.jstor.org/stable/2335207.

Lubin, M., Dowson, O., Garcia, J. D., Huchette, J., Legat, B., e Vielma, J. P. (2023). Jump 1.0: Recent improvements to a modeling language for mathematical optimization. *Mathematical Programming Computation*. In press.

Ripley, B. (2002). Time series in r 1.5.0. *R News*, 2:2–7.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

DOI: 10.59254/sbpo-2023-174882
https://proceedings.science/p/174882?lang=pt-br