ICIC Express Letters
Part B: Applications
Volume **15**, Number **2**, February **2024**

ICIC International ©2024 ISSN 2185-2766
pp. **207–216**

THE INTELLIGENT APPROACH OF AUTO-REGRESSIVE INTEGRATED MOVING AVERAGE WITH EXOGENOUS SEMANTIC (ARIMAXS) VARIABLES FOR COVID-19 INCIDENCE PREDICTION

Wanarat Juraphanthong[1] and Kraisak Kesorn[2,*]

[1]Computer Engineering Department
Industrial Technology Faculty
Pibulsongkram Rajabhat University
156 M.5 Plaichumpol, Muang, Phitsanulok 65000, Thailand
wanarat.j@psru.ac.th

[2]Computer Science and Information Technology Department
Science Faculty
Naresuan University
99 Moo 9, Thapo Sub-district, Muang District, Phitsanulok 65000, Thailand
*Corresponding author: kraisakk@nu.ac.th

Received May 2023; accepted July 2023

Abstract. *This paper presents the intelligent approach of the Auto-Regressive Integrated Moving Average with eXogenous Semantic (ARIMAXS) variables model, which represents a well-established extension of the ARIMAX model. The ARIMAX model is known to present challenges in interpreting exogenous covariates. The primary contribution of this study lies in leveraging semantic information encapsulated within an ontology to address this interpretability issue and enhance the predictive accuracy for COVID-19 incidences. By extending the specific variables associated with the underlying factors contributing to the COVID-19 epidemic, the intelligent approach incorporates these factors as semantic variables within the ARIMAX model. A comparative analysis was conducted against conventional methodologies such as ARIMA and ARIMAX, revealing that the intelligent ARIMAXS model demonstrates superior performance by achieving the lowest error rates. Moreover, environmental factors, including the number of tourists and air quality, emerge as significant semantic variables for effectively predicting COVID-19 incidences in this study.*

Keywords: Time series, ARIMA, ARIMAX, ARIMAXS, Semantic processing, Knowledge base, Ontology, Prediction, COVID-19

1. **Introduction.** The global population has encountered a substantial threat posed by the Coronavirus Disease 2019 (COVID-19) epidemic. As documented by the World Health Organization (WHO) [1], the cumulative number of COVID-19 deaths has exceeded 700 million since December 2019. In Thailand, the initial surge of COVID-19 cases exerted a significant impact on the healthcare sector due to the inadequacy of healthcare facilities. While the current situation remains under control, prudent preparedness for the ensuing phase is imperative, considering the ongoing mutations of the disease. The incidence of confirmed COVID-19 cases and associated fatalities is contingent upon a range of factors, encompassing both public health conditions and environmental considerations. Notably, risk factors such as heart or pulmonary disease, weakened immune systems, obesity, and diabetes contribute to the development of severe symptoms and increased mortality rates [2,3]. Furthermore, environmental aspects, including population density, tourist influx, and air quality, wield notable influence on disease transmission dynamics and the subsequent escalation of incidence rates [4].

DOI: 10.24507/icicelb.15.02.207

Time series analysis commonly utilizes the univariate ARIMA model [5], which relies on a single input variable. The ARIMAX model, an extended version of ARIMA, is a multivariate model that enables the inclusion of additional exogenous variables in the observed time series. However, in the ARIMAX model, the exogenous variables are considered covariates, with their coefficients assumed to affect only the exogenous data, rather than the observed time series itself. Consequently, this limitation hinders the interpretability and comprehension of the role and impact of these exogenous factors within the predictive process.

To overcome this limitation, the present study introduces a novel approach called Auto-Regressive Integrated Moving Average with eXogenous Semantic (ARIMAXS) variables. This approach expands the inclusion of exogenous variables in a manner that directly influences the interpretation of the observed data. Unlike traditional ARIMA and ARIMAX models, the ARIMAXS approach proposed in this research incorporates exogenous semantic variables derived from the COVID-19 ontology. The primary advantage of this approach is its enrichment of the model with contextual knowledge, enabling a more meaningful and interpretable analysis of the relationship between the exogenous factors and the observed time series. As a result, this enhancement holds the potential to improve the accuracy of COVID-19 incidence prediction. The experimental results demonstrated that the proposed ARIMAXS model is superior to the traditional approaches measured by Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

This paper is organized as follows. Section 2 provides a literature review that identifies existing studies in the field. Section 3 describes the proposed methods, encompassing the dataset and ontology, the framework and algorithm, as well as the evaluation metrics utilized. Section 4 presents and discusses the results of this study. Section 5 offers an in-depth discussion of the implications of the findings. Lastly, Section 6 concludes the study by summarizing the results and presenting suggestions for future research.

2. **Literature Review.**

2.1. **COVID-19 prediction with ARIMAX.**

The ARIMAX model is extensively employed across various disciplines, including education [6], engineering, and the healthcare field [7], with a specific emphasis on disease analysis and prediction. Notably, the ARIMAX model has been utilized in COVID-19 analyses through both traditional methods and hybrid models. Li et al. [8] introduced the concept of selecting relevant features for predicting COVID-19 cases by combining stepwise regression with the ARIMAX model to forecast short-term trends in the United States COVID-19 epidemic. Their research showcased accurate predictions of COVID-19 case numbers within a specific study area, accompanied by a 95% confidence interval, achieved through the implementation of stepwise regression feature selection. Building upon this notion of feature selection, Somyanonthanakul et al. [9] proposed an approach that merges ARIMAX with Association Rule Mining (ARM) to identify prognostic factors associated with short-term trends in the COVID-19 epidemic in the United States. ARIMAX effectively employed these factors for modeling and forecasting COVID-19 cases. The collaboration between ARM and ARIMAX resulted in a prediction model exhibiting lower error rates than alternative approaches. Nonetheless, it is crucial to acknowledge that this collaborative approach has limitations attributed to the usage of the a limited number of COVID-19 cases and clinical variables, consequently affecting the model's reliability.

The ARIMAX model was used to predict the number of COVID-19 cases in Jakarta, Indonesia [10]. Those authors suggested incorporating Google Trends data, specifically targeted searches, in addition to a daily dataset obtained from Jakarta's official COVID-19 website as external variables. The experimental results showed that ARIMAX achieved

a slightly lower error rate of 0.08% compared to the conventional ARIMA model. However, one notable limitation of this research was the absence of the proposed modifications to the ARIMAX model.

In a different study, Rahman and Chowdhury [11] hypothesized the significant role of meteorological factors in COVID-19 transmission across SAARC countries. The authors collected a daily dataset consisting of the number of confirmed COVID-19 cases, as well as various meteorological attributes such as minimum and maximum temperatures, relative humidity, surface pressure, daily precipitation, and maximum wind speed. To forecast the confirmed COVID-19 cases, all significant attributes were included as covariates in both the ARIMAX and XGBoost models. The findings shed light on the influence of diverse meteorological factors on different nations within the South Asian Association for Regional Cooperation (SAARC) region.

2.2. **Machine learning and a knowledge base model.** Sirichanya and Kraisak [12] conducted a comprehensive review of prior research that employed a knowledge base to enhance the performance of machine learning algorithms. Prominent examples include the semantic decision tree [13] and the semantic ARIMA [14], both of which proposed integrating a knowledge base into a conventional ARIMA model. Elsewhere, researchers have utilized knowledge bases for various tasks, such as data warehouse design [15,16].

To the best of our knowledge, no previous study has proposed combining a knowledge base with ARIMAX for data analysis and prediction, thus extending upon the groundwork established in our previous work [15]. This represents the primary innovation of our current research study, and further elucidation will be provided in subsequent sections.

3. Methodology.

3.1. **Dataset and COVID-19 ontology.** The dataset includes COVID-19 confirmed cases and environmental factors such as population, tourist movements, and air quality, which have an impact on the incidence of cases. We collected daily records from 77 provinces in Thailand, covering the period from April to December 2021, resulting in a total of 21,450 instances. The population, number of tourists, and air quality data were integrated with the daily COVID-19 case data. These datasets were obtained from the Open Government Data of Thailand [17], which were published by the Ministry of Public Health, Ministry of Interior, Ministry of Tourism and Sports, and Ministry of Natural Resources and Environment. The study region consisted of three provinces that had the highest number of reported cases in Thailand: Bangkok, Chiang Mai, and Chon Buri. To split the data, we allocated 70% for the training set and reserved the remaining 30% for the testing set. The COVID-19 ontology was adapted from the existing ontology developed by Sargsyan et al. [18] to suit the requirements of our specific task. We expanded certain entities and their successors within the original ontology, with a specific focus on the risk factors associated with COVID-19 transmission. Ultimately, the modified COVID-19 ontology consisted of 2,291 classes and 39,051 axioms, facilitating the semantic approach employed in our model.

3.2. **The proposed approach: ARIMAXS.** The proposed approach incorporates intelligent semantic reinforcement and prediction into the traditional ARIMAX model. The semantic information in the COVID-19 ontology was used as the pertinent knowledge base for data analysis and for determining the relevant lags of prediction. The framework of ARIMAXS is depicted in Figure 1.

3.2.1. *Semantic reinforcement.* The process of semantic reinforcement involves three main steps to improve the initial ARIMAX model using relevant data from a dataset, aided by an ontology structure. The first step, known as the semantic dataset generation stage, involves appending metadata from the COVID-19 ontology $\mathcal{O}$, which encapsulates semantic

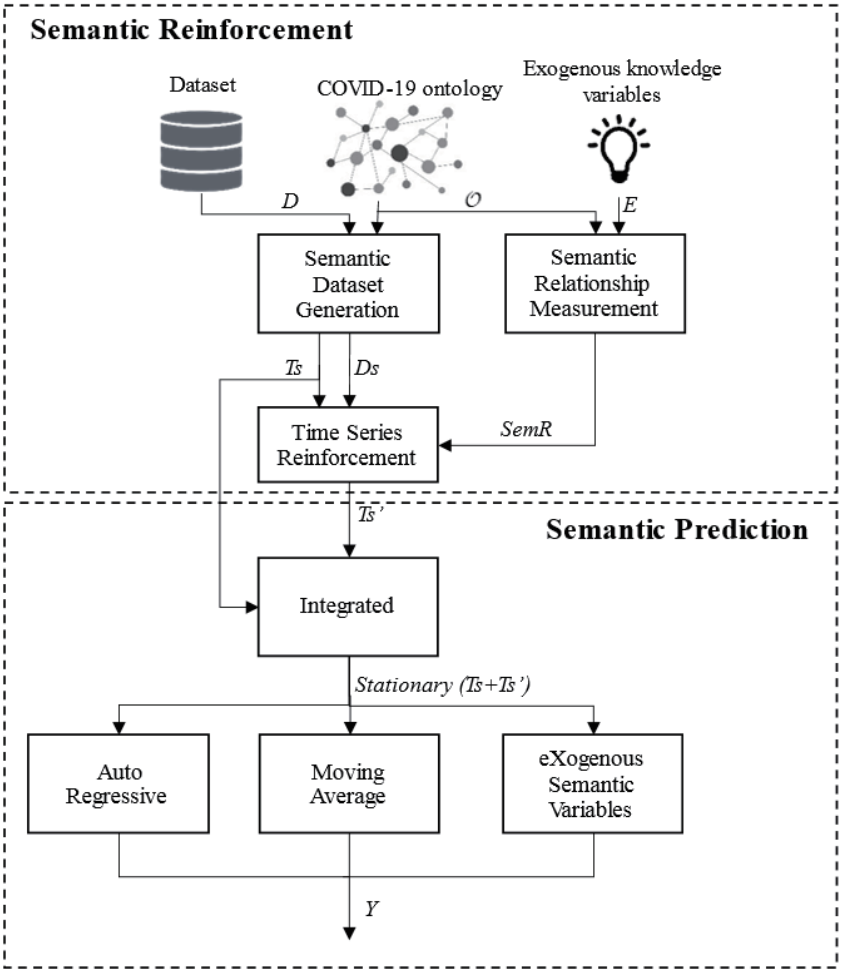

FIGURE 1. The framework of ARIMAXS

information, to dataset D. This dataset includes the original time series T. The subsequent phase is where we identify and use relevant data to enhance the initial ARIMAX model for semantic prediction. The relevance of the data is gauged by their semantic similarity to the original time series – if the semantic correlation is high, the data is considered relevant. The second main step, the semantic relationship measurement stage, involves calculating the Semantic Relationship ($SemR$) between two sets: the exogenous knowledge variable set E, and the COVID-19 entity set A. We determine this relationship by evaluating the distance of their Least Common Subsumer (LCS) and the maximum depth of the ontology. The final step, known as the time series reinforcement stage, involves computing the Semantic Distance ($SemD$). This is derived from the combination of the semantic relationship and Euclidean distance. The calculated semantic distance is then translated into a Semantic Similarity score ($SemS$) on a scale of 0 to 1. If the semantic similarity score is close to 1, the data is considered relevant to the original time series and is then used as semantic exogenous variables in the enhanced time series Ts'. The process of semantic reinforcement is outlined in Algorithm 1.

3.2.2. *Semantic prediction.* In our proposed methodology, our purpose was to refine the ARIMAX model, a well-established strategy for forecasting time series data, through the incorporation of a semantic (S) characteristic. The conventional ARIMAX model comprises the Auto-Regressive (AR), Integrated (I), Moving Average (MA), and eXogenous (X) variables. The infusion of a semantic element in our model facilitates a more sophisticated interpretation of the data, thereby expanding its potential analytical reach.

Algorithm 1: Semantic Reinforcement Algorithm

Input: Original time series T; Dataset D;
COVID-19 ontology $\mathcal{O}$; Threshold value η;
COVID-19 entity set A; Exogenous knowledge variables set E;

Output: Enhanced time series

Semantic Dataset Generation:

for each item **in** D **do**
  Add semantic metadata from $\mathcal{O}$ to D
end for
return Semantic dataset Ds **and** Semantic time series Ts

Semantic Relationship Measurement:

for each i **in** A **do**
  if all elements in E are the entity of $\mathcal{O}$ **then**
    Define the set $H = \{i\} \cup E$
    Compute $SemR_i = \frac{1}{|H|} \sum_{a \in H} \frac{dis(LCS(H), a)}{maxdepth(\mathcal{O})}$
  end if
end for
return Semantic relationship $SemR$

Time Series Reinforcement:

for each t **in** Ts **do**
  for each d **in** Ds **do**
    Compute $SemD(t, d) = \sum_{i \in A} SemR_i \cdot (t_i - d_i)^2$
    Compute $SemS$ from $SemD(t, d)$
    if $SemS < \eta$ **then**
      enhance d to Ts'
    end if
  end for
end for
return Ts'

During the Integration (I) phase, we employed a d-order difference approach to stabilize the time series. This crucial step is designed to ensure the stationary nature of the data in both the original and semantically enhanced time series, a prerequisite condition for yielding valid predictions. The AR facet of the model integrates p lags, representative of historical data points from the original time series (Ts), alongside r semantic lags, which denote points in time with semantic relevance from the enhanced time series (Ts'). The MA component of the model relies on q error lags derived from the predictive output (ε'), encompassing a combination of both original and semantic lags. In the final eXogenous Semantic (XS) variables section, we infuse the stationary eXogenous (X) variables into the model. These variables consider both the original p lags and the semantically enhanced r lags.

The culmination of this comprehensive data analysis is the forecasting of the time series Y. This prediction is calculated leveraging the semantic variables extracted from all the model's components. The semantic prediction process is detailed in Algorithm 2.

Algorithm 1 outlines the process of semantic reinforcement, which introduces the semantic (S) characteristics into our proposed model. By incorporating these semantic variables, the model expands its analytical capabilities. It goes beyond merely evaluating numerical or quantitative variables and brings in contextual knowledge, resulting in a deeper

Algorithm 2: Semantic Prediction Algorithm

Input: Original time series Ts; Enhanced time serie Ts';
Exogenous variables X;
p; d; q;

Output: Predicted time series Y

Integrated (I):

for each t **in** Ts **do**
    d-order difference of Ts_t and Ts'_r
end for
return d-order differenced Ts and Ts'

Auto-Regressive (AR) and Moving Average (MA) with eXogenous Semantic (XS) Variable:

for each t **in** Ts **do**
    Compute $Y_t = c + \sum_{i=1}^{p} \varphi_i \left(Ts_{t-i} + Ts'_r \right) + \sum_{i=1}^{q} \theta_i \varepsilon'_{t-i} + \sum_{i=1}^{p} \gamma_i (X_{t-i} + X_r)$;
    where c is a constant, φ is auto-regressive coefficients,
    θ is moving average coefficients, and γ is exogenous coefficients
end for
return Y

understanding of the time series dynamics. This incorporation is particularly useful in understanding scenarios such as the COVID-19 ontology.

Furthermore, the semantic prediction, as detailed in Algorithm 2, merges the semantic variables from all model components. This comprehensive integration bolsters the model's predictive performance which is achieved by considering not just the historical information from the original time series, but also the contextual knowledge extracted from the enhanced data. As a result of this all-encompassing analysis, the model's predictions are not only more accurate but also highly informative.

3.3. **Evaluation metrics.** We evaluated the forecasting performance of the ARIMAXS using standard metrics, including the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). These metrics are represented in Equations (1), (2), and (3), respectively.

$$MSE = \frac{1}{N} \times \sum_{i=1}^{N} (y - \hat{y})^2 \tag{1}$$

$$RMSE = \sqrt{\frac{1}{N} \times \sum_{i=1}^{N} (y - \hat{y})^2} \tag{2}$$

$$MAE = \frac{1}{N} \times \sum_{i=1}^{N} |y - \hat{y}| \tag{3}$$

where y is the actual value, $\hat{y}$ is the predicted value, and N is the number of data points.

4. **Experimental Results.** We conducted a comparison of our proposed ARIMAXS model with traditional ARIMA and ARIMAX models. For COVID-19 incidence forecasting, we integrated exogenous semantic variables into the models, which included factors such as population density, number of tourists, and air quality.

4.1. **Parameter estimation.** Every model we employed included specific parameters: p for the Auto-Regressive (AR) process, d for the Integrated (I) process, and q for the

Moving Average (MA) process. For the I process, we leveraged the Augmented Dickey-Fuller (ADF) unit root test [19] to confirm the stationary nature of the time series data. We utilized the d parameter in the d-order difference method to convert non-stationary series to stationary series. For the AR and MA processes, we identified the p and q parameters through the correlogram patterns of the Auto Correlation Function (ACF) and the Partial Auto Correlation Function (PACF) [20]. A comprehensive overview of the estimated parameters for all models can be found in Table 1.

Table 1. The parameter estimation of ARIMA, ARIMAX and ARIMAXS

Provinces	Parameter	ARIMA	ARIMAX	ARIMAXS_1	ARIMAXS_2	ARIMAXS_3	ARIMAXS_a
Bangkok	p	$[0,3]$	$[0,3]$	$[0,4]$	$[0,3]$	$[0,5]$	$[0,6]$
	d	1	1	1	1	1	1
	q	$[0,8]$	$[0,8]$	$[0,5]$	$[0,5]$	$[0,6]$	$[0,6]$
Chiang Mai	p	$[0,2]$	$[0,2]$	$[0,3]$	$[0,2]$	$[0,2]$	$[0,2]$
	d	0	0	0	0	0	0
	q	$[0,7]$	$[0,7]$	$[0,7]$	$[0,7]$	$[0,7]$	$[0,7]$
Chon Buri	p	$[0,9]$	$[0,9]$	$[0,10]$	$[0,11]$	$[0,10]$	$[0,10]$
	d	1	1	1	1	1	1
	q	$[0,6]$	$[0,6]$	$[0,11]$	$[0,11]$	$[0,11]$	$[0,11]$

In the case of the three provinces, we observed that both ARIMA and ARIMAX models – incorporating all exogenous variables (X_1, X_2, X_3 and X_a) – employ identical parameters as these models are based on the original time series. However, the ARIMAXS model, with its distinctive incorporation of each exogenous semantic variable, employs a range of parameters due to the differentiated enhanced time series that the model produces.

4.2. **Prediction performance.** The parameters specified in the previous section were used to determine the most appropriate parameters for the best-fitted model. Both the ARIMAXS and the original ARIMAX were constructed using individual exogenous variables such as population (X_1), tourists (X_2), and air quality (X_3), as well as multiple exogenous variables (X_a) that encompass all variables (X_1, X_2, X_3). We utilized ARIMA as the benchmark model for comparison purposes. The prediction performances of ARIMA, ARIMAX, and ARIMAXS are documented in Table 2.

Table 2. The prediction performances of ARIMA, ARIMAX and ARIMAXS

Provinces	ARIMA	X_1 = population		X_2 = tourist		X_3 = air quality		$X_a = X_{1,2,3}$	
		ARIMAX_1	ARIMAXS_1	ARIMAX_2	ARIMAXS_2	ARIMAX_3	ARIMAXS_3	ARIMAX_a	ARIMAXS_a
Bangkok	$(1,1,8)$	$(1,1,3)$	$(4,1,5)$	$(1,1,3)$	$(2,1,3)$	$(1,1,8)$	$(5,1,2)$	$(1,1,3)$	$(1,1,3)$
MSE	175,524.79	175,436.87	***155,386.74***	178,731.92	***161,753.93***	175,682.47	***150,131.11***	173,902.51	***148,577.71***
RMSE	418.95	418.85	***394.19***	422.77	***402.19***	419.15	***387.47***	417.02	***385.46***
MAE	241.63	241.22	***224.16***	247.61	***232.99***	241.91	***220.03***	240.81	***218.26***
Chiang Mai	$(1,0,1)$	$(1,0,1)$	$(1,0,1)$	$(1,0,1)$	$(1,0,1)$	$(1,0,1)$	$(1,0,1)$	$(1,0,1)$	$(1,0,1)$
MSE	62,669.31	62,114.55	***60,695.37***	61,981.98	***60,305.01***	62,264.35	***60,572.02***	63,178.448	***60,470.88***
RMSE	250.33	249.23	***246.36***	248.96	***245.57***	249.53	***246.11***	251.3532	***245.91***
MAE	80.59	78.3	***75.99***	74.44	***73.53***	76.21	***74.64***	78.1649	***73.55***
Chon Buri	$(7,1,3)$	$(7,1,3)$	$(4,1,4)$	$(7,1,3)$	$(4,1,8)$	$(7,1,3)$	$(4,1,8)$	$(7,1,3)$	$(4,1,8)$
MSE	9,633.32	10,579.69	***8,881.49***	9,485.76	***8,821.69***	9,654.49	***7,834.41***	10,453.07	***7,880.19***
RMSE	98.15	102.86	***94.24***	97.39	***93.92***	98.26	***88.51***	102.24	***88.77***
MAE	59.40	60.33	***56.47***	58.84	***56.72***	59.52	***51.64***	61.53	***52.11***

Our findings showed that the ARIMAXS model, with both individual and multiple exogenous variables, produced fewer errors compared to ARIMA and ARIMAX for the corresponding variables. For the Bangkok dataset, all exogenous variables used in ARIMAX only slightly reduced the prediction errors compared to the baseline models. Interestingly, the Mean Squared Error (MSE) of the ARIMA model was 175,524.79, whereas

the ARIMAX model with X_1, X_2, X_3, and X_a produced MSEs of 175,436.87, 178,731.92, 175,682.47, and 173,902.51, respectively.

In contrast, when employing all exogenous semantic variables in the ARIMAXS model, we observed a significant reduction in prediction errors. The ARIMAXS model reduced the MSE with X_1 to 155,386.74, X_2 to 161,753.93, X_3 to 150,131.11, and X_a to 148,577.71. In the Chiang Mai and Chon Buri datasets, the tourist numbers (an exogenous variable) led to a notable reduction in errors. The MSE of the ARIMA models for Chiang Mai was 62,669.31 and 9,633.32 for Chon Buri. The ARIMAX model with the X_2 variable could decrease the MSE to 61,981.98 and 9,485.76, whereas other variables resulted in a higher MSE.

However, the ARIMAXS model with all exogenous semantic variables had a significant impact in reducing prediction errors in both Chiang Mai and Chon Buri. In Chiang Mai, the MSE of the ARIMAXS model with X_1 decreased to 60,695.37, X_2 decreased to 60,305.01, and X_3 decreased to 60,572.02, while X_a decreased to 60,470.88. In Chon Buri, the corresponding MSE values decreased to 8,881.49, 8,821.69, 7,834.41, and 7,880.19, respectively. This successful reduction is attributed to the ability of our proposed model to semantically extend these variables, identify relevant data within the dataset, and use this information to enhance the prediction of incidences in the time series.

4.3. **Computational assessment.** To compare the computational performance of our proposed model with that of the baseline models, we estimated and compared the computational complexity costs of the ARIMA, ARIMAX, and ARIMAXS models. The results of this comparison are detailed in Table 3.

Table 3. The computational complexity cost of ARIMA, ARIMAX and ARIMAXS

	ARIMA	ARIMAX	ARIMAXS
Semantic reinforcement	–	–	$O(k) + O(a) + O(nk)$
Prediction/ Semantic prediction	$O(n)$	$O(n)$	$O(n+r)$

The complexity of the prediction element in both the ARIMA and ARIMAX models is denoted as $O(n)$, where n stands for the number of instances in the original time series. This is because these models predict precisely n points in time. In contrast, the complexity of the semantic prediction element in the ARIMAXS model, as calculated via Algorithm 2, extends to include r related points of time from the enhanced time series, making its complexity $O(n+r)$. Here, r signifies the number of relevant instances extended in the enhanced time series. The complexity of the semantic reinforcement element, as per Algorithm 1, is given by $O(k) + O(a) + O(nk)$. This represents the complexities of semantic dataset generation, semantic relationship measurement, and time series reinforcement processes, where k is the number of instances in the dataset and a is the number of entities in the COVID-19 ontology.

The findings suggest that the time series reinforcement process exhibits the highest complexity cost in this implementation version. Therefore, our future work will focus on conducting algorithmic optimization to achieve a more acceptable computational complexity cost for the overall process.

5. **Discussion.** This section further explores the results obtained through the use of exogenous variables in each model. Table 2 indicates that the original ARIMAX model, when applied with only the 'tourist' variable, succeeded in significantly reducing prediction error for Chiang Mai and Chon Buri. In contrast, the ARIMAXS model, employing all variables, managed to diminish prediction error across all three provinces. Our findings also showed that, in both Bangkok and Chon Buri, the 'air quality' attribute held

more significance than the 'tourist' attribute in terms of COVID-19 incidence. This is likely due to government-imposed travel restrictions, which resulted in consistently low tourist numbers in Bangkok and a constant number in Chon Buri. The 'tourist' attribute's impact varies between environments. In Chiang Mai, known for its natural beauty and northern culture, the 'tourist' attribute has a greater influence than the 'air quality' attribute, as compared to the built-up urban environment of Bangkok. Additionally, the 'population' attribute shows a marginal effect in provinces with high population density, such as Bangkok. Given these exogenous variables' significant impact, we can confidently state that our proposed method not only enhances prediction accuracy but also aids in determining the relative influence of these factors on COVID-19 incidence.

6. **Conclusions.** In this study, we compared three models (ARIMAXS, ARIMA, and ARIMAX) to forecast COVID-19 incidence. The ARIMAXS model, with reinforcement of exogenous semantic variables, outperformed the ARIMAX and ARIMA models. This improvement is due to the ontology-based knowledge that interprets exogenous covariates based on their semantic relationship. Utilizing exogenous semantic variables in the ARIMAXS model proved effective for predicting COVID-19 incidence. However, it is important to acknowledge a limitation of our approach. The reinforcement process introduces a high complexity cost due to the additional time required for evaluating data related to exogenous semantic factors. To address this limitation, future research endeavors may explore the implementation of adaptive techniques and optimization strategies for the reinforcement process. The objective of these efforts will be to enhance prediction performance while reducing computational complexity, thereby further improving the practical applicability of the ARIMAXS model.

Acknowledgement. This research was supported by Science Faculty, Naresuan University (NU), Grant No. R2565E060, and Thailand National Science, Research and Innovation (Fundamental Fund-NU: Grant No. R2566B035). The funder had no role in the study design, data collection, analysis, publication decision, or manuscript preparation. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation. Also, thanks to Mr Roy I. Morien of the Naresuan University Graduate School for his editing of the grammar, syntax and general English expression in this manuscript.

REFERENCES

[1] World Health Organization, *WHO Coronavirus (COVID-19) Dashboard*, https://covid19.who.int, Accessed on Mar. 04, 2023.

[2] Z. G. Dessie and T. Zewotir, Mortality-related risk factors of COVID-19: A systematic review and meta-analysis of 42 studies and 423,117 patients, *BMC Infectious Diseases*, vol.21, no.1, 855, DOI: 10.1186/s12879-021-06536-3, 2021.

[3] C. Kaeuffer et al., Clinical characteristics and risk factors associated with severe COVID-19: Prospective analysis of 1,045 hospitalised cases in North-Eastern France, March 2020, *Euro Surveill*, vol.25, no.48, 2000895, DOI: 10.2807/1560-7917.ES.2020.25.48.2000895, 2020.

[4] J. D. Kong, E. W. Tekwa and S. A. Gignoux-Wolfsohn, Social, economic, and environmental factors influencing the basic reproduction number of COVID-19 across countries, *PLoS ONE*, vol.16, no.6, e0252373, DOI: 10.1371/journal.pone.0252373, 2021.

[5] H. Alabdulrazzaq, M. N. Alenezi, Y. Rawajfih, B. A. Alghannam, A. A. Al-Hassan and F. S. Al-Anzi, On the accuracy of ARIMA based prediction of COVID-19 spread, *Results in Physics*, vol.27, 104509, DOI: 10.1016/j.rinp.2021.104509, 2021.

[6] W.-C. Chou, K. L. Lai and D.-F. Chang, Detecting multivariate series data with transfer function ARIMAX for teacher demand, *ICIC Express Letters, Part B: Applications*, vol.11, no.2, pp.129-136, DOI: 10.24507/icicelb.11.02.129, 2020.

[7] D.-F. Chang, C.-C. Chen and A. Chang, Forecasting with ARIMAX models for participating STEM programs, *ICIC Express Letters, Part B: Applications*, vol.11, no.2, pp.121-128, DOI: 10.24507/icicelb.11.02.121, 2020.

[8] Y. Li et al., Rapid prediction and evaluation of COVID-19 epidemic in the United States based on feature selection and improved ARIMAX model, *2021 2nd International Conference on Artificial Intelligence and Information Systems*, NY, USA, pp.1-8, DOI: 10.1145/3469213.3471327, 2021.

[9] R. Somyanonthanakul et al., Forecasting COVID-19 cases using time series modeling and association rule mining, *BMC Medical Research Methodology*, vol.22, no.1, 281, DOI: 10.1186/s12874-022-01755-x, 2022.

[10] B. S. Aji, Indwiarti and A. A. Rohmawati, Forecasting number of COVID-19 cases in Indonesia with ARIMA and ARIMAX models, *The 9th International Conference on Information and Communication Technology*, Yogyakarta, Indonesia, pp.71-75, DOI: 10.1109/ICoICT52021.2021.9527453, 2021.

[11] M. S. Rahman and A. H. Chowdhury, A data-driven extreme gradient boosting machine learning model to predict COVID-19 transmission with meteorological drivers, *PLoS ONE*, vol.17, no.9, e0273319, DOI: 10.1371/journal.pone.0273319, 2022.

[12] C. Sirichanya and K. Kraisak, Semantic data mining in the information age: A systematic review, *International Journal of Intelligent Systems*, vol.36, no.8, pp.3880-3916, DOI: 10.1002/int.22443, 2021.

[13] S. Chanmee and K. Kesorn, Exploiting a knowledge base for intelligent decision tree construction to enhance classification power, *Engineering and Applied Science Research*, vol.49, no.4, 2022.

[14] W. Juraphanthong and K. Kesorn, Time series data enrichment using semantic information for dengue incidence forecasting, *Science, Engineering and Health Studies*, 21050013, DOI: 10.14456/sehs.2021.50, 2021.

[15] N. Sanprasit, T. Titijaroonroj and K. Kesorn, A semantic approach to automated design and construction of star schemas, *Engineering and Applied Science Research*, vol.48, no.5, 2021.

[16] N. Sanprasit, K. Jampachaisri, T. Titijaroonroj and K. Kesorn, Intelligent approach to automated star-schema construction using a knowledge base, *Expert Systems with Applications*, vol.182, 115226, DOI: 10.1016/j.eswa.2021.115226, 2021.

[17] Digital Government Development Agency, *Open Government Data of Thailand*, https://data.go.th/, Accessed on Mar. 27, 2022.

[18] A. Sargsyan et al., The COVID-19 ontology, *Bioinformatics*, vol.36, no.24, pp.5703-5705, DOI: 10.1093/bioinformatics/btaa1057, 2021.

[19] D. A. Dickey and W. A. Fuller, Distribution of the estimators for autoregressive time series with a unit root, *Journal of the American Statistical Association*, vol.74, no.366a, pp.427-431, DOI: 10.1080/01621459.1979.10482531, 1979.

[20] G. E. P. Box, G. M. Jenkins, G. C. Reinsel and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th Edition, Wiley, Hoboken, New Jersey, 2015.