Iraqi Journal for Computer Science and Mathematics

Journal Homepage: http://journal.esj.edu.iq/index.php/IJCM

e-ISSN: 2788-7421 p-ISSN: 2958-0544

Soft Computing-Based Generalized Least Deviation Method Algorithm for Modeling and Forecasting COVID-19 using Quasilinear Recurrence Equations

Mostafa Abotaleb[1,*]*

[1]Department of System Programming, South Ural State University, 454080 Chelyabinsk, Russia

*Corresponding Author: Mostafa Abotaleb

DOI: https://doi.org/10.52866/ijcsm.2024.05.03.028
Received January 2024; Accepted May 2024; Available online August 2024

ABSTRACT: This study introduces an advanced algorithm based on the Generalized Least Deviation Method (GLDM) tailored for the univariate time series analysis of COVID-19 data. At the core of this approach is the optimization of a loss function, strategically designed to enhance the accuracy of the model's predictions. The algorithm leverages second-order terms, crucial for capturing the complexities inherent in time series data. Our findings reveal that by optimizing the loss function and effectively utilizing second-order model dynamics, there is a marked improvement in the predictive performance. This advancement leads to a robust and practical forecasting tool, significantly enhancing the accuracy and reliability of univariate time series forecasts in the context of monitoring COVID-19 trends.

Keywords: GLDM Algorithm; Time Series Forecasting; Loss Function Minimization; COVID-19 Time Series; Noise-Reduction Algorithm

1. INTRODUCTION

The COVID-19 pandemic has created an urgent demand for advanced analytical tools to predict its spread and assess potential public health interventions. Mathematical modeling, particularly using univariate time series analysis, provides a valuable framework for making these predictions and understanding the future trajectory of the epidemic. This research introduces a novel algorithm based on the Generalized Least Deviation Method (GLDM), specifically designed for epidemic data. This algorithm focuses on optimizing a well-defined loss function to improve forecasting accuracy.

In our model, the observed data at time t, v_t, is predicted based on past data points, v_{t-m}, using corresponding coefficients c_i and a stochastic error term δ_t. The functions f_i capture the relationships within the data. Our approach aims to minimize a loss function $L(\mathbf{c})$, which enhances the model's predictive performance by accurately adjusting the coefficients $\mathbf{c}$ based on past data. This optimization is achieved through a carefully developed algorithm that iteratively refines the coefficients, ensuring the model remains robust and adaptable to new data. By focusing on significant coefficients and systematically minimizing the loss function, our algorithm demonstrates exceptional capability in forecasting epidemiological trends. This makes it a critical tool for public health planning and response during the ongoing pandemic.

Time series modeling and forecasting are critical across economic, social, and environmental sectors, driving precise predictions that inform prevention, control, and strategic planning [1–11]. The pursuit of advanced predictive systems for time series analysis has become a focal point in scientific research, yielding an extensive body of literature. The ongoing quest for precision in forecasts and the continuous enhancement of algorithms keep predictive modeling at the forefront of research agendas. Given the unique attributes of each time series, no universal solution for forecasting exists; the choice of model is inherently dictated by the specific characteristics of the data under consideration, necessitating a tailored approach that adapts to new insights and methodologies [12–19].

Significant challenges have been encountered by researchers in forecasting real-time COVID-19 cases using traditional mathematical, statistical, and machine learning-based tools. In March 2020, studies using simple and efficient forecasting methods like the exponential smoothing model were conducted. These studies predicted cases ten days ahead with large confidence intervals. Despite the positive bias, the forecast error was found to be reasonable [20]. Forecasts using previously employed linear and exponential models were conducted for better preparation in terms of hospital bed availability, ICU admission estimation, resource allocation, emergency funding, and the proposal of strong containment measures [21]. Projections for ICU admissions in Italy were made for March 20, 2020. By the end of March 2020, ICU admissions and mechanical ventilation for critically ill patients reached their peak, shattering the health system of Lombardy, Italy [22]. Consequently, contemporary research spans a diverse array of models, including straightforward linear methods such as the Autoregressive Integrated Moving Average (ARIMA)[23–29].to Seasonal Autoregressive Integrated Moving Average (SARIMA)[30, 31], advanced nonlinear methods utilizing machine learning (ML) [32–38]. There's now a lot of knowledge about tracking vibration signals to check and predict the health and life of machines. Therefore, making these checks faster and more accurate, especially for special machines that work really hard, is very important. This is shown by[39, 40]. Often, understanding a machine's dynamics helps solve these issues. Finding the right math model to link a machine's condition with its diagnostic signs makes this easier. Models might include difference equations, phenomenological models, structural models, or regression models. The best model depends on the specific traits and behavior of the process being studied. For a long time, using statistics, neural networks, or math models for identification has been crucial across various fields. Today, these approaches are applied beyond industry, including efforts to predict the course of the Covid-19 pandemic, as illustrated by [41]. This work evaluates different well-known models' ability to forecast pandemic trends, develops software to run these methods, and conducts computational experiments with Covid-19 data. The authors find that their forecasting approach is versatile, applicable to various time series. Most predictions, particularly with large datasets, rely on various neural network models. For instance, [42] The study focuses on a neural network model designed for short-term predictions of ferrosilicon prices in Russia's domestic market. This model stands out for its accuracy and could support strategic decision-making at research institutes and metallurgical companies. The article outlines econometric models for assessing the metallurgical industry's economic indicators and forecasting ferrous metal production and its future growth. However, these models often seem like a "black box," providing answers without clear insight into how they work. To improve forecast quality, some researchers turn to cognitive modeling alongside neural networks[43, 44]. Given that the mentioned models focus on short-term forecasting, there's an urgent need to develop a mathematical approach for deriving high-quality quasi-linear difference equations.

As we delve deeper into the intricacies of epidemiological modeling, particularly within the realm of univariate time series analysis, the importance of mathematical precision and analytical rigor cannot be overstated. The coefficients $\mathbf{c} = \{c_1, c_2, \ldots, c_{n(m)}\}$, integral to our GLDM algorithm, play a pivotal role in quantifying the historical influence on current and future pandemic trends. These coefficients, through the process of loss function optimization, enable our algorithm to accurately model the dynamic nature of disease spread. This approach not only enhances the predictive capacity of our model but also emphasizes the significance of each variable in the univariate series, highlighting the nuanced interplay between past and predicted outcomes. By leveraging this algorithmic framework, we aim to provide a robust tool for forecasting epidemiological patterns, thereby offering a significant contribution to public health analytics and decision-making processes.

2. NOVELTY AND KEY DIFFERENCES

The suggested GLDM (Generalized Least Deviation Method) and classical regression, along with several other models, are used for modeling and forecasting data. This section discussed the key differences and novelties for these approaches.

2.1 GLDM: GENERALIZED LEAST DEVIATION METHOD

GLDM aims to minimize the deviation between observed and predicted values by optimizing a specifically defined loss function. This approach offers several novelties:

- **Objective**: GLDM formulates the objective as follows:

$$\min_{\mathbf{c}} L(\mathbf{c}) = \frac{1}{T}\sum_{t=1}^{T}(v_t - \hat{v}_t(\mathbf{c}))^2,$$

  where v_t represents the observed value, $\hat{v}_t(\mathbf{c})$ is the predicted value, and $\mathbf{c}$ represents the coefficients.

- **Model Structure**: GLDM utilizes a quasilinear recurrence equation model that captures the current value of the

variable being predicted as a combination of past values and their corresponding coefficients:

$$v_t = \sum_{i=1}^{p} c_i f_i(\{v_{t-m}\}_{m=1}^{q}) + \delta_t,$$

where f_i represents a function that incorporates higher-order terms and nonlinearity.

- **Robustness**: GLDM is known for its robustness against outliers and its ability to handle nonlinear patterns.
- **Forecasting Capability**: GLDM is specifically designed for time series forecasting and can effectively capture temporal dependencies in the data.
- **Advantages over Other Models**: GLDM offers several advantages over traditional and other modern models:
  - **Holt's Linear Trend Model**: GLDM can handle nonlinear trends more effectively than Holt's linear trend model by incorporating nonlinear terms and functions in its model structure.
  - **BATS and TBATS Models**: GLDM provides better robustness against outliers compared to BATS and TBATS models, and can capture nonlinear patterns more effectively.
  - **NNAR**: GLDM provides a more interpretable model compared to Neural Network Autoregression (NNAR), as it utilizes a combination of past values and coefficients which simplifies understanding the relationship between predictors and the forecasted variable.
  - **Classical Models**: GLDM improves upon classical models such as Simple Moving Average (SMA) and Exponential Smoothing (ES) by incorporating a more sophisticated model structure that captures higher-order terms and nonlinearity, enabling it to handle complex patterns and dependencies in the data more effectively.

2.2 ARIMA: AUTOREGRESSIVE INTEGRATED MOVING AVERAGE

ARIMA is a widely used model for time series forecasting. It combines autoregressive (AR), differencing (I), and moving average (MA) components [29]. The model equation can be expressed as:

$$v_t = d + \sum_{i=1}^{p} \phi_i v_{t-i} + \sum_{j=1}^{q} \theta_j \delta_{t-j} + \delta_t,$$

where d is a constant term, ϕ_i and θ_j are the autoregressive and moving average coefficients, and δ_t represents the error term.

Advantages of GLDM over ARIMA:

- GLDM can handle nonlinear patterns more effectively than ARIMA, as it incorporates higher-order terms and nonlinear functions in its model structure.
- GLDM is known for its robustness against outliers, which can be beneficial when dealing with data containing extreme observations.

2.3 HOLT'S LINEAR TREND MODEL

Holt's linear trend model is suitable for time series data with a linear trend. It consists of two components: the level (ℓ_t) and the trend (b_t). The model equations are given by [45]:

$$\ell_t = \alpha v_t + (1-\alpha)(\ell_{t-1} + b_{t-1}),$$
$$b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta) b_{t-1},$$
$$\hat{v}_{t+h} = \ell_t + h b_t,$$

where α and β are smoothing parameters, and $\hat{v}_{t+h}$ represents the forecasted value at time $t+h$.

Advantages of GLDM over Holt's Linear Trend Model:

- GLDM can handle nonlinear trends more effectively than Holt's linear trend model by incorporating nonlinear terms and functions in its model structure.
- GLDM's quasilinear autoregressive model captures temporal dependencies more explicitly, which can improve forecasting accuracy compared to Holt's linear trend model.

2.4 BATS AND TBATS MODELS

BATS (Box-Cox transformation, ARMA errors, Trend, and Seasonal components) and TBATS (Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend, and Seasonal components) models are designed to handle time series with multiple seasonal patterns, nonlinear trends, and transformations. The core equations for these models incorporate complex seasonal structures[46]:

$$v_t = \text{Box-Cox}(v_t, \lambda) + \sum_{i=1}^{p} \phi_i v_{t-i} + \sum_{j=1}^{q} \theta_j \delta_{t-j} + m_t + \delta_t,$$

where m_t represents the seasonal component modeled either through Fourier terms or specific seasonal ARMA processes.

Advantages of GLDM over BATS and TBATS Models:

- GLDM offers superior robustness against outliers compared to BATS and TBATS models, making it more reliable in the presence of extreme observations.
- Due to its a quasilinear recurrence equation structure, GLDM is more effective at capturing complex nonlinear patterns that can occur in highly dynamic time series data.

2.5 NNAR: NEURAL NETWORK AUTOREGRESSION

NNAR is a type of model that utilizes neural networks to forecast future values based on past data points. It is particularly suited for capturing nonlinear relationships within the data. The typical NNAR model structure can be described by the following function[47]

$$v_t = f(v_{t-1}, v_{t-2}, \ldots, v_{t-p}) + \delta_t,$$

where f represents a neural network function designed to map past values to a future prediction.

Advantages of GLDM over NNAR:

- GLDM provides greater interpretability compared to NNAR models. By using explicit mathematical functions and coefficients, GLDM allows for easier understanding and analysis of how inputs affect forecasts.
- The structured nature of GLDM, with its reliance on quasilinear terms and parameterized functions, offers better insights into temporal dependencies than the often "black-box" nature of neural networks.

2.6 CLASSICAL MODELS

Classical forecasting models such as the Simple Moving Average (SMA) and Exponential Smoothing (ES) provide baseline methodologies for time series analysis. These models generally employ simpler calculations, which can be described as follows:

Simple Moving Average (SMA):

$$v_t = \frac{1}{n} \sum_{i=t-n+1}^{t} v_i,$$

Exponential Smoothing (ES):

$$v_t = \alpha v_{t-1} + (1 - \alpha) v_{t-2},$$

where n is the number of terms to average, and α is the smoothing factor.

Advantages of GLDM over Classical Models:

- GLDM's advanced modeling capabilities allow it to handle data complexities and intricacies that classical models cannot, such as non-linearities and structural breaks in the data series.
- By optimizing a loss function that minimizes deviation between observed and predicted values, GLDM provides a more accurate and robust approach to forecasting, especially in scenarios where data exhibit volatility and irregular trends.

3. METHOD

Before delving into the specifics of our methodological approach for predicting daily COVID-19 cases, it is essential to underscore the mathematical and computational framework underpinning the Generalized Least Deviation Method (GLDM). Central to GLDM is the goal of minimizing the deviation between observed and predicted COVID-19 case counts, formalized as the optimization problem: minimize $L(\mathbf{c}) = \sum_{i=1}^{n} |v_i - \hat{v}_i(\mathbf{c})|$, where v_i represents the observed daily case numbers, $\hat{v}_i(\mathbf{c})$ the predicted case numbers derived from the model, and $\mathbf{c} = \{c_1, c_2, ..., c_{n(m)}\}$ the set of model coefficients. This method's efficacy stems from its robustness against outliers and its capability to yield dependable forecasts amidst non-linear disease spread patterns. Through the application of GLDM across varying datasets and model complexities, we aim to reveal the optimal model configuration that strikes a balance between forecasting accuracy and model complexity. This endeavor lays a solid groundwork for subsequent analysis and discussions concerning the model's performance and its suitability for predicting the course of the COVID-19 pandemic.

The initial stage of the forecasting procedure involves a *Time Series* dataset, denoted as $\{v_t\} \in \mathbb{R}_{t=1-m}^{T}$, where each v_t signifies a datum at time t, encapsulated within a period from 1 to T, with the initiation at an earlier point indexed by m.

Subsequent to the collection of time series data, the process incorporates a *GLDM Estimator algorithm*. GLDM, postulated as an acronym for Generalized Least Deviation Method, is postulated to calibrate the data, deducing a set of pivotal factors $\{c_1, c_2, \ldots, c_{n(m)}\} \in \mathbb{R}$. These factors, intrinsic real numbers, epitomize the inferred parameters obtained from the time series data.

These extracted factors are then harnessed by a *Predictor* mechanism to prognosticate future values. This predictor is designed to generate outputs encapsulating the *Forecasting Horizon* (FH) and prospective forward-looking values, indicative of the temporal scope and expected data points for this horizon respectively.

3.1 OVERVIEW OF GLDM MODEL APPLICATION

In our analysis, we apply the Generalized Least Deviation Method (GLDM) alongside the Weighted Least Deviation Method (WLDM) to effectively model and forecast COVID-19 time series data. This comprehensive approach starts with the initialization of data, proceeds through various estimation and optimization steps, and culminates in the generation of forecasts. Figure 1 provides a visual summary of these sequential steps, illustrating the methodical process employed to ensure accuracy and reliability in our predictions.

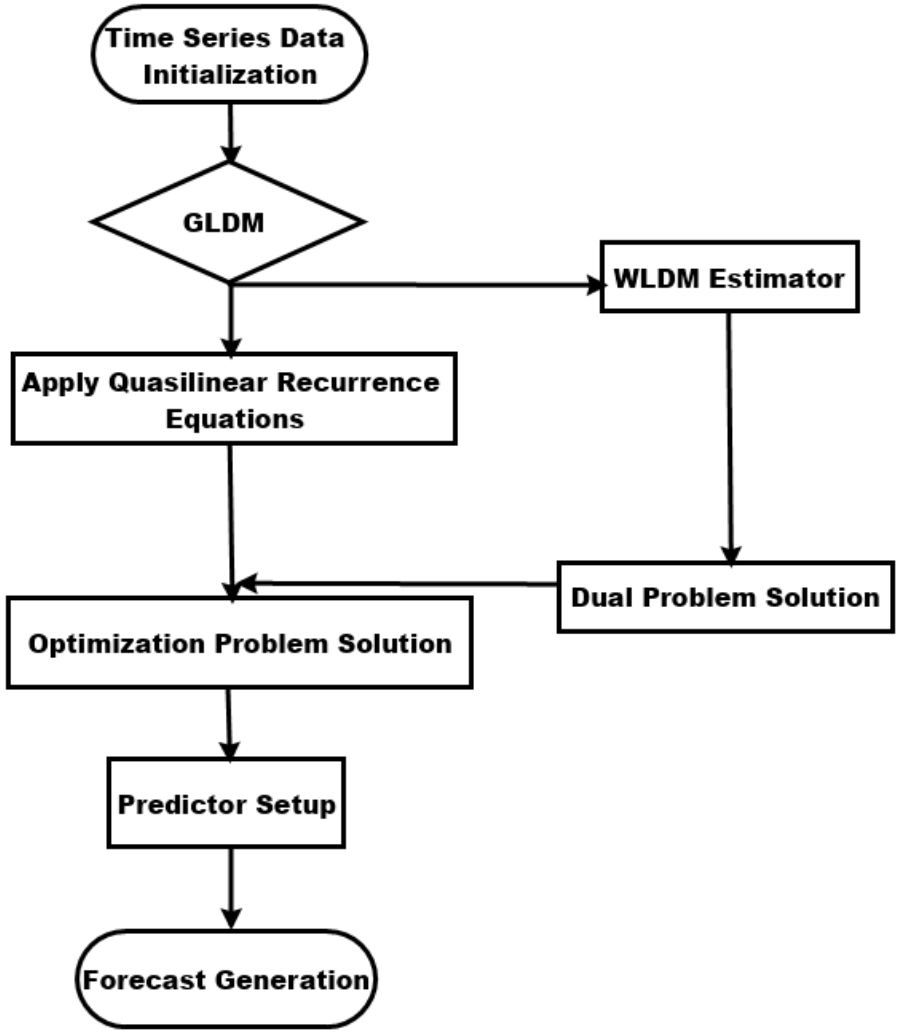

FIGURE 1. **Flowchart illustrating the steps involved in the application of the GLDM model to COVID-19 time series data.**

3.2 STEPS IN APPLYING THE GLDM MODEL

1. **Time Series Data Initialization**

$$\text{Initialization: } \{v_t\}_{t=1-m}^{T} \subseteq \mathbb{R} \tag{1}$$

2. **GLDM Estimator**

$$\text{Apply GLDM for initial estimation of } \mathbf{c} = \{c_1, c_2, \ldots, c_{n(m)}\} \tag{2}$$

3. **Apply Quasilinear Recurrence Equations**

$$v_t = \sum_{i=1}^{p(q)} c_i f_i \left(\{v_{t-m}\}_{m=1}^{q} \right) + \delta_t \tag{3}$$

4. **WLDM Estimator**

$$\text{Refine } \mathbf{c} \text{ using WLDM, focusing on minimizing deviations.} \tag{4}$$

5. **Dual Problem Solution**

$$\text{Solve the dual optimization problem to refine } \mathbf{c}. \tag{5}$$

6. **Optimization Problem Solution**

$$\text{Final optimization of } \mathbf{c} \text{ to minimize } L(\mathbf{c}). \tag{6}$$

7. **Predictor Setup**

$$\text{Configure the predictor with optimized } \mathbf{c} \text{ for future forecasting.} \tag{7}$$

8. **Forecast Generation**

$$\overline{v[t]}_{\tau} = \sum_{i=1}^{p(q)} c_i^* f_i \left(\{ \overline{v[t]}_{\tau-k} \}_{k=1}^{q} \right) \tag{8}$$

3.3 QUASILINEAR RECURRENCE EQUATIONS

Quasilinear recurrence equations represent a sophisticated class of difference equations that integrate linear and nonlinear dynamics to model the evolution of time series data. The general form of these equations is expressed as follows:

$$v_t = \sum_{i=1}^{p(q)} c_i f_i \left(\{ v_{t-m} \}_{m=1}^{q} \right) + \delta_t \tag{9}$$

where:

- v_t denotes the time series value at time t,
- c_i are the coefficients or parameters of the model,
- $f_i \left(\{ v_{t-m} \}_{m=1}^{q} \right)$ represents functions that depend on the past q values of the time series, embodying both linear and nonlinear interactions,
- δ_t is the stochastic error term, which accounts for the noise in the data.

The key characteristics of quasilinear recurrence equations are outlined as follows:

1. **Linearity in Parameters**: Despite potential nonlinearity in the functions f_i, the overall equation remains linear with respect to the coefficients c_i. This structure ensures that the prediction $\hat{v}_t$ is a linear amalgamation of the model terms $f_i \left(\{ v_{t-m} \}_{m=1}^{q} \right)$, where the coefficients c_i determine the weight of each term.
2. **Nonlinearity in Past Values**: The functions f_i allow for the inclusion of nonlinear relationships between the current value v_t and the past q values $\{ v_{t-m} \}_{m=1}^{q}$. This capability is essential for capturing complex dynamics that linear models fail to represent.
3. **Flexibility in Model Structure**: The model can be customized through the choice of functions f_i, which can be designed to capture various nonlinear patterns such as polynomial, trigonometric, or exponential forms. This flexibility enables the model to adapt effectively to the specific characteristics of the time series being analyzed.

The *order* of a quasilinear recurrence equation, denoted by q, indicates the number of past time steps included in the model. For example, a first-order model includes only the immediate past value v_{t-1}, while a second-order model also considers v_{t-2}. Models of higher order involve more past values, allowing for the capture of more detailed temporal dependencies but also increasing the risk of overfitting, particularly when the number of parameters is large relative to the amount of data available.

Estimation of the coefficients c_i typically involves optimizing a loss function, such as the sum of squared errors or the sum of absolute deviations. This optimization process aims to identify the best-fit coefficients that align closely with the observed time series data, thus enabling the model to generate accurate forecasts and provide deeper insights into the time series dynamics.

Quasilinear recurrence equations provide a powerful and interpretable tool for time series modeling and forecasting. Analyzing the estimated coefficients c_i offers insights into the importance of various factors influencing the evolution of the time series, including potential nonlinear interactions between past values.

3.4 PROBLEM NOTATION AND STATEMENT

The considered algorithm operates as follows (see Fig. 2).

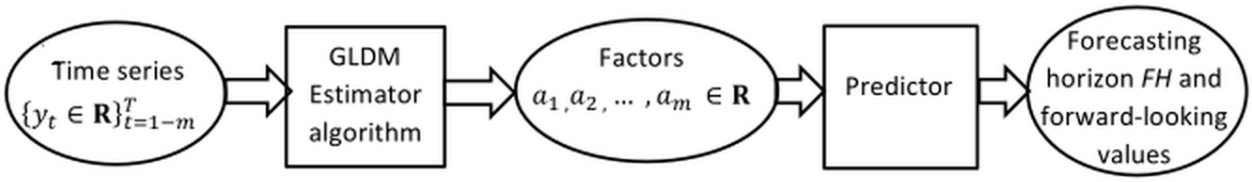

FIGURE 2. The framework for applying the model

Let's examine a time series for a specific tile. This approach can be extended to other tiles with adjustments to the parameters.

Linear autoregressive models offer a limited forecast range. Building suitable nonlinear models or neural networks might not always be feasible due to technical constraints. Quasilinear models, however, can extend the forecasting period. Let us implement our approach considered in [48] to determine the coefficients $c_1, c_2, c_3 \ldots, c_m \in \mathbb{R}$ of a m-th order quasilinear autoregressive model

$$v_t = \sum_{i=1}^{p(q)} c_i f_i(\{v_{t-k}\}_{k=1}^{q}) + \delta_t, \quad t = 1, 2, \ldots, T \tag{10}$$

Guided by recent updates on state variable values $\{v_t \in \mathbb{R}\}_{t=1-m}^{T}$ at specific time points t, where the functions

$$f_i : (\{v_{t-k}\}_{k=1}^{q}) \to \mathbb{R}, \quad \text{for } i = 1, 2, \ldots, p(q),$$

are predefined for $p(q)$ scenarios, and the sequence $\{\delta_t \in \mathbb{R}\}_{t=1}^{T}$ represents the set of unobservable errors.

The method under review focuses on identifying the coefficients of the difference equation outlined in (10). Employing the GLDM estimation algorithm presented in [48], this technique processes a time series $\{v_t \in \mathbb{R}\}_{t=-1-m}^{T}$ with a minimum length of $T + m \geq (1 + 3m + m^2)$. It aims to compute the coefficients $c_1, c_2, c_3, \ldots, c_m \in \mathbb{R}$ by addressing the corresponding optimization problem.

$$\sum_{t=1}^{T} \arctan \left| \sum_{i=1}^{p(q)} c_i f_i(\{v_{t-k}\}_{k=1}^{q}) - v_t \right| \to \min_{\{c_i\}_{i-1}^{p(q)} \subset \mathbb{R}} \tag{11}$$

The Cauchy distribution

$$F(\xi) = \frac{1}{\pi} \arctan(\xi) + \frac{1}{2}$$

possesses the highest entropy among distributions of random variables lacking both mean and variance. Consequently, the $\arctan(*)$ function is utilized as the loss function.

Let's consider a m-th order model with quadratic nonlinearity. Then the basic set $f_i(*)$ may contain the following functions

$$\begin{aligned} f_{(k)}(\{v_{t-k}\}_{k=1}^{q}) &= v_{t-k}, \\ f_{(kl)}(\{v_{t-k}\}_{k=1}^{q}) &= v_{t-k} \cdot v_{t-l}, \\ k = 1, 2, \ldots, q;\ l &= k, k+1, \ldots, q. \end{aligned} \tag{12}$$

Obviously, in this case $p(q) = 2q + C_q^2 = q(q+3)/2$, and the numbering of $f_{(*)}$ functions can be arbitrary. In particular, for $q = 2$ functions $f_{(*)}$ are the following

$$f_1 = v_1, \quad f_2 = v_2, \quad f_3 = v_1^2, \quad f_4 = v_2^2, \quad f_5 = v_1 \cdot v_2.$$

The model for this case looks like following:

$$v_t = (c_1 v_{t-1} + c_2 v_{t-2}) + \left(c_3 v_{t-1}^2 + c_4 v_{t-2}^2 + c_5 v_{t-1} v_{t-2}\right). \tag{13}$$

The predictor establishes a sequence indexed by $t = 1, 2, \ldots, T-1, T$, consisting of m-th order difference equations

$$\overline{v[t]}_\tau = \sum_{i=1}^{p(q)} c_i^* f_i\left(\{\overline{v[t]}_{-k}\}_{k=1}^{q}\right),$$
$$= t, t+1, t+2, t+3, \ldots, T-1, T, T+1, \ldots \quad (14)$$

for lattice functions $\overline{v[t]}$, where $\overline{v[t]}$ are forecasted values for v at the specific time t. By solving the Cauchy problem for this difference equation (14) with initial conditions

$$\overline{v[t]}_{t-1} = v_{t-1},\ \overline{v[t]}_{t-2} = v_{t-2},\ \ldots,\ \overline{v[t]}_{t-m} = v_{t-m} \quad t = 1, 2, \ldots, T-1, T, \quad (15)$$

we determine the function $\overline{v[t]}$ values.

This approach yields the collection $\overline{V} = \left\{\overline{v[t]}\right\}_{t=1}^{T}$ of feasible forecasts for v. These forecasts are then utilized to assess the probabilistic attributes of the anticipated v values.

3.5 EVALUATING BY GLDM

The task described in (11), namely the GLDM-estimation challenge, constitutes a problem with multiple local optima. GLDM estimations exhibit resilience against correlated data points within $\{v_t \in \mathbb{R}\}_{t=-1-m}^{T}$, and under proper configurations, they excel in scenarios where error probability distributions have tails heavier than those of a normal distribution, as discussed in [49]. These characteristics underscore the practicality of addressing the identification issue presented in (10) through the solutions provided by (11). Furthermore, by leveraging the connection between GLDM estimates and those obtained through the Weighted Least Deviation Method (WLDM), as explored by [50], we can efficiently tackle higher-dimensional instances of (11).

In the context of this study, we explore the GLDM estimation method as outlined in [51], integrating the approach with the WLDM estimation algorithm within the framework of the GLDM procedure.

The operational flow of the algorithm is depicted in Figure 3, with the following primary inputs:

- $S = \{S_t \in \mathbb{R}^N\}_{t \in T}$, representing the linear variety matrix;
- $\nabla_{\mathcal{L}}$, denoting the gradient projection of the objective function onto $\mathcal{L}$;
- weight coefficients $\{p_t \in \mathbb{R}^+\}_{t=1}^{T}$, for adjusting the significance of data points;
- and the specified state variables $\{v_t \in \mathbb{R}^+\}_{t=1-m}^{T}$, providing the initial conditions.

The algorithm iterates to converge to the optimal GLDM solution $A \in \mathbb{R}^{n(m)}$ alongside the residual vector $z \in \mathbb{R}^T$. Iteration ceases once the condition $\left(A^{(k)} = A^{(k-1)}\right)$ is met, indicating stability in the solution.

3.5.1. Evaluating by WLDM

Algorithm WLDM-estimator [52] receives a time series $\{v_t \in \mathbb{R}\}_{t=1-m}^{T}$ and weight factors $\{p_t \in \mathbb{R}^+\}_{t=1}^{T}$ as input data and calculates the factors

$$c_1, c_2, c_3 \ldots, c_{p(q)} \in \mathbb{R}$$

by solving the optimization problem

$$\sum_{t=1}^{T} p_t \cdot \left| \sum_{i=1}^{p(q)} c_i f_i(\{v_{t-k}\}_{k=1}^{q}) - v_t \right| \to \min_{\{c_i\}_{i=1}^{p(q)} \in \mathbb{R}^{p(q)}} \quad (16)$$

This algorithm's outline is depicted in Figure 4. The computational load of this algorithm is limited to $O(T^2)$, attributable to the straightforward nature of the permissible set: a cross-section of a T-dimensional cuboid (28) with a $(T - p(q))$-dimensional linear manifold (27).

The dual problem-solving algorithm (26)–(28) initiates the search for an optimal solution from zero, advancing in the direction of $\nabla_{\mathcal{L}}$. Should the trajectory intersect the boundary of the domain $\mathcal{T}$, the movement in that particular dimension is halted.

Upon achieving a result (w^*, R^*) through the gradient projection method [48], where w^* is the prime solution for (26)-(28), the optimum resolution for the problem (23)-(25) is given by

$$u_t^* = \frac{p_t + w_t^*}{2}, \quad v_t^* = \frac{p_t - w_t^*}{2}, \quad t = 1, 2, \ldots, T.$$

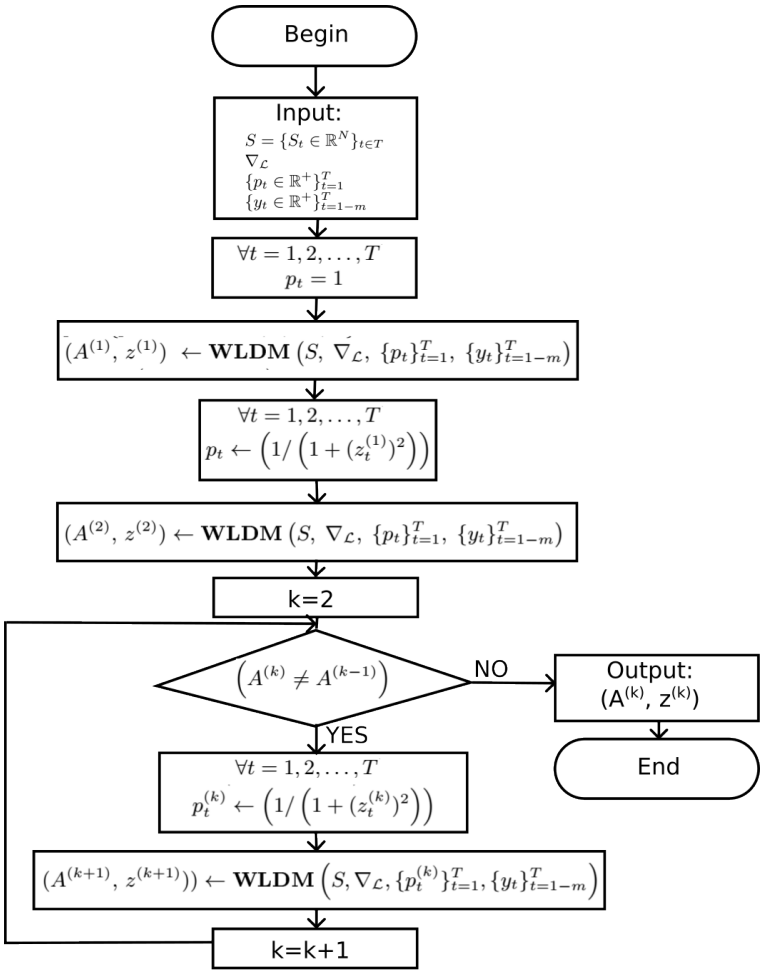

FIGURE 3. Illustration of the GLDM estimation algorithm.

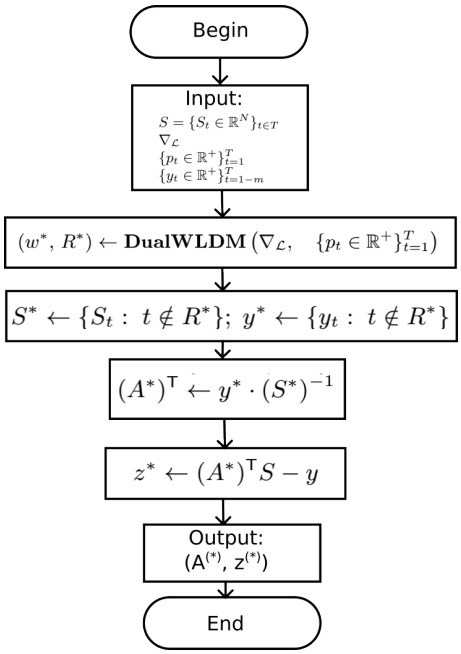

FIGURE 4. Workflow of the WLDM Estimation Algorithm

This outcome is derived from the complementarity principle for mutually dual tasks (20)–(22) and (23)–(25), suggesting that

$$v_t = \sum_{i=1}^{p(q)} \left[c_i f_i(\{v_{t-k}\}_{k=1}^{q}) \right] \qquad \forall t \notin R^*, \tag{17}$$

$$v_t = \sum_{i=1}^{p(q)} \left[c_i f_i(\{v_{t-k}\}_{k=1}^{q}) \right] + z_t^*, \quad \forall t in R^* : \; w_t^* = p_t, \tag{18}$$

$$v_t = \sum_{i=1}^{p(q)} \left[c_i f_i(\{v_{t-k}\}_{k=1}^{q}) \right] - z_t^*, \quad \forall t in R^* : \; w_t^* = -p_t. \tag{19}$$

Indeed, the solution set $(\{c_i^*\}_{i=1}^{p(q)}, z^*)$ to the system of linear algebraic equations given by (17)–(19) constitutes the dual optimal solution for problem (26)-(28) and simultaneously serves as the optimal solution for problem (16). This outcome substantiates the theorem presented in [53].

Theorem 3.1. *Let*

- w^* *be the optimal solution of the task* (26)-(28)*,*
- $(\{c_i^*\}_{i=1}^{p(q)}, z^*)$ *be the solution of a system of linear algebraic equations* (17)-(19)*.*

Then $\{c_i^*\}_{i=1}^{p(q)}$ *is the optimal solution to the task* (16)*.*

The main problem with the use of the WLDM-estimator is the absence of general formal rules for choosing weight coefficients. Consequently, this approach requires additional research.

This problem represents a convex piecewise linear optimization problem, and the introduction of additional variables reduces it to a linear programming problem.

The task (11), i.e. task of GLDM-estimation, is a concave optimization problem, and entering the additional variables reduces it to the following linear programming task

$$\sum_{t=1}^{T} p_t z_t \to \min_{\substack{(c_1,c_2,\ldots,c_{p(q)})\in \mathrm{R}^q, \\ (z_1,z_2,\ldots,z_T)\in \mathrm{R}^T}} \tag{20}$$

$$-z_t \le \sum_{i=1}^{p(q)} \left[c_i f_i(\{v_{t-k}\}_{k=1}^{q}) \right] - v_t \le z_t, \quad t = 1, 2, \ldots, T, \tag{21}$$

$$z_t \ge 0, \quad t = 1, 2, \ldots, T. \tag{22}$$

The task (20)–(22) has a canonical type with variables $p(q) + T$ and $3T$ inequality constraints including the conditions of non-negativity of $z_j,\ j = 1, 2, \ldots, T$.

The dual to (20) task is

$$\sum_{t=1}^{T} (u_t - v_t)\, v_t \to \max_{u,v\in \mathbb{R}^T}, \tag{23}$$

$$\sum_{t=1}^{T} c_i f_i(\{v_{t-k}\}_{k=1}^{q})\, (u_t - v_t) = 0,\ i = 1, 2, \ldots, p(q), \tag{24}$$

$$u_t + v_t = p_t, \quad u_t,\ v_t \ge 0, \quad t = 1, 2, \ldots, T. \tag{25}$$

Let us introduce variables $w_t = u_t - v_t,\ t = 1, 2, \ldots, T$. Conditions (25) imply that

$$u_t = \frac{p_t + w_t}{2}, \quad v_t = \frac{p_t - w_t}{2}, \quad -p_t \le w_t \le p_t, \quad t = 1, 2, \ldots, T.$$

So the optimal task (23)–(25) solution is equal to the optimal solution of task

$$\sum_{t=1}^{T} w_t \cdot v_t \to \max_{w \in \mathbb{R}^T}, \tag{26}$$

$$\sum_{t=1}^{T} f_i(\{v_{t-k}\}_{k=1}^{q}) \cdot w_t = 0,\ i = 1, 2, \ldots, p(q), \tag{27}$$

$$-p_t \le w_t \le p_t,\ t = 1, 2, \ldots, T. \tag{28}$$

Constraints (27) define $(T - p(q))$-dimensional linear variety $\mathcal{L}$ with $(p(q) \times T)$-matrix

$$S = \begin{bmatrix} f_1(\{v_{1-k}\}_{k=1}^{q}) & f_1(\{v_{2-k}\}_{k=1}^{q}) & \ldots & f_1(\{v_{T+1-k}\}_{k=1}^{q}) \\ f_2(\{v_{1-k}\}_{k=1}^{q}) & f_2(\{v_{2-k}\}_{k=1}^{q}) & \ldots & f_2(\{v_{T+1-k}\}_{k=1}^{q}) \\ \vdots & \vdots & \ddots & \vdots \\ f_{p(q)}(\{v_{1-k}\}_{k=1}^{q}) & f_{p(q)}(\{v_{2-k}\}_{k=1}^{q}) & \ldots & f_{p(q)}(\{v_{T+1-k}\}_{k=1}^{q}) \end{bmatrix}$$

Constraints (28) define T-dimensional parallelepiped $\mathcal{T}$.

The simple structure of the allowed set for task (26)–(28) representing the intersection of $(T - p(q))$-dimensional linear variety $\mathcal{L}$ (27) and T-dimensional parallelepiped $\mathcal{T}$ (28) allows to obtain its solution by an algorithm using the gradient projection of the objective function (26) (i.e. vector $\nabla = \{v_t\}_{t=1}^{T}$) on the allowed area $\mathcal{L} \cap \mathcal{T}$ defined by the constraints (27)–(28). The projection matrix on $\mathcal{L}$ is as following:

$$S_{\mathcal{L}} = E - S^T \cdot (S \cdot S^T)^{-1} \cdot S,$$

and gradient projection on $\mathcal{L}$ is equal to $\nabla_{\mathcal{L}} = S_{\mathcal{L}} \cdot \nabla$. Moreover, if the outer normal on any parallelepiped face forms a sharp corner with the gradient projection $\nabla_{\mathcal{L}}$, then movement by this face is halted.

The **DualWLDMSolver** Algorithm, as detailed in Algorithm 1, initiates the quest for the optimal solution from zero, progressing in the direction of the gradient $\nabla_{\mathcal{L}}$. Should the current position reach the boundary of the domain $\mathcal{T}$, the relevant movement coordinate is set to zero.

Algorithm 1 . DualWLDMSolver

Require: :
- $\nabla_{\mathcal{L}}$ ▹ Gradient projection
- $\{p_t \in \mathbb{R}^+\}_{t=1}^{T}$ ▹ Weight factors

Ensure: :
- $w^* = \arg\max_{w \in \mathbb{R}^T} \sum_{i=1}^{T} w_i \cdot v_i$ ▹ Optimal dual solution
- $R^* = \{t \in T : |w_t^*| = p_t\}$ ▹ Active restrictions

1: $w \leftarrow \{w_i = 0 :\ i = 1, 2, \ldots, T\};\ R \leftarrow \emptyset;\ g = \nabla_{\mathcal{L}}$
2: **while** $(\alpha_* \neq 0)$ **do**
3: $\quad\{(\alpha_*, t_*) \leftarrow \arg\max\{\alpha \ge 0 :\ -p_t \le w_t + \alpha g_t \le p_t\}\}$
4: $\quad w \leftarrow w + \alpha_* g;\ g_{t_*} \leftarrow 0;\ R := R \cup \{t_*\};$
5: **end while**
6: $w^* = w,\ R^* = R$
return (w^*, R^*)

The computational workload of this algorithm is limited to $O(T^2)$, owing to the straightforward nature of the permissible set: the intersection between a T-dimensional rectangular prism as described in (28) and a $(T - p(q))$-dimensional linear manifold referenced in (27). The derived solution set $(\{c_i^*\}_{i=1}^{p(q)}, z^*)$ from the system of linear equations (17)-(19) serves as the dual optimum for the issue delineated in (26)-(28) and as the prime solution for the dilemma outlined in (16). This outcome substantiates the theorem presented below.

Theorem 3.2. *Assuming w^* to be the prime resolution for the issue* (26)-(28) *and $(\{c_i^*\}_{i=1}^{p(q)}, z^*)$ as the outcome of the linear algebraic equation system* (17)-(19)*, then $(\{c_i^*\}_{i=1}^{p(q)}$ epitomizes the optimal solution for the problem* (16)*.*

Algorithm 2 . WLDM-estimator

Require: :
$S = \{S_t \in \mathbb{R}^N\}_{t\in T}$ ▷ The matrix of a linear subspace $\mathcal{L}$
$\nabla_{\mathcal{L}}$ ▷ Gradient projection on $\mathcal{L}$
$\{p_t \in \mathbb{R}^+\}_{t=1}^T$ ▷ Weight factors
$\{v_t \in \mathbb{R}^+\}_{t=1-m}^T$ ▷ Values of the given state variables
Ensure: :
$C^* \in \mathbb{R}^{p(q)}$ ▷ Optimal primal solution
$z^* \in \mathbb{R}^T$ ▷ Restrictions

$(w^*,\ R^*) \leftarrow$ **DualWLDMSoluter** $\left(\nabla_{\mathcal{L}},\quad \{p_t \in \mathbb{R}^+\}_{t=1}^T\right)$
$S^* \leftarrow \{S_t :\ t \notin R^*\};\ v^* \leftarrow \{v_t :\ t \notin R^*\}$ ▷ System (17) matrix
$(C^*)^{\mathsf{T}} \leftarrow v^* \cdot (S^*)^{-1}$ ▷ System (17) solution
$z^* \leftarrow (C^*)^{\mathsf{T}} S - v$ ▷ Find restrictions
return $(C^*,\ z^*)$

The above allows us to propose WLDM-estimator Algorithm 2. The main problem with the use of WLDM-estimator is the absence of general formal rules for choosing weight coefficients. Consequently, this approach requires additional research. The results established in [51], [54] allow us to reduce the problem of determining GLDM estimation to an iterative procedure with WLDM estimates.

3.5.2. GLDM estimation algorithm

The GLDM estimation task defined in (11) constitutes a concave optimization challenge. GLDM estimates demonstrate resilience against correlated data within $\{X_{jt} :\ t = 1, 2, \ldots, T;\ j = 1, 2, \ldots, N\}$ and, under optimal configurations, excel in accuracy for error probability distributions with tails heavier than those of a normal distribution [49]. This underlines the practicality of employing Algorithm (2) to address the identification issue outlined in (10). Insights from [51] enable simplifying the GLDM estimation into a series of steps utilizing WLDM estimates, detailed in Algorithm 3.

Algorithm 3 . GLDM-estimator

Require: :
$S = \{S_t \in \mathbb{R}^N\}_{t\in T}$ ▷ The matrix of a linear subspace $\mathcal{L}$
$\nabla_{\mathcal{L}}$ ▷ Gradient projection on $\mathcal{L}$
$\{p_t \in \mathbb{R}^+\}_{t=1}^T$ ▷ Weight factors
$\{v_t \in \mathbb{R}^+\}_{t=1-m}^T$ ▷ Values of the given state variables
Ensure: :
$C^* \in \mathbb{R}^{p(q)}$ ▷ Optimal GLDM solution
$z^* \in \mathbb{R}^T$ ▷ Residuals

$p \leftarrow \{p_t = 1 :\ t = 1, 2, \ldots, T\}$
$(C^{(1)},\ z^{(1)}) \leftarrow$
$\leftarrow$ **WLDMSolver** $\left(S,\ \nabla_{\mathcal{L}},\ \{p_t\}_{t=1}^T,\ \{v_t\}_{t=1-m}^T\right)$
for all $(t = 1, 2, \ldots T)$ **do**
    $p_t \leftarrow \left(1/\left(1 + (z_t^{(1)})^2\right)\right)$
end for
$(C^{(2)},\ z^{(2)}) \leftarrow$ **WLDMSolver** $\left(S,\ \nabla_{\mathcal{L}},\ \{p_t\}_{t=1}^T,\ \{v_t\}_{t=1-m}^T\right)$
$k \leftarrow 2$
while $\left(C^{(k)} \neq C^{(k-1)}\right)$ **do**
    for all $(t = 1, 2, \ldots T)$ **do**
        $p_t^{(k)} \leftarrow \left(1/\left(1 + (z_t^{(k)})^2\right)\right)$
    end for
    $((C, z)) \leftarrow$ **WLDMSolver** $\left(S, \nabla_{\mathcal{L}}, \{p_t^{(k)}\}_{t=1}^T, \{v_t\}_{t=1-m}^T\right)$
    $(C^{(k+1)},\ z^{(k+1)}) \leftarrow (C,\ z)$
    $k \leftarrow (k + 1)$
end while
$z^* \leftarrow z^{(k)},\quad (C^*) \leftarrow C^{(k)}$ ▷ Find restrictions
return $(C^*,\ z^*)$

Theorem 3.3. *The series* $\{\left(C^{(k)}, z^{(k)}\right)\}_{k=1}^{\infty}$, *generated through the GLDM-estimator Algorithm, is guaranteed to approach the global optimum* (c^*, z^*) *for the challenge presented in* (11).

The analysis of the **GLDM-estimator** Algorithm shows that its computational demand aligns with that of addressing primary and/or complementary WLDM issues, as defined in (16). Extensive computational tests have demonstrated that the average iteration count for the **GLDM-estimator** Algorithm correlates directly with the number of coefficients, $n(m)$, in the model equation.

Based on these observations, the computational load for applying the **GLDM-estimator** Algorithm in real-world scenarios is estimated to be

$$O(n(m)^3 T + n(m) \cdot T^2),$$

where $n(m)$ represents the number of model parameters and T denotes the number of time steps in the dataset. This formulation suggests that the complexity increases polynomially with the number of parameters and linearly with the square of the number of time steps, thus making it feasible for practical applications involving large datasets.

3.6 PREDICTOR

The forecasting mechanism constructs a sequence indexed by $t = 1, 2, \ldots, T-1, T$, utilizing m-th order difference equations (14) applied to lattice functions $\overline{v[t]}$, where $\overline{v[t]}_\tau$ predicts v_τ at each time t. We address the Cauchy problem for these difference equations (14), starting from the initial conditions (15), to determine the values of $\overline{v[t]}$. This approach results in the collection $\overline{V}_\tau = \{\overline{v[t]}_\tau\}_{t=1}^{T}$, encompassing all predictions for v_τ. Subsequently, this dataset aids in determining the probabilistic characteristics of v_τ, as detailed in Algorithm 4.

Algorithm 4 . Predictor

Require: :
$v = \{v_t \in \mathbb{R}^+\}_{t=1-m}^{T}$ ▹ Values of the given state variables
$c = \{c_i\}_{i=1}^{n(m)}$ ▹ Coefficients from the WLDM solution
Ensure: :
$PV[1:T][1:T] : PV[t][\tau] = \overline{v[t]}_\tau$ ▹ Forecast for v_τ at time t
ME ▹ Average prediction errors
MAE ▹ Average absolute prediction errors
$minFH$ ▹ Minimum reliable prediction horizon

while $(FH[Start] < m)$ **do**
  Start++;
  $PV[Start][0] = v[Start]$;
  $PV[Start][1] = v[Start + 1]$;
  for all $(t = Start + 2, \ldots, m)$ **do**
    $py = 0$;
    for all $j = 0, 1, \ldots, n$ **do**
      $result = G[j](PV[Start][t-1], PV[Start][t-2])$;
      $result = c[j] \times result$;
      $py+ = result$;
    end for
    $PV[Start][t] = py$;
    if $(|PV[Start][t] - v[Start + t]| > Threshold)$ **then**
      break;
    end if
  end for
  $FH[Start] = t$;
end while
$lastStart = t$;
$minFH = FH[Start]$;
$intminFHp = minFH$;
for all $t = 3, \ldots, Start$ **do**
  if $(minFH > FH[t])$ **then**
    $minFHp = FH[t]$;
  end if
end for
$minFH = (minFHp < minFH)minFHp : minFH$;
$ME = MAE = 0$; ▹ minFHp is the reasonable horizon for accuracy
for all $t = 3, \ldots, minFH$ **do**
  $MAE+ = |v[t + Start] - PV[Start][t]|$;
  $ME+ = (v[t + Start] - PV[Start][t])$;
end for
$MAE/ = minFH; ME/ = minFH$;
return $(MAE, ME, minFH)$

3.7 OPTIMIZING COEFFICIENTS

We present a general algorithm to optimize coefficients in quasilinear recurrence equations (QREs) of various orders for time series prediction. Our goal is to minimize the sum of absolute differences (SAD) between the predicted and actual values, which is a robust loss function suited for time series with outliers or non-normal error distributions.

Quasilinear recurrence equations incorporate past values of a time series to predict future values, allowing for both linear and non-linear dependencies. The model's complexity and potential for capturing intricate patterns increase with the order of the recurrence relation.

Given a time series $\{v_t\}_{t=1}^{N}$, a quasilinear recurrence equation of order n can be expressed as:

$$\hat{v}_t = \sum_{i=1}^{n} c_i v_{t-i} + \sum_{i=1}^{n} \sum_{j=i}^{n} c_{n+\binom{i}{2}+j-i} v_{t-i} v_{t-j}$$

where $\hat{v}_t$ is the predicted value at time t, and c_i are the coefficients to be optimized.

The following algorithm outlines the process to optimize the coefficients of a QRE:

The iterative approach ensures that coefficients are adjusted to closely model the underlying patterns in the time series while minimizing the impact of outliers or abnormal fluctuations. The use of SAD as a loss function enhances the robustness of the model against outliers.

- Overfitting, especially in higher-order models, where the model might capture noise rather than the underlying data pattern.
- Computational complexity increases with the order of the model, requiring more data to validate and stabilize the predictions.

Algorithm 5 General GLDM Model for Optimizing Coefficients

Define model order $n \in \{1, 2, 3\}$
Initialize coefficients $c_1, c_2, \ldots, c_{\frac{n(n+3)}{2}}$
Define maximum iterations M and convergence threshold ϵ
Initialize $SAD = \infty$ (Sum of Absolute Differences)
while $M > 0$ and change in $SAD > \epsilon$ **do**
    for each time point t from $n + 1$ to N **do**
        Calculate predicted value $\hat{v}_t$:
        $\hat{v}_t \leftarrow \sum_{i=1}^{n} c_i \cdot v_{t-i} + \sum_{i=1}^{n} \sum_{j=i}^{n} c_{n+\binom{i}{2}+j-i} \cdot v_{t-i} \cdot v_{t-j}$
    end for
    Calculate new SAD:
    $SAD_{new} \leftarrow \sum_{t=n+1}^{N} |v_t - \hat{v}_t|$
    if $SAD_{new} < SAD$ **then**
        $SAD \leftarrow SAD_{new}$
        Update coefficients $c_1, c_2, \ldots, c_{\frac{n(n+3)}{2}}$ to minimize SAD
    end if
    $M \leftarrow M - 1$
end while
Output the optimized coefficients and final SAD

3.8 ALGORITHMIC APPROACH TO LOSS FUNCTION OPTIMIZATION

The GLDM algorithm implements a systematic process to minimize the loss function, quantifying the prediction error for univariate time series data. The optimization's objective is to find the optimal set of model parameters, denoted $\mathbf{c}$, that minimizes the loss, leading to the most accurate predictions.

The loss function is defined as:

$$L(\mathbf{c}) = \frac{1}{T}\sum_{t=1}^{T}(v_t - \hat{v}_t(\mathbf{c}))^2, \tag{29}$$

where $L(\mathbf{c})$ represents the loss function, v_t are the actual observed values, $\hat{v}_t(\mathbf{c})$ is the predicted value based on the model parameters $\mathbf{c} = \{c_1, c_2, \ldots, c_n\}$, and T is the total number of observations. The model focuses on significant coefficients within $\mathbf{c}$ that have the greatest impact on forecast accuracy. This focus is crucial for simplifying the model and ensuring predictive reliability, especially in epidemiological studies.

The steps of the GLDM loss function optimization algorithm are outlined in Algorithm 6.

Algorithm 6 GLDM Loss Function Optimization

1: Initialize model parameters $\mathbf{c} = \{c_1, c_2, \ldots, c_n\}$ with small, random values.
2: Compute the initial loss $L(\mathbf{c})$ using the squared error.
3: **repeat**
4: Calculate the gradient of the loss function $\nabla_{\mathbf{c}} L(\mathbf{c})$.
5: Update the parameters in the direction of steepest descent: $\mathbf{c} \leftarrow \mathbf{c} - \alpha \nabla_{\mathbf{c}} L(\mathbf{c})$, where α is the learning rate.
6: Recompute the loss $L(\mathbf{c})$ to check for improvement.
7: **until** convergence is achieved, indicated by $\|\nabla_{\mathbf{c}} L(\mathbf{c})\| < \epsilon$ or a maximum number of iterations is reached.
8: **return** the optimized parameters $\mathbf{c}$.

The convergence of the algorithm is typically determined either by the gradient norm falling below a threshold ϵ, indicating that a local minimum has been reached, or by reaching a predefined maximum number of iterations. This iterative refinement is crucial for the GLDM model's ability to adapt to new data and provide accurate forecasts, particularly for epidemiological trends.

The simulations depicted in Figure 5 were executed using Python version 3.9.

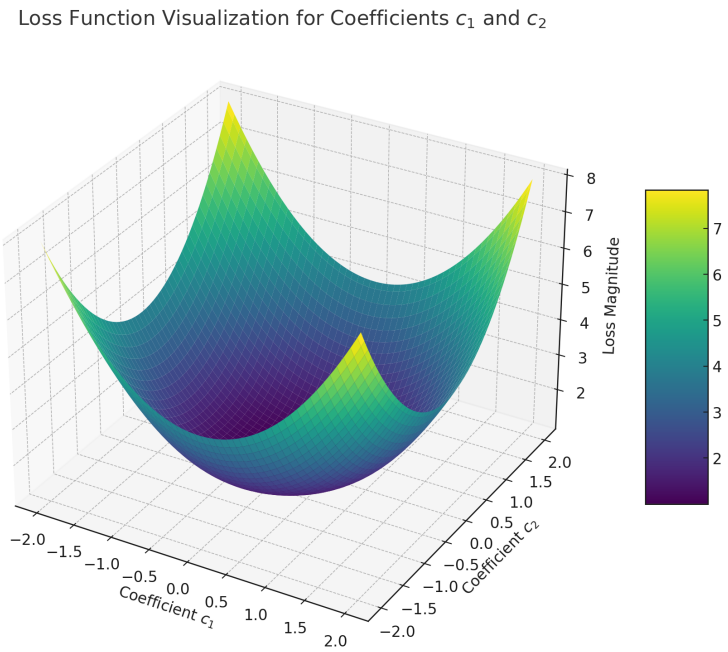

FIGURE 5. A 3D surface plot visualizing the loss function $L(c_1, c_2)$ for a first-order GLDM model, with the z-axis representing the loss magnitude in relation to the coefficient values on the x and y axes.

In the field of time series analysis, particularly within the domain of infectious disease modeling, the Generalized Least Deviation Method (GLDM) serves as a robust approach for developing predictive models. A pivotal aspect of GLDM involves understanding the influence of model coefficients on the accuracy of predictions. The relationship between these coefficients and the predictive error is visualized in Figure 5, which depicts a three-dimensional surface plot for a first-order GLDM model. The plot articulates how the loss magnitude, symbolized by $L(c_1, c_2)$, varies with the coefficients c_1 and c_2.

The graphical representation demonstrates a parabolic surface where the trough signifies the most favorable combination of c_1 and c_2 that minimizes loss, thereby reducing prediction error. In a first-order GLDM model, these coefficients are instrumental in determining the immediate past influence on subsequent forecasts.

Advancing to a second-order GLDM model, the complexity escalates as the number of coefficients increases to five, encompassing c_1 through c_5. This progression allows for a more nuanced depiction of the past two observations' impact, inclusive of their interactive effects, on the forecasted value. The addition of these coefficients escalates the dimensionality of the optimization endeavor, transforming the loss function into a more complex hypersurface within a multidimensional space.

While a direct visualization of a five-dimensional hypersurface is beyond our three-dimensional perception, Figure 5 conceptualizes the loss function $L(c_1, c_2, c_3, c_4, c_5)$ for a second-order model by simulating the impact of altering coefficients on the prediction error. The depth of color on the plot represents the magnitude of the loss, which is defined as:

$$L(\mathbf{c}) = \frac{1}{T}\sum_{t=1}^{T}(v_t - \hat{v}_t(\mathbf{c}))^2,$$

with darker shades symbolizing lower loss values. These regions correspond to the optimal coefficients where the model's predictions most closely align with observed data.

The predictive formula for the second-order GLDM model is:

$$\hat{v}_t = c_1 g_1(v_{t-1}) + c_2 g_2(v_{t-2}) + c_3 g_3(v_{t-1}^2) + c_4 g_4(v_{t-1}v_{t-2}) + c_5 g_5(v_{t-2}^2) + \epsilon_t,$$

where $\hat{v}_t$ predicts the outcome, $\mathbf{c} = \{c_1, c_2, c_3, c_4, c_5\}$ denotes the set of influential coefficients, g_j embodies the functions capturing relationships within the time series, and ϵ_t represents the stochastic error term. Identifying the most effective coefficient combination is fundamental to diminishing prediction error, a concept elegantly illustrated by the contours in the figure.

3.9 ERROR METRICS FOR MODELING DAILY COVID-19 CASES

The task of modeling daily COVID-19 cases requires rigorous mathematical and computational techniques, particularly within the domain of time series analysis. The precision and accuracy of these models are critical for guiding effective data-driven responses to the pandemic. To evaluate the performance of our predictive models, we utilize several key error metrics, each offering insights into various aspects of model accuracy and prediction bias. This section explores essential metrics such as Root Mean Square Error (RMSE), R-Squared (R^2), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Median Absolute Error, Mean Absolute Scaled Error (MASE), and Mean Bias Error (MBE). These metrics provide a robust framework for assessing our models, helping to highlight their strengths and pinpoint areas for improvement.

Understanding and applying these metrics is crucial for refining our predictive algorithms, thereby enhancing the accuracy of our daily COVID-19 case predictions. This effort highlights the vital role of integrating mathematics and computer science to tackle complex epidemiological challenges through advanced modeling techniques.

3.9.1. Root Mean Square Error (RMSE)

The RMSE is an essential metric for quantifying the accuracy of predictions, particularly useful in assessing how closely a model's predictions align with the actual observed numbers. It measures the average magnitude of errors between the predicted daily COVID-19 case counts ($\hat{v}_t$) and the actual reported figures (v_t). A lower RMSE value is indicative of a model that more accurately forecasts daily COVID-19 cases, thereby highlighting the model's predictive quality and reliability. This metric is indispensable for the iterative process of model refinement, aiming to minimize prediction errors and enhance the precision of daily COVID-19 case projections. The formula for calculating RMSE is provided as:

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(v_t - \hat{v}_t)^2} \tag{30}$$

where v_t represents the actual reported daily COVID-19 cases, $\hat{v}_t$ denotes the cases predicted by the model, and T is the total count of observations used in the model.

3.9.2. R-Squared (R^2)

The R^2 metric quantifies the fit of our predictive model to the observed daily COVID-19 case data. It calculates the proportion of variance in daily case numbers that can be predicted from our model, thereby offering insight into its explanatory power. The formula for R^2 is as follows:

$$R^2 = 1 - \frac{\sum_{t=1}^{T}(v_t - \hat{v}_t)^2}{\sum_{t=1}^{T}(v_t - \bar{v})^2} \tag{31}$$

where v_t represents the actual daily case numbers, $\hat{v}_t$ denotes the predicted case numbers by our model, $\bar{v}$ is the average of observed cases, and T is the total number of observed days. This metric is crucial for assessing the reliability of our predictions, with values closer to 1 indicating a model that accurately reflects observed trends in COVID-19 case numbers.

3.9.3. Mean Absolute Percentage Error (MAPE)

MAPE offers a standardized measure of prediction error expressed as a percentage, facilitating an intuitive understanding of model accuracy in forecasting daily COVID-19 cases. This metric is particularly valuable for comparing the performance of different models or for assessing improvement in model accuracy over time. By expressing errors in percentage terms, MAPE allows for a relative error comparison across datasets of varying scales. The formula for calculating MAPE, which reflects the average magnitude of errors between predicted and actual case numbers as a proportion of actual values, is given by:

$$\text{MAPE} = \frac{100\%}{T}\sum_{t=1}^{T}\left|\frac{v_t - \hat{v}_t}{v_t}\right| \tag{32}$$

where v_t represents the actual reported number of daily COVID-19 cases, $\hat{v}_t$ denotes the predicted case numbers, and T is the total number of observations. A lower MAPE value indicates a model with higher predictive accuracy, highlighting its effectiveness in closely estimating the real-world occurrence of daily COVID-19 cases.

3.9.4. Mean Absolute Error (MAE)

MAE is a crucial metric for evaluating the accuracy of predictions related to daily COVID-19 case numbers. It quantifies the average error magnitude across all predictions, disregarding the direction of these errors. This simplicity makes MAE particularly useful for understanding the general prediction error scale in forecasting daily COVID-19 cases, as it provides a direct average of absolute errors. The formula for MAE is given as follows, where lower values signify more accurate predictions:

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}|v_t - \hat{v}_t| \tag{33}$$

In this formula, v_t denotes the actual reported number of daily COVID-19 cases, $\hat{v}_t$ represents the predicted cases for the same day, and T is the total number of observed days. An effective model for predicting daily COVID-19 cases aims to minimize the MAE, reflecting closer alignment between the model's predictions and the observed case data.

3.9.5. Mean Error (ME)

ME provides a straightforward measure of the average bias in predictions of daily COVID-19 case numbers. Unlike MAE or MSE, ME takes into account the direction of the prediction errors, thus indicating whether the model tends to overestimate or underestimate the actual case counts. This metric is crucial for identifying systematic bias in predictive models, ensuring that forecasts neither consistently overshoot nor undershoot the real data. The formula for calculating ME is:

$$\text{ME} = \frac{1}{T}\sum_{t=1}^{T}(v_t - \hat{v}_t) \tag{34}$$

Here, v_t represents the actual reported number of daily COVID-19 cases, $\hat{v}_t$ is the number of cases predicted by the model for the corresponding day, and T is the total number of days included in the analysis. Positive ME values indicate an average overestimation of cases by the model, whereas negative values point to underestimation. The goal is to adjust the model to achieve an ME as close to zero as possible, indicating unbiased predictions.

4. RESULTS

The application of Generalized Least Deviation Method (GLDM) models to analyze COVID-19 infection and mortality data has provided invaluable insights into the dynamic nature of the pandemic across various regions. This section provides a focused analysis on the Samara Region and Russia, shedding light on the predictive accuracy and utility of GLDM models in these specific contexts.

Table 1 summarizes the length of the data collection periods for the Samara Region and Russia, which are critical for the models' analyses.

Table 1. Data collection length for the Samara Region and Russia.

No.	Region	Length
1	Samara Region	1003
2	Russia	882

Table 2 outlines the GLDM model's First-order coefficients for the Samara Region, with c_1 indicating a primary positive effect, and c_2 a represent minor adjustments within the model.

Table 2. First order GLDM Model Coefficients for COVID-19 Infection Cases in Samara Region

Coefficient	Value
c_1	1.0071
c_2	-1.2486 $\times 10^{-5}$

Table 3 outlines the GLDM model's second-order coefficients for the Samara Region, with c_1 indicating a primary positive effect, and c_2 a significant negative effect on the infection trend. The coefficients c_3 to c_5 represent minor adjustments within the model.

Table 3. Second Order GLDM Model Coefficients for COVID-19 Infection Cases in Samara Region

Coefficient	Value
c_1	1.2573
c_2	-0.2455
c_3	0.0002
c_4	0.0001
c_5	-0.0003

Figures 6 and 7 compare the actual COVID-19 infection data with first and second-order GLDM model projections for the Samara region, illustrating the models' close fit to the actual infection trends.

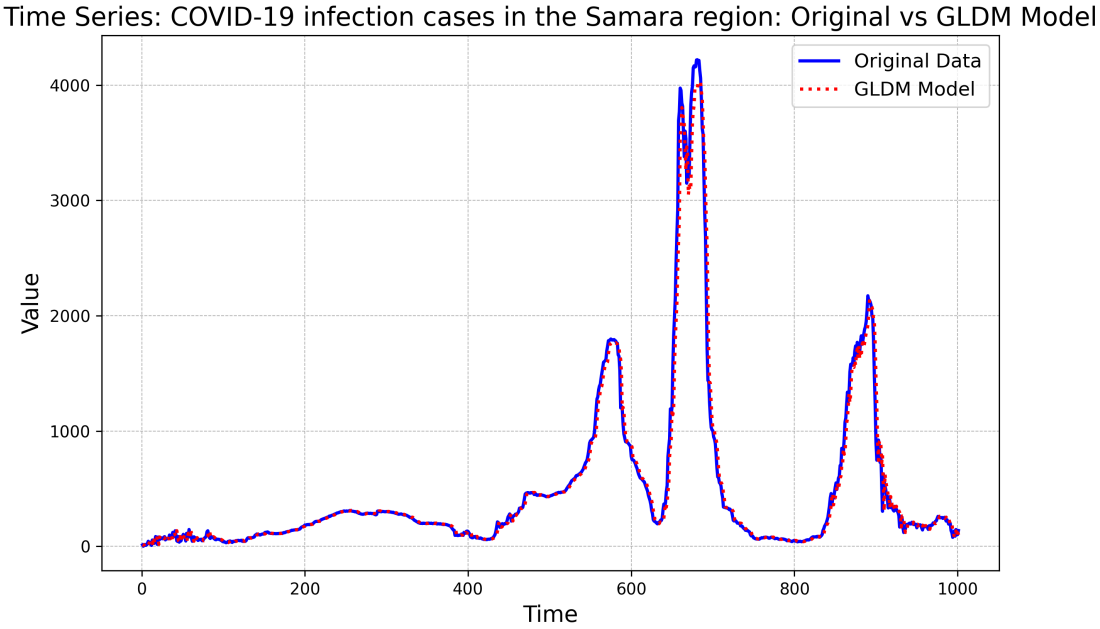

FIGURE 6. Comparing original and GLDM Model first-order predictions for COVID-19 cases in the Samara region: Time Series Analysis

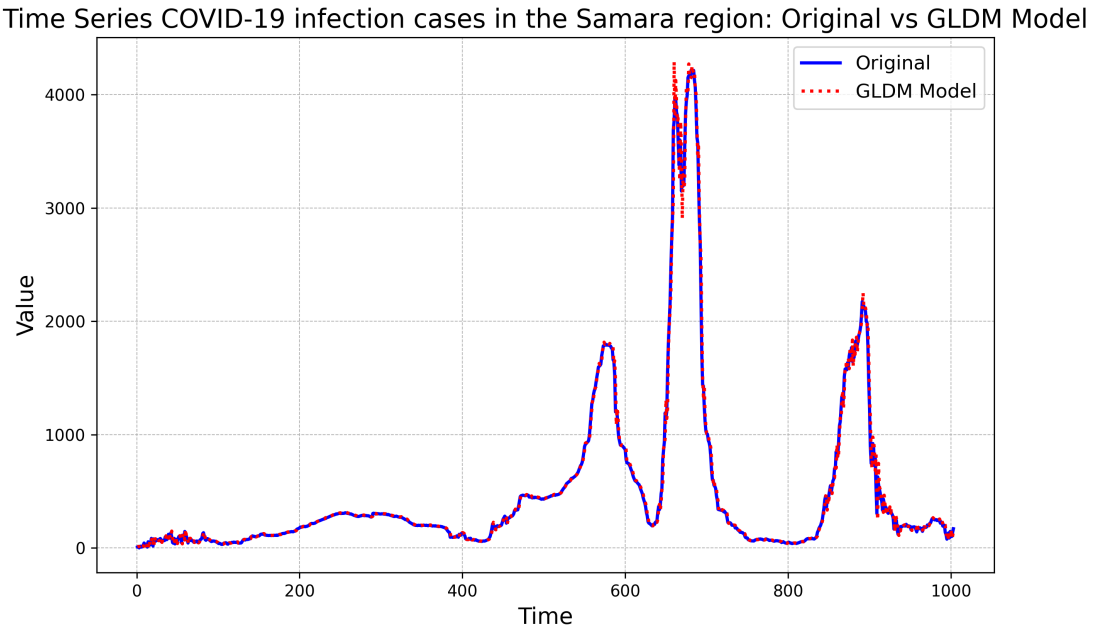

FIGURE 7. Comparing original and GLDM Model second-order predictions for COVID-19 cases in the Samara region: Time Series Analysis

Table 4 displays the first-order GLDM model coefficients for death cases in Russia, where c_1 equals 1, indicating a direct correlation with previous values, and c_2 is 0, showing no additional effect.

Table 4. **First Order GLDM Model Coefficients for Death Cases in Russia**

Coefficient	Value
c_1	1.0000
c_2	0.0000

Table 5 presents coefficients for a second-order GLDM model, with c_1 suggesting a significant effect on the death cases' trend, and c_2 showing additional influence. The remaining coefficients indicate minor adjustments to the model's predictions for death cases in Russia.

Table 5. **Second Order GLDM Model Coefficients for Death Cases in Russia**

Coefficient	Value
c_1	0.7265
c_2	0.2610
c_3	0.0020
c_4	0.0016
c_5	-0.0036

Table 6 lists the third-order GLDM coefficients, with their values shaping the death case trend in Russia.

Table 6. **Third Order GLDM Model Coefficients for Death Cases in Russia**

Coefficient	Value
c_1	0.5970
c_2	-0.3694
c_3	0.7396
c_4	0.0083
c_5	0.0101
c_6	-0.0009
c_7	-0.0185
c_8	0.0010
c_9	0.0000

Figures 8, 9, and 10 compare original COVID-19 death case data with first, second, and third-order GLDM model predictions for Russia, illustrating the model's fit.

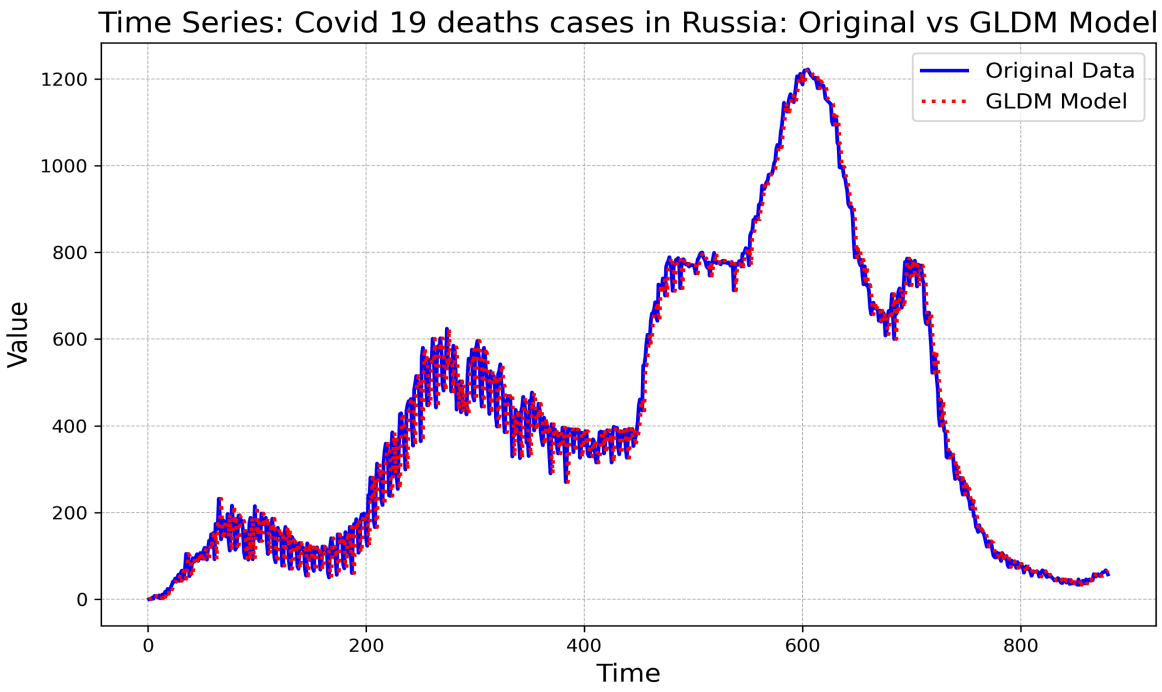

FIGURE 8. **Time Series: COVID-19 death cases in Russia with the GLDM Model (first order)**

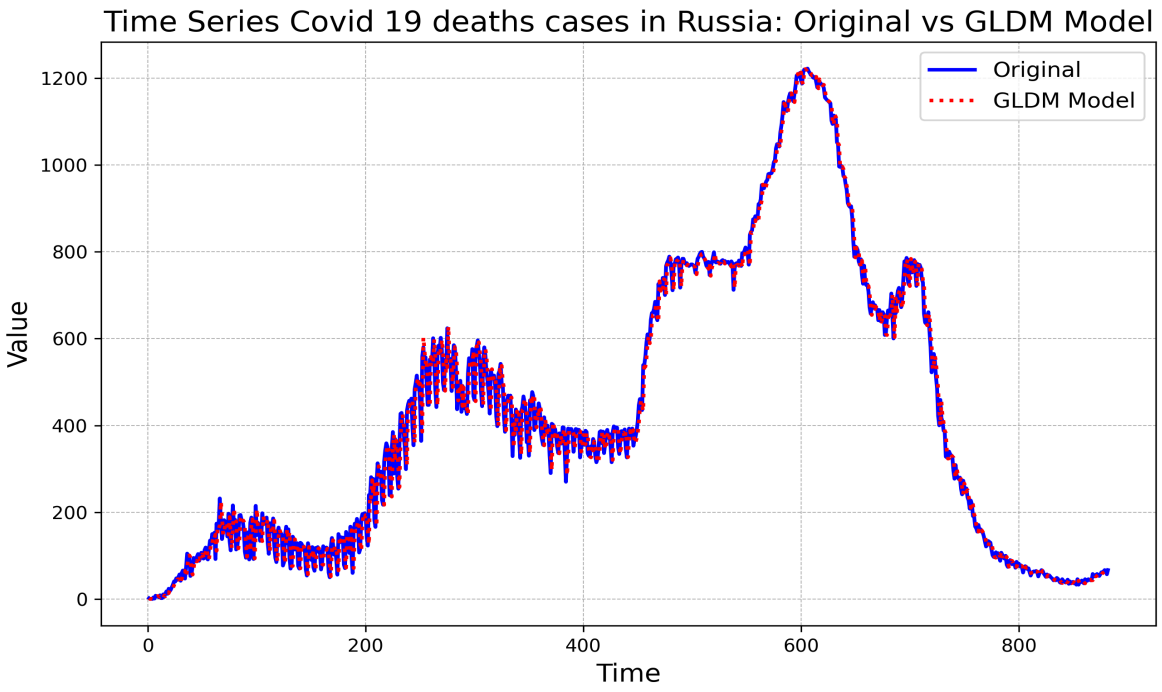

FIGURE 9. **Time Series: COVID-19 death cases in Russia with the GLDM Model (Second order)**

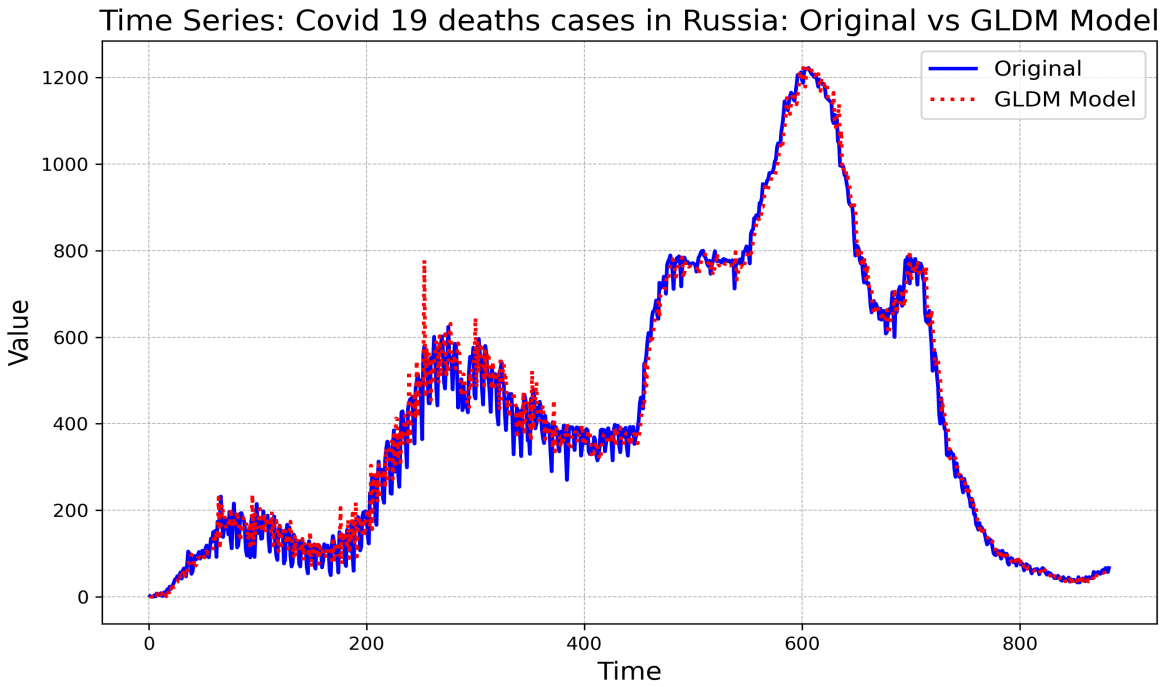

FIGURE 10. Time Series: COVID-19 death cases in Russia with the GLDM Model (third order)

The application of the Generalized Least Deviation Method (GLDM) models to COVID-19 infection cases in the Samara Region and mortality cases in Russia provides insightful analysis into the patterns of the pandemic's spread and impact. For the Samara Region, the performance metrics of the GLDM models reveal significant insights: the first-order model reported a Root Mean Square Error (RMSE) of 69.81, while the second-order model demonstrated improved accuracy with a reduced RMSE of 58.20, reflecting its enhanced capability in capturing the transmission dynamics of the virus. These findings are elaborated in Table 7

In the case of Russia's mortality data, both first and second-order GLDM models yielded close RMSE values of 33.78 and 33.31, respectively, suggesting their effectiveness in modeling death cases. However, the third-order model's higher RMSE of 41.43 may indicate overfitting, revealing that increased model complexity does not necessarily enhance forecast accuracy.

Comparative figures of model predictions against actual data for the Samara Region and Russia illustrate the GLDM models' remarkable fidelity in mirroring real-world occurrences. Such analytical depth affirms the critical role of these models in navigating the pandemic, providing a solid foundation for informed decision-making in public health strategy and resource distribution.

Detailed within Table 7 are the models' performance metrics, offering a nuanced view of their effectiveness through various statistical measures, including RMSE, R-squared, and others, for both the Samara Region and Russia.

Table 7. Error Matrix for COVID-19 Infection Death Cases in Russia and Specific Regions

Region	Order	RMSE	R-squared	MAPE	MAE	MSE	ME
Samara Region	First	69.81	0.9927	9.18	27.99	4872.80	6.99
	Second	58.20	0.9943	9.36	23.68	3387.41	-3.48
Russia	First	33.78	0.9896	11.12	22.24	1141.25	0.074
	Second	33.31	0.9898	10.96	22.22	1109.29	0.43
	Third	41.43	0.9843	13.47	29.37	1716.66	-2.79

Utilizing the Generalized Least Deviation Method (GLDM), Figures 12 and 11 display the fidelity of COVID-19 case and death rate modeling for the Samara Region and Russia, respectively. The radar diagrams underscore the method's

robustness in delineating the pandemic's patterns. Importantly, Figure 13 details the Sum of Absolute Differences (SAD), a loss function metric, to compare model orders. It reveals the superior accuracy of the second-order GLDM model, which registers a lower SAD value, denoting a more precise fit for the observed data in both regions. This suite of figures collectively accentuates the effectiveness of the second-order GLDM model, particularly in its loss function minimization, affirming its enhanced predictive quality across various epidemiological contexts.

In Figures 11 and 12, radar charts are presented, depicting the performance metrics of various algorithms. It is observed that the second order of GLDM outperforms the first and third orders for forecasting death cases in Russia. Additionally, in the context of COVID-19 infection cases in Samara, the second order exhibits superior performance over the first order.

FIGURE 11. Radar Diagrams for Goodness of Fit of GLDM Models for deaths COVID-19 Cases in Russia

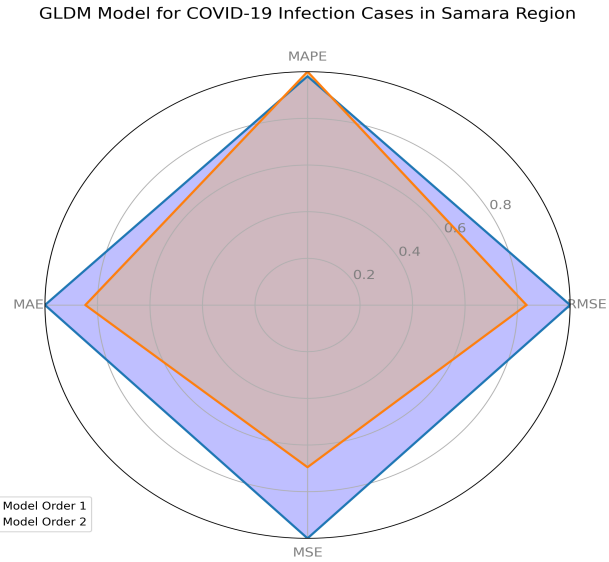

FIGURE 12. Radar Diagrams for Goodness of Fit of GLDM Models for COVID-19 Cases in Samara Region

A summary analysis of the Sum of Absolute Differences (SAD) for the Samara Region with respect to COVID-19 infection cases and for Russia with respect to COVID-19 death cases by model order is presented in Table 8. This analysis provides insight into the performance of different GLDM model orders. In the Samara Region, the Second Order model shows a lower SAD compared to the First Order model, suggesting a higher accuracy in modeling the infection trend. Conversely, in Russia, the Second Order model exhibits only a slight improvement in SAD over the First Order model, and the Third Order model displays a higher SAD, indicating that more complex models do not always yield better predictive performance. The table quantifies the accuracy of the GLDM models, allowing for a straightforward comparison of their effectiveness in different contexts.

Table 8. Sum of Absolute Differences for various regions by model order.

Region	Model Order	Sum of Absolute Differences
Samara Region	First Order	28044.96
	Second Order	23704.61
Russia	First Order	19591.00
	Second Order	19554.24
	Third Order	23238.35

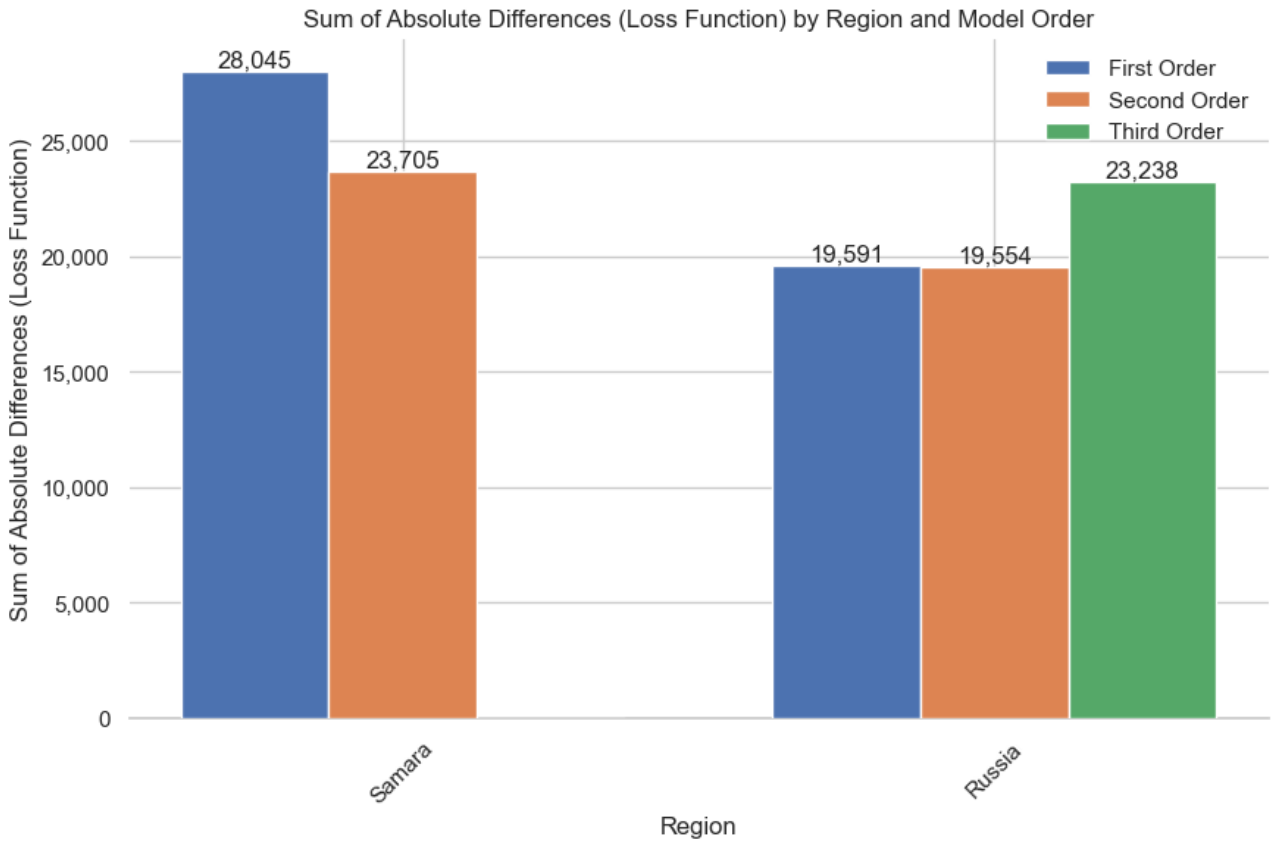

FIGURE 13. Sum of Absolute Differences (Loss Function) by Region and Model Order

5. COVID-19 QUASILINEAR RECURRENCE EQUATIONS

In the context of GLDM models, the order of the model corresponds to the number of previous time steps (*lags*) considered to predict the current value. The coefficients c_i are parameters that quantify the influence of each term in the model.

The first-order model takes into account only the immediate previous value to predict the current value. The equation is:

$$\hat{v}_t = c_1 \times v_{t-1} + (c_2 \times v_{t-1}^2) \tag{35}$$

where $\hat{v}_t$ is the predicted value at time t, v_{t-1} is the actual value at time $t-1$, c_1 is the coefficient for the previous value, and c_2 is the constant term or intercept of the model.

The second-order model includes not only the immediate previous value but also the value from two time steps ago. Additionally, it may consider interactions and non-linear effects of these values:

$$\begin{aligned} \hat{v}_t = {} & (c_1 \times v_{t-1}) + (c_2 \times v_{t-2}) \\ & + (c_3 \times v_{t-1}^2) + (c_4 \times v_{t-1} \cdot v_{t-2}) \\ & + (c_5 \times v_{t-2}^2) \end{aligned} \tag{36}$$

Here, c_3 represents the coefficient for the squared term of the previous value, c_4 is the coefficient for the interaction term, and c_5 is the coefficient for the squared term of the value from two time steps ago.

Extending to the third-order model, we include yet another time step back and additional interaction terms, enhancing the model's complexity and potential accuracy:

$$\begin{aligned}\hat{v}_t = &(c_1 \times v_{t-1}) + (c_2 \times v_{t-2}) + (c_3 \times v_{t-3}) \\ &+ (c_4 \times v_{t-1}^2) + (c_5 \times v_{t-1} \cdot v_{t-2}) + (c_6 \times v_{t-2}^2) \\ &+ (c_7 \times v_{t-1} \cdot v_{t-3}) + (c_8 \times v_{t-2} \cdot v_{t-3}) \\ &+ (c_9 \times v_{t-3}^2)\end{aligned} \tag{37}$$

This model includes terms up to the third previous time step and their interactions. For example, c_7 measures the interaction between the first and third previous values, while c_9 is the coefficient for the square of the third previous value.

Each of these models increases in complexity as the order rises, which may improve prediction accuracy but also raises the risk of overfitting, especially when the number of parameters becomes large relative to the amount of available data.

5.1 SAMARA REGION

The coefficients in the second-order GLDM equation are crucial for understanding the influence of past COVID-19 infection cases on future predictions. In our model for the Samara Region:

$$\hat{v}_t = (1.2573 \times v_{t-1}) + (-0.2455 \times v_{t-2}) + (0.0002 \times v_{t-1}^2) + (0.0001 \times v_{t-1} \cdot v_{t-2}) + (-0.0003 \times v_{t-2}^2), \tag{38}$$

the coefficient $c_1 = 1.2573$ is positive, indicating a direct relationship between the previous day's infection cases v_{t-1} and the predicted value $\hat{v}_t$. This suggests that an increase in cases from the previous day contributes positively to the forecast for the current day.

Conversely, the coefficient $c_2 = -0.2455$ is negative, which implies an inverse relationship for the infection cases two days prior v_{t-2}. A higher number of cases two days ago is associated with a lower prediction value for today, after controlling for the effect of the previous day. This may reflect factors such as interventions that were put in place after a spike in cases, or natural fluctuations in the spread of the virus.

The additional coefficients c_3 to c_5 further refine the model by capturing minor adjustments to the infection trend based on interactions and squared terms of the data from one and two days ago. This reflects the complexity of the disease transmission dynamics and allows the model to account for subtle changes in the data over time.

5.2 RUSSIA

In the context of GLDM models for epidemiological data, the coefficients signify how changes in past data points influence the forecasted value. For the COVID-19 death cases in Russia, our simplified second-order model uses five significant coefficients to capture the trend:

$$\hat{v}_t = (0.7265 \times v_{t-1}) + (0.2610 \times v_{t-2}) + (0.0020 \times v_{t-1}^2) + (0.0016 \times v_{t-1} \cdot v_{t-2}) + (-0.0036 \times v_{t-2}^2), \tag{39}$$

where $\hat{v}_t$ represents the predicted number of deaths at time t. Here, both coefficients are positive, which indicates that higher numbers of deaths in the past two days (v_{t-1} and v_{t-2}) are associated with a higher predicted number of deaths. The coefficient $c_1 = 0.7265$ suggests a stronger influence of the previous day's death count on the current prediction, while $c_2 = 0.2610$ suggests a slightly less, but still positive, influence from two days prior.

These positive coefficients reflect the persistence of the event; in this case, it could be indicative of continued transmission of the virus or other factors contributing to a sustained number of deaths. The absence of a negative coefficient means that there was no detected inverse relationship in the time frame considered for this model.

6. COMPARATIVE ANALYSIS AND SUPERIOR PERFORMANCE OF GLDM SECOND-ORDER MODEL FOR COVID-19 FORECASTING

The performance of various models for forecasting COVID-19 infection cases in the Samara Region and COVID-19 deaths in the Russian Federation is critically compared in Tables 9 and 10, focusing on two key error metrics: R-Squared (R^2) and Mean Absolute Percentage Error (MAPE). These tables provide a comprehensive overview of the accuracy and reliability of each model, underscoring the superior efficacy of the Generalized Least Deviation Method (GLDM) in comparison to other approaches.

Mean Absolute Percentage Error (MAPE) is a critical metric for evaluating COVID-19 forecasting models due to its mathematical and statistical properties. MAPE is defined as:

$$\text{MAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

where A_t represents the actual value and F_t represents the forecasted value at time t. This metric provides a percentage error, making it easily interpretable and facilitating intuitive understanding for stakeholders.

Statistically, MAPE adjusts for the scale of the data, allowing for consistent evaluation across different regions and time periods, which is particularly important given the heterogeneous nature of COVID-19 case numbers. MAPE's sensitivity to large errors ($|A_t - F_t|$) ensures that significant deviations in model predictions are effectively highlighted, providing a clear indication of model reliability.

By focusing on absolute percentage errors, MAPE mitigates the impact of outliers compared to squared error metrics such as Mean Squared Error (MSE), leading to a more robust assessment. This property is vital for accurate comparison between models and for ensuring that forecasts are reliable for public health planning, resource allocation, and policy-making.

The relative error measurement inherent in MAPE is particularly suited for dynamic and rapidly changing datasets, such as those encountered during the COVID-19 pandemic. Thus, MAPE is indispensable for evaluating the performance of predictive models in capturing the true trends and fluctuations of COVID-19 cases and deaths.

Table 9 details the performance of different models for COVID-19 infection cases in the Samara Region. The GLDM second-order model exhibits exceptional performance, demonstrated by its high R^2 value of 0.9943 and a notably low MAPE of 9.36%. This performance is particularly significant when compared to other models, such as the MLP model, which has an R^2 of 0.0814 and a MAPE of 197.1749%, and the SVM model, with an R^2 of 0.4098 and a MAPE of 12.5884%. The Auto ARIMA model and its hybrid variants also perform well, with R^2 values around 0.9956 and MAPEs close to 10%, but they do not surpass the precision of the GLDM second-order model. Similarly, the Exponential Smoothing model and the BATS and TBATS models, with R^2 values of approximately 0.9940 and MAPEs slightly above 10%, and the Prophet model, which has an R^2 of 0.6345 and a significantly higher MAPE of 179.7204%, also fall short in comparison. This illustrates the robustness of the GLDM second-order model in providing precise and reliable predictions for COVID-19 infection cases in the Samara Region.

The superior performance of the GLDM second-order model can be attributed to several mathematical advantages. Firstly, the model's ability to incorporate higher-order terms and nonlinearity enables it to capture complex, dynamic patterns and dependencies within the data that simpler models may overlook. Specifically, the GLDM second-order model for COVID-19 infection cases in the Samara Region is characterized by coefficients $c_1 = 1.2573$, $c_2 = -0.2455$, $c_3 = 0.0002$, $c_4 = 0.0001$, and $c_5 = -0.0003$. This set of coefficients allows the model to effectively capture the intricate dependencies and interactions between past infection rates. The GLDM's rigorous optimization process, which focuses on minimizing deviations between observed and predicted values, further enhances its predictive accuracy and robustness. Notably, the use of an arctangent-based loss function minimizes the impact of outliers, leading to more stable and reliable forecasts. Additionally, the GLDM's use of quasilinear recurrence equations allows for a more flexible and adaptable modeling approach, which is particularly beneficial for capturing the temporal dependencies and nonlinearities inherent in COVID-19 infection data.

Table 10 details the performance of different models for COVID-19 deaths in the Russian Federation. The GLDM second-order model demonstrates superior performance, evidenced by its high R^2 value of 0.9898 and a notably low MAPE of 10.96%. This performance is remarkable when compared to other models, such as the MLP model, which has an R^2 of 0.0446 and a MAPE of 167.1630%, and the SVM model, which shows an R^2 of 0.9742 and a MAPE of 17.3852%. The Auto ARIMA model and its hybrid variants also perform well, with R^2 values around 0.9917 and MAPEs just above 11%, but they do not surpass the accuracy of the GLDM second-order model. Similarly, the BATS and TBATS models, both with R^2 values of 0.9921 and MAPEs of 11.0584%, and the Prophet model, which has an R^2 of 0.9746 and a significantly higher MAPE of 53.7971%, also fall short in comparison. This illustrates the robustness of the GLDM second-order model in providing precise and reliable predictions for COVID-19 deaths in the region.

The mathematical advantages of the GLDM second-order model extend beyond its ability to incorporate higher-order terms. Its optimization framework, which strategically minimizes a well-defined loss function through the use of the arctangent, enhances the model's ability to adapt to new data and maintain high accuracy. The inclusion of second-order terms is crucial for capturing the complexities and nuances of time series data, making the GLDM particularly effective for modeling the nonlinear and stochastic nature of COVID-19 dynamics. This advanced methodological approach allows the GLDM second-order model to outperform other models, making it a critical tool for public health planning and intervention.

In summary, the GLDM second-order model not only demonstrates superior accuracy and reliability in forecasting COVID-19 infection cases in the Samara Region but also excels in predicting COVID-19 deaths in the Russian Federation,

outperforming a range of other predictive models. Its advanced methodological approach, which incorporates higher-order terms and robust optimization techniques, allows it to capture complex patterns and dependencies with greater precision. These optimized performance metrics make the GLDM second-order model an indispensable tool in the ongoing efforts to model and understand the dynamics of the COVID-19 pandemic, providing critical insights for public health strategies and interventions.

Table 9. **Error Metrics (R-Squared and MAPE) for Various Models for COVID-19 Infection Cases in Samara Region**

Model	R-Squared	MAPE (%)
MLP model	0.0814	197.1749
SVM model	0.4098	12.5884
Auto ARIMA model	0.9956	9.7693
Exponential Smoothing model	0.9946	9.5721
BATS model	0.9940	10.1551
TBATS model	0.9940	10.1551
Prophet model	0.6345	179.7204
Hybrid autoARIMA+ES	0.9955	10.8418
Hybrid autoARIMA+Polynomial	0.9956	9.7832
GLDM Second Order	**0.9943**	**9.3600**

Table 10. **Error Metrics (R-Squared and MAPE) for Various Models for COVID-19 Deaths in Russian Federation**

Model	R-Squared	MAPE (%)
MLP model	0.0446	167.1630
SVM model	0.9742	17.3852
Auto ARIMA model	0.9917	11.0454
Exponential Smoothing model	0.9898	11.1647
BATS model	0.9921	11.0584
TBATS model	0.9921	11.0584
Prophet model	0.9746	53.7971
Hybrid autoARIMA+ES	0.9915	11.2055
Hybrid autoARIMA+Polynomial	0.9917	11.0749
GLDM Second Order	**0.9898**	**10.9600**

7. DISCUSSION

This study presented an advanced algorithm based on the Generalized Least Deviation Method (GLDM), aimed at improving the predictive accuracy of COVID-19 time series data. The findings from the application of this algorithm demonstrate a marked enhancement in forecasting performance, which can be primarily attributed to the optimization of a specially designed loss function and the utilization of second-order model dynamics.

7.1 MATHEMATICAL FORMULATION OF GLDM

The GLDM algorithm is formulated to minimize a loss function $L(c)$ defined as:

$$L(c) = \sum_{t=1}^{T} (v_t - \hat{v}_t(c))^2 ,$$

where v_t represents the observed data at time t, $\hat{v}_t(c)$ is the predicted value, and c are the coefficients optimizing the model.

7.2 INCORPORATION OF SECOND-ORDER DYNAMICS

The model incorporates second-order dynamics through the equation:

$$\hat{v}_t = c_1 f_1(v_{t-1}) + c_2 f_2(v_{t-2}) + \ldots + c_n f_n(v_{t-n}) + \delta_t,$$

where f_i are functions modeling the relationship within the data, capturing complex patterns inherent in the COVID-19 time series data.

8. CONCLUSION

In summary, the second-order GLDM models applied to COVID-19 data from the Samara Region and Russia have provided valuable insights into the dynamics of infection cases and death counts, respectively.

For the Samara Region, the coefficients in the GLDM equation reveal the intricate relationship between past infection cases and future predictions. The positive coefficient $c_1 = 1.2573$ signifies a direct relationship with the previous day's infection cases (v_{t-1}), while the negative coefficient $c_2 = -0.2455$ indicates an inverse relationship with infection cases two days prior (v_{t-2}). Additionally, the supplementary coefficients c_3 to c_5 capture minor adjustments to the infection trend based on interactions and squared terms of the data from one and two days ago. Together, these coefficients refine the model, enabling it to account for subtle changes in the infection data over time.

Similarly, for Russia, the GLDM model for COVID-19 death cases reveals significant coefficients that influence the prediction accuracy. The positive coefficients $c_1 = 0.7265$ and $c_2 = 0.2610$ highlight the persistent influence of the previous day's death count and deaths from two days prior, respectively. Additionally, the coefficients c_3 to c_5 capture additional nuances and adjustments to the death trend based on interactions and squared terms of the data from one and two days ago.

In summary, our application of the Generalized Least Deviation Method (GLDM) to the univariate time series data of COVID-19 has yielded predictive insights with significant accuracy. By carefully calibrating the model to emphasize the most influential coefficients, we have optimized the loss function to achieve an effective balance between model complexity and predictive capability. Our results demonstrate that the simplified second-order model, with its significant coefficients, provides a robust predictive framework for the time series data under study. The minimized loss function, which is central to the efficiency of the GLDM algorithm, indicates that our model's refinement process successfully enhanced its forecasting accuracy. This outcome not only validates the model's application but also reinforces the importance of precision in the selection of model parameters for epidemic tracking.

Overall, the absence of negative coefficients in both models suggests no detected inverse relationships within the considered timeframe, indicating the persistence of the observed trends. These findings underscore the importance of GLDM models in capturing the complex dynamics of infectious disease transmission and death counts, providing valuable insights for epidemic tracking and public health interventions.

9. FUTURE RESEARCH

Future research should extend the scope of current modeling efforts by incorporating multivariate time series data, which could unveil more complex interdependencies and potential causalities within the spread of COVID-19. Further development of the GLDM algorithm to adaptively select significant coefficients could yield a more dynamic model, responsive to shifts in data trends. Additionally, exploring the integration of other loss functions might reveal alternative methods to refine the predictive accuracy of such epidemiological models. Studies could also examine the robustness of the model in other epidemiological scenarios, potentially offering a versatile tool for public health forecasting. Advanced statistical techniques, such as machine learning algorithms, could be employed to assess the GLDM's performance against other predictive models, thus contributing to the broader compendium of epidemiological modeling literature.

ACKNOWLEDGMENT

The study is supported by the Russian Science Foundation regional grant number 23-21-10009.

FUNDING

The study is supported by the Russian Science Foundation regional grant number 23-21-10009.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] R. Pandian, D. Shanthi, and N. Selvaganesh, "An articulate heart attack detection system using mine blast optimization (mbo) based multilayer perceptron neural network (mlpnn) model," *Iraqi Journal For Computer Science and Mathematics*, vol. 4, no. 2, pp. 143–155, 2023.

[2] A. Abdalrada, A. F. Neamah, and H. Murad, "Predicting diabetes disease occurrence using logistic regression: An early detection approach," *Iraqi Journal For Computer Science and Mathematics*, vol. 5, no. 1, pp. 160–167, 2024.

[3] A. J. A. Alshareefi *et al.*, "The early warning of financial failure for iraqi banks based on robust adaptive lasso logistic regression," *Iraqi Journal For Computer Science and Mathematics*, vol. 5, no. 1, pp. 112–124, 2024.

[4] M. A. Hameed, E. T. Yassen, and W. M. Jasim, "Enhancement methods for energy consumption prediction in smart house based on machine learning," *Iraqi Journal For Computer Science and Mathematics*, vol. 4, no. 4, pp. 92–99, 2023.

[5] E. Abbas and B. Al-Sarray, "Particle swarm optimization for penalize cox models in long-term prediction of breast cancer data," *Iraqi Journal For Computer Science and Mathematics*, vol. 4, no. 4, pp. 215–224, 2023.

[6] V. Suganya, S. Selvi, N. Ashokkumar, S. Prema, *et al.*, "An automated lion-butterfly optimization (lbo) based stacking ensemble learning classification (selc) model for lung cancer detection," *Iraqi Journal For Computer Science and Mathematics*, vol. 4, no. 3, pp. 87–100, 2023.

[7] G. S. Mohammed, S. Al-Jamabi, and T. Abbas, "Main challenges (generation and returned energy) in a deep intelligent analysis technique for renewable energy applications," *Iraqi J. Comput. Sci. Math.*, vol. 4, no. 3, pp. 34–47, 2023.

[8] M. A. Kadhim and A. M. Radhi, "Heart disease classification using optimized machine learning algorithms," *Iraqi Journal For Computer Science and Mathematics*, vol. 4, no. 2, pp. 31–42, 2023.

[9] N. Selvaganesh, D. Shanthi, and R. Pandian, "A novel biased probability neural network (bpnn) and regularized extreme learning machine (relm) based hearing loss prediction system," *Iraqi Journal For Computer Science and Mathematics*, vol. 4, no. 2, pp. 56–71, 2023.

[10] R. Qamar, "Gradient techniques to predict distributed denial-of-service attack," *Iraqi Journal For Computer Science and Mathematics*, vol. 3, no. 2, pp. 55–71, 2022.

[11] Z. R. Mohsin, "Investigating the use of an adaptive neuro-fuzzy inference system in software development effort estimation," *Iraqi Journal For Computer Science and Mathematics*, vol. 2, no. 2, pp. 18–24, 2021.

[12] M. S. I. Alsumaidaie, K. M. A. Alheeti, and A. K. Alaloosy, "Intelligent detection of distributed denial of service attacks: A supervised machine learning and ensemble approach," *Iraqi Journal for Computer Science and Mathematics*, vol. 4, no. 3, pp. 12–24, 2023.

[13] S. Packeer and D. Kannangara, "Detection of pedophilia content online: A case study using telegram," *Iraqi Journal For Computer Science and Mathematics*, vol. 3, no. 2, pp. 72–77, 2022.

[14] A. Y. Fahad, A. H. Battal, and A. Yaseen, "Estimating long-run elasticity between crude oil consumption, real oil price, and real gdp in global markets," *Iraqi Journal for Computer Science and Mathematics*, vol. 4, no. 2, pp. 109–117, 2023.

[15] M. Mijwil, A. K. Faieq, and M. Aljanabi, "Early detection of cardiovascular disease utilizing machine learning techniques: Evaluating the predictive capabilities of seven algorithms," *Iraqi Journal For Computer Science and Mathematics*, vol. 5, no. 1, pp. 263–276, 2024.

[16] M. Mijwil, M. Aljanabi, *et al.*, "Towards artificial intelligence-based cybersecurity: The practices and chatgpt generated ways to combat cybercrime," *Iraqi Journal For Computer Science and Mathematics*, vol. 4, no. 1, pp. 65–70, 2023.

[17] M. Mijwil, I. E. Salem, and M. M. Ismaeel, "The significance of machine learning and deep learning techniques in cybersecurity: A comprehensive review," *Iraqi Journal For Computer Science and Mathematics*, vol. 4, no. 1, pp. 87–101, 2023.

[18] K. Arora, M. M. Mijwil, *et al.*, "Novel energy optimized ldpc codes for next-generation mimo ofdm systems," *Iraqi Journal For Computer Science and Mathematics*, vol. 4, no. 4, pp. 1–12, 2023.

[19] K. Aggarwal, M. M. Mijwil, A.-H. Al-Mistarehi, S. Alomari, M. Gök, A. M. Z. Alaabdin, S. H. Abdulrhman, *et al.*, "Has the future started? the current growth of artificial intelligence, machine learning, and deep learning," *Iraqi Journal for Computer Science and Mathematics*, vol. 3, no. 1, pp. 115–123, 2022.

[20] F. Petropoulos and S. Makridakis, "Forecasting the novel coronavirus covid-19," *PloS one*, vol. 15, no. 3, p. e0231236, 2020.

[21] G. Grasselli, A. Pesenti, and M. Cecconi, "Critical care utilization for the covid-19 outbreak in lombardy, italy: early experience and forecast during an emergency response," *Jama*, vol. 323, no. 16, pp. 1545–1546, 2020.

[22] G. Grasselli, A. Zangrillo, A. Zanella, M. Antonelli, L. Cabrini, A. Castelli, D. Cereda, A. Coluccello, G. Foti, R. Fumagalli, *et al.*, "Baseline characteristics and outcomes of 1591 patients infected with sars-cov-2 admitted to icus of the lombardy region, italy," *Jama*, vol. 323, no. 16, pp. 1574–1581, 2020.

[23] Q. Mehmood, M. Sial, M. Riaz, and N. Shaheen, "Forecasting the production of sugarcane in pakistan for the year 2018-2030, using box-jenkin's methodology.," 2019.

[24] R. Jamil, "Hydroelectricity consumption forecast for pakistan using arima modeling and supply-demand analysis for the year 2030," *Renewable Energy*, vol. 154, pp. 1–10, 2020.

[25] J. J. Selvaraj, V. Arunachalam, K. V. Coronado-Franco, L. V. Romero-Orjuela, and Y. N. Ramírez-Yara, "Time-series modeling of fishery landings in the colombian pacific ocean using an arima model," *Regional Studies in Marine Science*, vol. 39, p. 101477, 2020.

[26] M. Wang, "Short-term forecast of pig price index on an agricultural internet platform," *Agribusiness*, vol. 35, no. 3, pp. 492–497, 2019.

[27] F. Petropoulos, E. Spiliotis, and A. Panagiotelis, "Model combinations through revised base rates," *International Journal of Forecasting*, vol. 39, no. 3, pp. 1477–1492, 2023.

[28] P. Wegmüller and C. Glocker, "Us weekly economic index: Replication and extension," *Journal of Applied Econometrics*, vol. 38, no. 6, pp. 977–985, 2023.

[29] T. Kufel, "Arima-based forecasting of the dynamics of confirmed covid-19 cases for selected european countries," *Equilibrium. Quarterly Journal of Economics and Economic Policy*, vol. 15, no. 2, pp. 181–204, 2020.

[30] A. Guizzardi, F. M. E. Pons, G. Angelini, and E. Ranieri, "Big data from dynamic pricing: A smart approach to tourism demand forecasting," *International Journal of Forecasting*, vol. 37, no. 3, pp. 1049–1060, 2021.

[31] J. R. García, M. Pacce, T. Rodrigo, P. R. de Aguirre, and C. A. Ulloa, "Measuring and forecasting retail trade in real time using card transactional data," *International Journal of Forecasting*, vol. 37, no. 3, pp. 1235–1246, 2021.

[32] C. Katris and M. G. Kavussanos, "Time series forecasting methods for the baltic dry index," *Journal of Forecasting*, vol. 40, no. 8, pp. 1540–1565, 2021.

[33] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "Deepar: Probabilistic forecasting with autoregressive recurrent networks," *International journal of forecasting*, vol. 36, no. 3, pp. 1181–1191, 2020.

[34] J. Huber and H. Stuckenschmidt, "Daily retail demand forecasting using machine learning with emphasis on calendric special days," *International Journal of Forecasting*, vol. 36, no. 4, pp. 1420–1438, 2020.

[35] Y. Li, H. Bu, J. Li, and J. Wu, "The role of text-extracted investor sentiment in chinese stock price prediction with the enhancement of deep learning," *International Journal of Forecasting*, vol. 36, no. 4, pp. 1541–1562, 2020.

[36] S. Smyl and N. G. Hua, "Machine learning methods for gefcom2017 probabilistic load forecasting," *International Journal of Forecasting*, vol. 35, no. 4, pp. 1424–1431, 2019.

[37] W. Chen, H. Xu, L. Jia, and Y. Gao, "Machine learning model for bitcoin exchange rate prediction using economic and technology determinants," *International Journal of Forecasting*, vol. 37, no. 1, pp. 28–43, 2021.

[38] H. Calvo-Pardo, T. Mancini, and J. Olmo, "Granger causality detection in high-dimensional systems using feedforward neural networks," *International Journal of Forecasting*, vol. 37, no. 2, pp. 920–940, 2021.

[39] A. Panyukov and A. Tyrsin, "350. stable parametric identification of vibratory diagnostics objects.," *Journal of Vibroengineering*, vol. 10, no. 2, 2008.
[40] A. Tyrsin, "Robust construction of regression models based on the generalized least absolute deviations method," *Journal of Mathematical Sciences*, vol. 139, pp. 6634–6642, 2006.
[41] T. Makarovskikh and M. Abotaleb, "Comparison between two systems for forecasting covid-19 infected cases," in *Computer Science Protecting Human Society Against Epidemics: First IFIP TC 5 International Conference, ANTICOVID 2021, Virtual Event, June 28–29, 2021, Revised Selected Papers 1*, pp. 107–114, Springer, 2021.
[42] M. Ponce and A. Sandhel, "covid19. analytics: An r package to obtain, analyze and visualize data from the coronavirus disease pandemic," *arXiv preprint arXiv:2009.01091*, 2020.
[43] R. Panchal and B. Kumar, "Forecasting industrial electric power consumption using regression based predictive model," in *Recent Trends in Communication and Electronics*, pp. 135–139, CRC Press, 2021.
[44] D. Yakubova, "Econometric models of development and forecasting of black metallurgy of uzbekistan," *Asian Journal of Multidimensional Research (AJMR)*, vol. 8, no. 5, pp. 310–314, 2019.
[45] A. Sulasikin, Y. Nugraha, J. Kanggrawan, and A. L. Suherman, "Forecasting for a data-driven policy using time series methods in handling covid-19 pandemic in jakarta," in *2020 IEEE International Smart Cities Conference (ISC2)*, pp. 1–6, IEEE, 2020.
[46] P. Mishra, M. Abotaleb, K. Karakaya, A. Mostafa, H. Yonar, H. T. A. Blbas, U. H. Rahman, and S. Das, "State of the art in covid-19 in the saarc countries and china using bats, tbats, holt's linear and arima model," *J. Agric. Biol. Appl. Stat*, vol. 1, no. 1, pp. 1–24, 2022.
[47] I. Demir and M. Kirisci, "Forecasting covid-19 disease cases using the sarima-nnar hybrid model," *Universal Journal of Mathematics and Applications*, vol. 5, no. 1, pp. 15–23, 2022.
[48] A. Panyukov, T. Makarovskikh, and M. Abotaleb, "Forecasting with using quasilinear recurrence equation," in *International Conference on Optimization and Applications*, pp. 183–195, Springer, 2022.
[49] T. Makarovskikh, A. Panyukov, and M. Abotaleb, "Using general least deviations method for forecasting of crops yields," in *International Conference on Mathematical Optimization Theory and Operations Research*, pp. 376–390, Springer, 2023.
[50] J. Pan, H. Wang, and Q. Yao, "Weighted least absolute deviations estimation for arma models with infinite variance," *Econometric Theory*, vol. 23, no. 5, pp. 852–879, 2007.
[51] A. V. Panyukov and Y. A. Mezaal, "Stable estimation of autoregressive model parameters with exogenous variables on the basis of the generalized least absolute deviation method," *IFAC-PapersOnLine*, vol. 51, no. 11, pp. 1666–1669, 2018.
[52] A. V. Panyukov and Y. A. Mezaal, "Improving of the identification algorithm for a quasilinear recurrence equation," in *Advances in Optimization and Applications: 11th International Conference, OPTIMA 2020, Moscow, Russia, September 28–October 2, 2020, Revised Selected Papers 11*, pp. 15–26, Springer, 2020.
[53] A. Panyukov, T. Makarovskikh, and M. Abotaleb, "Forecasting with using quasilinear recurrence equation," in *International Conference on Optimization and Applications*, pp. 183–195, Springer, 2022.
[54] A. V. Panyukov *et al.*, "Stable identification of linear autoregressive model with exogenous variables on the basis of the generalized least absolute deviation method," - . : , vol. 11, no. 1, pp. 35–43, 2018.