Ferdowsi University of Mashhad

Iranian Journal of Numerical Analysis and Optimization

Volume 14, Number 2

June 2024

Serial Number: 29

Ferdowsi University of Mashhad, Iran

In the Name of God

Iranian Journal of Numerical Analysis and Optimization (IJNAO)

This journal is authorized under the registration No. 174/853 dated 1386/2/26 (2007/05/16), by the Ministry of Culture and Islamic Guidance.

Volume 14, Number 2, June 2024

ISSN-Print: 2423-6977, **ISSN-Online:** 2423-6969

Publisher: Faculty of Mathematical Sciences, Ferdowsi University of Mashhad

Published by: Ferdowsi University of Mashhad Press

Printing Method: Electronic

Address: Iranian Journal of Numerical Analysis and Optimization

Faculty of Mathematical Sciences, Ferdowsi University of Mashhad

P.O. Box 1159, Mashhad 91775, Iran.

Tel. : +98-51-38806222 , **Fax:** +98-51-38807358

E-mail: ijnao@um.ac.ir

Website: http://ijnao.um.ac.ir

This journal is indexed by:

- SCOPUS
- ZbMATH Open
- ISC
- DOAJ
- SID
- Civilica
- Magiran
- Mendeley
- Academia.edu
- Linkedin

- The Journal granted the International degree by the Iranian Ministry of Science, Research, and Technology.

Iranian Journal of Numerical Analysis and Optimization

Volume 14, Number 2, June 2024

Ferdowsi University of Mashhad - Iran

Iranian Journal of Numerical Analysis and Optimization

This journal is published under the auspices of Ferdowsi University of Mashhad

* Full Professor

** Associate Professor

We would like to acknowledge the help of Miss Narjes khatoon Zohorian in the preparation of this issue.

Letter from the Editor–in–Chief

I would like to welcome you to the Iranian Journal of Numerical Analysis and Optimization (IJNAO). This journal has been published two issues per year and supported by the Faculty of Mathematical Sciences at the Ferdowsi University of Mashhad. The faculty of Mathematical Sciences with the centers of excellence and the research centers is well-known in mathematical communities in Iran.

The main aim of the journal is to facilitate discussions and collaborations between specialists in applied mathematics, especially in the fields of numerical analysis and optimization, in the region and worldwide. Our vision is that scholars from different applied mathematical research disciplines pool their insight, knowledge, and efforts by communicating via this international journal. In order to assure the high quality of the journal, each article is reviewed by subject-qualified referees. Our expectations for IJNAO are as high as any well-known applied mathematical journal in the world. We trust that by publishing quality research and creative work, the possibility of more collaborations between researchers would be provided. We invite all applied mathematicians especially in the fields of numerical analysis and optimization to join us by submitting their original work to the Iranian Journal of Numerical Analysis and Optimization.

We would like to inform all readers that the Iranian Journal of Numerical Analysis and Optimization (IJNAO), has changed its publishing frequency from "Semiannual" to a "Quarterly" journal since January 2023. The four journal issues per year will be published in the months of March, June, September, and December. One of our goals is to continue to improve the speed of both the review and publication processes, while try continuing to publish the best available international research in numerical analysis and optimization, with the high scientific and publication standards that the journal is known for.

Ali R. Soheili

Editor-in-Chief

Contents

Quantum solutions of a nonlinear Schrödinger equation 330
S. Arfaoui and A. Ben Mabrouk

Discontinuous Galerkin approach for two-parametric convection-diffusion equation with discontinuous source term 347
K. R. Ranjan and S. Gowrisankar

High order second derivative multistep collocation methods for ordinary differential equations 367
S. Fazeli

Jarratt and Jarratt-variant families of iterative schemes for scalar and system of nonlinear equations 391
O. Ogbereyivwe, E.J. Atajeromavwo and S.S. Umar

An innovative particle physics optimization algorithm for efficient test case minimization in software testing 417
U. Jaiswal and A. Prajapati

Dynamics of Cholera Pathogen Carriers and Effect of Hygiene Consciousness in Cholera Outbreaks 449
S.F. Abubakar and M.O. Ibrahim

Drug release from a two-layer stent coating considering the viscoelastic property of the arterial wall: A mathematical and numerical study 475
E. Azhdari and A. Emami

Mathematical analysis and forecasting of controlled Spatio-temporal dynamics of the EG.5 Virus 500
E.M. Moumine, O. Balatif and M. Rachik

Optimizing natural gas liquids (NGL) production process: A multi-objective approach for energy-efficient operations using genetic algorithm and artificial neural networks 522
N. Sabeghi, B. Kiani Talaei, R. Ghanbari and K. Ghorbani-Moghadam

A new bi-level data envelopment analysis model to evaluate the Human Development Index 545
E. Hajinezhad, H. Hajinezhad and M.R. Alirezaee

Modeling individual mobility's impact on COVID-19 transmission: Insights from a two-patch SEIR-V approach . . 580
M. Bouziane, M.A. Boubekeur, M.E.B. Keddar and O. Belhamiti

Goursat problem in Hyperbolic partial differential equations with variable coefficients solved by Taylor collocation method 613
F. Birem, A. Boulmerka, H. Laib and C. Hennous

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 2, 2024, pp 330–346
https://doi.org/10.22067/ijnao.2023.84855.1324
https://ijnao.um.ac.ir/

Research Article

Quantum solutions of a nonlinear Schrödinger equation

S. Arfaoui[1,2] and A. Ben Mabrouk[1,3,4,*]

Abstract

In the present paper, we precisely conduct a quantum calculus method for the numerical solutions of PDEs. A nonlinear Schrödinger equation is considered. Instead of the known classical discretization methods based

*Corresponding author

Received 13 October 2023; revised 9 December 2023; accepted 27 December 2023

Sabrine Arfaoui
[1]Laboratory of Algebra, Number Theory and Nonlinear Analysis, Department of Mathematics, Faculty of Sciences, University of Monastir, Avenue of the Environment, 5019 Monastir, Tunisia.
[2]Institut Supérieur d'Informatique du Kef, Université de Jendouba, 5 Rue Saleh Ayech - 7100 Kef, Tunisia. e-mail: sabrine.arfaoui@issatm.rnu.tn

Anouar Ben Mabrouk
[1] Laboratory of Algebra, Number Theory and Nonlinear Analysis, Department of Mathematics, Faculty of Sciences, University of Monastir, Avenue of the Environment, 5019 Monastir, Tunisia.
[3] Department of Mathematics, Higher Institute of Applied Mathematics and Computer Science, University of Kairouan, Street Assad Ibn Alfourat, Kairouan 3100, Tunisia.
[4] Department of Mathematics, Faculty of Sciences, University of Tabuk, King Faisal Road, Tabuk 47512, Saudi Arabia. e-mail: anouar.benmabrouk@fsm.rnu.tn

How to cite this article
Arfaoui, S. and Ben Mabrouk, A., Quantum solutions of a nonlinear Schrödinger equation. *Iran. J. Numer. Anal. Optim.*, 2024; 14(2): 330-346. https://doi.org/10.22067/ijnao.2023.84855.1324

on the finite difference scheme, Adomian method, and third modified versions, we consider a discretization scheme leading to subdomains according to q-calculus and provide an approximate solution due to a specific value of the parameter q. Error estimates show that q-calculus may produce efficient numerical solutions for PDEs. The q-discretization leads effectively to higher orders of convergence provided with faster algorithms. The numerical tests are applied to both propagation and interaction of soliton-type solutions.

AMS subject classifications (2020): Primary 35C08, 35Q55; Secondary 65L80, 81Q05.

Keywords: NLS equation; Quantum calculus; Numerical solution; Error estimates.

1 Introduction

The present paper is devoted essentially to the development of a numerical scheme to approximate the solution of a Nonlinear Schrödinger (NLS) equation in a quantum calculus framework. The aim crosses in fact the restriction to the resolution of such an equation and goes further to show that q-calculus may provide a good framework for numerical solutions of PDEs in general. It is well known that a major literature on the numerical solutions of PDEs is based on classical methods, such as finite difference, finite elements and volumes, Fourier analysis, and recently wavelets. See [7, 5, 9, 10, 11, 12, 16, 19, 28].

The NLS equation is strongly linked to the modeling of real physical phenomena, such as Newton's laws and energy conservation in classical mechanics, the behavior of dynamical systems, the description of a particle in a nonrelativistic setting in quantum mechanics, and so on.Therefore, the NLS equation attracted researchers from both theoretical and applied mathematics and physics. See [16, 19, 22, 28].

Originally, Schrödinger's stated a linear form describing a moving particle according to the model equation

$$\Delta\psi + \frac{8\pi^2 m}{\hbar^2}\,(E - V(x))\,\psi = 0, \tag{1}$$

where ψ is known as a wave function, m is the particle mass, $\hbar$ is the Planck's constant, E is the energy, and V is a potential (see [13, 20, 23, 25, 29]).

Based upon the analogy between mechanics and optics, Schrödinger applied a perturbation method to show an equivalence between his wave function in mechanics and Heisenberg's matrix. This gave rise next to the time dependent model

$$i\hbar\psi_t = -\frac{\hbar^2}{2m}\,\Delta\psi + V(x)\psi - \gamma|\psi|^2\psi \quad \text{in } \mathbb{R}^N, \tag{2}$$

known as the cubic NLS equation. Next, different variants and forms have been developed and investigated by researchers in different fields (see [1, 4, 17, 24, 26]).

The present paper is devoted to the development of a numerical method based on q-calculus to approximate the solution of a reduced NLS equation in $\mathbb{R}$ written on the form

$$\begin{cases} iu_t + \Delta u + f(u) = 0, & (x,t) \in \Omega \times (t_0, +\infty), \\ u(x,t_0) = u_0(x), & x \in \Omega, \\ \dfrac{\partial u}{\partial n}(x,t) = 0, & (x,t) \in \partial\Omega \times (t_0, +\infty). \end{cases} \tag{3}$$

We consider a domain Ω in $\mathbb{R}$, and t_0 is a real parameter fixed as the initial time, u_t is the first order partial derivative in time, Δ is the Laplace operator, and $\dfrac{\partial}{\partial n}$ is the outward normal derivative. Moreover, $\partial\Omega$ is the boundary of Ω. Also, $u = u(x,t)$ and $u_0 = u_0(x)$ are complex valued functions. In addition, f is a nonlinear function of u assumed to be at least continuous.

In [15], the stationary solutions of problem (3) have been studied using direct methods issued from the equation on the whole space. See also [14]. In [6], a Lyapunov–Sylvester method has been applied to solve numerical NLS and Heat equations.

The organization of the present work will be as follows. In section 2, the q-calculus essential tools will be reviewed. Section 3 is devoted to the presentation of our main method. The discrete quantum version of a cubic NLS equation will be developed with the necessary analysis of convergence,

stability, solvability, and consistency. Section 4 is the subject of numerical experimentation due to our theoretical part. We conclude afterward.

2 Quantum calculus toolkit

One of the interesting fields of extensions of real analysis is the so-called q-theory, which provides some discrete and/or some refinement of continuous analysis in subspaces such as $\mathbb{R}_q$, which is composed of the discrete grid $\pm q^n$, $n \in \mathbb{Z}$, $q \in (0,1)$. Recall that for all $x \in \mathbb{R}^*$, there exists a unique $n \in \mathbb{Z}$ such that $q^{n+1} < |x| \leq q^n$, which guarantees some density of the set $\mathbb{R}_q$ in $\mathbb{R}$.

This section aims to introduce some basic concepts of q-theory. We present some definitions, notations, and properties of q-derivatives, which will be useful later. Backgrounds on q-theory may be found in [3, 21] and the references therein.

For $0 < q < 1$, denote

$$\mathbb{R}_q = \{\pm q^n, \ \ n \in \mathbb{Z}\} \ \text{ and } \ \widetilde{\mathbb{R}}_q^+ = \mathbb{R}_q^+ \bigcup \{0\}.$$

We propose in this section to recall two basic functions that are applied almost everywhere in q-theory and its applications. See, for example, [18].

Definition 1. The q-derivative of a function is defined by

$$D_q f(x) = \begin{cases} \dfrac{f(x) - f(qx)}{(1-q)x}, & x \neq 0, \\ f'(0)\ , & else, \end{cases}$$

provided that f is differentiable at 0.

The operator D_q is the q-analogue of the classical derivative, as indeed, if f is differentiable, we get

$$\lim_{q \longrightarrow 1} D_q f(x) = \frac{df(x)}{dx}.$$

Many concepts of derivatives and integration rules have been extended for the case of q-calculus.

The only drawback of the q-calculus is the fact that they remain applied and investigated especially in harmonic functional analysis for the major part

of the literature. A first step ahead has been conducted by Koornwinder and Swarttouw when studying Jackson's third q-Bessel function. Their work motivates researchers to develop different q-differential operators. Recently, q-calculus returns to take place in PDEs. Indeed, consider, for example, an elliptic equation

$$\Delta u + f(u, x) = 0, \tag{4}$$

where f is a suitable function, generally nonlinear in u. We may search for a numerical q-approximation by considering a grid points in $\mathbb{R}_q$ instead of finite difference/finite elements used usually. In q-theory, we already have a q-analog of the Laplace operator expressed as

$$\Delta_q u(x) = \frac{qu(q^{-1}x) - (1+q)u(x) + u(qx)}{x^2}.$$

For $x = q^n$ in $\mathbb{R}_q^+$, we get from (4),

$$u_{n+1} - (1+q)u_n + qu_{n-1} = -(1-q)^2 q q^{2n} f_n,$$

where $u_n = u(q^n)$ and f_n is some discretization of $f(u, x)$. We thus obtain a recursive equation permitting to compute u_n recursively. More about applications of q-calculus in partial differential equations may be found in [3]. A widely known example in q-theory is the Bessel type equation

$$\begin{cases} \Delta_q u(x) = -\lambda^2\, u(x), \\ u(0) = 1,\ u'(0) = 0, \end{cases}$$

($\lambda \in \mathbb{C}$), which has as unique solution a modified q-Bessel function. In the present paper, we will exploit the q-calculus to develop numerical solutions of some PDEs.

3 The discrete NLS equation

In this section, we develop the details of our numerical method. For this aim, we fix $\Omega = [0, 1]$ and $t_0 = 0$. Fix also a time step $l_k = (1-q)q^k$, and for $k \in \mathbb{N}$, we denote t_k the kth instant. For $n \in \mathbb{N}$, we denote $x_n = q^n$, and $h_n = (1-q)q^n$ the nonuniform space step. Denote also $u_n^k = u(t_k, x_n)$ the

net function and U_n^k its numerical approximation (the solution of the discrete problem). We discretize problem (3) as follows,

$$i\frac{U_n^{k+1}-U_n^k}{l_k}+\frac{2q}{1+q}\frac{qU_{n-1}^k-(1+q)U_n^k+U_{n+1}^k}{h_n^2}+f(U_n^k)=0. \tag{5}$$

By setting for $n,k\in\mathbb{N}$,

$$\delta_n^k=\frac{2q}{1+q}\frac{l_k}{h_n^2},\qquad \text{and}\qquad \sigma_n^k=\frac{2q}{1+q}\frac{l_k}{h_n^2}i=\delta_n^k i,$$

the discrete problem (5) becomes

$$U_n^{k+1}=q\sigma_n^kU_{n-1}^k+(1-(1+q)\sigma_n^k)U_n^k+\sigma_n^kU_{n+1}^k+F_n^k, \tag{6}$$

where $F_n^k=il_kf(U_n^k)$. Now, denote for $k\in\mathbb{N}$,

$$U^k=(U_n^k)_{n\in\mathbb{N}},$$

the infinite vector of the numerical solution at the time k. Denote also

$$\beta_n^k=1-(1+q)\sigma_n^k.$$

We get the following dynamical infinite matrix-vector system:

$$U^{k+1}=A_kU^k+F^k, \tag{7}$$

where A_k is the infinite tridiagonal matrix with elements

$$A_k=\begin{pmatrix}
1-\sigma_0^k & \sigma_0^k & 0 & \dots & \dots & \dots & \dots & \dots \\
q\sigma_1^k & \beta_1^k & \sigma_1^k & 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\
0 & q\sigma_2^k & \beta_2^k & \sigma_2^k & 0 & \ddots & \ddots & \ddots & \vdots \\
0 & 0 & q\sigma_3^k & \beta_3^k & \sigma_3^k & 0 & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & q\sigma_n^k & \beta_n^k & \sigma_n^k & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots
\end{pmatrix}.$$

Infinite (especially tridiagonal) matrices are met in many fields and have been applied widely. Finite forms are met in PDEs, such as finite difference methods, in numerical analysis. See, for instance, [2, 8]. In mathematics, infinite

tridiagonal matrices are initially related to the so-called Jacobi operators (see [27]).

Note in problem (7) that a main difference with classical methods, such as the finite difference method, is the possibility to relax one assumption on boundary conditions. We only need such an assumption for one extremity of the domain Ω.

Now, observe that, for each k, we get

$$|1-\sigma_0^k|^2 = 1+|\delta_0^k|^2 > |\delta_0^k|^2 = |\sigma_0^k|^2,$$

and similarly,

$$|\beta_n^k|^2 = 1+(1+q)^2|\delta_n^k|^2 > (1+q)^2|\delta_n^k|^2 = (1+q)^2|\sigma_n^k|^2,$$

which means that the matrix A_k is a dominant-diagonal matrix, which guarantees the solvability of our discrete scheme and leads to the following theorem.

Theorem 1. The numerical problem (7) is uniquely solvable, for $k \geq 2n+1$.

In terms of the classical numerical schemes, such as the finite difference, the assumption $k > n$ replaces the assumption $l = o(h^2)$, where l and h are the time and space steps for the finite difference scheme.

To investigate the stability of the numerical scheme, we propose to apply the Lyapunov criterion for stability, which states that a dynamical system $\mathcal{L}(U_k, U_{k-1}, \dots) = 0$ is stable in the Lyapunov sense if, for any bounded initial solution U_0, the solution U_n remains bounded for all $n \geq 0$ uniformly on n. Here, we will precisely prove the following result.

Lemma 1. The solution U^k is bounded independently of k whenever the initial solution U^0 is bounded.

Proof. We will proceed by recurrence on k. Assume firstly that $\|U^0\| \leq \eta$, for some η positive. It follows from the assumption $k \geq 2n+1$ that

$$|\sigma_n^k| \leq \frac{2q}{(1+q)(1-q)}.$$

Therefore, using system (6), for $k = 0$, we obtain

$$U_n^1 = q\sigma_n^0 U_{n-1}^0 + (1-(1+q)\sigma_n^0)U_n^0 + \sigma_n^0 U_{n+1}^0 + F_n^0.$$

As U^0 is bounded, and the nonlinear function f is continuous, we deduce that U^1 is bounded. So, assume that U^k is bounded independently of k. Using system (6), we get

$$|U_n^{k+1}| \leq \frac{1+3q}{1-q}|U_n^k| + |F_n^k|.$$

Using the recurrence hypothesis and again the continuity of the nonlinear function F, we deduce that

$$|U_n^{k+1}| \leq \frac{2(1+q)}{1-q}C,$$

where $C > 0$ is a constant independent of k.

The consistency of the proposed method is evaluated by means of the local truncation error arising from the discretization scheme. Assuming that the solution u is sufficiently regular, we get the principal part as

$$\mathcal{L}(u)(x,t) = \frac{il_k}{2}u_{tt} + \frac{(1-q)h_n}{3q}u_{xxx} + o(l_k + h_n). \tag{8}$$

It is clearly observable that the truncation operator $\mathcal{L}(u)$ goes to 0 as n, k go to ∞. This yields that the quantum numerical scheme is consistent at a minimum order 1 in time and space.

To finish with the convergence of the numerical method, we apply the Lax–Richtmyer equivalence criterion, which states that for consistent numerical approximations, stability and convergence are equivalent. We thus obtain the following lemma. □

Lemma 2. As the numerical scheme is consistent and stable, it is then convergent.

Indeed, recall here that we have already proved in (8) that the used scheme is consistent. Next, Lemma 1 yields the stability of the scheme. Consequently, the Lax equivalence theorem guarantees the convergence.

4 Numerical implementations

We propose in this experimental part to develop numerical examples to validate the theoretical results developed in the previous sections. We will use an L_2 discrete norm to evaluate the error between the exact solutions and the numerical ones as

$$\|X\|_2 = \Big(\sum_i |X_i|^2\Big)^{1/2},$$

for any vector (series) $X = (X_i)$ eventually in $L^2(\mathbb{C})$. Denote u^k the net function $u(x, t^k)$ and U^k the numerical solution. We propose to compute the discrete error

$$\text{Er} = \max_k \|U^k - u^k\|_2 \tag{9}$$

on the grid (x_n), $n \geq 0$.

We take for the rest $f(u) = |u|^2 u$, which gives the original cubic NLS equation. We next take the classical soliton-type solution

$$u(x,t) = \sqrt{\frac{2a}{q_s}} \exp\Big(i\big(\frac{1}{2}cx - \theta t + \varphi\big)\Big) sech\Big(\sqrt{a}(x - ct) + \phi\Big),$$

where a, q_s, c, $\theta = \dfrac{c^2}{4} - a$, φ and ϕ are some appropriate constants. For t fixed, this function decays exponentially as $|x| \to \infty$. It is a soliton-type disturbance, which travels with speed c and with a-governed amplitude. See [5, 6, 10, 11, 12, 20, 22, 28, 29].

4.1 Propagation of a single soliton

In the first experimentation, we focus on a single-soliton-type particle. The computations are done for $0 \leq x \leq 1$ and $0 \leq t \leq 1$. We fix the q parameter to many different values according to the closeness to 0 or to 1. So, let $q \in \{q_i = \frac{i}{8},\ i = 1, \ldots, 7\}$. We also fix the soliton parameters $a = 0.01$, $q_s = 1$, $c = 0.1$, and the phase parameters $\varphi = \phi = 0$. Figures 1 and 2 illustrate two cases of the numerical solution for the propagation of a single soliton issued from our quantum numerical scheme.

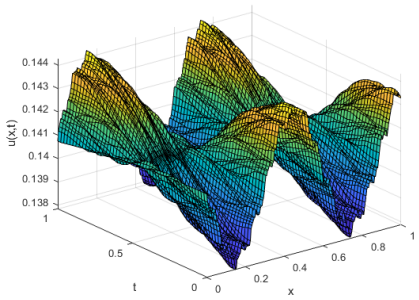


Figure 1: Propagation of a single soliton, for $q = \frac{1}{8}$.

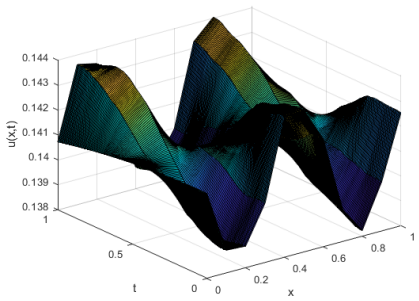


Figure 2: Propagation of a single soliton, for $q = \frac{3}{8}$.

Tables 1 and 2 illustrate the error estimates between the numerical solution and the exact one for different values of the quantum parameter q, and for different values of the maximal time index K. The space grid is fixed to N number of points. Truncating in an order N for practice feasibility of the system (7), we denote TrU_N^k the truncated vector

$$TrU_N^k = [U_0^k, U_1^k, \ldots, U_N^k]^T,$$

and TrF_N^k the truncated vector

$$TrF_N^k = [F_0^k, F_1^k, \ldots, F_N^k]^T,$$

where the upper script ^{T} is for the transpose. We denote similarly, TrA_k^N the truncated matrix

$$TrA_k^N = \begin{pmatrix} 1-\sigma_0^k & \sigma_0^k & 0 & \dots & \dots & \dots \\ q\sigma_1^k & \beta_1^k & \sigma_1^k & 0 & \ddots & \ddots \\ 0 & q\sigma_2^k & \beta_2^k & \sigma_2^k & 0 & \ddots \\ 0 & 0 & q\sigma_3^k & \beta_3^k & \sigma_3^k & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & q\sigma_N^k & \beta_N^k \end{pmatrix}.$$

The system (7) will be approximated by

$$TrU_N^{k+1} = TrA_k^N TrU_N^k + TrF_N^k.$$

Table 1: Error estimates for a single soliton for $N = 20$

				$K = 10$			
q	q_1	q_2	q_3	q_4	q_5	q_6	q_7
Er	$1.17\,10^{-5}$	$1.28\,10^{-5}$	$2.41\,10^{-4}$	$2.52\,10^{-4}$	$2.18\,10^{-4}$	$3.11\,10^{-4}$	$2.87\,10^{-4}$
				$K = 15$			
q	q_1	q_2	q_3	q_4	q_5	q_6	q_7
Er	$2.01\,10^{-6}$	$2.32\,10^{-6}$	$3.01\,10^{-5}$	$2.84\,10^{-5}$	$2.71\,10^{-5}$	$3.01\,10^{-5}$	$2.77\,10^{-5}$
				$K = 20$			
q	q_1	q_2	q_3	q_4	q_5	q_6	q_7
Er	$1.27\,10^{-7}$	$1.44\,10^{-7}$	$1.23\,10^{-6}$	$2.15\,10^{-6}$	$2.53\,10^{-6}$	$2.01\,10^{-5}$	$2.76\,10^{-5}$

Table 2: Error estimates for a single soliton for $N = 50$

				$K = 10$			
q	q_1	q_2	q_3	q_4	q_5	q_6	q_7
Er	$1.17e-5$	$1.28e-5$	$2.41e-4$	$2.52e-4$	$2.18e-4$	$3.11e-4$	$2.87e-4$
				$K = 15$			
q	q_1	q_2	q_3	q_4	q_5	q_6	q_7
Er	$2.01e-6$	$2.32e-6$	$3.01e-5$	$2.84e-5$	$2.71e-5$	$3.01e-5$	$2.77e-5$
				$K = 20$			
q	q_1	q_2	q_3	q_4	q_5	q_6	q_7
Er	$1.27e-7$	$1.44e-7$	$1.23e-6$	$2.15e-6$	$2.53e-6$	$2.01e-5$	$2.76e-5$

It is notable from Tables 1 and 2 that the numerical quantum discrete scheme converges with good error estimates. This encourages to application such a discretization idea for solving more complicated problems.

4.2 Interaction of two solitons

We consider here two solitons traveling with the same speed but in opposite directions in order to obtain the interaction phenomenon. The computations are done as in the previous experimentation on the space domain $\Omega_x = [0, 1]$ and a time space also $I_t = [0, 1]$. We fix the q parameter as previously to $q \in \{q_i = \frac{i}{8},\ i = 1, \ldots, 7\}$. We also fix the soliton parameters as follows:

- For the first soliton, we put $a = 1$, $q_s = 2$, $c = 4$, $\varphi = 0$, and $\phi = 15$.
- For the second soliton, we put $a = 2,25$, $q_s = 2$, $c = -4$, $\varphi = 0$, and $\phi = -7,5$.

As in the previous experimentation, Tables 3 and 4 illustrate the error estimates between the numerical solution and the exact one for different values of the quantum parameter q, and for different values of the maximal time index K. The space grid is fixed to a number N of points.

Table 3: Error estimates for two-interacted solitons for $N = 20$

$K = 10$							
q	q_1	q_2	q_3	q_4	q_5	q_6	q_7
Er	$2.27e-5$	$2.45e-5$	$2.51e-4$	$2.76e-4$	$3.14e-4$	$3.44e-4$	$4.02e-4$
$K = 15$							
q	q_1	q_2	q_3	q_4	q_5	q_6	q_7
Er	$2.11e-6$	$2.18e-6$	$2.05e-5$	$2.61e-5$	$2.85e-5$	$3.12e-5$	$4.05e-5$
$K = 20$							
q	q_1	q_2	q_3	q_4	q_5	q_6	q_7
Er	$1.92e-7$	$2.02e-7$	$2.15e-6$	$2.44e-6$	$3.21e-6$	$3.32e-5$	$3.876e-5$

Table 4: Error estimates for two-interacted solitons for $N = 50$

				$K = 10$			
q	q_1	q_2	q_3	q_4	q_5	q_6	q_7
Er	$125e-5$	$1.34e-5$	$1.44e-4$	$2.21e-4$	$2.43e-4$	$2.31e-4$	$2.15e-4$
				$K = 15$			
q	q_1	q_2	q_3	q_4	q_5	q_6	q_7
Er	$1.15e-6$	$1.27e-6$	$1.32e-5$	$1.17e-5$	$1.95e-5$	$2.02e-5$	$2.33e-5$
				$K = 20$			
q	q_1	q_2	q_3	q_4	q_5	q_6	q_7
Er	$1.16e-7$	$1.31e-7$	$2.13e-6$	$2.23e-6$	$2.27e-6$	$2.47e-5$	$2.55e-5$

As in the previous case, we notice from Tables 3 and 4 that the numerical quantum discretization yielded a very close approximated solution to the exact one. This is clearly shown by the error estimates, where the maximum error is estimated by 10^{-4} over all the values of the quantum parameter q in two Tables 3 and 4. This finding motivates the use of a quantum numerical scheme for more general and/or complicated PDEs.

Remark 1. We know that in the case of classical finite difference and finite element methods, the error is generally estimated in the power of the space step $h = \Delta x$, which in turn is related to the size N of the mesh or, equivalently, the number of the discretization points. In these cases, we obtain an error to the order h^α or equivalently, the order $N^{-\alpha}$, for some exponent α. In the new q-quantum case, we obtain an error to the order of $q^{\alpha N}$, which has the form of a geometric sequence with a ratio q^α in the interval $(0, 1)$, which therefore converges to 0 more rapidly than the previous sequence $N^{-\alpha}$. We then gain a quick and more precise scheme. These facts make it unnecessary to concretely develop the numerical comparisons, as we know in advance that the classical schemes will give higher error, lower speed of convergence, and higher running time. Mathematically, it is easy to see that for $\alpha, \beta > 0$, it holds that for all $\varepsilon > 0$, there exisits $N_0 \in \mathbb{N}$, such that $q^{\beta N} N^\alpha \leq \varepsilon$, for all $N \geq N_0$. This means that, the q-quantum scheme, even at a lower order $\beta < \alpha$ is better than the classical schemes. In other words, the q-quantum scheme does not necessitate calibrating the discretization to get higher orders for convergence, consistency, and good stability, as it is done for the

classical finite difference and finite element methods with implicit method, Crank–Nicolson, centering/decentering method, higher regularity order finite elements bases, and so on.

5 Conclusion

In the present paper, the principal aim was to test the efficiency of the quantum calculus in the approximation of the solutions of PDEs. As a prototypical example, we applied q-calculus to derive a numerical scheme for the well-known cubic NLS equation. As expected, the q-calculus yielded good approximations illustrated by low error estimates. The findings in the present paper make therefore good motivation to continue to exploit quantum calculus for the numerical (and also exact) solutions of different types of PDEs. Comparisons with other models, such as finite difference, finite volumes, and also wavelets as recent developments in mathematical analysis are fascinating and motivating future extensions. Compared to the classical finite difference scheme method, we may conclude theoretically that, in fact, the present quantum scheme is more efficient, as it is based on geometric sequences of time and space steps, which surely converge more rapidly than arithmetic discretizations. Therefore, we expect that involving or including hybrid schemes may induce the best results. Finally, an interesting question raised from the present work may be formulated as follows: Given an infinite matrix that is truncated in an order n. We know that at most in $\mathbb{C}$, any truncation has at most n eigenvalues. What can we expect for the original linear operator defined by means of the infinite matrix? This gives rise to possible chaotic behavior as a future study of the present case of matrices, which are issued from parabolic, hyperbolic PDEs.

Acknowledgment

The authors would like to thank the editor of IJNAO and the anonymous reviewers for their insightful comments and suggestions which helped to improve our work.

References

[1] Ablowitz, M.J., Prinari, B. and Trubatch, A.D. *Discrete and continuous nonlinear Schrödinger systems*, Cambridge Univ. Press, Cambridge, 2004.

[2] Akheizer, N.I. *The classical moment problem and some related questions in analysis,* Oliver & Boyd, 1965.

[3] Annaby, M.H.and Mansour, Z.S. *q-fractional calculus and equations*, Lecture Notes in Mathematics 2056, Editors: J.-M. Morel and B. Teissier, Springer 2012.

[4] Avron, J., Herbst, I. and Simon, B. *Schrödinger operators with electromagnetic fields, III. Atoms in homogeneous magnetic field*, Commun. Math. Phys. 79 (1981), 529–572.

[5] Ben Mabrouk, A. and Ayadi, M. *A linearized finite-difference method for the solution of some mixed concave and convex nonlinear problems*, Appl. Math. Comput. 197 (2008), 1–10.

[6] Ben Mabrouk, A. and Ayadi, M. *Lyapunov type operators for numerical solutions of PDEs*, Appl. Math. Comput. 204 (2008), 395–407.

[7] Ben Mabrouk, A., Ben Mohamed, M.L. and Omrani, K. *Finite difference approximate solutions for a mixed sub-superlinear equation*, Appl. Math. Comput. 187 (2007), 1007–1016.

[8] Berezanskii, Y.M. *Expansion in eigenfunction of self-adjoint operators*, AMS, Providence, RI, 1968.

[9] Bezia, A., Ben Mabrouk, A. and Betina, K. *Lyapunov-Sylvester operators for* $(2+1)$*-Boussinesq equation*, Electron. J. Differ. Equ. 2016 (2016), 1–19.

[10] Bratsos, A.G. *A linearized finite-difference method for the solution of the nonlinear cubic Schrödinger equation*, Comm. Appl. Analysis 4(1) (2000), 133–139.

[11] Bratsos, A.G. *A linearized finite-difference scheme for the numerical solution of the nonlinear cubic Schrödinger equation*, Korean J. Comput. Appl. Math. 8(3) (2001), 459–467.

[12] Bratsos, A.G., Tsituras, Ch. and Natsis, D.G. *Linearized numerical schemes for the Boussinesq equation*, Appl. NUm. Anal. Comp. Math. 2(1) (2005), 34–53.

[13] Byeon, J. and Wang, Z.Q. *Standing waves with a critical frequency for nonlinear Schrödinger equations*, Arch. Ration. Mech. Analysis 165 (2002), 295–316.

[14] Chteoui, R. and Ben Mabrouk, A. *A generalized Lyapunov-Sylvester computational method for numerical solutions of NLS equation with singular potential*, Anal. Theory Appl. 33 (2017), 333–354.

[15] Chteoui, R., Ben Mabrouk, A. and Ounaiess, H. *Existence and properties of radial solutions of a sub-linear elliptic equation*, J. Part. Diff. Eq. 28 (1) (2015), 30–38.

[16] Delfour, M., Fortin, M. and Payre, G. *Finite difference solutions of a non-linear Schrödinger equation*, J. Computa. Phys. 44 (1981), 277–288.

[17] Floer, A. and Weinstein, A. *Nonspreading wave packets for the cubic Schrödinger equation with a bounded potential*, J. Funct. Anal. 69 (1986), 397–408.

[18] Gasper, G. and Rahman, M. *Basic Hypergeometric serie*, Second edition, Cambridge University Press, 2004.

[19] Grillakis, M.G. *On nonlinear Schrödinger equations*, Commun. Partial. Differ. Equations. 25 (2000), 1827–1844.

[20] Hasegawa, A. and Kodama, Y. *Solitons in optical communications*, Academic Press, San Diego, 1995.

[21] Kac, V. and Cheung, P. *Quantum calculus*, Springer, 2002.

[22] Lamb, G.L. *Elements of soliton theory*, Wiley, 1980.

[23] Malomed, B.A. *Variational methods in nonlinear fiber optics and related fields*, Progress in Optics 43 (2002), 69–191.

[24] Oh, Y.G. *Existence of semi classical bound states of nonlinear Schrödinger equations with potentials of the class* (V_a), Commun. Partial Differ. Equ. 13 (1988), 1499–1519.

[25] Onorato, M., Osborne, A.R., Serio, M. and Bertone, S. *Freak waves in random oceanic sea states*, Phys. Rev. Lett. 86 (2001), 5831–5834.

[26] Sulem, C. and Sulem, P.-L. *The nonlinear Schrödinger equation: Self-focusing and wave collapse*, Applied Mathematical Sciences, Vol. 139, Springer-Verlag, New York, 1999.

[27] Teschl, G. *Jacobi operators and completely integrable nonlinear lattices*, Mathematical Surveys and Monographs, 72. American Mathematical Society, Providence, RI, 2000.

[28] Twizell, E.H., Bratsos, A.G. and Newby, J.C. *A finite-difference method for solving the cubic Schrödinger equation*, Computation of nonlinear phenomena. Math. Comput. Simulation 43 (1997), 67–75.

[29] Zakharov, V.E. *Collapse and self-focusing of Langmuir waves*, Handbook of Plasma Physics, (M. N. Rosenbluth and R. Z. Sagdeev, eds.), vol. 2 (A. A. Galeev and R. N. Sudan, eds.) 81–121, Elsevier, 1984.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 2, 2024, pp 347–366
https://doi.org/10.22067/ijnao.2024.84456.1316
https://ijnao.um.ac.ir/

Research Article

Discontinuous Galerkin approach for two-parametric convection-diffusion equation with discontinuous source term

K. R. Ranjan*, and S. Gowrisankar

Abstract

In this article, we explore the discontinuous Galerkin finite element method for two-parametric singularly perturbed convection-diffusion problems with a discontinuous source term. Due to the discontinuity in the source term, the problem typically shows a weak interior layer. Also, the presence of multiple perturbation parameters in the problem causes boundary layers on both sides of the boundary. In this work, we develop the nonsymmetric discontinuous Galerkin finite element method with interior penalties to handle the layer phenomenon. With the use of a typical Shishkin mesh, the domain is discretized, and a uniform error estimate is obtained. Numerical experiments are conducted to validate the theoretical conclusions.

*Corresponding author

Received 15 August 2023; revised 1 January 2024; accepted 1 January 2024

Kumar Rajeev Ranjan
Department of Mathematics, National Institute of Technology Patna, Patna, India. e-mail: kumarr.phd18.ma@nitp.ac.in

S. Gowrisankar
Department of Mathematics, National Institute of Technology Patna, Patna, India. e-mail: s.gowri@nitp.ac.in

How to cite this article
Ranjan, K. R. and Gowrisankar, S., Discontinuous Galerkin approach for two-parametric convection-diffusion equation with discontinuous source term. *Iran. J. Numer. Anal. Optim.*, 2024; 14(2): 347-366. https://doi.org/10.22067/ijnao.2024.84456.1316

AMS subject classifications (2020): Primary AMS 65L70; Secondary AMS 65L20, AMS 65L6.

Keywords: Convection-diffusion problem; The NIPG methods; Shishkin mesh; Interior layers; Uniform convergence.

1 Introduction

In this article, we consider a two-parametric singularly perturbed problem with a discontinuous source term

$$\begin{cases} -\varepsilon_1 u'' + \varepsilon_2 b(x)u' + c(x)u = f, \quad x \in \Omega = (0,1), \\ u(0) = u(1) = 0, \end{cases} \tag{1}$$

where $0 < \varepsilon_1,\ \varepsilon_2 \ll 1$. In addition, the coefficient functions, $b(x)$ and $c(x)$, are assumed to sufficiently smooth such that $b(x) \geq b_0 > 0$, $c(x) \geq c_0 > 0$, and

$$\gamma^2 = c(x) - \frac{1}{2}\varepsilon_2 b'(x) \geq \gamma_0 > 0, \tag{2}$$

where b_0, c_0, and γ_0 are positive constants and the source function, $f(x)$, has discontinuity of type I at some interior point say $x = d$. Moreover, we also assume that $[f](d) \leq C$, where $[f](d) = f(d+) - f(d-)$. Problems of these types have several real-world applications, like transport phenomena in chemistry, biology, financial mathematics, and so on. Several numerical techniques are available in the literature on finite difference methods and finite element methods; for example, see [1, 10, 14] for the finite element method (FEM) and [3] for the finite difference method (FDM). Moreover, few articles [7, 4] are available that deal with two-parametric singularly perturbed boundary value problems, in which the authors have used various difference schemes to establish uniform convergence. In [11], authors have studied the superconvergence of non-symmetric interior penalty Galerkin FEM for two-parametric singularly perturbed boundary value problems.

The FEM is one of the efficient numerical methods among all other numerical methods to compute numerical solutions of differential equations. Kadalbajoo and Yadaw [5] established a second-order convergence of FEM

in maximum norm on Shishkin mesh for two-parametric convection-diffusion problem. Brdar and Zarin [2] achieved uniform convergence of FEM in the energy norm on Bakhvalov mesh, while Zarin [12] has proved uniform convergence on exponentially graded mesh with FEM. In spite of these works, no one has considered the two-parametric convection-diffusion problem with discontinuous coefficients or source terms, which also predominately appears in the modeling of real-world problems.

On the other way, the nonsymmetric interior penalty Galerkin (NIPG) method uses penalty parameters to enforce both the continuity of the solution and boundary conditions. Another interesting factor about the NIPG method is to handle the unstructured mesh, discontinuous coefficients, and local mesh refinements. Furthermore, an NIPG method can be made stable and coercive in the sense of convergence by taking some user-chosen penalty parameters. Moreover, in other variants of the discontinuous Galerkin method like symmetric interior penalty Galerkin method and incomplete interior penalty Galerkin method, we have to choose penalty parameters, and it is a herculean task to establish its coercivity.

Therefore, in this present work, we consider a two-parametric convection-diffusion problem with a discontinuous source term and establish a second-order convergence (superconvergence) of the NIPG method in the norm $\|\cdot\|_{\varepsilon}$ induced by the corresponding bilinear form.

The article is arranged in the following way: In Section 2, we study the continuous problem and the bounds for different components of the exact solution. Section 3 is all about the bilinear form with respect to the NIPG method, its properties, and the existence and uniqueness of the weak solution. Construction of the Shishkin-type mesh, error analysis, and establishment of convergence results are introduced in Section 4. The numerical results and discussions are kept in Section 5 that which supports our theoretical conclusions.

Note: Throughout this paper, we have taken C as a generic constant, which is free from perturbation parameters ε_1, ε_2, and the discretization parameter, N.

2 The continuous problem and bounds for the solutions

This section is designed to discuss the solution of the continuous problem (1), their smooth and layer components, and their respective bounds. In general, the solution to the reduced problem (1) does not satisfy the boundary conditions. Therefore we have boundary and interior layers in the solution. To handle these boundary and interior layers, let p_0 and p_1 be two solutions of the characteristic equation:

$$-\varepsilon_1 p^2(x) + \varepsilon_2 b(x)p(x) + c(x) = 0.$$

Moreover, $p_0(x) < 0$ and $p_1(x) > 0$ describe the boundary layers at $x = 0$ and $x = 1$, respectively. Let

$$\lambda_l = -\max_{x\in[0,1]} p_0(x), \quad \lambda_r = \min_{x\in[0,1]} p_1(x),$$

where λ_l and λ_r are defined as

$$\lambda_l = \min_{x\in[0,1]} \frac{-\varepsilon_2 b(x) - \sqrt{\varepsilon_2^2 b(x)^2 + 4\varepsilon_1 c(x)}}{2\varepsilon_1},$$
$$\lambda_r = \min_{x\in[0,1]} \frac{-\varepsilon_2 b(x) + \sqrt{\varepsilon_2^2 b(x)^2 + 4\varepsilon_1 c(x)}}{2\varepsilon_1}.$$

In that case, the problems are categorized on the basis of the relationship between the perturbation parameters:

Case *I*. When $\varepsilon_1 \ll \varepsilon_2 = 1$, then $\lambda_l = \mathcal{O}(\varepsilon_1^{-1})$ and $\lambda_r = \mathcal{O}(1)$. It is the case of singular perturbation problem with a only one perturbation parameter.

Case *II*. When $\varepsilon_1 \ll \varepsilon_2 \ll 1$, then $\lambda_l = \mathcal{O}(\varepsilon_2\varepsilon_1^{-1})$ and $\lambda_r = \mathcal{O}(\varepsilon_1^{-1})$. Clearly one can see that $\lambda_l < \lambda_r$, which results a stronger boundary layer at $x = 1$ than the boundary layer at $x = 0$.

Case *III*. When $\varepsilon_2^2 \ll \varepsilon_1 \ll 1$, then λ_l and λ_r both are of order $\mathcal{O}(\varepsilon_1^{-1/2})$, and problem behaves like reaction diffusion problems.

Further details about the category of the problems can be found in [11]. In this work, we discuss the a posteriori error estimation for the problem that falls under the category of case II.

Theorem 1. Let $p \in (0,1)$ be arbitrary and let

$$\varepsilon_2 m \left\|b'\right\|_{L^\infty(\Omega)} \leq k(1-p).$$

Then, the following bounds for the derivatives of the exact solution holds:

$$\left|u^k(x)\right| \leq C\big(1+\lambda_l^k e^{-p\lambda_l x}+\lambda_l^k e^{-p\lambda_l(d-x)}+\lambda_r^k e^{-p\lambda_r(x-d)}+\lambda_r^k e^{-p\lambda_r(1-x)}\big), \quad x \in \Omega, \text{ for } 0 \leq k \leq m.$$

Proof. The proof can be found in [6]. □

This motivates us to decompose the exact solution u of (1) into smooth and layer components in the following way:

$$u = S + E_{l,1} + E_{l,2} + E_{r,1} + E_{r,2}$$

such that the bounds satisfy the following relations:

$$|S^k(x)| \leq C, \tag{3a}$$
$$|E_{l,1}^k(x)| \leq C\lambda_l^k e^{-p\lambda_l x}, \tag{3b}$$
$$|E_{l,2}^k(x)| \leq C\lambda_l^k e^{-p\lambda_l(d-x)}, \tag{3c}$$
$$|E_{r,1}^k(x)| \leq C\lambda_r^k e^{-p\lambda_r(x-d)}, \tag{3d}$$
$$|E_{r,2}^k(x)| \leq C\lambda_r^k e^{-p\lambda_r(1-x)}, \tag{3e}$$

for $x \in \Omega$ and $0 \leq k \leq m$.

3 The NIPG method

Let $\mathcal{T}_N = I_j = (x_{j-1}, x_j)$, $j = 1, 2, \cdots, N$ be a partition of the domain Ω. Define the broken Sobolev space of order k, $H^k(\Omega, \mathcal{T}_N) := \big\{v : v \in L^2(\Omega), v\big|_{I_j} \in H^k(I_j), \text{ for all } I_j \in \mathcal{T}_N\big\}$. Definitions of the Sobolev norm and seminorm are as follows:

$$\|v\|_{k,\mathcal{T}_N} := \left(\sum_{j=1}^{N} \|v\|_{k,I_j}^2\right)^{1/2} \quad |v|_{k,\mathcal{T}_N} := \left(\sum_{j=1}^{N} |v|_{k,I_j}^2\right)^{1/2}.$$

We define the finite element space V^N that are related to the partition $\mathcal{T}_N$ as follows:

$$V^N = \{v : v \in L^2(\Omega), v\big|_{I_j} \in P^1(I_j), \text{ for all } I_j \in \mathcal{T}_N\}.$$

Here, $P^1(I_j)$ is the space of all polynomials of degree at most one on I_j. It is to note that the elements in V^N are allowed to be discontinuous across the boundaries of each element I_j.

The corresponding weak formulation of (1) can be read as follows: Find $u_N \in V^N$ such that

$$B(u_N, v_N) = l(v_N), \quad \text{for all } v_N \in V^N, \tag{4}$$

where

$$B(u, v) = B_1(u, v) + B_2(u, v).$$

Also, $B_1(u, v)$, $B_2(u, v)$, and $l(v)$ are defined for all $u, v \in H^1(\Omega, \mathcal{T}_N)$ as

$$\begin{aligned} B_1(u,v) &= \sum_{j=1}^{N} \int_{I_j} \varepsilon_1 u'v'dx + \sum_{j=1}^{N} \varepsilon_1 \big([u(x_j)]\{v'(x_j)\} - \{u'(x_j)\}[v(x_j)]\big) \\ &\quad + \sum_{j=1}^{N} \sigma_j [u(x_j)][v(x_j)], \\ B_2(u,v) &= \sum_{j=1}^{N} \int_{I_j} \big(\varepsilon_2 b(x)u' + c(x)u\big)dx + \sum_{j=1}^{N-1} b(x_j)[u(x_j)]v(x_j^-), \\ l(v) &= \sum_{j=1}^{N} \int_{I_j} fvdx. \end{aligned}$$

Here, σ_j's, $j = 0, 1, 2, \cdots, N$, are discontinuity-penalization parameters associated with the grid points x_j. Jumps and averages are defined by

$$\begin{aligned} [v(x_j)] &= v(x_j^+) - v(x_j^-), \qquad j = 1, 2, \cdots, N-1, \\ \{v(x_j)\} &= \frac{1}{2}\big(v(x_j^+) - v(x_j^-)\big), \quad j = 1, 2, \cdots, N-1, \end{aligned}$$

where $v(x_j^+) = \lim_{x \to x_j^+} v(x)$ and $v(x_j^-) = \lim_{x \to x_j^-} v(x)$, and at the boundary nodes these are defined as

$$[v(x_0)] = v(x_0), \quad \{v(x_0)\} = v(x_0), \quad [v(x_N)] = -v(x_N), \quad \{v(x_N)\} = v(x_N).$$

These notations can be seen in [8, 9, 15]. For any $v \in H^2(\Omega, \mathcal{T}_N)$, the norm is defined by

$$\|v\|_\varepsilon^2 = \varepsilon_1 \sum_{i=1}^{N} h_i \sum_{j=1}^{M} w_j v'(x_{i,j})^2 + \sum_{j=1}^{N} \|\gamma v\|_{L^2(I_j)}^2 + \sum_{j=0}^{N} \left(\frac{1}{2} b(x_j) + \sigma_j\right) [v(x_j)]^2.$$

Here, $x_{i,j}$ are the Gaussian points in the elements $I_j = (x_{j-1}, x_j)$, and $w_j > 0$ are the weights for M points Gaussian quadrature rule. The coercivity of bilinear form can be established by a simple calculation as in [8] as

$$B(v, v) = \|v\|_\varepsilon^2 .$$

Using coercivity, one can establish the result,

$$\|u_N\|_\varepsilon \leq C \|f\|_{L^2(\Omega)} , \tag{5}$$

where u_N is the solution of weak formulation (4). More details can be found in [9]. Furthermore, the result (5) implies the uniqueness of the solution of (4). Rank-nullity theorem in a finite-dimensional space and uniqueness give the existence of the solution.

Lemma 1. Let u and u_N be the solutions of (1) and (4), respectively. Then the bilinear form defined in (4) satisfies Galerkin orthogonality property

$$B(u - u_N, v_N) = 0, \quad \text{for all } v_N \in V^N. \tag{6}$$

Proof. Detailed proof can be found in [11]. □

4 Error analysis on Shishkin-type mesh

In this section, we introduce Shishkin-type mesh for the discretizing the domain of definition. Then, we apply the NIPG method on the introduced mesh to avail the convergence result theoretically. Proceeding in this way,

the domain, $\overline{\Omega}$ is divided into N subintervals as follows: Let $\Omega_{l,1}$, $\Omega_{l,2}$, $\Omega_{l,3}$, $\Omega_{r,1}$, $\Omega_{r,2}$, and $\Omega_{r,3}$ be subdomains such that

$$\Omega_{l,1} = [0, \delta_0], \quad \Omega_{l,2} = [\delta_0, d - \delta_0], \quad \Omega_{l,3} = [d - \delta_0, d], \quad \Omega_{r,1} = [d, d + \delta_1],$$
$$\Omega_{r,2} = [d + \delta_1, 1 - \delta_1], \quad \Omega_{r,3} = [1 - \delta_1, 1].$$

for some δ_0 and δ_1 defined by

$$\delta_0 = \min\left\{\frac{d}{8}, \frac{\delta}{\lambda_l}\ln N\right\}, \quad \delta_1 = \min\left\{\frac{1-d}{8}, \frac{\delta}{\lambda_r}\ln N\right\}.$$

Here, δ is a user defined parameter. Therefore, the mesh length of each subdomain can be defined as

$$h_j = \begin{cases} \frac{8}{N}\frac{\delta}{\lambda_l}\ln N, & \text{for } \Omega_{l,1} \cup \Omega_{l,3}, \\ \frac{4(d-2\delta_0)}{N}, & \text{for } \Omega_{l,2}, \\ \frac{8}{N}\frac{\delta}{\lambda_r}\ln N, & \text{for } \Omega_{r,1} \cup \Omega_{r,3}, \\ \frac{4(1-d-2\delta_1)}{N}. & \text{for } \Omega_{r,2}. \end{cases} \tag{7}$$

The following bounds can be obtained for the above step sizes:

$$h_j \leq \begin{cases} CN^{-1}\ln N\lambda_l^{-1}, & \text{for } \Omega_{l,1} \cup \Omega_{l,3}, \\ CN^{-1}, & \text{for } \Omega_{l,2} \cup \Omega_{r,2}, \\ CN^{-1}\ln N\lambda_r^{-1}, & \text{for } \Omega_{r,1} \cup \Omega_{r,3}. \end{cases} \tag{8}$$

Lemma 2. Let S^I be the piecewise Lagrange's interpolation of smooth component S of the exact solution u of (1). Then, we have

$$\left\|S - S^I\right\|_\varepsilon \leq CN^{-2}.$$

Proof. Using the definition of jump for the smooth part of the solution, as $S - S^I$ is continuous across interelement boundaries. So, $[(S - S^I)(x_j)] = 0$, it gives

$$\left\|S - S^I\right\|_\varepsilon^2 = \varepsilon_1 \sum_{i=1}^{N} h_i \sum_{j=1}^{M} \left((S - S^I)'(x_{i,j})\right)^2 + \sum_{j=1}^{N} \int_{I_j} \gamma^2 (S - S^I)^2 dx. \tag{9}$$

Now, we have to calculate the bounds for the two factors separately. We make the use of result on interpolation error estimate in [13]

$$\varepsilon_1 \sum_{i=1}^{N} h_i \sum_{j=1}^{M} \left((S-S^I)'(x_{i,j})\right)^2 \le C\varepsilon_1 \sum_{j=1}^{N} h_j^3 |S|^2_{H^2(I_j)} \le C \sum_{j=1}^{N} h_j^5$$
$$\le C\big(N(N^{-1}\ln N\lambda_l^{-1})^5 + N^{-4}\big). \tag{10}$$

Here, we have used $\lambda_r^{-1} \le \lambda_l^{-1}$ and bounds on the mesh lengths from (8). To establish the result in Lemma 2, we have to find out the essential bound for second term in (9)

$$\sum_{j=1}^{N} \left\|\gamma(S-S^I)\right\|_{L^2(I_j)} \le C \sum_{j=1}^{N} h_j^2 \int_{x_{j-1}}^{x_j} |S''(x)|dx \le C \sum_{j=1}^{N} h_j^3$$
$$\le C\big(N(N^{-1}\ln N\lambda_l^{-1})^3 + N^{-2}\big), \tag{11}$$

using (10) and (11) in (9). □

Lemma 3. Let $E_{l,i}$ and $E_{r,i}$ for $i = 1, 2$ be the piecewise Lagrange's interpolations of left and right layer components E_l and E_r, respectively. Then, we have the following estimation:

$$\left\|(E_{l,i} - E^I_{l,i})^s\right\|_{L^\infty(\Omega\setminus\Omega_{l,i})} \le C\lambda_l^s N^{-2}, \qquad i = 1, 2, s = 0, 1, 2.$$
$$\left\|(E_{r,i} - E^I_{r,i})^s\right\|_{L^\infty(\Omega\setminus\Omega_{r,i})} \le C\lambda_r^s N^{-2}, \qquad i = 1, 2, s = 0, 1, 2.$$

Proof. We will prove only the estimation for $E_{l,1} - E^I_{l,1}$, and other estimations for remaining part will be same.

Making the use of the Cauchy–Schwarz inequality gives us

$$\left\|(E_{l,1} - E^I_{l,1})^s\right\|_{L^\infty(\Omega\setminus\Omega_{l,1})} \le \left\|(E_{l,1})^s\right\|_{L^\infty(\Omega\setminus\Omega_{l,1})} + \left\|(E^I_{l,1})^s\right\|_{L^\infty(\Omega\setminus\Omega_{l,1})}$$
$$\le 2\left\|(E_{l,1})^s\right\|_{L^\infty(\Omega\setminus\Omega_{l,1})}.$$

Here, we have used the stability property for special interpolant from Lemma 3.2 in [15]. Now we have to bound only $\left\|(E_{l,1})^s\right\|_{L^\infty(\Omega\setminus\Omega_{l,1})}$.

$$\left\|(E_{l,1})^s\right\|_{L^\infty(\Omega\setminus\Omega_{l,1})} \le C\lambda_l^s \max_{x\in[0,\delta_0]} e^{-p\lambda_l x} \le C\lambda_l^s N^{-p\delta} \le C\lambda_l^s N^{-2}.$$

Here we have used $p\delta \ge 2$, which is the required estimate, we had to prove. For estimations in the second part of the Lemma 2, we apply the same path of approach as above. So we have done the proof of lemma. □

Theorem 2. Let E^I denote the piecewise Lagrange's interpolations of layer components of solution E. Therefore, the interpolation error of the layer component satisfies

$$\left\|E - E^I\right\|_\varepsilon \leq C(N^{-1}\ln N)^2.$$

Proof. To establish the bound in this theorem, we express errors in terms of left boundary layer, right boundary layer, and interior layer components; that is,

$$\left\|E - E^I\right\|_\varepsilon \leq \left\|E_{l,1} - E^I_{l,1}\right\|_\varepsilon + \left\|E_{l,2} - E^I_{l,2}\right\|_\varepsilon + \left\|E_{r,1} - E^I_{r,1}\right\|_\varepsilon + \left\|E_{r,2} - E^I_{r,2}\right\|_\varepsilon. \tag{12}$$

We derive the bounds for all the four terms in the right hand side of (12). Taking the first term from (12) and using coercivity of bilinear form from Section 3, we have

$$\left\|E_{l,1} - E^I_{l,1}\right\|^2_\varepsilon = \varepsilon_1 \sum_{i=1}^{N} h_i \sum_{i=1}^{M} \left((E_{l,1} - E^I_{l,1})'(x_{i,j})\right)^2 + \sum_{j=1}^{N} \int_{I_j} \gamma^2 (E_{l,1} - E^I_{l,1})^2 dx. \tag{13}$$

Now, we estimate the first term from (13) over the fine mesh region $\Omega_{l,1}$, using classical result on interpolation error estimates

$$\begin{aligned}
\varepsilon_1 \sum_{i=1}^{N/8} h_i \sum_{j=1}^{M} \left((E_{l,1} - E^I_{l,1})'(x_{i,j})\right)^2 &\leq C\varepsilon_1 \sum_{j=1}^{N/8} h_j^3 \left(\int_{x_{j-1}}^{x_j} |E'''_{l,1}| dx \right)^2 \\
&\leq C\varepsilon_1 \sum_{j=1}^{N/8} h_j^3 \left(\int_{x_{j-1}}^{x_j} |E'''_{l,1}|^2 dx \right) \left(\int_{x_{j-1}}^{x_j} dx \right) \\
&\leq C\varepsilon_1 \sum_{j=1}^{N/8} h_j^4 \int_{x_{j-1}}^{x_j} \lambda_l^4 e^{-2p\lambda_l x} dx \\
&\leq C(N^{-1}\ln N)^4. \qquad (14)
\end{aligned}$$

$$\begin{aligned}
\sum_{j=1}^{N/8} \int_{I_j} \gamma^2 (E_{l,1} - E^I_{l,1})^2 dx &\leq C \sum_{j=1}^{N/8} h_j^4 \int_{x_{j-1}}^{x_j} |E''_{l,1}|^2 dx \\
&\leq C \sum_{j=1}^{N/8} h_j^4 \int_{x_{j-1}}^{x_j} \lambda_l^4 e^{-2p\lambda_l x} dx
\end{aligned}$$

$$\leq C(N^{-1}\ln N)^4. \tag{15}$$

Using (14) and (15), we have established the first part of (12) for the finer part of the domain $\Omega_{l,1}$. In the same manner, we can establish the bound for $\left\|E_{l,2}-E^I_{l,2}\right\|_\varepsilon$, $\left\|E_{r,1}-E^I_{r,1}\right\|_\varepsilon$, and $\left\|E_{r,2}-E^I_{r,2}\right\|_\varepsilon$ over $\Omega_{l,2}$, $\Omega_{r,1}$ and $\Omega_{r,2}$, respectively. Therefore, we can obtain

$$\left\|E_{l,2}-E^I_{l,2}\right\|_{\varepsilon,\Omega_{l,2}} \leq C(N^{-1}\ln N)^2, \tag{16a}$$

$$\left\|E_{r,1}-E^I_{r,1}\right\|_{\varepsilon,\Omega_{r,1}} \leq C(N^{-1}\ln N)^2, \tag{16b}$$

$$\left\|E_{r,2}-E^I_{r,2}\right\|_{\varepsilon,\Omega_{r,2}} \leq C(N^{-1}\ln N)^2. \tag{16c}$$

Now, we estimate $\left\|E_{l,1}-E^I_{l,1}\right\|_\varepsilon$ over $\Omega\setminus\Omega_{l,1}$. In this domain, we will derive the bound for the error in maximum norm. Using the result in Lemma 3

$$\varepsilon_1\sum_{i=N/8+1}^{N} h_i\sum_{j=1}^{M} w_j(E_{l,1}-E^I_{l,1})'(x_{i,j})^2 \leq C\varepsilon_1\sum_{j=N/8+1}^{N} h_j\lambda_l^2N^{-4} \leq CN^{-4}.$$

Again, bound on the second term may be calculated using triangular inequality and stability property as

$$\begin{aligned}\left\|E_{l,1}-E^I_{l,1}\right\|^2_{L^2(\Omega\setminus\Omega_{l,1})} &\leq \|E_{l,1}\|^2_{L^2(\Omega\setminus\Omega_{l,1})} + \leq \left\|E^I_{l,1}\right\|^2_{L^2(\Omega\setminus\Omega_{l,1})}\\ &\leq 2\|E_{l,1}\|^2_{L^2(\Omega\setminus\Omega_{l,1})}\\ &\leq C\sum_{j=N/8+1}^{N}\int_{x_{j-1}}^{x_j} e^{-2p\lambda_l x}dx \leq CN^{-4}.\end{aligned}$$

With the help of above two estimations a conclusion can be made

$$\left\|E_{l,1}-E^I_{l,1}\right\|^2_{\varepsilon,\Omega\setminus\Omega_{l,1}} \leq CN^{-4}. \tag{17}$$

Similar to (17), we can estimate the other remaining terms, and using (14)–(17), we conclude the bound in Theorem 2. □

Theorem 3. Let u^I be the piecewise Lagrange's interpolation of the exact solution u of (1) on Gaussian points. Then interpolation error $\zeta = u - u^I$ satisfies

$$\|\zeta\|_\varepsilon \leq C(N^{-1}\ln N)^2.$$

Proof. Using the solution decomposition and the estimations from Lemma 2 and Theorem 2, we can conclude that

$$\|\zeta\|_{\varepsilon} \leq \left\|S - S^I\right\|_{\varepsilon} + \sum_{k=l,r} \sum_{i=1}^{2} \left\|E_{k,i} - E_{k,i}^I\right\|_{\varepsilon} \leq C(N^{-1} \ln N)^2.$$

□

Now, we conclude the bound for closeness error by means of Galerkin orthogonality and coercivity. Therefore, we need the following theorem.

Theorem 4. Let u and u_N be the exact and discrete solution of (1) and its weak formulation (4), respectively. For u^I, the piecewise Lagrange's interpolation of u, let $\chi = u^I - u_N$ satisfies

$$\|\chi\|_{\varepsilon} \leq C(N^{-1} \ln N)^2.$$

Proof. Coercivity and Galerkin orthogonality give us

$$\|\chi\|_{\varepsilon}^2 = B(\chi, \chi) = -B(\zeta, \chi). \tag{18}$$

From the definition, ζ is continuous on interelement boundaries. That is, $[\zeta(x_j)] = 0$. So, $B_1(\zeta, \chi)$ can be reduced to

$$B_1(\zeta, \chi) = \sum_{j=1}^{N} \varepsilon_1 \int_{x_{j-1}}^{x_j} \zeta' \chi' dx - \sum_{j=1}^{N} \varepsilon_1 \{\zeta'(x_j)\}[\chi(x_j)]. \tag{19}$$

Now, we estimate the first term in (19),

$$\begin{aligned} \sum_{j=1}^{N} \varepsilon_1 \int_{x_{j-1}}^{x_j} \zeta' \chi' dx &\leq \left(\sum_{j=1}^{N} \varepsilon_1 \int_{x_{j-1}}^{x_j} (\zeta')^2 dx \right)^{1/2} \left(\sum_{j=1}^{N} \varepsilon_1 \int_{x_{j-1}}^{x_j} (\chi')^2 dx \right)^{1/2} \\ &\leq C(N^{-1} \ln N)^2 \|\chi\|_{\varepsilon}. \end{aligned}$$

Second term in (19) can be estimated using Cauchy–Schwarz inequality and the exact choice of discontinuity-penalization parameter $\sigma_j = N$,

$$\begin{aligned} \sum_{j=1}^{N} \varepsilon_1 \{\zeta'(x_j)\}[\chi(x_j)] &\leq \left(\sum_{j=0}^{N} \frac{\varepsilon_1}{\sigma_j} \{\zeta'(x_j)\}^2 \right)^{1/2} \left(\sum_{j=0}^{N} \sigma_j [\chi(x_j)]^2 \right)^{1/2} \\ &\leq C(N^{-1} \ln N)^2 \|\chi\|_{\varepsilon}. \end{aligned}$$

Using both these inequalities in (19), gives us

$$B_1(\zeta,\chi) \le C(N^{-1}\ln N)^2 \left\|\chi\right\|_\varepsilon . \tag{20}$$

Last but not the least is to estimate $B_2(\zeta,\chi)$,

$$\sum_{j=1}^{N}\int_{x_{j-1}}^{x_j}\big(\varepsilon_2 b(x)\zeta' + c(x)\zeta\big)\chi dx = \sum_{j=1}^{N}\varepsilon_2\int_{x_{j-1}}^{x_j} b(x)\zeta\chi' dx + \sum_{j=1}^{N}\int_{x_{j-1}}^{x_j}\big(c(x)-\varepsilon_2 b'(x)\big)\zeta\chi dx.$$

We estimate the two terms of right side of above equation separately,

$$\begin{aligned}&\left|\sum_{j=1}^{N}\varepsilon_2\int_{x_{j-1}}^{x_j} b(x)\zeta\chi' dx\right| \\ &\quad\le C\left(\sum_{i=1}^{3}\left\|\zeta\right\|_{L^\infty(\Omega_{l,i})}\left\|\chi'\right\|_{L^1(\Omega_{l,i})} + \sum_{i=1}^{3}\left\|\zeta\right\|_{L^\infty(\Omega_{r,i})}\left\|\chi'\right\|_{L^1(\Omega_{r,i})}\right) \\ &\quad\le C(N^{-1}\ln N)^2\left\|\chi\right\|_\varepsilon .\end{aligned}$$

Here, we have used

$$\left\|\chi'\right\|_{L^1(\Omega_{l,i})} \le C(\lambda_l^{-1}\ln N)^{1/2}\left\|\chi'\right\|_{L^2(\Omega_{l,i})} \le C(\varepsilon_1^{-1}\lambda_l^{-1}\ln N)^{1/2}\left\|\chi\right\|_\varepsilon$$

and results of Lemma 3.

Finally, the last term can be bounded as

$$\left|\sum_{j=1}^{N}\int_{x_{j-1}}^{x_j}\big(c(x)-\varepsilon_2 b'(x)\big)\zeta\chi dx\right| \le C\left\|\zeta\right\|_{L^2(\Omega)}\left\|\chi\right\|_{L^2(\Omega)} \le C(N^{-1}\ln N)^2\left\|\chi\right\|_\varepsilon .$$

Keeping in view of above two inequalities, we have

$$B_2(\zeta,\chi) \le C(N^{-1}\ln N)^2\left\|\chi\right\|_\varepsilon . \tag{21}$$

From (18), (20) and (21), we get the result of Theorem 4. □

Theorem 5. Let u and u_N be the solutions of continuous problem (1) and the discrete problem (4), respectively. Then the discretization error satisfies the following estimates:

$$\|u - u_N\|_\varepsilon \leq C(N^{-1} \ln N)^2.$$

Proof. Applying triangular inequality on the estimation from Theorems 3 and 4, we get the desired result. □

5 Numerical results and their assessments

In this section, we consider two test problems to validate the theoretical estimation established in the previous section.

Example 1. In this context, we have the first test problem:

$$\begin{aligned} -\varepsilon_1 u'' + \varepsilon_2(3x^2 + 2)u' + 7u &= f(x), \\ u(0) = u(1) &= 0, \end{aligned}$$

where $f(x)$ is defined by

$$f(x) = \begin{cases} 3x^2 + 2, & x \leq d, \\ 0.5x, & x > d, \end{cases}$$

and $d = 0.7$.

The above test problem is a singularly perturbed convection-diffusion problem with two parameters and discontinuity of type I in the source term at an interior point $x = d$ of the domain. The solution of the above test problem exhibits boundary layers at both end points $x = 0,\ 1$. In addition, a layer at the point of discontinuity of source term $x = d$ in the interior of the domain is recorded. The curve of the computed solution along with the exact solution is provided in Figure 1. The solutionprofile shows both the boundary layers at $x = 0$ and $x = 1$ and also the interior layer at $x = 0.7$. Here, 0.7 is the point of discontinuity for $f(x)$ in Example 1. The exact choice of discontinuity-penalization parameter used for this is $\sigma_j = N,$ for all j, which are discussed in Section 4.

Error and rate of convergence are examined for various values of discretization parameter N and the singular perturbation parameter ε_1. Two parameters ε_1 and ε_2 are related to each other by the relation $\varepsilon_1 \leq \varepsilon_2^2$. For the sake of convenience in our computation, we take the general case $\varepsilon_1 = \varepsilon_2^2$.

The exact solution to the above test problem is not known. Hence for the computational purpose, we determine errors for DG-norm by $\|u_N - u_{2N}\|_{\mathrm{DG}}$. We apply the double mesh principle to deduce the rate of convergence and errors. Therefore, the following formula is applied

$$R_\varepsilon^N = \log_2 \left(\frac{\|u_N - u_{2N}\|_{\mathrm{DG}}}{\|u_{2N} - u_{4N}\|_{\mathrm{DG}}} \right),$$

and the consequence is mentioned in Tables 1 and 2. From these tables, it is clear that the scheme is convergent with an order of convergence up to 2. In the study of [11], we see that if we change our domain discretization from Shishkin mesh to Bakhvalov mesh and Bakhvalov mesh to exponential mesh, then the rate of convergence increases gradually. In this regard, we achieve comparatively better order of convergence, and numerically it will be achieved to 2.

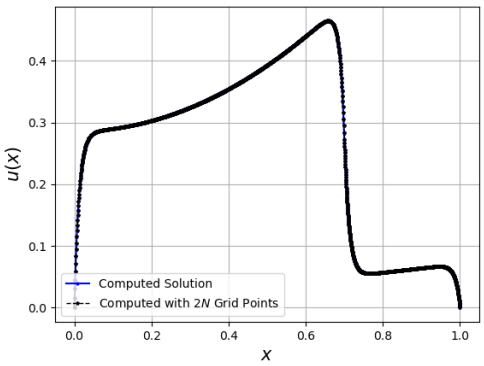


Figure 1: Computed and exact solution for Example 1 for $N = 2048$ and $\varepsilon = 10^{-6}$.

Table 1: $\|u_N - u_{2N}\|_{\mathrm{DG}}$ errors for Example 1

	Number of intervals (N)					
ε_2	64	128	256	512	1024	2048
10^{-2}	2.7829E-03	1.3574E-03	6.7059E-04	3.3331E-04	1.6617E-04	8.2964E-05
10^{-3}	3.7937E-03	1.8284E-03	8.9695E-04	4.4416E-04	2.2100E-04	1.1023E-04
10^{-4}	5.3400E-03	2.5120E-03	1.2140E-03	5.9633E-04	2.9548E-04	1.4707E-04
10^{-5}	7.7506E-03	3.5276E-03	1.6647E-03	8.0656E-04	3.9677E-04	1.9676E-04

Table 2: Convergence rates R_ε^N for Example 1

	Number of intervals (N)				
ε_2	128	256	512	1024	2048
10^{-2}	0.8902	0.9406	0.9705	0.9855	0.9928
10^{-3}	1.1069	0.8868	0.9140	0.9545	0.9774
10^{-4}	1.7920	1.5795	1.0695	0.9100	0.9347
10^{-5}	1.4870	1.8389	1.8514	1.5312	1.0518

Example 2. The second test problem in this context is taken a

$$\begin{aligned} -\varepsilon_1 u'' + \varepsilon_2(7x+5)u' + 3u &= f(x), \\ u(0) = u(1) &= 0, \end{aligned}$$

where $f(x)$ is defined by

$$f(x) = \begin{cases} x, & x \le d, \\ 2-x, & x > d, \end{cases}$$

and $d = 0.7$.

The test problem given above in Example 2 is also regarded as two-parametric convection-diffusion problem with discontinuous source term. Due to the lack of an exact solution to the above problem, we use the double mesh principle to conclude the error and rate of convergence of the obtained numerical solutions. The rate of convergence is obtained by the following formula:

$$P_\varepsilon^N = \log_2 \left(\frac{\|u_N - u_{2N}\|_{\mathrm{DG}}}{\|u_{2N} - u_{4N}\|_{\mathrm{DG}}} \right).$$

The same choice of discontinuity penalization parameter as in the previous example is imposed here. The best possible relation between ε_1 and ε_2 is $\varepsilon_1 = \varepsilon_2^2$. Now, the error $\|u_N - u_{2N}\|_{\mathrm{DG}}$ and the rate of convergence P_ε^N for the computed solution for the test problem in Example 2 with different values of discretization parameter N are given below in Tables 3 and 4, respectively.

The second-order convergence appears to be achieved up to the logarithmic factor from Table 4.

Table 3: $\|u_N - u_{2N}\|_{\mathrm{DG}}$ errors for Example 2

	Number of intervals (N)					
ε_2	64	128	256	512	1024	2048
10^{-2}	9.8446E-03	3.0845E-03	9.9633E-04	3.3445E-04	1.1108E-04	3.9958E-05
10^{-3}	4.0731E-02	2.0132E-02	8.8730E-03	3.5715E-03	1.4107E-03	5.3149E-04
10^{-4}	4.1807E-02	2.0444E-02	8.8440E-03	3.5894E-03	1.3849E-03	5.3839E-04
10^{-5}	4.1841E-02	2.0351E-02	8.8344E-03	3.5877E-03	1.4237E-03	5.2625E-04

Table 4: Convergence rates P_{ε}^{N} for Example 2

	Number of intervals (N)				
ε_2	128	256	512	1024	2048
10^{-2}	1.6742	1.6303	1.5748	1.5901	1.4751
10^{-3}	1.8166	1.8819	1.6128	1.3401	1.4083
10^{-4}	1.9320	1.8089	1.3009	1.3739	1.5630
10^{-5}	1.9397	1.8239	1.6000	1.5333	1.5359

5.1 Discussion

Several articles are available in the literature that deal with the convergence and superconvergence of numerical solutions for problems with one and two small parameters. However, in all the articles, problems with continuous coefficients and source terms are discussed. In [11], authors have established the superconvergence result of the discontinuous Galerkin method for a two-parametric singular perturbation problem in which all the coefficients are assumed to be continuous. In that, authors have used three types of meshes usual Shishkin mesh, Bakhvalov mesh, and exponentially graded mesh for domain discretization and shown their rate of convergence of order two by

talking linear elements in its finite element space. The convergence results on both the Bakhvalov mesh and exponentially graded mesh have a sharper rate of convergence, that is $\mathcal{O}(N^{-2})$, than the rate of convergence obtained on the usual Shishkin mesh. In the present work, we attempt to obtain the convergence result for two-parametric problem in which the source term is discontinuous. The discontinuous Galerkin method is applied, and the second-order convergence has been obtained up to the logarithmic factor.

6 Conclusion

In this article, we proposed and discussed uniformly convergent discontinuous Galerkin FEM for two-parametric singularly perturbed problems with a discontinuous source term. The problem usually exhibits a weak interior layer because of the discontinuous source term and the presence of the multiple perturbation parameters gives rise to boundary layers on both sides of the boundary. In order to address the layering phenomenon, we developed the non-symmetric discontinuous Galerkin FEM with interior penalties through this work. With the use of a typical Shishkin mesh, the domain is discretized, and a uniform error estimate was obtained. The numerical experiments were conducted to validate the theoretical conclusions.

Acknowledgements

Authors are grateful to there anonymous referees and editor for their constructive comments.

References

[1] Babu, A.R. and Ramanujam, N. *The SDFEM for singularly perturbed convection-diffusion problems with discontinuous source term arising in the chemical reactor theory*, Int. J. Comput. Math. 88(8) (2011), 1664–1680.

[2] Brdar, M. and Zarin, H. *A singularly perturbed problem with two parameters on a Bakhvalov-type mesh*, J. Comput. Appl. Math. 292(1) (2016), 307–319.

[3] Cen, Z. *A hybrid difference scheme for a singularly perturbed convection-diffusion problem with discontinuous convection coefficient*, Appl. Math. Comput. 169(1) (2005), 689–699.

[4] Das, P. and Mehrmann, V. *Numerical solution of singularly perturbed convection-diffusion-reaction problems with two small parameters*, BIT Numer. Math. 56(1) (2016), 51–76.

[5] Kadalbajoo, M.K. and Yadaw, A.S. *Parameter-uniform Ritz-Galerkin finite element method for two parameter singularly perturbed boundary value problems*, Int. J. Pure Appl. Math. 55(2) (2009), 287–300.

[6] Linß, T. *Layer-adapted meshes for reaction-convection-diffusion problems*, Springer, 2009.

[7] Linß, T. and Roos, H.G. *Analysis of a finite-difference scheme for a singularly perturbed problem with two small parameters*, J. Math. Anal. Appl. 289(2) (2004), 355–366.

[8] Ranjan, K.R. and Gowrisankar, S. *Uniformly convergent NIPG method for singularly perturbed convection diffusion problem on Shishkin type meshes*, Appl. Numer. Math. 179(4) (2022); 125–148.

[9] Ranjan, K.R. and Gowrisankar, S. *NIPG method on Shishkin mesh for singularly perturbed convection–diffusion problem with discontinuous source term*, Int. J. Comput. Methods 20(2) (2023), 2250048.

[10] Roos, H.G. and Zarin, H. *The streamline-diffusion method for a convection–diffusion problem with a point source*, J. Comput. Appl. Math. 150(1) (2003), 109–128.

[11] Singh, G. and Natesan, S. *Study of the NIPG method for two–parameter singular perturbation problems on several layer adapted grids*, J. Appl. Math. Comput. 63(1) (2020), 683–705.

[12] Zarin, H. *Exponentially graded mesh for a singularly perturbed problem with two small parameters*, Appl. Numer. Math. 120 (2017), 233–242.

[13] Zhang, J., Ma, X. and Lv, Y. *Finite element method on Shishkin mesh for a singularly perturbed problem with an interior layer*, Appl. Math. Lett. 121 (2021), 107509.

[14] Zhu, P. and Xie, S. *Higher order uniformly convergent continuous/discontinuous Galerkin methods for singularly perturbed problems of convection-diffusion type*, Appl. Numer. Math. 76 (2014), 48–59.

[15] Zhu, P., Yang, Y. and Yin, Y. *Higher order uniformly convergent NIPG methods for 1-D singularly perturbed problems of convection-diffusion type*, Appl. Math. Model. 39(22) (2015), 6806–6816.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 2, 2024, pp 367–390
https://doi.org/10.22067/ijnao.2024.85789.1358
https://ijnao.um.ac.ir/

Research Article

High order second derivative multistep collocation methods for ordinary differential equations

S. Fazeli*,

Abstract

In this paper, we introduce second derivative multistep collocation methods for the numerical integration of ordinary differential equations (ODEs). These methods combine the concepts of both multistep methods and collocation methods, using second derivative of the solution in the collocation points, to achieve an accurate and efficient solution with strong stability properties, that is, A-stability for ODEs. Using the second-order derivatives leads to high order of convergency in the proposed methods. These methods approximate the ODE solution by using the numerical solution in some points in the r previous steps and by matching the function values and its derivatives at a set of collocation methods. Also, these methods utilize information from the second derivative of the solution in the collocation methods. We present the construction of the technique and discuss the analysis of the order of accuracy and linear stability properties. Finally,

*Corresponding author

Received 15 December 2023; revised 16 January 2024; accepted 17 January 2024

Somayyeh Fazeli
Marand Technical Faculty, University of Tabriz, Tabriz, Iran. e-mail: fazeli@tabrizu.ac.ir

How to cite this article
Fazeli, S., High order second derivative multistep collocation methods for ordinary differential equations. *Iran. J. Numer. Anal. Optim.*, 2024; 14(2): 367-390. https://doi.org/10.22067/ijnao.2024.85789.1358

some numerical results are provided to confirm the theoretical expectations. A stiff system of ODEs, the Robertson chemical kinetics problem, and the two-body Pleiades problem are the case studies for comparing the efficiency of the proposed methods with existing methods.

AMS subject classifications (2020): 65R20.

Keywords: Collocation; Linear stability; Ordinary differential equation; Second derivative methods.

1 Introduction

This paper is devoted to introducing a new class of collocation methods by using the second derivative of the solution for the numerical solution of ordinary differential equations (ODEs) in the form

$$\begin{cases} y' = f(y(t)), \quad t \in [t_0, T], \\ y(t_0) = y_0, \end{cases} \tag{1}$$

where the function $f : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ is sufficiently smooth, $y_0 \in \mathbb{R}^d$ is a known initial value, and the dimensionality of the system is shown by d.

One-step collocation methods for ODEs have been analyzed by many authors (see [3] and references therein contained). As it is well known, the collocation methods is a technique that the given equation is transformed to an algebraic equation by representing the solution as a combination of basic functions belonging to a chosen finite dimensional space, usually a piecewise algebraic polynomial, which enforces the integral equation at the chosen collocation points.

Multistep methods play a crucial role in the numerical integration of ODEs due to their ability to obtain the approximate solution of a high order of convergence. Lie and Nørsett [14] first introduced the idea of multistep collocation methods. In comparison with classical collocation methods, these methods depend on more parameters while computational cost does not exceed them, and these methods have high convergence order and strong stability properties.

In comparison with one-step methods, in view of computational cost and precision of approximation, the multistep methods are more efficient [13].

Using higher order derivative can be a helpful way to achieve the integration methods will be more stable with high order convergency in both onestep and multistep methods. Firstly, a class of second-order derivative formulas is created, and the stability of these formulas was investigated by Enright [7]. Following this work, successful implementation of second-order derivatives has been developed in many literatures [4, 5, 6]. Also, second derivative two-step collocation methods have been recently constructed in [8]. These methods and all other traditional second derivative methods can be considered as special cases of the class of second derivative general linear methods (see, for instance, [1, 2, 4]).

We aim to provide a second derivative extension of the multistep collocation methods (SDMCMs) that use the second derivatives of the solution in the collocation points. These methods combine the advantages of both multistep collocation methods and second derivative methods to obtain methods with higher order accuracy and desired stability properties. In other word, we derive a special case of multistep collocation methods, which fixed number r of previous time steps and first and second derivative of the solution in m collocation points are used in the approximation of the solution in every subinterval. The advantages of these methods are their ability to handle stiff ODEs and their superior accuracy compared to multistep collocation methods while the computational cost does not exceed.

Theoretical concepts of new proposed methods, SDMCMs, is discussed in details in Section 2. In Section 3, the order condition of the proposed method is obtained. Section 4 is dedicated to analyzing the linear stability properties of SDMCMs with respect to the linear basic test equation. The proposed methods with two and three steps and one and two collocation parameters are constructed in Section 5. Finally, Section 6 presents several numerical examples in order to validate the effectiveness of proposed methods and experimentally verify the order of convergence.

2 Construction of method

Let us divide the interval $I := [t_0, T]$ to the finite set of subintervals $[t_n, t_{n+1}]$ for $n = 0, 1, \ldots, N-1$ by introducing a uniform mesh h in the form

$$I_h = \{t_0 + nh |\ n = 0, 1, \ldots, N,\ h > 0, \quad Nh = T - t_0\}.$$

Also, consider the abscissa vector $c = [c_1, c_2, \ldots, c_m]^T$ and define the collocation parameters by $t_{n,i} := t_n + c_i h$. The multi-step continuous approximation to the solution of (1) is defined by

$$\begin{cases} P(t_n + sh) = \displaystyle\sum_{k=0}^{r-1} \varphi_k(s) y_{n-k} + h \sum_{j=1}^{m} \Big(\psi_j(s) f(P(t_{n,j}))\Big) \\ \qquad\qquad + h^2 \displaystyle\sum_{j=1}^{m} \Big(\chi_j(s) g(P(t_{n,j}))\Big), \\ y_{n+1} = P(t_{n+1}), \end{cases} \tag{2}$$

with $s \in (0, 1]$ and the function g is the second derivative of unknown function $y(t)$ and is defined by $g(\cdot) = f'(\cdot) f(\cdot)$. Also, for the implementation of the method, the approximate solution in the initial interval $[t_0, t_{r-1}]$ is needed where can be computed by an arbitrary method with sufficiently high order convergency . The method is defined by the coefficient functions $\varphi_k(s),\ \psi_j(s),\ \chi_j(s),\ j = 1, \ldots, m,\ k = 0, 1, \ldots, r-1$, which are polynomials of degree $2m + r - 1$. The interpolation conditions for $P(t_n + sh)$ at the points t_{n-i}, lead to

$$P(t_{n-i}) = y_{n-i}, \quad i = 0, 1, \ldots, r-1, \tag{3}$$

and collocating the both sides of (1) and its derivative in the collocation points $t_{n,i}$ is written in the form

$$P'(t_{n,i}) = f(P(t_{n,i})), \quad P''(t_{n,i}) = g(P(t_{n,i})), \quad i = 1, 2, \ldots, m. \tag{4}$$

The coefficient polynomials of method are determined by conditions (3) and (4), which lead to

$$\varphi_k(-i) = \delta_{ik},\ \ \psi_j(-i) = 0, \quad \chi_j(-i) = 0, \tag{5}$$

and

$$\begin{aligned} &\varphi_k'(c_i)=0, \quad \psi_j'(c_i)=\delta_{ij}, \quad \chi_j'(c_i)=0,\\ &\varphi_k''(c_i)=0, \quad \psi_j''(c_i)=0, \quad \chi_j''(c_i)=\delta_{ij}, \end{aligned} \tag{6}$$

where δ_{ij}, $i,j=1,2,\ldots,m$, is the usual Kronecker delta. The construction of these polynomials is obtained by the Hermite–Birkhoff interpolation [18]. To find the unique polynomial that satisfies these conditions, the polynomials $\varphi_k(s)$, $\psi_j(s)$, and $\chi_j(s)$ can be formulated in the form

$$\varphi_k(s)=\sum_{i=0}^{2m+r-1}\Phi_i^{[k]}\frac{s^i}{i!}, \quad \psi_j(s)=\sum_{i=0}^{2m+r-1}\Psi_i^{[j]}\frac{s^i}{i!}, \quad \chi_j(s)=\sum_{i=0}^{2m+r-1}\chi_i^{[j]}\frac{s^i}{i!}. \tag{7}$$

Now, the result of setting $s=-k,\ k=0,1,\ldots,r-1$, in (7) and setting $s=c_i,\ i=1,2,\ldots,m$ in the first and second derivatives of the polynomials, leads to a linear system for coefficient of the polynomials. The coefficient matrix $A\in\mathbb{R}^{(2m+r)\times(2m+r)}$ is given by

$$A=\begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0\\ 1 & \frac{(-1)^1}{1!} & \frac{(-1)^2}{2!} & \frac{(-1)^3}{3!} & \ldots & \frac{(-1)^{2m+r-1}}{(2m+r-1)!}\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & \frac{(-r+1)^1}{1!} & \frac{(-r+1)^2}{2!} & \frac{(-r+1)^3}{3!} & \ldots & \frac{(-r+1)^{2m+r-1}}{(2m+r-1)!}\\ 0 & 1 & \frac{c_1^1}{1!} & \frac{c_1^2}{2!} & \ldots & \frac{c_1^{2m+r-2}}{(2m+r-2)!}\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 1 & \frac{c_m^1}{1!} & \frac{c_m^2}{2!} & \ldots & \frac{c_m^{2r+m-2}}{(2r+m-2)!}\\ 0 & 0 & 1 & \frac{c_1^1}{1!} & \ldots & \frac{c_1^{2m+r-3}}{(2m+r-3)!}\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 1 & \frac{c_m^1}{1!} & \ldots & \frac{c_m^{2m+r-3}}{(2m+r-3)!} \end{bmatrix},$$

and the vectors in the right hand of linear system of equations are defined by $\mathbf{u}_r^{[k]}\in\mathbb{R}^r,\ k=1,2,\ldots,r$, $\mathbf{v}_m^{[j]}\in\mathbb{R}^m,\ j=1,2,\ldots,m$ as

$$(\mathbf{u}_r^{[k]})_i = \begin{cases} 0, & i \neq k, \\ 1, & i = k, \end{cases} \qquad (\mathbf{v}_m^{[j]})_i = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}$$

By solving the systems

$$\begin{aligned} A\Phi^{[k]} &= [\mathbf{u}_r^{[k+1]}, \mathbf{0}_m, \mathbf{0}_m]^T, \; k = 0, 1, \ldots, r-1, \\ A\Psi^{[j]} &= [\mathbf{0}_r, \mathbf{v}_m^{[j]}, \mathbf{0}_m]^T, \qquad j = 1, 2, \ldots, m, \\ A\chi^{[j]} &= [\mathbf{0}_r, \mathbf{0}_m, \mathbf{v}_m^{[j]}]^T, \qquad j = 1, 2, \ldots, m, \end{aligned} \tag{8}$$

the coefficients of the polynomials are determined.

Now, the conditions for applying the poising condition [9] are established. Poising condition is a criterion that determines whether the given set of data points can be interpolated using the Hermite–Birkhoff interpolation. By using this condition, it can be shown that the solution of the interpolation problems can be determined uniquely. For this purpose, the following theorems, which are needed to check the uniqueness of the solution of the Hermite–Birkhoff interpolation, are mentioned [9]:

Let k and n be natural numbers, and define the matrix

$$E = \|\epsilon_{ij}\|, \quad i = 1, \ldots, k, \; j = 0, 1, \ldots, n-1,$$

as a matrix with k rows and n columns having elements

$$\epsilon_{ij} = 0 \;\text{ or }\; 1,$$

which are such that

$$\sum_{i,j} \epsilon_{ij} = n.$$

We shall also assume that the matrix E has no zero rows. Also, suppose that $\{x_i\}_{i=1}^k$ is an increasing real number as

$$x_1 < x_2 < \cdots < x_k.$$

Also, the set of ordered pairs is considered by

$$e = \{(i, j) \mid \; \epsilon_{ij} = 1\}.$$

The interpolation problem is described by the reals x_i and the "incidence matrix" E in the form

$$f^{(j)}(x_i) = y_i^{(j)} \quad \text{for} \quad (i,j) \in e. \tag{9}$$

The matrix E has a structure where each row corresponds to a different interpolation point, and the columns represent the derivatives at those points. It is appropriate to see (9) from the point of view of a Hermite–Birkhoff interpolation problem which we show it in an abbreviated form by HB-problem.

Definition 1. When the conditions

$$P(x) \in \pi_{n-1}, \quad P^{(j)}(x_i) = 0, \quad \text{for all } (i,j) \in E,$$

are established, then the HB-problem (9) is poised, and as a result, $P(x) \equiv 0$, equivalently, the matrix E is called poised if the related interpolation problem has a unique solution for any set of constants $y_i^{(j)}$, in which the problem is independent of choosing of the ordered interpolation points $x_l, x_2, \ldots, x_k$.

Define

$$\tilde{m}_j = \sum_{i=1}^{k} \epsilon_{i,j}, \quad \tilde{M}_l = \sum_{j=0}^{l} \tilde{m}_j, \quad j,l = 0,1,\ldots,n-1.$$

It is shown by Schoenberg [18] that a necessary condition for poising of E is

$$\tilde{M}_l \geq l+1, \quad l = 0,1,\ldots,n-1,$$

and these inequalities are recognized as Polya conditions.

Definition 2. Let the incidence matrix E be with k rows. Let f_i be the column index of the first one that appears in the row i. Moreover, E is called a pyramid matrix if, for each i, $\epsilon_{ij} = 1$ implies $\epsilon_{ij'} = 1$ for $f_i \leq j' \leq j$, and there is some value of $1 \leq i \leq k$ so that $f_1 \geq f_2 \geq \cdots \geq f_i$ and $f_i \leq f_{i+1} \leq \cdots \leq f_k$.

The necessary condition for poising the matrix E with respect to the ordering points $x_1 < x_2 < \cdots < x_k$, is declared by Ferguson [9] in the next theorem.

Theorem 1. If E is a pyramid matrix with k rows, satisfying the Polya conditions, then E is poised with respect to the ordering $x_1 < x_2 < \cdots < x_k$.

Now, we show that our interpolation problems (5)–(6) have unique solution or equivalently the conditions hold true for the given interpolation problem. For these problems, the interpolation points can be considered by

$$-r+1 < -r+2 < \cdots < -1 < 0 < c_1 < c_2 < \cdots < c_m.$$

So, the matrix E with $m+r$ rows and $2m+r$ columns can be defined by

$$E = \begin{bmatrix} 1\,0\,0\,0\cdots 0 \\ 1\,0\,0\,0\cdots 0 \\ \vdots\,\vdots\,\vdots \quad\quad \vdots \\ 1\,0\,0\,0\cdots 0 \\ 0\,1\,1\,0\cdots 0 \\ \vdots\,\vdots\,\vdots \quad\quad \vdots \\ 0\,1\,1\,0\cdots 0 \end{bmatrix}.$$

Thus, we have

$$\tilde{m}_0 = r, \quad \tilde{m}_1 = m, \quad \tilde{m}_2 = m, \quad \tilde{m}_3 = 0, \ \ldots, \tilde{m}_{n-1} = 0,$$

$$\tilde{M}_0 = r, \quad \tilde{M}_1 = r+m, \quad \tilde{M}_3 = 2m+r, \ \ldots, \ \tilde{M}_{n-1} = 2m+r = n.$$

By considering the numbers $\tilde{M}_j$, since

$$\tilde{M}_0 >= 1, \quad \tilde{M}_1 >= 2, \quad \tilde{M}_3 >= 3, \ \ldots, \ \tilde{M}_{n-1} >= n.$$

the Polya condition is satisfied for the matrix E. Also, by considering

$$f_1 = 0, \ f_2 = 0, \ \ldots, \ f_r = 0, \ f_{r+1} = 1, \ \ldots, \quad f_{r+m} = 1,$$

according Definition 2, one can see that the matrix E is a pyramid matrix. Thus, by Theorem 1, it is concluded that E is poised with respect to the ordering points $-r+1 < -r+2 < \cdots < -1 < 0 < c_1 < c_2 < \cdots < c_m$. Thus, satisfying these conditions guarantees a unique and smooth interpolating

polynomial that satisfies the given data (5)–(6). In the other word, the matrix A is nonsingular.

When a nonlinear ODE is integrated by implicit (first derivative) methods, the approximation of the stage values leads to solving a nonlinear algebraic system of equations, which can be solved by Newton's iterative methods. Thus, the Jacobian matrix $\partial f/\partial y$ is usually computed. Therefore, computing the second derivative function $g := (\partial f/\partial y)f(y)$ does not impose any additional computational cost. It should be mentioned that in the application of these methods, the Jacobian matrix $\partial g/\partial y$ is approximated by $(\partial f/\partial y)^2$, which is a piecewise constant approximation.

3 Continuous order conditions

In this section, we investigate continuous order conditions for the method (2). We know that the collocation polynomial $P(t_n + sh)$ provides a uniform approximation of order p to $y(t_n + sh)$, $s \in [0, 1]$. For this end, the approximate solution $P(t_n + sh)$ is replaced by $y(t_n + sh)$ in (2), where $y(t)$ is the exact solution of (1). Then by subtracting the both sides of the obtained relation, the discretization error is obtained in the form

$$\begin{aligned}\varepsilon(t_n + sh) =& y(t_n + sh) - \sum_{k=0}^{r-1} \varphi_k(s) y(t_{n-k}) - h \sum_{j=1}^{m} \psi_j(s) f(y(t_{n,j})) \\ & - h^2 \sum_{j=1}^{m} \chi_j(s) g(y(t_{n,j})),\end{aligned} \tag{10}$$

where $s \in [0, 1]$ and $n = 1, 2, \ldots, N - 1$. By expanding local discretization error in Taylor series around the point t_n and collecting terms with the same powers of h we have the following theorem.

Theorem 2. Assume that the function $f(y)$ is sufficiently smooth. Then the method (2) has uniform order p if the following conditions are satisfied:

$$\begin{cases} \sum_{k=0}^{r-1} \varphi_k(s) = 1, \\ \sum_{k=0}^{r-1} k\varphi_k(s) + \sum_{j=1}^{m} \psi_j(s) = s, \\ \sum_{k=0}^{r-1} \frac{(-k)^i}{i!}\varphi_k(s) + \sum_{j=1}^{m} \left(\psi_j(s)\frac{c_j^{i-1}}{(i-1)!} + \chi_j(s)\frac{c_j^{i-2}}{(i-2)!} \right) = \frac{s^i}{i!}, \end{cases} \tag{11}$$

where $s \in [0,1]$, $i = 2, 3, \ldots, p$. Moreover, the local discretization error (10) takes the form

$$\varepsilon(t_n + sh) = h^{p+1} C_p(s) y^{(p+1)}(t_n) + \mathcal{O}(h^{p+2}), \tag{12}$$

in which

$$\begin{aligned} C_p(s) =& \frac{s^{p+1}}{(p+1)!} - \sum_{k=0}^{r-1} \frac{(-k)^{p+1}}{(p+1)!}\varphi_k(s) \\ & - \sum_{j=1}^{m} \left(\psi_j(s)\frac{(c_j)^p}{p!} + \chi_j(s)\frac{c_j^{p-1}}{(p-1)!} \right). \end{aligned} \tag{13}$$

The set of order conditions (11) leads to a linear system of $p+1$ equations in $2m+r$ unknowns. The unknowns of this system are the coefficient polynomials of the proposed method.

Therefore, considering p at most equal to $2m+r-1$ guarantees that the linear system (11) is compatible. This leads to the following result.

Remark 1. The maximum attainable uniform order of convergence for the new method (2) is $2m+r-1$.

In the following, the necessary conditions for zero-stability of the new method are investigated. For this end, the method is applied to the simple ODE $y' = 0$, where the recurrence relation is obtained in the form

$$y_{n+1} = \sum_{k=0}^{r-1} \theta_k y_{n-k},$$

where $\theta_k = \varphi_k(1)$. Thus, the characteristic polynomial of this recurrence relation is

$$p(\omega) = \sum_{k=0}^{r-1} \theta_k \omega^k,$$

Therefore, the constructed method is zero-stable if and only if $-1 < \theta_i \leq 1$ for $i = 1, 2, \ldots, r-1$.

4 Linear stability analysis

We now focus our attention on the linear stability properties of the SDMCMs (2) with respect to the standard basic test equation of Dahlquist

$$y' = \lambda y, \tag{14}$$

where λ is a complex number with negative real part. Let us consider the vectors and matrices

$$\begin{aligned}
&A = [\psi_j(c_i)]_{i,j=1}^m \in \mathbb{R}^{m\times m}, && \overline{A} = [\chi_j(c_i)]_{i,j=1}^m \in \mathbb{R}^{m\times m},\\
&v^T = [\psi_i(1)]^T \in \mathbb{R}^m, && w^T = [\chi_i(1)]^T \in \mathbb{R}^m,\\
&(\varphi(c))_{j,i} = [\varphi_j(c_i)] \in \mathbb{R}^{m\times r}, && j = 0, 1, \ldots, r-1, \quad i = 1, 2, \ldots, m,\\
&P(t_n + ch) = [P(t_n + c_i h)]_{i,j=1}^m.
\end{aligned} \tag{15}$$

Theorem 3. Applying the new method (2) to the basic test equation (14) leads to the recurrence relation in the form

$$\begin{bmatrix} y_{n+1} \\ y_n^{(r)} \end{bmatrix} = \begin{bmatrix} M_{11}(z) & 0 \\ \mathbf{I}_r & \mathbf{0}_{r,1} \end{bmatrix} \begin{bmatrix} y_n \\ y_{n-1}^{(r)} \end{bmatrix}, \tag{16}$$

where $z := \lambda h$ and

$$\begin{aligned}
M_{11}(z) &= \varphi(1) + \left(zv^T + z^2 w^T\right) Q^{-1} \varphi(c) \in \mathbb{R}^{1\times r},\\
Q &= \mathbf{I}_m - zA - z^2 \overline{A} \in \mathbb{R}^{m\times m}.
\end{aligned}$$

Proof. First, the new constructed method (2) is applied to the test equation (14), and then both sides of the obtained equation are collocated at points $c_i, \quad i = 1, \ldots, m$. Thus the algebraic relation is obtained in the form

$$P(t_n + c_i h) = \sum_{k=0}^{r-1} \varphi_k(c_i) y_{n-k} + \lambda h \sum_{j=1}^{m} \psi_j(c_i) P(t_{n,j}) + \lambda^2 h^2 \sum_{j=1}^{m} \chi_j(c_i) P(t_{n,j}), \tag{17}$$

for $i = 1, \ldots, m$. Also, the approximated solution in the next step is obtained by

$$y_{n+1} = \sum_{k=0}^{r-1} \varphi_k(1) y_{n-k} + \lambda h \sum_{j=1}^{m} \psi_j(1) P(t_{n,j}) + \lambda^2 h^2 \sum_{j=1}^{m} \chi_j(1) P(t_{n,j}), \tag{18}$$

for $n = r-1, r, \ldots, N-1$. By the notations introduced in (15), the relations (17)–(18) can be written in the matrix form as

$$\begin{cases} P(t_n + ch) = \varphi(c)\mathbf{y}_n^{(r)} + zAP(t_n + ch) + z^2\overline{A}P(t_n + ch), \\ \qquad y_{n+1} = \varphi(1)\mathbf{y}_n^{(r)} + zv^T P(t_n + ch) + z^2 w^T P(t_n + ch). \end{cases} \tag{19}$$

Thus by setting $Q = \mathbf{I} - zA - z^2\overline{A}$, the first relation in (19) leads to

$$P(t_n + ch) = Q^{-1}\varphi(c)\mathbf{y}_n^{(r)},$$

where substituting it in (19) for computing y_{n+1} leads to

$$y_{n+1} = \Big(\varphi(1) + (zv^T + z^2 w^T)Q^{-1}\varphi(c)\Big)\mathbf{y}_n^{(r)}. \tag{20}$$

This relation is equivalent by the first row of (16) and the proof is complete. □

The coefficient matrix in (16) is shown by $R(z) \in \mathbb{C}^{(r+1)\times(r+1)}$, and it is called the stability matrix of the method. So the stability function is obtained by

$$p(\omega, z) = \det(\omega I - R(z)). \tag{21}$$

The stability properties of (2) with respect to the standard test equation (14) are dependent on the corresponding stability function $p(\omega, z)$.

Now, by multiplying the stability function (21) by its denominator, the stability polynomial of the method is obtained, which will be shown by the same symbol $p(w, z)$. Therefore, the stability properties of the corresponding methods depend on the obtained polynomial $p(w, z)$. The region of absolute

stability, $\mathcal{A}$, is defined to be a region of the complex z-plane such that the roots of $p(\omega, z) = 0$ lie within the unit circle whenever z lies in the interior of the region. To obtain the region of absolute stability, we use the boundary locus method [15]. Inserting $w = e^{i\theta}$, the roots of stability polynomial describe the bound of stability region.

5 Construction of A-stable SDMCMs

In this section, we focus our attentions on two steps and three steps continuous methods, $(r = 2, 3)$ with one and two collocation parameters. The polynomials of the methods are constructed by solving the linear system (8). Also, one can construct these polynomials by solving the system of order conditions (11) by $p = 2m + r - 1$. Then, by considering the stability matrix (16), collocation parameters in the interval $[0, 1]$ are determined in order to achieve A-stability.

5.1 Construction of SDMCMs with $r = 2$ and $m = 1$

In this subsection, we first investigate two step new methods with $r = 2$ and one collocation parameter of order $p = 2m + r - 1 = 3$. The coefficients of these methods are

$$\varphi_0(s) = \frac{s^3 - 3cs^2 + 3c^2 s + 3c^2 + 3c + 1}{3c^2 + 3c + 1},$$

$$\varphi_1(s) = -\frac{s(s^2 - 3cs + 3c^2)}{3c^2 + 3c + 1},$$

$$\psi_1(s) = \frac{(-s^2 + 3cs + 3c + 1)s}{3c^2 + 3c + 1},$$

$$\chi_1(s) = \frac{s((2c + 1)s^2 - (3c^2 + 1)s - 3c^2 - 2c)}{2(3c^2 + 3c + 1)}.$$

The stability polynomial of the method is written in the form

$$p(w, z) = \sum_{i=1}^{3} p_i(z) w^i$$

$$
\begin{aligned}
=&\frac{1}{6c^2+6c+2}\Big(\Big((c^4+2c^3+c^2)z^2+(-4c^3-6c^2-2c)z+6c^2+6c+2\Big)w^3\\
&+\Big((-2c^4+4c^2-2)z^2+(8c^3-8c)z-12c^2-4\Big)w^2\\
&+\Big((c^4-2c^3+c^2)z^2+(-4c^3+6c^2-2c)z+6c^2-6c+2\Big)w\Big).
\end{aligned}
$$

By performing an advanced search based on the boundary locus method, we find that there are some A-stable methods within this class of methods, for example, when the collocation parameter c lies in the interval $[0.578, 1]$, the obtained method is the A-stable of order 3. For example, by choosing $c = 1$, an A-stable method of order 3 is obtained where the polynomials of methods are in the form

$$
\varphi_0(s) = \frac{s^3-3s^2+3s+7}{7}, \quad \varphi_1(s) = \frac{-s^3+3s^2-3s}{7},
$$
$$
\psi_1(s) = \frac{-s^3+3s^2+4s}{7}, \quad \chi_1(s) = \frac{3s^3-2s^2-5s}{14}.
$$

5.2 Construction of SDMCMs with $r = 2$ and $m = 2$

In this subsection, we describe the construction of the method with $r = 2$ and two collocation parameters. The polynomials of method are of degree 5 and can be obtained by solving the system of order condition (11) with $p = 2m + r - 1 = 5$. The stability polynomial of the method is obtained in the form

$$
p(w, z) = \sum_{i=1}^{3} p_i(z) w^i,
$$

where $p_i(z)$ are polynomials of degree 2. By an extensive computer searching, we could not found A-stable methods for $c_1, c_2 \in [0, 1]$, but in some cases, an extensive stability region (near to A-stability) can be observed.

For example, choosing $c = [\frac{1}{2}, 1]$ leads to method of order 5 with unbounded stability region, where its stability region is plotted in Figure 1, and its coefficients are determined by

$$
\varphi_0(s) = 1 + \frac{15}{182}s - \frac{45}{182}s^2 + \frac{5}{14}s^3 - \frac{45}{182}s^4 + \frac{6}{91}s^5,
$$

$$\varphi_1(s) = \frac{-15}{182}s + \frac{45}{182}s^2 - \frac{5}{14}s^3 + \frac{45}{182}s^4 - \frac{6}{91}s^5,$$
$$\psi_1(s) = \frac{-124}{91}s + \frac{372}{91}s^2 - \frac{4}{7}s^3 - \frac{356}{91}s^4 + \frac{192}{91}s^5,$$
$$\psi_2(s) = \frac{415}{182}s - \frac{699}{182}s^2 + \frac{3}{14}s^3 + \frac{757}{182}s^4 - \frac{198}{91}s^5,$$
$$\chi_1(s) = \frac{-209}{182}s + \frac{263}{182}s^2 + \frac{5}{14}s^3 - \frac{283}{182}s^4 + \frac{62}{91}s^5,$$
$$\chi_2(s) = \frac{-149}{364}s + \frac{265}{364}s^2 - \frac{3}{28}s^3 - \frac{281}{364}s^4 + \frac{43}{91}s^5.$$

Also, the stability polynomial of this method is

$$\begin{aligned} p(\omega, z) = \ & \tfrac{\omega}{2912}(18z^4 - 150z^3 + 727z^2 - 2120z + 2912)\omega^2 \\ & + \tfrac{\omega}{2912}\left((-72z^2 - 768z - 2944)\omega + z^2 + 8z + 32\right). \end{aligned}$$

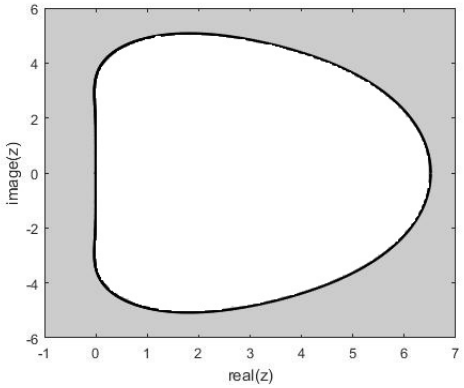


Figure 1: Stability region (Shaded region) of method with $r = 2$, $m = 2$ and $c = [\frac{1}{2}, 1]$.

5.3 Construction of SDMCMs with $r = 3$ and $m = 1$

We now consider 3-step SDMCMs with one collocation parameter of order $p = 4$. Solving the system of order condition (11) corresponding to $p = 4$, we obtain the family of methods of order 4 depending on the parameter c. We apply the boundary locus method on the stability polynomial $p(w, z)$, in

order to determine the values of the free parameters $c \in [0,1]$ achieving A-stability. We obtained when c belongs in $[0.619,1]$, the constructed methods are A-stable. Choosing, for example, for $c = \frac{7}{10}$, we obtain the three-step formula of uniform order $p = 4$ with coefficients given by

$$\varphi_0(s) = 1 + \frac{7}{10}s - \frac{81}{170}s^2 - \frac{11}{170}s^3 + \frac{19}{170}s^4,$$
$$\varphi_1(s) = -\frac{4}{5}s + \frac{42}{85}s^2 + \frac{12}{85}s^3 - \frac{13}{85}s^4,$$
$$\varphi_2(s) = \frac{1}{10}s - \frac{3}{170}s^2 - \frac{13}{170}s^3 + \frac{7}{170}s^4,$$
$$\psi_1(s) = \frac{2}{5}s + \frac{39}{85}s^2 - \frac{1}{85}s^3 - \frac{6}{85}s^4,$$
$$\chi_1(s) = \frac{-1}{5}s - \frac{29}{170}s^2 + \frac{8}{85}s^3 + \frac{11}{170}s^4.$$

5.4 Construction of SDMCMs with $r = 3$ and $m = 2$

We now present the construction for SDMCMs of uniform convergence order 6 with $r = 3$ and $m = 2$. First, by solving the order conditions (11), the polynomials of a method of degree 6 are determined depending on the parameters c_1 and c_2. The stability polynomial of the method is obtained in the form

$$p(w,z) = \sum_{i=1}^{4} p_i(z) w^i,$$

where $p_i(z)$ are polynomials of degree 2. We apply the boundary locus method on the stability polynomial $p(w,z)$, in order to determine the values of the free parameters $c_1, c_2 \in [0,1]$ achieving A-stability. The range of (c_1, c_2) in the domain $[0,1] \times [0,1]$, which leads to A-stable methods, is plotted in Figure 2.

For example, by choosing $c = [\frac{1}{2}, 1]$, an A-stable method of order $p = 6$ is obtained, where coefficients are

$$\varphi_0(s) = \frac{13216s^6 - 21564s^5 - 33123s^4 + 101259s^3 - 87639s^2 + 32517s + 219758}{219758},$$

$$\varphi_1(s) = \frac{-14672s^6 + 25536s^5 + 30786s^4 - 103768s^3 + 91308s^2 - 34104s}{219758},$$

$$\varphi_2(s) = \frac{1456s^6 - 3972s^5 + 2337s^4 + 2509s^3 - 3669s^2 + 1587s}{219758},$$

$$\psi_1(s) = \frac{77504s^6 + 20400s^5 - 305456s^4 + 70768s^3 + 253872s^2 - 65248s}{109879},$$

$$\psi_2(s) = \frac{-83384s^6 - 11604s^5 + 323186s^4 - 120143s^3 - 211887s^2 + 159662s}{109879},$$

$$\chi_1(s) = \frac{2744s^6 - 680s^5 - 11128s^4 + 8296s^3 + 7520s^2 - 8480s}{9989},$$

$$\chi_2(s) = \frac{33320s^6 + 12944s^5 - 116167s^4 + 33872s^3 + 76025s^2 - 53638s}{219758}.$$

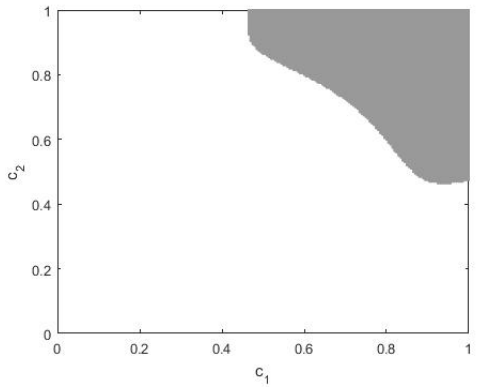


Figure 2: Acceptable (c_1, c_2) pairs for A-stability of SDMCMs with $r = 3,\ m = 2$.

6 Numerical eexperiments

In this section, we present numerical results arising from the application of the SDMCMs in order to confirm the theoretical expectations, show the efficiency and accuracy of the new method, and validate the order of these methods in the integration of stiff systems. In what follows, we describe the details of the implemented methods:

- **Method 1:** A-stable SDMCM of convergence order 3 with $r = 2,\ m = 1$, and collocation parameter $c = 1$.

- **Method 2:** SDMCM of convergence order 5 with $r = 2,\ m = 2$, and collocation parameters $c_1 = \frac{5}{10},\ c_2 = 1$.
- **Method 3:** A-stable SDMCM of convergence order 4 with $r = 3,\ m = 1$, and collocation parameter $c = 1$.
- **Method 4:** A-stable SDMCM of convergence order 6 with $r = 3,\ m = 2$, and collocation parameters $c = [\frac{1}{2}, 1]$.

Also, we compared the results with the method in [6, 8], by the implementation of the method.

- **Method 5:** Second derivative two-step collocation method with $m = 1$, $p = 4m = 4$, $c = \frac{2}{3}$ and $q = 1$, introduced in [8].
- **Method 6:** Two-step almost collocation method with $m = 2$, $p = 2m = 4$, $c = [\frac{1}{2}, \frac{9}{10}]$, introduced in [6].

Computational experiments are done by applying the methods to the following problems:

P1. The nonlinear stiff ODE

$$\begin{cases} y_1'(t) = -10004y_1(t) + 10000y_2(t)^2, & y_1(0) = 1, \\ y_2'(t) = y_1(t) - y_2(t)(1 + y_2(t)^3), & y_2(0) = 1, \end{cases}$$

with $t \in [0, 1]$. The exact solution is $[y_1(t) \quad y_2(t)]^T = [e^{-4t} \quad e^{-t}]^T$.

P2. The Robertson chemical kinetics problem [13]

$$\begin{cases} y_1'(t) = -0.04y_1(t) + 10^4 y_2(t)y_3(t), & y_1(0) = 1, \\ y_2'(t) = 0.04y_1(t) - 10^4 y_2(t)y_3(t) - 3 \times 10^7 y_2(t)^2, & y_2(0) = 0, \\ y_3' = 3 \times 10^7 y_2(t)^2, & y_3(0) = 0, \end{cases}$$

with $t \in [0, 1000]$.

P3. The Pleiades problem [16] is a celestial mechanics problem of seven stars in the plane of coordinates x_i and y_i and masses $m_i = i$ for $i = 1, 2, \ldots, 7$. The problem consists of a system of 14 special second-order differential equations rewritten to a first-order form, thus providing a

system of ODEs of dimension 28. The formulation and data have been taken from [11]. The problem is in the form

$$z'' = f(z), \quad z(0) = z_0, \quad z'(0) = z'_0,$$

with $z \in \mathbb{R}^{14}$ and $0 \leq t \leq 3$. By considering $z = (x^T, y^T)^T$, $x, y \in \mathbb{R}^7$, the function $f : \mathbb{R}^{14} \to \mathbb{R}^{14}$ is given by $f(z) = f(x, y) = (f^{(1)}(x, y), f^{(2)}(x, y))^T$, where $f^{(1)}, f^{(2)} : \mathbb{R}^{14} \to \mathbb{R}^7$ are

$$f_i^{(1)} = \sum_{j \neq i} m_j (x_j - x_i)/r_{ij}^{3/2},$$
$$f_i^{(2)} = \sum_{j \neq i} m_j (y_j - y_i)/r_{ij}^{3/2}, \quad i = 1, 2, \ldots, 7,$$

with $m_i = i$ and

$$r_{i,j} = (x_i - x_j)^2 + (y_i - y_j)^2.$$

We convert this problem to a first-order form by defining $w = z'$, which leads to a system of 28 nonlinear differential equations of the form

$$\begin{pmatrix} z' \\ w' \end{pmatrix} = \begin{pmatrix} w \\ f(z) \end{pmatrix},$$

where $(z^T, w^T)^T \in \mathbb{R}^{28}$ and $0 \leq t \leq 3$. The initial values are

$$\begin{pmatrix} z_0 \\ w_0 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \\ x'_0 \\ y'_0 \end{pmatrix}, \quad \begin{cases} x_0 = (3, 3, -1, -3, 2, -2, 2)^T, \\ y_0 = (3, -3, 2, 0, 0, -4, 4)^T, \\ x'_0 = (0, 0, 0, 0, 0, 1.75, -1, 5)^T, \\ y'_0 = (0, 0, 0, -1.25, 1, 0, 0)^T. \end{cases}$$

Table 1 presents the reference solution at the end of the integration interval, which is reported in [16].

In our numerical experiments, we apply Method 1– Method 4 to the given test problems with the fixed step sizes $h = T/N$ for several integer values of N. The global error of the methods at the endpoint of the interval of integration is listed by Err. Also, to verify theoretical results on the order of accuracy, we compute a numerical estimate to the order of accuracy of the methods by the formula $\log_2(\|Err_h(x)\|/\|Err_{h/2}(x)\|)$, where $Err_h(x)$ and $Err_{h/2}(x)$

Table 1: Reference solution at the end of the integration interval

x_1 =0.3706139143970502	y_1 = -3.943437585517392
x_2 = 3.237284092057233	y_2 = -3.271380973972550
x_3 = -3.222559032418324	y_3 =5.225081843456543
x_4 = 0.6597091455775310	y_4 =-2.590612434977470
x_5 = 0.3425581707156584	y_5 = 1.198213693392275
x_6=1.562172101400631	y_6 = -0.2429682344935824
x_7 = -0.7003092922212495	y_7 = 1.091449240428980
x_1' = 3.417003806314313	y_1' = −3.741244961234010
x_2' = 1.354584501625501	y_2' = 0.3773459685750630
x_3' = −2.590065597810775	y_3' = 0.9386858869551073
x_4' = 2.025053734714242	y_4' = 0.3667922227200571
x_5' = −1.155815100160448	y_5' = −0.3474046353808490
x_6' = −0.8072988170223021	y_6' = 2.344915448180937
x_7' = 0.5952396354208710	y_7' = −1.947020434263292

are the errors at the endpoint of the interval of integration corresponding to the step sizes h and $h/2$, respectively. we observed that the expected order was achieved. Also, by comparing results by methods introduced in [6, 8], one can see that the results are compatible with Method 5 and the solution by this new method is more precise than the results of Method 4 while the computational cost for the new method is less.

In problems P2, the reference solutions reported in [10] are used for computing the global error. The obtained results are reported in Table 2.

Table 2: The results of the methods for problem **P1**.

	N	4	8	16	32	64
Method 1	Err	2.22×10^{-4}	3.40×10^{-5}	4.64×10^{-6}	6.04×10^{-7}	7.71×10^{-8}
	$p(h)$		2.71	2.87	2.94	2.97
Method 2	Err	2.12×10^{-7}	8.38×10^{-9}	2.93×10^{-10}	9.66×10^{-12}	3.10×10^{-13}
	$p(h)$		4.66	4.84	4.92	4.96
Method 3	Err	1.95×10^{-5}	1.92×10^{-6}	1.39×10^{-7}	9.30×10^{-9}	5.99×10^{-10}
	$p(h)$		3.35	3.78	3.91	3.96
Method 4	Err	3.96×10^{-9}	1.01×10^{-10}	1.93×10^{-12}	3.33×10^{-14}	5.46×10^{-16}
	$p(h)$		5.31	5.70	5.86	5.93
Method 5	Err	1.17×10^{-3}	6.14×10^{-5}	2.58×10^{-6}	1.01×10^{-7}	5.96×10^{-9}
	$p(h)$		4.26	4.57	4.67	4.59
Method 6	Err	1.51×10^{-3}	1.74×10^{-4}	1.49×10^{-5}	1.08×10^{-6}	7.08×10^{-8}
	$p(h)$		3.11	3.54	3.79	3.93

Table 3: The results of SDMCMs for problem **P2**

	N	500	750	1000	1250	1500
Method 1	Err	4.41×10^{-5}	1.69×10^{-6}	8.22×10^{-6}	4.62×10^{-6}	2.85×10^{-7}
	$p(h)$		2.37	2.50	2.58	2.69
Method 2	Err	3.41×10^{-7}	7.41×10^{-8}	2.17×10^{-8}	8.07×10^{-9}	4.46×10^{-9}
	$p(h)$		3.76	4.27	4.42	4.64
Method 3	Err	1.99×10^{-6}	8.24×10^{-7}	3.99×10^{-7}	2.16×10^{-7}	7.81×10^{-8}
	$p(h)$		3.07	3.25	3.36	3.54
Method 4	Err	7.86×10^{-8}	1.67×10^{-8}	5.26×10^{-9}	1.58×10^{-9}	5.42×10^{-10}
	$p(h)$		4.84	4.97	5.37	5.90

Table 4: The results of SDMCMs for problem **P3**

	h	$\frac{1}{2} \times 10^{-3}$	$\frac{1}{4} \times 10^{-3}$	$\frac{1}{8} \times 10^{-3}$	$\frac{1}{16} \times 10^{-3}$
Method 1	Err	1.99×10^{-1}	2.49×10^{-2}	3.13×10^{-3}	3.90×10^{-4}
	$p(h)$		2.99	2.99	3.00
Method 2	Err	7.32×10^{-4}	2.31×10^{-5}	7.21×10^{-7}	2.47×10^{-8}
	$p(h)$		4.98	4.99	5.00
Method 3	Err	5.39×10^{-2}	2.52×10^{-3}	1.24×10^{-4}	6.67×10^{-4}
	$p(h)$		4.41	4.34	4.22
Method 4	Err	2.15×10^{-5}	3.71×10^{-7}	6.02×10^{-9}	9.46×10^{-11}
	$p(h)$		5.86	5.94	5.99

7 Conclusion

We have introduced a new family of SDMCMs. The constructed r-step method with m collocation parameters has a uniform order $2m + r - 1$. Examples of SDMCMs up to order 6 with the property of A-stability were constructed. The accuracy and efficiency of constructed methods were verified by solving some stiff problems. SDMCMs are efficient in solving stiff

problems, which are confirmed by numerical experiments. Future work will be in the investigation of G-simplecticity of SDMCMs [12, 17].

References

[1] Abdi, A. *Construction of high order quadratically stable second derivative general linear methods for the numerical integration of stiff ODEs*, J. Comput. Appl. Math. 303 (2016), 218–228.

[2] Abdi, A. and Hojjati, G. *Implementation of Nordsieck second derivative methods for stiff ODEs*, Appl. Numer. Math. 94 (2015), 241–253.

[3] Brunner, H. *Collocation methods for Volterra integral and related functional equations*, Cambridge University Press, 2004.

[4] Butcher, J.C. and Hojjati, G. *Second derivative methods with RK stability*, Numer. Algorithms. 40 (2005), 415–429.

[5] Cash, J.R. *Second derivative extended backward differentiation formulas for the numerical integration of stiff systems*, SIAM J. Numer. Anal. 18 (1981), 21–36.

[6] D'Ambrosio, R., Ferro, M., Jackiewicz, Z. and Paternoster, B. *Two-step almost collocation methods for ordinary differential equations*, Numer. Algorithms. 53 (2010), 195–217.

[7] Enright, W.H. *Second derivative multistep methods for stiff ordinary differential equations*, SIAM. J. Numer. Anal. 7 (1974), 321–331.

[8] Fazeli, S. and Hojjati, G. *Second derivative two-step collocation methods for ordinary differential equations*, Appl. Numer. Math. 156 (2020), 514–527.

[9] Ferguson, D. *Some interpolation theorems for polynomials*, J. Approx. Theory. 9 (1973), 327–348.

[10] Hairer, E.: https://www.unige.ch/~hairer/testset/testset.html.

[11] Hairer, E., Norsett, S.P. and Wanner, G. *Solving ordinary differential equations I: Nonstiff problems*, Springer-Verlag. second revised edition, 1993.

[12] Hairer, C.L.E. and Wanner, G. *Geometric numerical integration. Structure-preserving algorithms for ordinary differential equations*, Second Edition. Springer Series in Computational Mathematics 31, Springer-Verlag, 2006.

[13] Hairer, E. and Wanner, G. *Solving Ordinary Differential Equations II: Stiff and Differential Algebraic Problems*, Springer, Berlin, 2010.

[14] Lie, I. and Nørsett, S.P. *Superconvergence for multistep collocation*, Math. Comp. 52 (1989), 65–79.

[15] Liniger, W. and Willoughby, R.A. *Efficient numerical integration of stiff systems of ordinary differential equations*, Technical Report RC-1970, Thomas J. Watson Research Center, Yorktown Heihts, New York, 1976.

[16] Mazzia, F. and Magherini, C. *Test set for initial value problem solvers*, University of Bari. Italy, 2006.

[17] McLachlan, R.I. and Quispel, G.R.W. *Geometric integrators for ODEs*, J. Phys. A: Math. Gen. 39 (2006), 5251–5285.

[18] Schoenberg, I.J. *On Hermite-Birkhoff interpolation*, J. Math. Anal. Appl. 16 (1966), 538–543.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 2, 2024, pp 391–416
https://doi.org/10.22067/ijnao.2024.84415.1314
https://ijnao.um.ac.ir/

Research Article

Jarratt and Jarratt-variant families of iterative schemes for scalar and system of nonlinear equations

O. Ogbereyivwe*, E. J. Atajeromavwo and S. S. Umar

Abstract

This manuscript puts forward two new generalized families of Jarratt's iterative schemes for deciding the solution of scalar and systems of nonlinear equations. The schemes involve weight functions that are based on bi-variate rational approximation polynomial of degree two in both its numerator and denominator. The convergence study conducted on the

*Corresponding author

Received 12 September 2023; revised 29 November 2023; accepted 17 February 2024

Oghovese Ogbereyivwe
Department of Mathematics and Statistics, Delta State University of Science and Technology, Delta State, Nigeria. e-mail: ogho2015@gmail.com, oghovese.ogbereyivwe@dspz.edu.ng

Edafe John Atajeromavwo
Department of Software Engineering, Delta State University of Science and Technology, Delta State, Nigeria. e-mail: edafejohn2006@yahoo.com, edafejohn72@gmail.com

Salisu Shehu Umar
Department of Statistics, Federal Polytechnic, Auchi, Nigeria. e-mail: ssumar85@yahoo.com

How to cite this article
Ogbereyivwe, O., Atajeromavwo, E.J. and Umar, S.S., Jarratt and Jarratt-variant families of iterative schemes for scalar and system of nonlinear equations. *Iran. J. Numer. Anal. Optim.*, 2024; 14(2): 391-416. https://doi.org/10.22067/ijnao.2024.84415.1314

schemes, indicated that they have convergence order (CO) four in scalar space and retain the same number of CO in vector space. The numerical experiments conducted on the schemes when used to decide the solutions of some real-life nonlinear models show that they are good challengers of some well-known and robust existing iterative schemes.

AMS subject classifications (2020): Primary 65H05; Secondary 41A25.

Keywords: Jarratt scheme, Iterative scheme, Rational approximation polynomial, Nonlinear equation, System of nonlinear equation.

1 Introduction

A challenging task in the area of numerical computing is that of determining the solution of a scalar nonlinear equation $\Gamma(x) = 0$ and the system of nonlinear equation $\Gamma(X) = \mathbf{0}$. This is because, plethora of real life situations or systems, are often modeled into either scalar nonlinear or system of nonlinear equations. For instance, More [8] presented a compiled physical problems that are mostly modeled in the form of system of nonlinear equations. Problems in chemical equilibrium were modeled into nonlinear equations and studied by Shacham and Kehat [15]. Systems of nonlinear equations were applied to modeled problems in kinematics syntheses, neurophysiology, economics, and combustion engineering, which were considered and studied by Grosan and Abraham [3]. The nonlinear global positioning system problems modeled into nonlinear systems, were extensively exploited and solved by Yaseen, Zafar, and Alsulami [19]. More literature on physical occurrences or phenomena that are modeled into systems of nonlinear equations can be found in [6, 7, 9].

In order to have better insight of the problems that are modeled into nonlinear equations, the solutions of the models are usually desired to be known. Unfortunately, there is no general algebraic formula for solving all types of nonlinear equations. Consequent upon this, iterative schemes are designed to fill this gap. While many researchers have developed several schemes for dealing with the solutions of scalar nonlinear or system of nonlinear equations,

some of the schemes suffered some setbacks such as low convergence order (CO) and efficiency, non-convergence, divergence from true solution, breakdown, or inability to directly extend the schemes for solving scalar nonlinear equations to solving multidimensional nonlinear system and still conserving their CO.

The Jarratt scheme (JS) [4] presented as

$$x_{i+1} = x_i - \left[\frac{\Gamma'(x_i) + 3\Gamma'(y_i)}{6\Gamma'(y_i) - 2\Gamma'(x_i)}\right]\frac{\Gamma(x_i)}{\Gamma'(x_i)}, \qquad i = 0, 1, 2, \ldots, \tag{1}$$

where $y_i = x_i - \frac{2}{3}\frac{\Gamma(x_i)}{\Gamma'(x_i)}$ is one of the famous classical iterative scheme that is efficient, converges to solution with high CO, and can directly be extended to multidimensional form and still retain its CO. Many JS families have been put forward in literature. One of such families is that due to Behl, Kanwar, and Sharma [1] and presented as follows:

$$x_{i+1} = x_i - \frac{\left[(q_1^2 - 22q_1q_2 - 27q_2^2)\Gamma'(x_i) + 3(q_1^2 + 10q_1q_2 + 5q_2^2)\Gamma'(y_i)\right]\Gamma(x_i)}{2\left[q_1\Gamma'(x_i) + 3q_2\Gamma'(y_i)\right]\left[3(q_1+q_2)\Gamma'(y_i) - (q_1+5q_2)\Gamma'(x_i)\right]}, \tag{2}$$

where $q_i \in \Re$ such that $q_1 \neq q_2$ and $q_1 \neq -3q_2$. Kanwar, Kumar, and Behl [5] modified the family of the iterative scheme in (2) to deal with problems in multivariate case. Their scheme reads as follows:

$$\begin{aligned} Y_i &= X_i - \frac{2}{3}M(X_i) \\ X_{i+1} &= X_i - \frac{1}{2}\left[(q_1 I + 3q_2 N(X_i))\left(3(q_1+q_2)N(X_i) - (q_1+5q_2)I\right)\right]^{-1} \\ &\quad \times \left[(q_1^2 - 22q_1q_2 - 27q_2^2)I + 3(q_1^2 + 10q_1q_2 + 5q_2^2)N(X_i)\right]M(X_i), \\ N(X_i) &= \Gamma'(X_i)^{-1}\Gamma'(Y_i), \qquad M(X_i) = \Gamma'(X_i)^{-1}\Gamma'(X_i), \end{aligned} \tag{3}$$

where I is an identity matrix corresponding to the nonlinear system dimension.

In [2], another JS family was designed and analyzed using complex dynamics techniques to investigate whether behind the JS in the developed family. There are other JS variants that are computationally better in stability and numerical performance. After analysis, their work concluded that JS performed best among all the concrete members of the developed family.

Deriving motivation from the schemes put forward in (2) and (3), we propose two new families of Jarratt and Jarratt-variant schemes for solving nonlinear equations in this paper. The new families of schemes were derived based on the bi-variate rational approximation polynomial of degree two in both its numerator and denominator. The derived families of the schemes are of high CO and optimal as conjectured by Traub [18].

This paper is arranged as follows: Section 2 is meant for the development of the two families and their theoretical convergence analysis. Furthermore, some typical members of the families are noted in this section. Section 3 represents the extension of the families to multivariate cases and their convergence criteria. In Section 4, the application of some identified members of the developed families of iterative schemes to solve some real-life models is presented. Finally, the concluding remarks are made in Section 5.

2 The iterative schemes

We begin by acknowledging the following fundamental concepts.

Definition 1. [18, 10] Let x_* be the solution of $\Gamma(x) = 0$ and $d_i = |x_i - x_*|$ the error at ith iteration point of an iterative scheme $\psi(x_i)$. If an equation of the form $d_{i+1} = \Omega d_i^{\eta} + O(d_i^{\eta+1})$ can be derived from $\psi(x_i)$ via the Taylor expansions of the functions $\Gamma(\cdot)$ and functions derivatives $\Gamma'(\cdot)$ it contains, then d_{i+1} is known as the iterative scheme asymptotic error equation, η is CO, and Ω is asymptotic constant.

Definition 2. [18, 10] The efficiency of an iterative scheme $\psi(x_i)$ is the value $\eta^{\frac{1}{T}}$, where T is the sum of all different functions assessment in one iteration cycle.

Definition 3. [14] The computational CO (η_{coc}) of an iterative scheme $\psi(x_i)$ is the value obtained by using the formula

$$\eta_{coc} \approx \frac{\log \omega_1}{\log \omega_2}, \tag{4}$$

where $\omega_1 = \left|\frac{\Gamma(x_i)}{\Gamma(x_{i-1})}\right|$, $\omega_2 = \left|\frac{\Gamma(x_{i-1})}{\Gamma(x_{i-2})}\right|$, and x_{i-2}, x_{i-1} together with x_i are last three consecutive iteration results of $\psi(x_i)$.

The formulas in Definitions 1, 2, and 3 can easily be modified to handle cases in vector form.

2.1 The Jarratt's family (JF)

We consider the weighted scheme presented as follows:

$$x_{i+1} = x_i - P_{2,2}\left[\Gamma'(x_i), \Gamma'(y_i)\right] \frac{\Gamma(x_i)}{\Gamma'(x_i)}, \tag{5}$$

where

$$\begin{aligned} &P_{2,2}\left[\Gamma'(x_i), \Gamma'(y_i)\right] \\ &\quad = \frac{a_2\left[\Gamma'(x_i) - \Gamma'(y_i)\right]^2 + \Gamma'(y_i)\left[a_1\left(\Gamma'(x_i) - \Gamma'(y_i)\right) + \Gamma'(y_i)\right]}{a_4\left[\Gamma'(x_i) - \Gamma'(y_i)\right]^2 + \Gamma'(y_i)\left[a_3\left(\Gamma'(x_i) - \Gamma'(y_i)\right) + \Gamma'(y_i)\right]} \end{aligned} \tag{6}$$

is a bi-variate rational approximation polynomial of degree two in its numerator and denominator, $a_i (i = 1, 2, 3, 4) \in \Re$ are free parameters such that not all $a_i = 0$ and $y = x_i - \frac{2}{3}\frac{\Gamma(x_i)}{\Gamma'(x_i)}$. To determine the contribution of the scheme (5), some conditions are set on the parameters so as to ensure its convergence when implemented to solve scalar nonlinear equations.

Theorem 1. Suppose that $\Gamma : D \subset \Re \to \Re$ is a scalar function that is sufficiently differentiable in D such that $\Gamma'(\cdot) \neq 0$ in D. Furthermore, let $x_* \in D$ be a simple solution of Γ. If x_0 is in the neighborhood of x_*, then the sequence of solution approximations $\{x_i\}_{i\geq 0}, (x_n \in D)$ produced by the family of the scheme in (5), will converge to x_* with CO four subject to the conditions that $a_3 = \frac{4a_1 - 3}{4}$ and $a_4 = \frac{3 - 12a_1 + 16a_2}{16}$.

Proof. Applying the Taylor's expansion on $\Gamma(x)$ and $\Gamma'(x)$ around the solution x_* and setting $x = x_i$, then

$$\Gamma\left(x_i\right) = \Gamma'\left(x_*\right)\left[d_i + \sum_{n=2}^{4} c_n d_i^n + O\left(d_i^5\right)\right], \qquad i{=}0, 1, 2, \ldots, \tag{7}$$

and

$$\Gamma'\left(x_i\right) = \Gamma'\left(x_*\right)\left[1 + \sum_{n=2}^{4} n c_n d_i^{n-1} + O\left(d_i^5\right)\right], \tag{8}$$

holds for $c_n = \frac{1}{n!}\frac{\Gamma^{(n)}(x_*)}{\Gamma'(x_*)}, n \geq 2$.

Using the expansions in (7) and (8), the expression for y_i is as follows:

$$y_i = x_i - \frac{2}{3}\frac{\Gamma(x_i)}{\Gamma'(x_i)} = \frac{1}{3}d_i + \frac{2c_2}{3}d_i^2 - \frac{3(c_2^2 - c_3)}{3}d_i^3 + \frac{2(4c_2^2 - 7c_2c_3 + 3c_4)}{3}d_i^4 + \frac{4(4c_2^4 - 10c_2^2c_3 + 3c_3^2 + 5c_2c_4 - 2c_5)}{3}d_i^5 + O(d_i^6). \tag{9}$$

Now, the Taylor's expansion for $\Gamma(y_i)$ is

$$\Gamma(y_i) = \frac{1}{3}d_i + \frac{7c_2}{9}d_i^2 + \left(\frac{8c_2^2}{9} + \frac{37c_3}{27}\right)d_i^3 + \frac{2(10c_2^3 - 16c_2c_3 + 9c_4)}{9}d_i^4 + \frac{4\left(12c_2^4 - 27c_2^2c_3 + 8c_3^2 + 12c_2c_4 - 6c_5\right)}{9}d_i^5 + O(d_i^6), \tag{10}$$

and for $\Gamma'(y_i)$, follow the next equation:

$$\Gamma'(y_i) = 1 + \frac{2}{3}d_i + \frac{4c_2^2 - c_3}{3}d_i^2 + \left(\frac{-8c_2^3}{3} + 4c_2c_3\right)d_i^3 + \frac{4(4c_2^4 - 8c_2c_3 + 2c_3^2 + 3c_2c_4)}{3}d_i^4 + \frac{4\left(8c_2^5 - 20c_2^3c_3 + 9c_2c_3^2 + 10c_2^2c_4 - 3c_3c_4 - 4c_2c_5\right)}{3}d_i^5 + O(d_i^6). \tag{11}$$

The use of the expansions in (7), (8), and (11), enables the deduction of the expansion of the the weight function $P_{2,2}\left[\Gamma'(x_i), \Gamma'(y_i)\right]$ in (6) as follows:

$$\begin{aligned} &P_{2,2}\left[\Gamma'(x_i), \Gamma'(y_i)\right]\frac{\Gamma(x_i)}{\Gamma'(x_i)} \\ &= d_i + \left(\frac{-3 + 4a_1 - 4a_3}{3}\right)d_i^2 \\ &\quad + \frac{1}{9}\left(\begin{array}{r} 2(9 + 8a_a + 16a_3 + 8a_3^3 - 8a_1(2 + a_3) - 8a_4)c_2^2 \\ + 3(-3 + 4a_1 - 4a_3)c_3 \end{array}\right)d_i^3 \\ &\quad + \frac{1}{27}\left(\begin{array}{r} 4(-27 - 49a_3 - 52a_3^2 - 16a_3^3 - 4a_2(13 + 4a_3) \\ + \cdots + 27(-3 + 4a_1 - 4a_3)c_4 \end{array}\right)d_i^4 \\ &\quad + \frac{1}{81}\left((8(81 + 127a_3 + 210a_3^2 + 144a_3^3 + 32a_3^4 - \cdots + 9(4 - 5a_1 + 5a_3)c_5\right)d_i^5 \\ &\quad + O(d_i^6). \end{aligned} \tag{12}$$

The substitution of (12) into (5) yields

$$\begin{aligned} x_{i+1} =& x_i - P_{2,2}\left[\Gamma'(x_i), \Gamma'(y_i)\right] \frac{\Gamma(x_i)}{\Gamma'(x_i)} \\ =& x_* - d_i - d_i + \left(\frac{-3 + 4a_1 - 4a_3}{3}\right) d_i^2 \\ &+ \left(\frac{2((9 + 8a_2 + 16a_3 + 8a_3^3 - 8a_1(2 + a_3) - 8a_4)c_2^2 + 3(-3 + 4a_1 - 4a_3)c_3}{9}\right) d_i^3 \\ &+ \left(\frac{(4(-27 - 49a_3 - 52a_3^2 - 16a_3^3 - \cdots + 27(-3 + 4a_1 - 4a_3)c_4}{27}\right) d_i^4 \\ &+ \left(\frac{(8(81 + 127a_3 + 210a_3^2 + 144a_3^3 - \cdots + 9(4 - 5a_1 + 5a_3)c_5}{81}\right) d_i^5 + O(d_i^6). \end{aligned} \tag{13}$$

The expression in (13) will reduce to equation of order four if the next set of equations hold:

$$\begin{aligned} &-3 + 4a_1 - 4a_3 = 0, \\ &9 + 8a_2 + 16a_3 + 8a_3^3 - 8a_1(2 + a_3) - 8a_4 = 0. \end{aligned} \tag{14}$$

The equations in (14) are satisfied when

$$\begin{aligned} a_3 &= \frac{4a_1 - 3}{4}, \\ a_4 &= \frac{3 - 12a_1 + 16a_2}{16}. \end{aligned} \tag{15}$$

Consequently, by the application of (15) in (13), it reduces to

$$\begin{aligned} x_{i+1} =& x_* + \left(\frac{(10 - 4a_1 + 16a_2)c_2^3 - 9c_2c_3}{9}\right) d_i^4 \\ &+ \frac{1}{108}\left(\begin{array}{l} 4(-121 + 16a_1^2 + a_1(48 - 64a_2) - 208a_2)c_2^4 \\ \quad - 72(-13 + 4a_1 - 16a_2)c_2^2c_3 - 324c_2c_4 + 27(-8c_3^2 + c_5) \end{array}\right) d_i^5 \\ &+ O\left(d_i^6\right). \end{aligned} \tag{16}$$

By Definition 1, the equation in (16) is the error equation of the scheme in (5) and therefore, has minimum CO four. □

Remark 1. When the conditions in (15) are substituted into the scheme in (5) and after some simplification, a new two-free-parameter family of the Jarratt scheme (JF) is presented as follows:

$$x_{i+1} = x_i + P_{2,2}\left[\Gamma'(x_i), \Gamma'(y_i)\right] \frac{\Gamma(x_i)}{\Gamma'(x_i)}, \tag{17}$$

where $P_{2,2}\left[\Gamma'(x_i), \Gamma'(y_i)\right]$ expression is

$$P_{2,2}\left[\Gamma'(x_i), \Gamma'(y_i)\right] = \left[\frac{16\left(a_2\left[\Gamma'(x_i)-\Gamma'(y_i)\right]^2+\Gamma'(y_i)\left[a_1\left(\Gamma'(x_i)-\Gamma'(y_i)\right)+\Gamma'(y_i)\right]\right)}{\left(-3+12a_1-16a_2\right)\Gamma'(x_i)^2+2\left(9-20a_1+16a_2\right)\Gamma'(x_i)\Gamma'(y_i)+\left(-31+28a_1-16a_2\Gamma'(y_i)^2\right)}\right] \quad (18)$$

The JF require one $\Gamma(\cdot)$ and two $\Gamma'(\cdot)$ evaluations in one cycle of iteration and by Definition 2, the family has $E.I = 1.5874$.

2.1.1 The JF Particular cases

Here, for some specific values of a_1 and a_2 imposed on the scheme in (17), a particular case of the developed JF can be put forward. Some typical cases are provided next.

Case 1: For $a_1 = \frac{1}{4}$ and $a_2 = 0$, the famous JS is rediscovered as follows:

$$x_{i+1} = x_i + \left[\frac{\Gamma'(x_i)+3\Gamma'(y_i)}{2\Gamma'(x_i)-6\Gamma'(y_i)}\right]\frac{\Gamma(x_i)}{\Gamma'(x_i)}. \quad (19)$$

Case 2: For $a_1 = \frac{3}{4}$ and $a_2 = \frac{3}{8}$, a new IS (JF_1) is put forward as follows:

$$x_{i+1} = x_i - \left[\frac{5}{8}+\frac{3\Gamma'(x_i)^2}{8\Gamma'(y_i)^2}\right]\frac{\Gamma(x_i)}{\Gamma'(x_i)}. \quad (20)$$

Case 3: For $a_1 = a_2 = 1$, a new scheme (JF_2) is formed as follows:

$$x_{i+1} = x_i - 16\left[\frac{\Gamma'(x_i)^2-\Gamma'(x_i)\Gamma'(y_i)+\Gamma'(y_i)^2}{7\Gamma'(x_i)^2-10\Gamma'(x_i)\Gamma'(y_i)+19\Gamma'(y_i)^2}\right]\frac{\Gamma(x_i)}{\Gamma'(x_i)}. \quad (21)$$

Case 4: For $a_1 = a_2 = 0$, a new scheme (JF_3) is constructed as follows:

$$x_{i+1} = x_i - \left[\frac{16\Gamma'(y_i)^2}{3\Gamma'(x_i)^2-18\Gamma'(x_i)\Gamma'(y_i)+31\Gamma'(y_i)^2}\right]\frac{\Gamma(x_i)}{\Gamma'(x_i)}. \quad (22)$$

2.2 The Jarratt-variant family (JVF)

Here an iterative scheme that is a variant of the Jarratt's scheme family put forward in (17) is presented as follows:

$$x_{i+1} = x_i - R_{2,2}\left[\Gamma'(x_i), \Gamma'(y_i)\right] \frac{\Gamma'(x_i)}{\Gamma'(x_i)}, \tag{23}$$

where

$$R_{2,2}\left[\Gamma'(x_i), \Gamma'(y_i)\right] = \left[\frac{\Gamma'(y_i)^2 + b_1\Gamma'(x_i)\Gamma'(y_i) + b_2\Gamma'(x_i)^2}{\Gamma'(y_i)^2 + b_3\Gamma'(x_i)\Gamma'(y_i) + b_4\Gamma'(x_i)^2}\right], \tag{24}$$

is a degree two rational polynomial and $b_i (i = 1, 2, 3, 4) \in \Re$ are free parameters and not all equal zero. Next, the convergence condition for the IS (23) is investigated.

Theorem 2. Under the hypothesis made on the function $\Gamma(x)$ in Theorem 1, the IS in (23) will be of CO four when $b_2 = \frac{5+b_1}{3}$, $b_3 = 2(1 + b_1)$ and $b_4 = \frac{-1-2b_1}{3}$ provided $b_1 \neq -2$.

Proof. Assume that the expansions in (7)–(11) holds. Then

$$\begin{aligned}
&R_{2,2}\left[\Gamma'(x_i), \Gamma'(y_i)\right] \frac{\Gamma'(x_i)}{\Gamma'(x_i)} \\
&= \left(\frac{1 + b_1 + b_2}{1 + b_3 + b_4}\right) d_i \\
&\quad - \frac{1}{3\theta^2}\left(3 + 7b_3 - b_2(5 + b_3 - 3b_4) + 11b_4 + b_1(-1 + 3b_3 + 7b_4)\right) d_i^2 \\
&\quad + \frac{1}{9\theta^3}\begin{pmatrix} 2((9 + 34b_3 + 33b_3^2 + 42b_4 + 90b_3b_4 + 65b_4^2 - b_2(15 + 7b_3^2 \\ + 38b_4 - 9b_4^2 + 6b_3(5 + b_4)) + b_1(-7 + 9b_3^2 - 6b_4 + 33b_4^2 \\ + b_3(-6 + 34b_4)))c_2^2 + 3(-3 - 7b_3 + b_1(1 - 3b_3 - 7b_4) \\ + b_2(5 + b_3 - 3b_4) - 11b_4)(1 + b_3 + b_4)c_3) \end{pmatrix} d_i^3 \\
&\quad + \frac{1}{27\theta^4}\begin{pmatrix} -(-4(27 + 130b_3 + 231b_3^2 + 144b_3^3 + \cdots - b_2(19 + 22b_3^3 + \cdots \\ 23b_3^2(5 + 3b_4) + \cdots + 6b_2(5 + b_3 - 3b_4) - 11b_4)(1 + b_3 + b_4)^2c_4)) \end{pmatrix} d_i^4 \\
&\quad + O(d_i^5).
\end{aligned} \tag{25}$$

where $\theta = 1 + b_3 + b_4$.

Our expectation here, is to reduce the error equation in (25) to order four. This suffices to solving the following set of equations:

$$\begin{aligned} b_1 + b_2 - b_3 - b_4 &= 0, \\ 4b_1 - b_3 + 3b_4 &= -3, \\ 2b_1 + 3b_4 &= -1. \end{aligned} \tag{26}$$

The solution set satisfying (26) is

$$b_2 = \frac{5 + b_1}{3}, \quad b_3 = 2(1 + b_1), \quad b_4 = \frac{-1 - 2b_1}{3}. \tag{27}$$

The substitution of (27) in (25) produces

$$R_{2,2}\left[\Gamma'(x_i), \Gamma'(y_i)\right] \frac{\Gamma'(x_i)}{\Gamma'(x_i)} = d_i + \left(\frac{c_2((10 + 3b_1)c_2^2 - 3(2 + b_1)c_3)}{2(2 + b_1)} \right) d_i^4 + O(d_i^5). \tag{28}$$

When (28) is substituted in (23), we have

$$x_{i+1} = x^* + \left(\frac{c_2((10 + 3b_1)c_2^2 - 3(2 + b_1)c_3)}{3(2 + b_1)} \right) d_i^4 + O(d_i^5). \tag{29}$$

From Definition 1, the equation (29) indicates that the IS in (23) has CO four. □

Remark 2. When the parameters b_2 and b_3 are as defined in (27), then the scheme in (23) will reduce to a new one-parameter JVF of schemes that are of CO four and presented as follows:

$$x_{i+1} = x_i - \left[\frac{3\Gamma'(y_i)^2 + 3b_1\Gamma'(x_i)\Gamma'(y_i) + (5 + b_1)\Gamma'(x_i)^2}{3\Gamma'(y_i)^2 + 6(1 + b_1)\Gamma'(x_i)\Gamma'(y_i) - (1 + 2b_1)\Gamma'(x_i)^2} \right] \frac{\Gamma(x_i)}{\Gamma'(x_i)}. \tag{30}$$

Again, the family in (30) requires three assessment of functions in an iteration cycle. Therefore, $E.I = 1.5874$.

2.2.1 Some JV family concrete members

Here, for specific values for b_1, a particular case of (30) can be put forward. Consider the following cases:

Case 1: Put $b_1 = -\frac{10}{3}$ in (30). A new scheme (JVF_1) of CO four is obtained as follows:

$$x_{i+1} = x_i - \left[\frac{5\Gamma'(x_i)^2 - 30\Gamma'(x_i)\Gamma'(y_i) + 9\Gamma'(y_i)^2}{17\Gamma'(x_i)^2 - 42\Gamma'(x_i)\Gamma'(y_i) + 9\Gamma'(y_i)^2} \right] \frac{\Gamma(x_i)}{\Gamma'(x_i)}. \tag{31}$$

Case 2: For $b_1 = -5$ in (30), a new scheme (JVF_2) of CO four is determined as follows:

$$x_{i+1} = x_i - \left[\frac{\Gamma'(y_i)\left(\Gamma'(y_i) - 5\Gamma'(x_i)\right)}{\Gamma'(y_i)^2 - 8\Gamma'(x_i)\Gamma'(y_i) + 3\Gamma'(x_i)^2}\right]\frac{\Gamma(x_i)}{\Gamma'(x_i)}. \tag{32}$$

3 Extension of the families to *n*-dimensional form

In this section, the developed JF and JVF schemes were extended to solving the n-dimensional system of nonlinear equations $\Gamma(X) = \mathbf{0}$, where $\Gamma : \Theta \subseteq \Re^n \to \Re^n$ describes the dimension of the nonlinear system. Next, we show that the JF and JVF retain their CO when extended to solve n-dimensional nonlinear systems of equations. We define $X_* = \left(x_*^{(1)}, x_*^{(2)}, \ldots, x_*^{(n)}\right)^T$ as the solution of the $\Gamma(X) = \mathbf{0}$, and for an initial iteration starting point $X_0 = \left(x_0^{(1)}, x_0^{(2)}, \ldots, x_0^{(n)}\right)^T$ close to X_*, the equivalence of the developed families of schemes in n-dimensional form are as follows:

Scheme (17) (nJF):

$$X_{i+1} = X_i - A^{-1}B\left[M(X_i)\right], \tag{33}$$

where

$$A = \begin{bmatrix} \beta_1 I \\ +\,2\beta_2 N(X_i) \\ +\,\beta_3 N(X_i)^2 \end{bmatrix}^{-1}, \qquad B = \begin{bmatrix} -16 \begin{pmatrix} a_4(I - N(X_i))^2 + \\ (a_1\left(N(X_i) - N(X_i)^2\right) \\ + \qquad N(X_i)^2) \end{pmatrix} \end{bmatrix}, \tag{34}$$

$\beta_1 = 12a_1 - 16a_2 - 3$, $\beta_2 = -20a_1 + 16a_2 + 9$, $\beta_3 = 28a_1 - 16a_2 - 31$,
and
Scheme (23) (nJVF):

$$X_{i+1} = X_i - G^{-1}H\left[M(X_i)\right], \tag{35}$$

where

$$G = \begin{bmatrix} 3N(X_i)^2 \\ +\,6(1+b_1)N(X_i) \\ -\,(1+2b_1)I \end{bmatrix}^{-1}, \qquad H = \begin{bmatrix} 3N(X_i)^2 \\ +\,3b_1 N(X_i) \\ +\;\;(5+b_1)I \end{bmatrix}, \tag{36}$$

respectively.

We note that the $(i+1)$th iteration error of (33) and (35) is $E_{i+1} = \Omega(E_i^{\eta}) + O\left(E_i^{\eta+1}\right)$, is known as the error equation and η is the CO. Observe that $E_i^{\eta} = \left(E_i^{(1)}, E_i^{(2)}, \ldots, E_i^{(n)}\right)^{\eta}$. Next, we prove that the families of methods in (17) and (23) retain their CO in $\Re^n$ space.

Theorem 3. Assume that $\Gamma : \Delta \subseteq \Re^n \longrightarrow \Re^n$ is Frechet differentiable and that $\Gamma'(\cdot) \neq \mathbf{0}$ in the convex set Δ contains the solution X_* of $\Gamma(X) = \mathbf{0}$. For an initial guess X_0 close to X_*, the scheme in (33) will form a sequence $\{X_i\}_{i=0}^{\infty}$ of numerical results that converges to X_* with order four.

Proof. The Taylor's expansion of $\Gamma(X)$ around X_i up to the fourth order is

$$\Gamma(X) = \Gamma(X_i) + \sum_{k=1}^{4} \frac{1}{k!}\Gamma^{(k)}(X_i)(X - X_i)^k + O\left(\|X - X_i\|^5\right), \tag{37}$$

where $\Gamma^{(k)}(\cdot)$ is the kth-Frechet derivative of $\Gamma(\cdot)$.

Let $E_i = X_i - X_*$ be error at ith iteration point. Set $X = X_*$ in (37). Then

$$\Gamma(X_i) = \sum_{k=1}^{4} \left[(-1)^{k+1}\frac{1}{k!}\Gamma^{(k)}(X_i)\,(E_i)^k\right] + O\left(\|E_k\|^5\right). \tag{38}$$

Pre-multiplying the equation in (38) by $\Gamma(X_i)^{-1}$, we obtain

$$M(X_i) = E_i + \sum_{k=2}^{4} \left[(-1)^{k+1}\frac{1}{k!}\Gamma(X_i)^{-1}\Gamma^{(k)}(X_i)\,(E_i)^k\right] + O\left(\|E_i\|^5\right). \tag{39}$$

By applying (38) into the first step of the iteration cycle $Y_i = X_i - \frac{2}{3}M(X_i)$, we get

$$Y_i - X_i = \frac{2}{3}\left(-E_i + \sum_{k=2}^{4} \left[(-1)^{k}\frac{1}{k!}\Gamma(X_i)^{-1}\Gamma^{(k)}(X_i)\,(E_i)^k\right] + O\left(\|E_i\|^5\right)\right). \tag{40}$$

Consequently,

$$\begin{aligned}
&(Y_k - X_i)^2 \\
&= \frac{4}{9}E_i^2 - \frac{4}{9}\Gamma(X_i)^{-1}\Gamma''(X_i)E_i^3 \\
&\quad + \frac{1}{27}\Gamma(X_i)^{-1}\left[4\Gamma'''(X_i) + 3\Gamma''(X_i)\Gamma(X_i)^{-1}\Gamma''(X_i)\right]E_i^4 + O\left(\|E_i\|^5\right),
\end{aligned} \tag{41}$$

$$(Y_i - X_i)^3 = -\frac{8}{27}E_i^3 + \frac{4}{9}\Gamma(X_i)^{-1}\Gamma''(X_i)E_i^4 + O\left(\|E_i\|^5\right), \tag{42}$$

and

$$(Y_i - X_i)^4 = -\frac{16}{81}E_i^4 + O\left(\|E_i\|^5\right). \tag{43}$$

The Taylor's expression of $\Gamma(Y_i)$ around X_i is obtained as follows:

$$\Gamma'(Y_i) = \sum_{k=1}^{4}\left[\frac{1}{k!}\Gamma^{(k)}(X_k)\left(Y_i - X_i\right)^{k-1}\right] + O\left(\|Y_i - X_i\|^5\right). \tag{44}$$

A little simplification on (33), while taking note that $M(X_i) = \Gamma'(X_i)\Gamma(X_i)$, we have

$$A\Gamma'(X_i)E_{i+1} = A\Gamma'(X_i)E_i - B\Gamma(X_i). \tag{45}$$

Using (38) and (44), the next expansions were obtained:

$$\begin{aligned}
&A\Gamma'(X_i)E_i \\
&= -16\Gamma'''(X_i)E_i + \frac{8}{3}(11 - 4a_1)\Gamma'(X_i)^2\Gamma''(X_i)E_i^2 \\
&\quad + \frac{8}{9}\left(4(5a_1 - 2)(4 + a_2)\Gamma''(X_i)^2 + (4a_1 - 11)\Gamma'(X_i)\Gamma'''(X_i)\right) \\
&\quad + \frac{4}{81}\begin{pmatrix} 9(28a_1 - 16a_2 - 131)\Gamma'''(X_i)^3 + 3(92a_1 - 32a_2 \\ -161)\Gamma''(X_i)\Gamma'(X_i)\Gamma'''(X_i) + (4a_1 - 11)\Gamma'(X_i)^2\Gamma^{(iv)}(X_i) \end{pmatrix} E_i^4 \\
&\quad + O\left(\|E_i\|^5\right)
\end{aligned} \tag{46}$$

and

$$\begin{aligned}
&B\Gamma(X_i)E_i \\
&= -16\Gamma'''(X_i)E_i + \frac{8}{3}(11 - 4a_1)\Gamma'(X_i)^2\Gamma''(X_i)E_i^2 \\
&\quad + \frac{8}{9}\left(4(5a_1 - 2)(4 + a_2)\Gamma''(X_i)^2 + (4a_1 - 11)\Gamma'(X_i)\Gamma'''(X_i)\right) \\
&\quad + \frac{2}{81}\begin{pmatrix} 108(5a_1 - 4a_2 - 6)\Gamma'''(X_i)^3 + 24(23a_1 - 8a_2 - 38) \\ \times\Gamma''(X_i)\Gamma'(X_i)\Gamma'''(X_i) + (32a_1 - 91)\Gamma'(X_i)^2\Gamma^{(iv)}(X_i) \end{pmatrix} E_i^4 \\
&\quad + O\left(\|E_i\|^5\right).
\end{aligned} \tag{47}$$

When (46) and (47) are substituted into the right hand side of (45), we get

$$A\Gamma'(X_i)E_{i+1} = \frac{2}{27}\begin{bmatrix} 6(2a_1 - 8a_2 - 5)\Gamma'''(X_i) - \Gamma'(X_i)^2\Gamma^{(iv)}(X_i) \\ + 18\Gamma'(X_i)\Gamma''(X_i)\Gamma'''(X_i) \end{bmatrix} E_i^4 + O\left(\|E_i\|^5\right). \tag{48}$$

The error equation in (48) implies that the IS in (33) is of CO four. □

Theorem 4. Assume that the hypothesis on the function $\Gamma(X)$ holds as in Theorem 3. Then the scheme in (35) will produce sequence of numerical results that converge to X_* with CO four.

Proof. The proof follows the manner of the proof of Theorem **??**. Consequently, its error equation in $\Re^n$ space is

$$G\Gamma'(X_i)E_{i+1} = \frac{1}{54}\begin{bmatrix} 9(10 + 3b_1)\Gamma''(X_i)^3 - 18(2 + b_1)\Gamma'(X_i)\Gamma''(X_i)\Gamma'''(X_i) \\ + (2 + b_1)\Gamma'(X_i)^2\Gamma^{(iv)}(X_i) \end{bmatrix} E_i^4 + O\left(\|E_i\|^5\right). \tag{49}$$

□

4 Numerical experiments results

This section presents the experimentation of some concrete forms of the developed JF and JVF on the solution of nonlinear and systems of nonlinear equations. The numerical results obtained were compared with the results of some existing competitors that are also Jarratt's variants. To verify the theoretical CO of the JF, JV, nJF, and nJVF obtained in Sections 2 and 3, we use the computational CO (η_{coc}) given in Definition 3.

4.1 Numerical experiments with scalar nonlinear models

This subsection provides some numerical experiments conducted on the developed schemes when applied to solve some physical problems modeled into nonlinear equations.

Model application 1 (Chemical engineering, [11]). To determine the concentration of a chemical mixture at time x in a mixed reactor, it can be obtained by using the nonlinear equation

$$\Gamma_1(x) = -0.75\exp(-0.05x) + 1 = 0. \tag{50}$$

The nonlinear equation solution is $-5.7536414490\ldots$.

Model Application 2 (Chemical Reactor, [11, 12]). In a chemical reactor, the fractional species conversion x can be determined by the nonlinear equation

$$\Gamma_2(x) = \frac{8x^2\,(4-x)^2}{(2-x)(6-3x)^2} - 0.186 = 0. \tag{51}$$

The solution to the nonlinear equation is $0.2777575428\ldots$.

Model Application 3 (Projectile model, [17]). The nonlinear model governing an electron movement between two parallel plates is

$$\Gamma_3(x) = \frac{\pi}{4} + x - 0.5\cos x = 0. \tag{52}$$

Its solution is $-0.3094661392\ldots$.

Model Application 4 (Structural engineering, [17]). The nonlinear model that described stresses encountered by finite structures in underground is:

$$\Gamma_4(x) = -\frac{1}{4} + \frac{\sin x\cos x + x}{\pi} = 0. \tag{53}$$

The nonlinear model solution is $0.4160444988\ldots$.

Model Application 5 (Population model, [16, 11]). The population dynamics is described by using the differential equation

$$P'(x) = \delta P(x) + \mu, \tag{54}$$

where $P(x)$ is time x population, δ is population birth rate, and μ represent rate of immigration. The model solution is

$$P(x) = P_0\exp(\delta x) + \frac{\mu}{\delta}\left(\exp(\delta x) - 1\right), \tag{55}$$

where P_0 is initial population. A problem may arise to find to find the birth rate δ at the end of first year given that $P_0 = 100000, \mu = 435000$, and 1564000 is population added at the end of the year. This amounts to solving

the nonlinear equation

$$\Gamma_5(\delta) = P(\delta) = -\frac{435000}{\delta}\left(\exp(\delta) - 1\right) - 1000000\exp(\delta) + 1564000 = 0. \quad (56)$$

The solution of the nonlinear equation in (56) is $0.1009979296\ldots$.

The obtained numerical experimentation results when the developed JF and JVF members were applied to solve the nonlinear models above were presented in Tables 1 and 2. The computation results in terms of the number of iterations (N), residual error $|\Gamma(x)|$, the total number of arithmetic operations with functions N_T, and computational order of convergence η_{coc}, were compared with results by some existing schemes taken from [1] (see (2)) using $q_1 = 0, q_2 = 0$ and denoted as (BS), while from [5](see (3)), taking $q_1 = 50, q_2 = \frac{1}{10}$ and denoted as (KS). The mpmath-PYTHON was utilized in designing computational program and $\Gamma(x_i) \leq 10^{-100}$ used as program terminal criterion. To reduce truncation error, all computations were set to 1000 decimal places accuracy. The expression $S.Te{-}U$ represents $S.T \times 10^{-U}$ where $S, T, U \in \Re$.

Table 1: Numerical results on models 1-2

Model	IS	x_0	N	N_T	$\lvert\Gamma(x_{i+1})\rvert$	η_{coc}	x_0	N	N_T	$\lvert\Gamma(x_{i+1})\rvert$	η_{coc}
	BS		5	80	3.3e-111	4.0		6	96	1.3e-323	4.0
	KS		4	84	1.5e-180	4.0		4	84	1.5e-136	4.0
	JS		4	60	2.7e-178	4.0		4	60	7.5e-134	4.0
	JF_1		4	64	8.4e-148	4.0		4	64	7.9e-104	4.0
	JF_2		4	84	5.7e-117	4.0		5	105	5.6e-295	4.0
$\Gamma_1(x)$	JF_3	5.0	4	76	1.5e-138	4.0	10.0	5	95	7.1e-307	4.0
	JVF_1		4	92	1.1e-169	4.0		4	92	5.8e-126	4.0
	JVF_2		4	80	9.3e-197	4.0		4	80	9.9e-135	4.0
	BS		5	80	2.4e-101	4.0		5	80	5.1e-101	4.0
	KS		5	105	3.3e-269	4.0		5	105	1.2e-196	4.0
	JS		5	75	5.4e-257	4.0		5	75	1.1e-193	4.0
	JF_1		5	80	2.2e-187	4.0		5	80	2.1e-147	4.0
	JF_2		5	105	1.8e-113	4.0		5	105	2.4e-103	4.0
$\Gamma_2(x)$	JF_3	0.1	5	95	5.8e-139	4.0	.78	4	76	7.6e-141	4.0
	JVF_1		5	115	1.7e-241	4.0		5	115	9.5e-173	4.0
	JVF_2		5	100	2.0e-139	4.0		5	100	2.2e-278	4.0

Table 2: Numerical results on models 3-5

Model	IS	x_0	N	N_T	$\lvert\Gamma(x_{i+1})\rvert$	η_{coc}	x_0	N	N_T	$\lvert\Gamma(x_{i+1})\rvert$	η_{coc}
	BS		4	64	1.6e-187	4.0		4	64	7.5e-154	4.0
	KS		4	84	7.8e-289	4.0		4	84	1.6e-193	4.0
	JS		4	60	2.0e-286	4.0		4	60	6.0e-193	4.0
	JF_1		4	64	2.1e-251	4.0		4	64	1.5e-179	4.0
	JF_2		4	84	4.3e-219	4.0		4	84	5.8e-162	4.0
$\Gamma_3(x)$	JF_3	-0.7	4	76	2.2e-250	4.0	0.5	4	76	3.0e-222	4.0
	JVF_1		4	92	6.2e-275	4.0		4	92	1.2e-188	4.0
	JVF_2		4	80	2.4e-304	4.0		4	80	5.9e-222	4.0
	BS		4	64	2.6e-192	4.0		5	80	1.9e-131	4.0
	KS		4	84	9.6e-203	4.0		4	84	3.8e-106	4.0
	JS		4	60	1.2e-202	4.0		4	60	2.6e-105	4.0
	JF_1		4	64	2.5e-200	4.0		5	80	1.3e-372	4.0
	JF_2		4	84	4.6e-196	4.0		5	105	2.2e-307	4.0
$\Gamma_4(x)$	JF_3	0.0	4	76	1.2e-209	4.0	0.9	5	95	1.7e-388	4.0
	JVF_1		4	92	4.9e-202	4.0		4	92	7.0e-103	4.0
	JVF_2		4	80	2.4e-202	4.0		4	80	1.5e-178	4.0
	BS		5	80	9.7e-218	4.0		6	96	2.0e-258	4.0
	KS		5	105	4.4e-391	4.0		5	105	1.7e-149	4.0
	JS		5	75	3.7e-387	4.0		5	75	2.6e-147	4.0
	JF_1		5	80	6.2e-304	4.0		6	96	1.2e-415	4.0
	JF_2		5	105	1.3e-227	4.0		6	126	1.2e-259	4.0
$\Gamma_5(x)$	JF_3	1.5	5	95	1.4e-418	4.0	3.0	5	95	1.7e-131	4.0
	JVF_1		5	115	4.2e-351	4.0		5	115	3.7e-124	4.0
	JVF_2		5	100	1.4e-132	4.0		5	100	3.8e-272	4.0

4.2 Numerical experiments with nonlinear multivariate models

This subsection put forward the experiments performed on some specific forms of the JF and JV to solve some nonlinear multivariate models. In all computational experiments, 100 digits floating point arithmetic and $\|\Gamma(X_i)\|_\infty < 10^{-25}$ as halting criteria were utilized in the computational program designed in sympy PYTHON environment. Methods performance was compared based on the number of iterations (N), function residual error $\|\Gamma(X_{i+1})\|$, computational order of convergence η_{coc}, efficiency E, and total

approximate computational cost TCC. In this case, the total computational cost TCC of a scheme is estimated as follows:

$$TCC = \Gamma'_C + \Gamma_C + PMM_C + PMV_C + PSM_C + MI_C + ASM_C + ASV_C, \quad (57)$$

where $\Gamma'_C, \Gamma_C, PMM_C, PMV_C, PSM_C, MI_C, ASM_C, ASV_C$ are approximate computational cost of evaluations in the matrix Γ', vector Γ, product of matrix with matrix (PMM), product of matrix with vector (PMV), product of scalar with matrix (PSM), matrix inverse MI, addition or subtraction of matrices ASM, and addition or subtraction of vectors (ASV), respectively. Here, we assumed that the cost of evaluating arithmetic operations of functions in matrices and vectors is equal. The costs associated with these operations are presented in Table 3.

Table 3: Cost of matrices operations

Operations	Cost
Γ'	m^2
Γ	m
PMM	$2m^3 - m^2$
PMV	$2m^2 - m$
PSM	m^2
MI	$\frac{2m^3}{3}$
ASM	m^2
ASV	m

In Table 4, information on the approximate number of computational operations evaluations in one cycle of thevarious schemes are given, while Table 5 provides the respective scheme's total approximate computational cost TCC expressed as a function of the dimension m of the system of equations.

Table 4: Schemes cost of matrices operations evaluations

Methods	Γ'	Γ	PMM	PMV	PSM	MI	ASM	ASV
nBS	2	1	1	2	3	2	2	2
nKS	2	1	4	2	7	2	3	2
nJS	2	1	1	2	3	2	2	2
nJF_1	2	1	3	2	2	2	1	2
nJF_2	2	1	4	2	4	2	4	2
nJF_3	2	1	4	2	4	2	2	2
nJV_1	2	1	4	2	6	2	4	2
nJF_2	2	1	4	2	3	2	3	2

Table 5: Methods order and TCC

Methods	Order	TCC
nBS	4	$\frac{10}{3}m^3+11m^2+m$
nKS	4	$\frac{28}{3}m^3+12m^2+m$
nJS	4	$\frac{10}{3}m^3+10m^2+m$
nJF_1	4	$\frac{22}{3}m^3+6m^2+m$
nJF_2	4	$\frac{28}{3}m^3+10m^2+m$
nJF_3	4	$\frac{28}{3}m^3+8m^2+m$
nJV_1	4	$\frac{28}{3}m^3+12m^2+m$
nJF_2	4	$\frac{28}{3}m^3+8m^2+m$

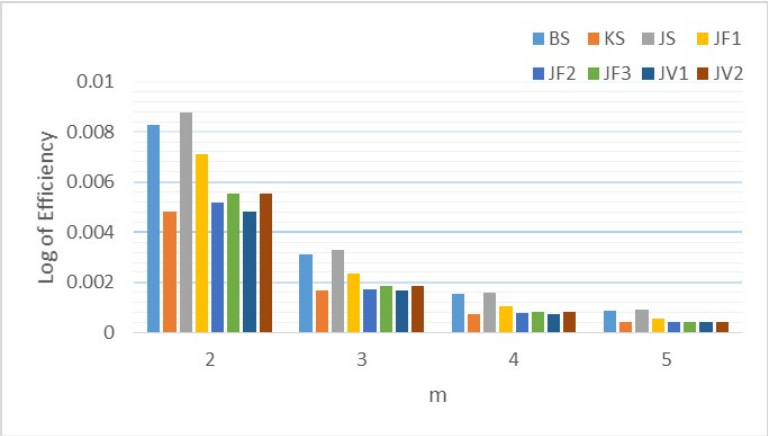


Figure 1: Schemes efficiency dynamics for $m = 2, \ldots, 5$

In line with Definition 1, we estimate the efficiencies of the schemes by $E = \eta^{\frac{1}{TCC}}$. Figures 1 and 2 show the dynamics of the schemes efficiencies as the size m changes. From Figures 1 and 2, we observe that the schemes efficiencies follow the order $JS > BS > JF_1 > JF_3 \geq JV_2 > JF_2 > JV_1 \geq KS$ when $m = 2$ to $m = 5$, and it retains this results for the case $m = 10$ to $m = 30$. The numerical results in Tables 6 and 7 verify these results. Although the compared schemes BS and KS have good efficiencies, they failed to converge to solutions for most problems tested. This is an indication that they have low stability compared to the developed schemes.

Model Application 6 (Chemical equilibrium [7]). The following set of nonlinear equations $\Gamma_6(X) = \mathbf{0}$ describes the chemical equilibrium system that involves the combination of carbon oxide (x_1), oxygen (x_2), Hydrogen

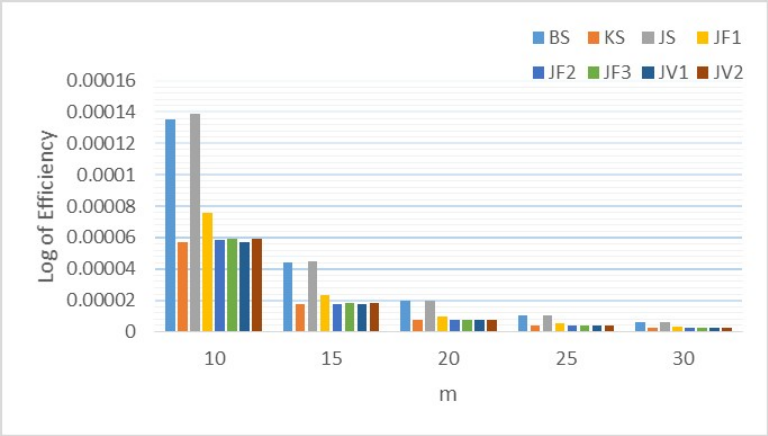


Figure 2: Schemes efficiency dynamics for $m = 10, \ldots, 30$

(x_3), Nitrogen (x_4), and x_5 a factor depicting the moles number of a product formed for each mole of consumed propane.

$$\begin{aligned}
x_1 + x_1x_2 - 3x_5 &= 0,\\
x_1 + 2x_1x_2 + x_2x_3^2 + \rho_8x_2 - \rho x_5 + \rho_{10}x_2^2 + \rho_7x_1x_3 + \rho_9x_2x_4 &= 0,\\
2x_2x_3^2 + 2\rho_5x_3^2 - 8x_5 + \rho_6x_3 + \rho_7x_2x_3 &= 0,\\
\rho_9x_2x_4 + 2x_4^2 - 4\rho_5 &= 0,\\
x_1\left(x_2+1\right) + \rho_{10}x_2^2 + x_2x_3^2 + \rho_8x_2 + \rho_5x_3^2 &\\
+ x_4^2 - 1 + \rho_6x_3 + \rho_7x_2x_3 + \rho_9x_2x_4 &= 0,
\end{aligned} \tag{58}$$

where $\rho = 10, \rho_{10} = 0.193, \rho_6 = \frac{0.002597}{\sqrt{40}}, \rho_7 = \frac{0.003448}{\sqrt{40}}, \rho_8 = \frac{0.00001799}{40}, \rho_9 = \frac{0.0002155}{\sqrt{40}}, \rho_{10} = \frac{0.00003846}{40}$. The model exact solution X_* in the domain $\Delta = (-1,1) \times (33.5, 35.5) \times (-1,1) \times (-0.8, 1.8) \times (-1,1)$ approximated to 20 decimal places using the initial start points $X_0^{(1)} = (0.6, 33.6, 0.6, 1.5, -0.7)^T$ and $X_0^{(2)} = (-0.2, 33.0, 0.9, 1.0, -0.3)^T$ is $X_* = (0.0031\ldots, 34.5979\ldots, 0.0650\ldots, 0.8593\ldots, 0.0369\ldots)^T$.

Model Application 7 (Combustion problem [9]). The modeled combustion problem at a temperature of 3000^0C in system of nonlinear equations $\Gamma_7(X) = \mathbf{0}$ was given as follows:

$$-(1.0e{-}15) + x_2 + 2x_6 + x_9 + 2x_{10} = 0, \quad -(3.0e{-}5) + x_3 + x_8 = 0,$$
$$-(5.0e{-}5) + x_1 + x_3 + 2x_5 + 2x_8 + x_9 + x_{10} = 0,$$
$$-(5.0e{-}5) + x_4 + 2x_7 = 0, \qquad -x_1^2 + (0.5140437e{-}7)x_5 = 0,$$
$$-2x_2^2 + (0.100006932e{-}6)x_6 = 0, \qquad -x_4^2 + (0.7816278e{-}15)x_7 = 0,$$
$$-x_1x_3 + (0.1496236e{-}6)x_8 = 0, \qquad -x_1x_2 + (0.6194411e{-}7)x_9 = 0,$$
$$-x_1x_2^2 + (0.2089296e{-}14)x_{10} = 0.$$

The solution to the model is

$X_* = (1.47e{-}7, 2.26e{-}7, 1.51e{-}5, 6.27e{-}11,$
$4.20e{-}7, 1.01e{-}6, 4.99e{-}6, 1.48e{-}5, 5.37e{-}7, 3.60e{-}6)^T$, when $X_0^{(1)} = (0.1,$ $0.4, 0.2, 0.3, 0.1, 0.6, 0.7, 0.5, 0.1, 0.4)^T$ and $X_0^{(2)} = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1,$ $0.1, 0.1, 0.1, 0.1)^T$ were used as iteration starting points.

Model Application 8 (Economics model [3]). Consider the modeling in economics that can be scaled up to n-dimensions and reads as follows:

$$\sum_{i=1}^{n-1} x_i = -1, \qquad \left[x_j + \sum_{i=1}^{n-j-1} x_i x_{i+j}\right] x_n = -1, \quad 1 \le j \le n-1. \tag{59}$$

Here, we set $n = 4$ to generate a system of nonlinear equations $\Gamma_8(X) = \mathbf{0}$ and used $X_0^{(1)} = \cos\left(\frac{1}{j}\right)$ and $X_0^{(2)} = \left(\frac{1}{j+4}\right)$ as starting points in two different applications. The model solutions is
$X_* = (1.4464\ldots, 1.9730\ldots, -4.4194\ldots, 0.2262\ldots)^T$.

Application 9 (Boundary value problem (BVP) [13]). Let the BVP

$$\Gamma'' + \theta^2 (\Gamma)^2 + 1 = 0, \quad \Gamma(0) = 0, \quad \Gamma(1) = 0. \tag{60}$$

The interval $[0, 1]$ is partitioned such that $x_0 = 0 < x_1 < x_2 < \cdots < x_{m-1} < x_m$, $x_{i+1} = x_i + h$, $h = \frac{1}{m}$.

Define $\Gamma_0 = \Gamma(x_0) = 0$, $\Gamma_1 = \Gamma(x_1) \cdots \Gamma_{m-1} = \Gamma(x_{m-1})$,$\Gamma_m = \Gamma(x_m) =$ 1. By discretizing the problem using finite difference such that

$$\Gamma' = \frac{\Gamma_{i+1} - \Gamma_{i-1}}{2h}, \qquad \Gamma'' = \frac{\Gamma_{i+1} - 2\Gamma_i + \Gamma_{i-1}}{h^2}, \tag{61}$$

the problem becomes

$$\Gamma_{i+1} - 2\Gamma_i + \Gamma_{i-1} + \frac{\theta^2}{4}(\Gamma_{i+1} - \Gamma_{i-1}) + h^2 = 0, \quad i = 1, 2, 3, \ldots, m-1. \tag{62}$$

For $m = 10$ and $\theta = 2$, the problem was solved by using $X_0 = \{0, 0, \ldots, 0\}^T$ as initial start point and applying $\|\Gamma(X_i)\| < 10^{-5}$ as stopping condition. The problem solution is

$X_* = (0.605\ldots, 0.101\ldots, 0.128\ldots, 0.144\ldots, 0.149\ldots, 0.144\ldots,$
$0.128\ldots, 0.101\ldots, 0.605\ldots)^T$.

The computation results obtains on Applications 6 to 9 are presented in Tables 6 and 7.

Table 6: Numerical results on Applications 6 -7

Model	IS	X_0	N	$\|\Gamma(x_{i+1})\|$	η_{coc}	TCC	E
	nBS			Failed			
	nKS			Failed			
	nJS		3	7.4e-07	4.0	2015	1.00069
	nJF_1		4	1.1e-22	4.0	4287	1.00032
	nJF_2		4	2.6e-18	4.0	5687	1.00024
$\Gamma_6(X)$	nJF_3	$X_0^{(1)}$	4	1.0e-24	4.0	5487	1.00025
	$nJVF_1$		3	1.4e-09	4.0	4415	1.00031
	$nJVF_2$		3	9.7e-08	4.0	4115	1.00031
	nBS			Failed			
	nKS			Failed			
	nJS		4	1.7e-17	4.0	2687	1.00052
	nJF_1		4	4.2e-19	4.0	4287	1.00032
	nJF_2		4	2.0e-15	4.0	5687	1.00024
$\Gamma_6(X)$	nJF_3	$X_0^{(2)}$	4	3.8e-17	4.0	5487	1.00025
	$nJVF_1$		3	3.8e-07	4.0	4415	1.00031
	$nJVF_2$		4	4.3e-21	4.0	5487	1.00025
	nBS			Failed			
	nKS			Failed			
	nJS		31	1.8e-25	4.0	50850	1.00002726
	nJF_1		32	1.0e-25	4.0	90123	1.00001538
	nJF_2		33	3.4e-25	4.0	122045	1.00001135
$\Gamma_7(X)$	nJF_3	$X_0^{(1)}$	32	5.0e-25	4.0	115211	1.00001203
	$nJVF_1$		29	2.0e-25	4.0	110094	1.00001259
	$nJVF_2$		31	1.6e-25	4.0	111610	1.00001242

4.3 Results discussion

From the results in Tables 1, 2, 6, and 7, the new schemes put forward solved all the presented models with good precision and consistent CO. The η_{coc} obtained for all the developed schemes are in agreement with the theoretical order of convergence derived in Sections 2 and 3. On the other hand, the methods of Behl, Kanwar, and Sharma [1] (nBS) and Kanwar, Kumar, and

Table 7: Numerical results on Applications 7 -9

Model	IS	X_0	N	$\|\Gamma(x_{i+1})\|$	η_{coc}	TCC	E
	nBS			Failed			
	nKS			Failed			
	nJS		29	2.2e-25	4.0	47570	1.00002914
	nJF_1		31	2.2e-25	4.0	87306	1.00001588
	nJF_2		33	1.1e-25	4.0	122045	1.00001136
$\Gamma_7(X)$	nJF_3	$X_0^{(2)}$	31	4.6e-25	4.0	111610	1.00001242
	$nJVF_1$		26	4.6e-26	4.0	98705	1.00001404
	$nJVF_2$		29	1.4e-29	4.0	104410	1.00001328
	nBS			Failed			
	nKS			Failed			
	nJS		9	8.8e-10	4.0	3396	1.00041
	nJF_1		6	6.2e-21	4.0	3416	1.00040
	nJF_2		6	4.5e-21	4.0	4568	1.00030
$\Gamma_8(X)$	nJF_3	$X_0^{(1)}$	10	1.2e-22	4.0	7293	1.00019
	$nJVF_1$		6	2.0e-11	4.0	4760	1.00029
	$nJVF_2$		9	6.1e-07	4.0	6564	1.00021
	nBS			Failed			
	nKS			Failed			
	nJS		16	2.3e-18	4.0	6037	1.00023
	nJF_1		5	1.6e-07	4.0	2847	1.00049
	nJF_2		16	2.0e-17	4.0	12181	1.00011
$\Gamma_8(X)$	nJF_3	$X_0^{(0)}$	26	3.1e-09	4.0	18963	1.00007
	$nJVF_1$		17	1.0e-10	4.0	13487	1.00010
	$nJVF_2$		9	3.9e-07	4.0	6564	1.00021
	nBS			Failed			
	nKS		6	1.3e-05	4.0	45252	1.00003064
	nJS		6	1.3e-05	4.0	18036	1.00007687
	nJF_1		6	1.3e-05	4.0	33588	1.00004127
	nJF_2		6	1.3e-05	4.0	44289	1.00003130
$\Gamma_9(X)$	nJF_3	$X_0^{(0)}$	6	1.3e-05	4.0	43308	1.00003201
	$nJVF_1$		6	1.3e-05	4.0	45252	1.00003064
	$nJVF_2$		6	1.3e-05	4.0	43308	1.00003201

Behl [5] (nKS) failed to converge in most of the problems used for computational experiments. This is an indication that the developed families of iterative schemes are more stable. Furthermore, results show that the classical JS performed better than all its variants in the two developed families in terms of convergence and efficiency. This coincides with the findings of Cordero, Segura, and Terregros [2], that among the members of JS families, the JS is the best.

Conclusion

We have successfully developed two new families of JS and its variants to decide the solution of scalar nonlinear and system of nonlinear equations. The parameters contained in the structures of the developed families can assume any real values subject to some simple conditions to obtain infinitely many special forms that have CO four when utilized to solve problems in both scalar and vector space. In fact, the famous Jarrat IS (JS) is a special form of one of the families. The computational experiments conducted on the new schemes show that they are good competitors to some existing schemes that are also JS variants.

For further work, the complex dynamical and chaotic behavior of the developed schemes can be considered.

Acknowledgment

The authors appreciate the observations, corrections, and suggestions by the Editors and anonymous reviewers. Their contributions led to the improvement of the initial state of this manuscript.

References

[1] Behl, R., Kanwar, V. and Sharma, K.K. *Optimal equi-scaled familie of Jarratt's method for solving equations*, Int. J. Comput. Math. 90 (2013) 408–422.

[2] Cordero, A., Segura, E. and Terregrosa, J.R. *Behind Jarratt's Steps: Is Jarratt's scheme the Best version of itself?*, Discrete Dynamics in Nature and Society, 2013.

[3] Grosan, C. and Abraham, A. *A new approach for solving nonlinear equations systems*, IEEE Trans. Syst. Man Cybern. A Syst. Hum. 38 (2008) 698–714.

[4] Jarratt, P. *Some fourth-order multipoint iterative methods for solving equations*, Math. Comput. 20 (1966) 434–437.

[5] Kanwar, V., Kumar, S. and Behl, R. *Several new families of Jarratt's method for for solving systems of nonlinear equations*, Appl. Appl. Math. 8 (2013) 701–716.

[6] Lin, Y., Bao, L. and Jia, X. *Convergence analysis of a variant of the Newton method for solving nonlinear equations*, Comput. Math. Appl. 59 (2010) 2121–2127.

[7] Meintjes, K. and Morgan, A.P. *Chemical equilibrium systems as numerical test problems*, ACM Trans. Math. Softw. 16 (1990) 143–151.

[8] More, J.J. *A collection of nonlinear model problems*, Computational solution of nonlinear systems of equations (Fort Collins, CO, 1988), 723–762, Lectures in Appl. Math., 26, Amer. Math. Soc., Providence, RI, 1990.

[9] Morgan, A.P. *Solving polynomial systems using continuation for scientific and engineering problems*, Englewood Cliffs, Prentice-Hall, 1987.

[10] Ogbereyivwe, O. and Izevbizua, O. *A three-parameter class of power series based iterative method for approximation of nonlinear equations*, Iran. J. Numer. Anal. Optim. 13(2) (2023) 157–169.

[11] Ogbereyivwe, O. and Ojo-Orobosa, V. *Family of optimal two-step fourth order iterative method and its extension for solving nonlinear equations*, J. Interdiscip. Math. 24(5) (2022) 1347–1365.

[12] Ogbereyivwe, O., Umar, S.S. and Izevbizua, O. *Some high-order convergence modifications of the Housedholder method for nonlinear equations*, Commun. Nonlinear Anal. 5 (2023) 1–16.

[13] Ortega, J.M. and Rheinboldt, W.C. *Iterative solution of nonlinear equations in several variables*, Academic Press, New York, 1970.

[14] Petkovic, M.S. *Remarks on "On a general class of multipoint root-finding methods of high computational efficiency"* , SIAM J. Numer. Anal. 49(3) (2011) 1317–1319.

[15] Shacham, M. and Kehat, E. *An iteration method with memory for the solution of a non-linear equation*, Chem. Eng. Sci. 27(11) (1972) 2099–2101.

[16] Sharma, R., Kumar, S. and Singh, H. *A new class of derivative-free root solvers with increasing optimal convergence order and their complex dynamics*, SeMA J. 80 (2) (2023) 333–352.

[17] Sivakumar, P. and Jayaraman, J. *Some new higher order weighted Newton methods for solving nonlinear equations with applications*, Comput. Math. Appl. 24(2) (2023) 59.

[18] Traub. J.F. *Iterative methods for the solution of equations*, Prentice-Hall, New Jersey, 1964.

[19] Yaseen, S., Zafar, F. and Alsulami, H.A. *An efficient Jarratt-type iterative method for solving nonlinear Global Positioning System problem*, Axioms, 12(6) (2023) 562.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 2, 2024, pp 417–448
https://doi.org/10.22067/ijnao.2024.85066.1333
https://ijnao.um.ac.ir/

Research Article

An innovative particle physics optimization algorithm for efficient test case minimization in software testing

U. Jaiswal*, and A. Prajapati

Abstract

Software testing is a crucial step in the development of software that guarantees the dependability and quality of software products. A crucial step in software testing is test case minimization, which seeks to minimize the number of test cases while ensuring maximum coverage of the system being tested. It is observed that the existing algorithms for test case minimization still suffer in efficiency and precision. This paper proposes a new optimization algorithm for efficient test case minimization in software testing. The proposed algorithm is designed on the base parameters of the metaheuristic algorithms, inspired by scientific principles. We evaluate the performance of the proposed algorithm on a benchmark suite of test cases from the

*Corresponding author

Received 1 November 2023; revised 16 December 2023; accepted 7 January 2024

Updesh Jaiswal
Department of Computer Science Engineering and Information Technology, Jaypee Institute of Information Technology Noida, India. e-mail: updesh1984@gmail.com

Amarjeet Prajapati
Department of Computer Science Engineering and Information Technology, Jaypee Institute of Information Technology Noida, India. e-mail: amarjeetnitkkr@gmail.com

How to cite this article
Jaiswal, U. and Prajapati, A., An innovative particle physics optimization algorithm for efficient test case minimization in software testing. *Iran. J. Numer. Anal. Optim.*, 2024; 14(2): 417-448. https://doi.org/10.22067/ijnao.2024.85066.1333

literature. Our experimental results show that the proposed algorithm is highly effective in reducing the number of test cases while maintaining high coverage of the system under test. The algorithm outperforms the existing optimization algorithms in terms of efficiency and accuracy. We also conduct a sensitivity analysis to investigate the effect of different parameters on the performance of the proposed algorithm. The sensitivity analysis results show that the performance of the algorithm is robust to changes in the parameter values. The proposed algorithm can help software testers reduce the time and effort required for testing while ensuring maximum coverage of the system under test.

AMS subject classifications (2020): Mathematics Education 97-XX; Computer Science 97P10, 97P40, 97P80.

Keywords: Metaheuristic optimization; Test cases minimization; Software engineering; Nature inspired algorithm.

1 Introduction

Software testing is a crucial step in the creation of software that tries to guarantee the dependability and quality of software products. Test case reduction, which seeks to lower the number of test cases necessary to obtain maximum coverage of the system under test, is one of the crucial jobs in software testing. The cost and work associated with software testing can be considerably decreased by test case minimization, which can also increase the process's effectiveness and efficiency.

Test case minimization includes selecting a subset of test cases from a bigger set of test cases to realize the same scope. The chosen subassemblies must be able to distinguish all absconds and disappointments of the framework beneath the test. The test case diminishment preparation is frequently done physically, which can be time-consuming and error-prone.

To computerize the method of test case minimization, different optimization calculations have been created [21]. These calculations utilize scientific and computational methods to produce an ideal subset of test cases that accomplish the most extreme scope with the least number of test cases. A few of

the well-known optimization calculations utilized for test case minimization incorporate genetic algorithm (GA) [2], Ant colony optimization (ACO) [32], particle swarm optimization (PSO) [14], and simulated annealing. The work presented in [17] gives a description of the use of search-based techniques in test suite reduction with a detailed systematic review.

Whereas reduction in the testing time, lower testing expenses, increased testing efficiency, and higher-quality software products are some of the advantages of test case minimization. Test case minimization can also help identify redundant and unnecessary test cases, which can be removed from the test suite to further improve the efficiency of the testing process.

The existing optimization algorithms for test case minimization have some limitations in terms of efficiency and accuracy. Therefore, there is a need for an innovative optimization algorithm that can overcome these limitations, provide a more efficient, and accurate solution for test case minimization in software testing.

The utilization of metaheuristic calculations for test case minimization can offer assistance to overcome the impediments of conventional optimization calculations, such as thorough look and covetous calculations. Metaheuristic calculations are planned to handle complex optimization issues with expansive look spaces and nonlinear objective capacities, making them well-suited for test case minimization. However, based on the particular needs and demands of the software development project, the goals of test case minimization utilizing metaheuristic algorithms can change. The goals can incorporate lessening the number of test cases, diminishing the testing time, making strides in the proficiency of the testing handle, moving forward the quality of the program item, and lessening the fetching of testing.

This gives rise to the motivation to work on the optimization of the test cases in software engineering. This inspiration for test case minimization utilizing metaheuristic calculations is to optimize the choice of test cases to decrease the number of test cases, guaranteeing that the chosen test cases accomplish the most extreme scope of the framework beneath the test. By lowering the cost and labor needed for software testing, these fewer test cases increase the efficacy and efficiency of the testing process.

The motivation behind the development of optimization algorithms for software testing stems from the need to address the challenges posed by the complexity of modern software, the demand for efficient resource utilization, the drive for higher software quality, the acceleration of development cycles, and the imperative to reduce testing costs. These algorithms offer a strategic and automated approach to testing that aligns with the dynamic nature of contemporary software development practices.

As software applications become increasingly intricate, with intricate interactions and dependencies, traditional manual testing methods struggle to provide adequate coverage and efficiency. Optimization algorithms aim to address these challenges by automating and improving the test case selection process. One key motivation is the need for effective resource utilization. Software testing can be resource-intensive, requiring significant time and human effort. Optimization algorithms help streamline this process by intelligently selecting a subset of test cases that provide maximum coverage and effectiveness. This efficiency not only saves time but also ensures that testing resources are allocated to the most critical areas of the software.

With such a motivation, the main contributions of the proposed metaheuristic algorithm for test case minimization are as follows:

1. The proposed algorithm produces an ideal subset of test cases that accomplishes the greatest coverage with the least number of test cases.

2. The algorithm is outlined to handle complex optimization issues with expansive look spaces and nonlinear objective capacities, making it more proficient than conventional optimization algorithms.

3. The proposed algorithm takes into account various factors such as code coverage, test suite size, and execution time to generate an optimal test suite that achieves maximum coverage with the minimum number of test cases.

4. The proposed algorithm can be customized to fulfill the particular demands and needs of the software development project.

5. The proposed algorithm is evaluated on a benchmark suite of test cases from the literature.

The rest of the paper is arranged in the subsequent section, highlighting important modules as the background-related work is presented in Section 2. The description of the proposed metaheuristic algorithm is presented in Section 3. In Section 4, the experimentation results of the proposed algorithm are presented. Later, a comparative study is presented in Section 5. Finally, the paper gives the key summarization of the work and is presented in Section 6.

2 Literature review

In the past, there has been significant work done on test case minimization using metaheuristic algorithms in the field of software engineering. There are several works on test case prioritization using particle swarm optimization, where the authors propose a test case prioritization technique using PSO. The PSO algorithm is used to select the most critical test cases for early execution, reducing the testing time and effort required [33, 31, 20]. However, there are many other works that focus on the reverse aspect, that is, generation of the test cases using the PSO algorithm. One such work is demonstrated in [29], where the PSO-based search technique is applied with an improved combined function to generate test cases for the critical paths. Another work that gives the comparative study on metaheuristic algorithms for test case generation is presented in [28]. The authors present the details of using the various optimization algorithms and their fitness functions to generate the test cases.

Some of the authors have also focused on the method of hybridization to minimize the number of test cases. One such work is proposed in [9], where the authors used the collective information of the two metaheuristic algorithms to minimize the feature description size. Another work by the authors of [8] presents the study on the test case selection, and reduction using Quantum PSO. Also, the work using the PSO for regression testing and its test case minimization is presented in [34]. This paper proposes a comparative analysis approach for test case selection for regression testing using PSO. The PSO algorithm is used to optimize the selection of test cases

based on various criteria, such as coverage and fault detection capability, while minimizing the size of the test suite.

One of the works published on test case prioritization used the PSO with the modified conditional criteria for test case minimization and selection is [25]. A recent work [15] presents the multi-objective modified PSO algorithm for the reduction of the test suite. In an overall assessment of the work done on PSO-based test case selection, it is found that the authors claimed that in terms of increasing the effectiveness and efficiency of the testing process, PSO had demonstrated encouraging results. Additionally, while retaining complete coverage of the system under test, PSO algorithms can considerably minimize the number of test cases needed for software testing. Indeed, in critical analysis, it is observed that PSO can become trapped in local optima, limiting its ability to explore the entire search space. This can lead to suboptimal solutions and reduce the effectiveness of the test suite. Also, PSO can become computationally expensive as the number of test cases increases. This can limit its scalability and make it impractical for larger software systems. In addition, the performance of PSO is highly sensitive to the choice of parameters, such as the number of particles, the maximum number of iterations, and the inertia weight. Poor parameter settings can lead to suboptimal performance and slow convergence. Thus, the PSO may not be found to be suitable for all types of software systems or test scenarios. Its effectiveness may depend on the nature of the system under test and the objectives of the testing process.

Some of the authors have used the ACO algorithm for the minimization of the test cases. One such work is reported in [24], where the authors used the complete graph to represent all the test cases in the test suite. An empirical analysis of using the ant systems for multi-faceted test case prioritization is presented in [26]. The authors validated their proposed approach on a freely available widely used benchmark dataset, called software infrastructure repository. Also, there is a work from early 2017, where the authors used the ACO approach for regression test case prioritization [11]. The authors here used the epistatic test case segment approach for multiobjective search-based regression test case prioritization reflects the correlation between genes in

the evolution process. Furthermore, the approach is validated on the three benchmarks and found satisfactory on industry standards.

In the literature, there are some other optimization algorithms as well that were used for the minimization and prioritization of the test cases [16]. On this, the work on using test case minimization with GA is presented in [30]. The creators utilized the center concept of the GA to construct an algorithm called *Test Reduce* to discover an optimized and negligible set of test cases for Internet applications. The modified strategy for test case minimization in the COTS method using the GA was employed by the authors of one of the research works [10]. To maximize fitness values in test suit development, the authors applied the GA in the boundary value analysis and partitioning testing.

Besides this, one of the papers proposes a hybrid approach that combines GA and Fuzzy logic for the minimization of the test cases. The authors used the GA to filter out and find the set of candidate test cases. Whereas, the fuzzy logic is used to select the optimal subset of test cases that minimize the size of the test suite while maximizing the fault detection ability [19].

In many other works, the literature suggests the use of the various other metaheuristic algorithms that were used in the optimization of the test data in a variety of applications. On this, the use of the gravitational search algorithm is also found useful in one of the articles for the test case minimization [6]. The authors here utilized the discrete and combinatorial gravitational look calculation to illuminate the test case prioritization and minimization. The try is carried out in a controlled environment with the benchmark dataset, and the results are compared with the GA. In comparison, it is found valuable within the ideal choice of the test cases.

On a similar aspect, the data optimization and prioritization using the novel BAT algorithm is proposed in [7]. The authors used the two-step procedure to enhance the work and modified the traditional BAT algorithm with the inclusion of the new objective function. The algorithm is tested on the dataset with different evaluation criteria. Later, the author presented the comparative analysis with the existing algorithms and found that the proposed approach is performing well.

Some of the authors used the hybrid approach as well for the minimization of the test cases. This hybridization is seen in many aspects by collaborating the multiple metaheuristic algorithms to form a new objective function. On a similar theme, the authors of [9] showed the hybridization of the two metaheuristic algorithms named molecule swarm and positioned firefly calculation for the program test case minimization. Some of the authors also used the standalone firefly algorithm [18] for the test case minimization. However, on comparative, analysis of these two variants suggests that the hybrid algorithm results are better than the use of the standalone method.

The writing moreover depicts the utilization of a half-breed technique for fault detection and combinatorial optimization procedures for configuration-aware basic testing [1] and the minimization utilizing the test code closeness and evolutionary search [27].

In an overall assessment of test case minimization, it is found that it is an important problem in software engineering that aims to reduce the size of the test suite while maintaining its effectiveness in detecting faults in the software system. To address this issue, analysts have connected different metaheuristic algorithms, such as GA, PSO, and ACO, to optimize the test suite. Some authors also explored the design framework of the novel design of the metaheuristic algorithm for software mutation testing in a controlled environment [4].

The previous contributions on test case minimization using metaheuristic algorithms have shown promising results in improving the efficiency and effectiveness of the testing process as also discussed in [12] presenting with a detailed decade review. These approaches have minimized the count of test cases that were actually required for software testing while maintaining maximum coverage of the system under test. The contributions have also highlighted the importance of considering the limitations of these algorithms, such as limited scalability, sensitivity to parameter settings, and limited diversity of solutions.

The comparative study of the related works used for the software test case minimization is presented in Table 1. The table presents the summarized results of the reviewed works and highlights the advantages and the limitations of the studies.

Table 1: Comparative study of the related works on software test case minimization

Reference	Optimization Approach	Advantages	Limitations
[16, 30, 10, 19]	Genetic Algorithms (GAs)	- Automatically explores diverse test cases	- May get stuck in local minima, especially for complex landscapes
		- Handles discrete search spaces	- High computational cost for large-scale problems
		- Captures interactions between test cases	- Difficult to interpret results
[33, 31, 20, 34, 29]	Particle Swarm Optimization (PSO)	- Efficient for continuous and discrete spaces	- Sensitive to parameter tuning
		- Simplicity in implementation	- Convergence to suboptimal solutions
		- Ability to escape local minima	-Slow convergence
[7]	BAT optimization	- Global search capability	- Slow convergence for certain problem landscapes
[9]	Molecule Swarm and Firefly Optimization	- Versatility in handling various objectives	- Parameter sensitivity
[18]	Firefly Optimization	- Provides a probabilistic framework	-Slow convergence
[24, 26, 11]	Ant Colony Optimization (ACO)	- Inspired by natural foraging behavior	- Limited scalability for large problem instances
		- Robustness to changes in the environment	- Convergence highly dependent on parameters
		- Handles discrete and combinatorial problems	-Slow convergence
[28, 9]	Hybrid algorithms and Comparative studies	- Synergy between exploration and exploitation	- Increased complexity in implementation
		- Improved global search capability	- Inherit limitations of individual algorithms
		- Enhanced convergence characteristics	-Slow convergence

3 Proposed metaheuristic optimization algorithm

The main requirement for an optimization algorithm in software test case minimization is to reduce the number of test cases required to test a system while still ensuring that the system's functionality is thoroughly tested. This is important because testing is a crucial step in software development, and reducing the number of test cases required can save time and resources. In addition to reducing the number of test cases, the optimization algorithm should also ensure that the remaining test cases provide maximum coverage of the system's functionality. This means that the algorithm should be able to identify the most critical and relevant test cases, rather than just selecting them randomly or based on simple criteria such as code coverage.

Furthermore, the optimization algorithm should be able to handle large and complex software systems and be efficient enough to provide results in

a reasonable amount of time. This is important because testing can be a time-consuming process, and any optimization algorithm used should not significantly increase the overall testing time.

Based on this, here we developed a new metaheuristic optimization algorithm called the *Physics-informed particle decay Algorithm* (PPDA), based on the nature of the physics to study the decay of the particles. The proposed approach is new in terms of idea and objective function.

3.1 Physics-informed particle decay algorithm (PPDA)

The PPDA is a metaheuristic optimization algorithm that is inspired by principles from physics, specifically regarding stability and particle decay. PPDA simulates the behavior of particles, where each particle represents a potential solution to the optimization problem being solved. The basic flowchart of the proposed PPDA algorithm is shown in Figure 1.

The concept of stability is important in PPDA, as it helps to ensure that the particles converge toward a good solution. The particles in PPDA are attracted toward the energy valleys, which represent the best solutions to the optimization problem. The valley itself represents a stable state, while the particles represent the energy fluctuations that occur in the system. The proposed algorithm uses the concept of stability to ensure that the particles move and converge toward the best solution.

The different modes of decay in PPDA are also important for the algorithm's performance. In physics, particle decay refers to the process by which a particle breaks down into smaller particles or other forms of energy. In the proposed algorithm, particle decay is used to simulate the process by which particles move toward each other.

There are several modes of particle decay in PPDA, each of which represents a different behavior of the particles. For example, the mode of "attractive decay" causes the particles to move toward each other and converge to the best solution. On the other hand, the mode of "repulsive decay" causes the particles to move away from each other and explore different areas of the search space.

3.1.1 Background preliminaries for PPDA

The base of the proposed algorithm is particle physics. In particle physics, the process of colliding two particles to generate new particles is known as particle collision. Particle collisions are used to study the properties of subatomic particles, such as quarks and leptons, and to probe the fundamental forces of nature. In a particle collision, two particles are accelerated to very high speeds using particle accelerators, such as the large hadron collider at CERN. When the particles collide, they can create a shower of new particles that are detected by particle detectors.

The collision itself is a physical reaction that involves the exchange of energy and momentum between the colliding particles. Depending on the energy of the colliding particles and the angle of the collision, different types of particles can be generated. One type of particle collision is known as elastic scattering, in which the colliding particles bounce off each other without generating any new particles. Inelastic collisions, on the other hand, can generate new particles, such as mesons and baryons.

The collision can also generate antiparticles, which are particles that have the same mass as their corresponding particles but have opposite charges. For example, a collision between a proton and an antiproton can generate a shower of particles, including pions, kaons, and other mesons.

Thus, we can see the two forms of the particles in the universe, that is, *stable* and *unstable* particles. Stable particles are particles that do not decay into other particles, while unstable particles decay into other particles after a certain amount of time.

Stable particles include fundamental particles such as electrons, neutrinos, and photons, as well as composite particles such as protons and neutrons. These particles have lifetimes that are much longer than the age of the universe and do not decay spontaneously. On the other hand, unstable particles, have lifetimes that are shorter than the age of the universe and decay into other particles through various physical processes.

Furthermore, as a fundamental feature of these particles, unstable particles emit energy through disintegration or decay. When an unstable particle decays, it transforms into one or more other particles, releasing energy in the

process. The decay of an unstable particle can occur through various physical processes, depending on the nature of the particle and the forces that govern its behavior. During the decay process, the unstable particle loses energy, which is released in the form of radiation, such as γ rays, or particles, such as electrons or positrons. This energy can be measured and used to study the properties of the unstable particle and the forces that govern its behavior.

The decay rate of an unstable particle is described by its half-life, which is defined as the amount of time it takes for half of the particles in a sample to decay. The half-life of a particle can range from fractions of a second to billions of years, depending on its nature.

However, determining the stability of a particle involves considering various factors, including the count of neutrons (Ne) and protons (Ze) in the nucleus, and the Ne/Ze ratio is one of the most crucial factors. The balance between the strong nuclear force, which binds protons and neutrons together, and the electromagnetic force, which repels positively charged protons, determines the stability of a nucleus. If the number of neutrons and protons is such that the strong force is greater than the electromagnetic force, the nucleus is stable. However, if the strong force is not strong enough to overcome the electromagnetic force, the nucleus is unstable and may undergo radioactive decay.

The Ne/Ze ratio is a key aspect in determining a nucleus's stability. It is the ratio of the number of neutrons to the number of protons in the nucleus. In general, a stable nucleus has a Ne/Ze ratio that is close to 1. If the Ne/Ze ratio is too high or too low, the nucleus is unstable and may undergo radioactive decay. For light nuclei, the stable Ne/Ze ratio is close to 1:1. However, as the atomic number increases, the stable Ne/Ze ratio increases slightly. In addition to the Ne/Ze ratio, the size of the nucleus also plays a role in its stability. Larger nuclei tend to be less stable than smaller nuclei and may undergo radioactive decay to become more stable.

There are three sorts of emissions that define the particle's level of stability during the decay process, that is, α *decay,* β *decay*, and γ *decay* [22]. All three decay particles are emitted from the nucleus, which can result in a change in the Ne/Ze ratio of the nucleus. In α decay, a nucleus emits an α particle, which consists of two protons and two neutrons. This emission

reduces the number of protons (Ze) and neutrons (Ne) in the nucleus by 2 and 4, respectively. As a result, the Ne/Ze ratio of the nucleus decreases, making it more stable.

In β decay, a neutron in the nucleus is changed into a proton or vice versa, and the nuclear material emits an electron (β particle), antineutrino, or positron (positron emission). Depending on the type of β decay, this alters the number of protons and neutrons in the nucleus, causing the Ne/Ze ratio to either rise or fall.

In γ decay, a nucleus in an excited state emits a γ ray, which is a high-energy photon. Moreover, γ decay does not change the number of protons or neutrons in the nucleus, and thus, it does not directly affect the Ne/Ze ratio. However, γ decay can occur as a result of α or β decay, which may change the Ne/Ze ratio.

Within the proposed approach, we have utilized the standards of the decay process through different particles that can serve as a starting point for the planning of the modern metaheuristic algorithm in which the arrangement candidates' execution can be improved by utilizing the thought that particles have the inclination to reach a steady position.

3.1.2 Mathematical foundation of PPDA

The mathematical structure of the proposed optimization algorithm uses the previously defined concept and preliminaries. The first step of the proposed algorithm is the initialization, wherein in the search space, the solution candidates (A_i) are assumed to be particles with changing degrees of solidness and expected to be a particular locale of the universe as appeared in supposed to be particles with varying degrees of stability and assumed to be a specific region of the universe as shown in the following equation (1):

$$A = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_i \\ \vdots \\ A_n \end{bmatrix} = \begin{bmatrix} A_1^1 & A_1^2 & \dots & A_1^j & \dots & A_1^m \\ A_2^1 & A_2^2 & \dots & A_2^j & \dots & A_2^m \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ A_i^1 & A_i^2 & \dots & A_i^j & \dots & A_i^m \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ A_n^1 & A_n^2 & \dots & A_n^j & \dots & A_n^m \end{bmatrix}, \tag{1}$$

where the symbolic representation of the used variables is defined as the total number of particles given by variable n, the problem dimension is represented by m, A_i^j is deciding the beginning position of the ith candidate for jth decision variable.

Moreover, the mathematical form of the decision variable is shown by the following equation (2):

$$A_i^j = A_{i,\min i}^j + R_m(A_{i,\max i}^j - A_{i,\min i}^j), \tag{2}$$

where $A_{i,\min i}^j$ and $A_{i,\max i}^j$ speaks to the lower and upper bounds of the jth variable within the ith candidate and the R_m is the irregular generator within the extend $[0, 1]$.

The core difference between neutron-rich and neutron-poor particles is then calculated using the objective function set for each particle utilizing the analysis with the Enrichment Level of Neutron (NEL) of the particles. NEL of a particle can be calculated using the formula given in the following equation (3):

$$NEL = \frac{(Ne - Ze)}{(Ne + Ze)}, \tag{3}$$

where Ne represents the number of neutrons in the nucleus, and Ze represents the number of protons in the nucleus.

The NEL value indicates the neutron excess or deficiency in the nucleus relative to the number of protons. If the NEL value is positive, it indicates a neutron excess, meaning there are more neutrons than protons in the nucleus. A positive NEL value suggests a tendency towards neutron-richness. If the NEL value is negative, it indicates a neutron deficiency, meaning there are fewer neutrons than protons in the nucleus. A negative NEL value suggests a tendency towards proton-richness.

Whereas the following equation (4) gives the computation of the core difference between neutron-rich and neutron-poor particles using NEL:

$$CD = \frac{\sum_{i=1}^{n} NEL_i}{n}, \tag{4}$$

where CD is the core difference between neutron-rich and neutron-poor particles, NEL_i is the NEL for ith particle, and n is the total number of particles.

Thus, if the NEL of a particle is discovered to be more than the CD, that is $(NEL_i > CD)$, the associated particle is assumed to have the next Ne/Ze proportion, and the rot handle is expected to be utilizing α, β, or γ plans.

Next, in the third phase, we will predict the particle's stability levels based on the objective function evaluated using the following equation (5):

$$L_i = \frac{NEL_i - B_L}{W_L - B_L}, \qquad i \in [1, n], \tag{5}$$

where L_i is the ith particle stability level, B_L is the particle's best stability level, W_L is the particle's worst stability level, and n is the total number of particles in the system.

Furthermore, to calculate the stability of the system (SS), we generate a value in the interval of 0 to 1, probably with a random selection. In case the predicted particle stability level (L_i) is higher than the selected SS, that is, $(L_i > SS)$ then α and γ decay are to be considered.

Based on the theory of particle physics for α decay, it is observed that the α rays were emitted to enhance the particle's stability. Using the same theory in the proposed algorithm as a position-updating technique to generate the new solution particle, two α values were generated, that is, α_1 and α_2. The generated α_i values are given by the following equation (6):

$$\alpha_1 \in [1, r], \qquad \alpha_2 \in [1, \alpha_1], \tag{6}$$

where α_1 is for the number of emitted rays, r is a random number in between 1 to n, and α_2 is the number of to-be-emitted α rays. These emitted rays are considered as the decision variables to generate the new particle, which is given by the following equation (7):

$$A_i^{new_\alpha} = A_i(A_{B_L}(A_i^j)), \tag{7}$$

where $i \in [1, n],\ j = \alpha_2,\ A_i^{new_\alpha}$ is the new particles generated in the sample space, A_i is the ith particle's current position in the sample space, A_{B_L} is the particle's position with the leading soundness level, and A_i^j is the jth choice particle for the transmitted beam.

Aside from the α decay, there is also a γ decay in which γ rays are emitted to strengthen the stability of the excited particles. In the proposed algorithm, this is considered as another particle's position-updating process to generate the new particles. On a similar pattern, like for α decay, here also we use two values of γ, that is, γ_1 and γ_2. Here, γ_1 signifies the number of emitted photons, and γ_2 in the range of [1, γ_1] denotes the information on the photons to be examined in the particles. Thus, these photons in the particles are considered to be the decision variable in the sample space, where the old particle is replaced with the neighboring particle. Finally, the distance between the particle of interest and other leftover particles is calculated as we do in the K-nearest neighbor algorithm. Here we select the nearest particle having the value of $k = 1$ and is given by the following equation (8):

$$\min(Dist_i^k) = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}, \tag{8}$$

where $i \in [1, n],\ j \in [1, n-1],\ Dist_i^k$ is the distance between the ith particle and the neighbor k particles, and the pair (X_i, Y_i) is the particles coordinate in the sample space.

Using the above information, the second particle generating process is given by using the following equation (9):

$$A_i^{new_\gamma} = A_i(A_N(A_i^j)), \tag{9}$$

where $i \in [1, n],\ j = \gamma_2,\ A_i^{new_\gamma}$ is the new particles generated in the sample space, A_i is the ith particle's current position in the sample space, A_N is the ith particles neighborhood position vector, and A_i^j is the jth decision particle in space (emitted ray).

It is to be noted here that if the particle's stability level (L_i) is lower than the SS, that is, ($L_i < SS$), then the β decay is then thought to occur with increasingly unstable particles with lower stability levels.

Now, the mathematical representation of the particle position update via reaching towards the stability band is given by the following equation (10):

$$A_i^{Update_1} = A_i + \frac{(a.A_{B_L} - b.A_C)}{L_i}, \tag{10}$$

where $i \in [1, n]$, $A_i^{Update_1}$ and A_i are the updated and current positions of the ith particle, a and b are the two numbers randomly chosen from the interval of $[0, 1]$ and controls the particle's movement, and A_C is the position of the center of the particle in space and is given by the following equation (11):

$$A_C = \frac{\sum_{i=1}^{n} A_i}{n}. \tag{11}$$

Another position overhauling preparation is carried out for particles utilizing β rot, in which a controlled development towards the particle or a candidate with the leading solidness level (A_{B_L}) and a neighboring particle (A_N) is performed, while the particle's stability level has no bearing on the movement process. This strategy points to extending the misuse and investigation levels of the calculation which is given by the following equation (12):

$$A_i^{Update_2} = A_i + (c.A_{B_L} - d.A_N), \tag{12}$$

where $i \in [1, n]$, $A_i^{Update_2}$ and A_i are the updated and current positions of the ith particle, c, and d are the two numbers randomly chosen from the interval of $[0, 1]$ and control the particle's movement, and A_{B_L}, A_N have their usual meaning as discussed earlier.

When the NEL of a particle is discovered to be lower than the CD, that is, $(NEL_i < CD)$, this indicates that the particle has a lower value of Ne/Ze ratio. It means that the particle tends to emit positrons in order to come closer to the stability band. Using a similar concept in the proposed algorithm, the new particle position is updated by using the following equation (13):

$$A_i^{update} = A_i + e, \tag{13}$$

where $i \in [1, n]$, A_i^{Update} and A_i are the updated and current positions of the ith particle, e is the numbers randomly chosen from the interval of $[0, 1]$ and control the particle's movement.

Finally, after the execution of the proposed algorithm completely, at the end, we are with any one of the solution steps, that is, (1) if ($NEL_i > CD$), the algorithms return two newly generated particle positions for each of the particles as given by (10) and (12), and (2) if ($NEL_i < CD$), the algorithms return a single new particle position as given by (13). Hence, in an overall analysis, three processes for updating positions are included in the algorithm. Whereas one position upgrading method happens in arrangement candidates, where abuse is fulfilled, two of these methods happen in choice factors, where the investigation is performed. This technique's problematic aspect is that while the exploration phase may direct the program to locally optimal answers, the other phase aims to fine-tune the prior solutions to find the best candidate globally.

3.1.3 Algorithmic design of the proposed methodology

The description of the proposed algorithm in the pseudo-code is presented in Algorithm 1. The given code provides the detailed flow of the algorithm step-by-step. At the end of the algorithm, the return is in the form of the particle positioning with the best stability prediction.

4 Experimentation results

4.1 Dataset for experimentation

To test the performance of the proposed algorithm we used the dataset from the open-source platform, which was having the test cases of size 100. The structure of the dataset is composed of several attributes, which were given as follows:

- Test ID
- Test cases
- Pre-Conditions
- Precedence

Algorithm 1: Pseudo-code of the proposed approach

Input: Initialize the population of particles randomly within the search space

Input: Evaluate the fitness of each particle in the population using NEL

Output: Particle with best stability level

Data: Set the global best position and fitness value as the position and fitness of the best particle.

while $iteration - count < Max_{iteration}$ **do**

- Compute the CD of the particle
- Compute the particle's best stability level A_{B_L}
- **for** $i \in [1, n]$ **do**
  - Compute the stability of the ith particle (L_i)
  - Compute the NEL_i of the ith particle
  - **if** $NEL_i > CD$ **then**
    - Compute the SS of the particle
  - **if** $L_i > SS$ **then**
    - Generate α_1 and α_2
    - **for** $j \in [1, \alpha_2]$ **do**
      - compute $A_i^{new_\alpha} = A_i(A_{B_L}(A_i^j))$
    - Generate γ_1 and γ_2
    - Determine neighboring particle (A_N)
    - **for** $j \in [1, \gamma_2]$ **do**
      - compute $A_i^{new_\gamma} = A_i(A_N(A_i^j))$
    - **else if** $L_i < SS$ **then**
      - Find the center of the particle (A_C)
      - compute $A_i^{Update_1} = A_i + \frac{(a.A_{B_L} - b.A_C)}{L_i}$
      - Find the neighboring particle (A_N)
      - compute $A_i^{Update_2} = A_i + (c.A_{B_L} - d.A_N)$
    - **if** $NEL_i < CD$ **then**
      - compute $A_i^{update} = A_i + e$;

Return the particle with stability level (A_{B_L})

- Pass/Fail

The downloaded dataset is in the form of the comma separated file having attribute information. The sample structure of the used dataset is given in Figure 2.

4.2 Parameters value for optimization algorithms

The various optimization algorithms that were used in the experimentation will be having some parameter values that is shown in Table 2.

Table 2: Optimization algorithms used in experimentation with their parameter values

Algorithm	**Parameter 1**	**Value 1**	**Parameter 2**	**Value 2**	**Parameter 3**	**Value 3**	**Parameter 4**	**Value 4**
ACO	Alpha	0.5	Beta	1	Evaporate	0.1	-	-
GA	Population	100	Crossover	0.8	Mutation	0.05	-	-
PSO	Swarm	100	Inertia wt	0.7	Personal Coff.	2.0	Social Coff.	1.5
GSK	Population	100	Knowledge factor	0.5	Knowledge ratio	0.9	Knowledge rate	10
WSO	Moment	0.7	Modality	0.6	Movement	0.3	Search Coff.	0.5
TOA	Team size	100	T factor	0.4	random num	0.2	-	-
CHIO	Population	100	Reproduction rate	0.4	Case age	0.3	-	-

4.3 Result computation

To evaluate the performance of the proposed algorithm here ten benchmark test functions were used. These functions are known as Bukin function, cross-in-tray function, Easom function, Goldstein-price function, Himmelblau's function, Levi function, six-hump camel function, Schaffer's function, Weierstrass function, and Xin-She Yang function. These test functions cover a range of characteristics, such as multimodality, irregularity, and high dimensionality, allowing researchers to assess the performance of metaheuristic algorithms in different problem landscapes.

It's worth noting that each test function has its own mathematical formulation and specific properties, making them suitable for evaluating different aspects of optimization algorithms. Also, each of the test functions is con-

sidered to have a fixed dimension of size 50. The set value of the global best is selected as 0 for all the test functions to make uniformity.

Here, we have utilized 100 partitioned optimization runs to calculate the cruel, the standard deviation, and the vital number of objective work assessments in order to assess the performance. The method utilizes a halting criterion with a tolerance of 1×10^{-12} for the globally optimal values of the alleged problems and a maximum of 50,000 evaluations of objective functions. To evaluate the performance of the proposed algorithm, all the tests were conducted with a fixed population of size 50. The iteration-wise convergence of the proposed algorithm with respect to the time taken is shown in Figure 3. The theoretical convergence of the NEL and the CD with +ve and -ve values on the iteration count of 50 is shown in Figure 4.

The fitness values recorded by using the dataset with the proposed algorithm are given in Table 3. The table shows the result of the calculation of fitness values computed on the basis of per epoch size.

Table 3: Result of the proposed algorithm on the dataset giving fitness values

Iteration count	**Test Case**	**Fitness function**
1	10	5.9488e(-010)
5	18	7.1491e(-010)
10	27	1.7217e(-009)
15	38	4.1091e(-009)
20	46	6.0999e(-009)
25	55	6.1999e(-009)
30	62	8.2628e(-007)
35	70	8.2101e(-007)
40	85	9.998e(-007)
45	93	9.9974e(-007)
50	100	9.998e(-007)

The experimentation result of the proposed algorithm on the selected ten benchmark test functions is presented in Table 4. All the benchmark test functions on the proposed algorithm are evaluated by using the three parameters, that is, mean, standard deviation, and the best value.

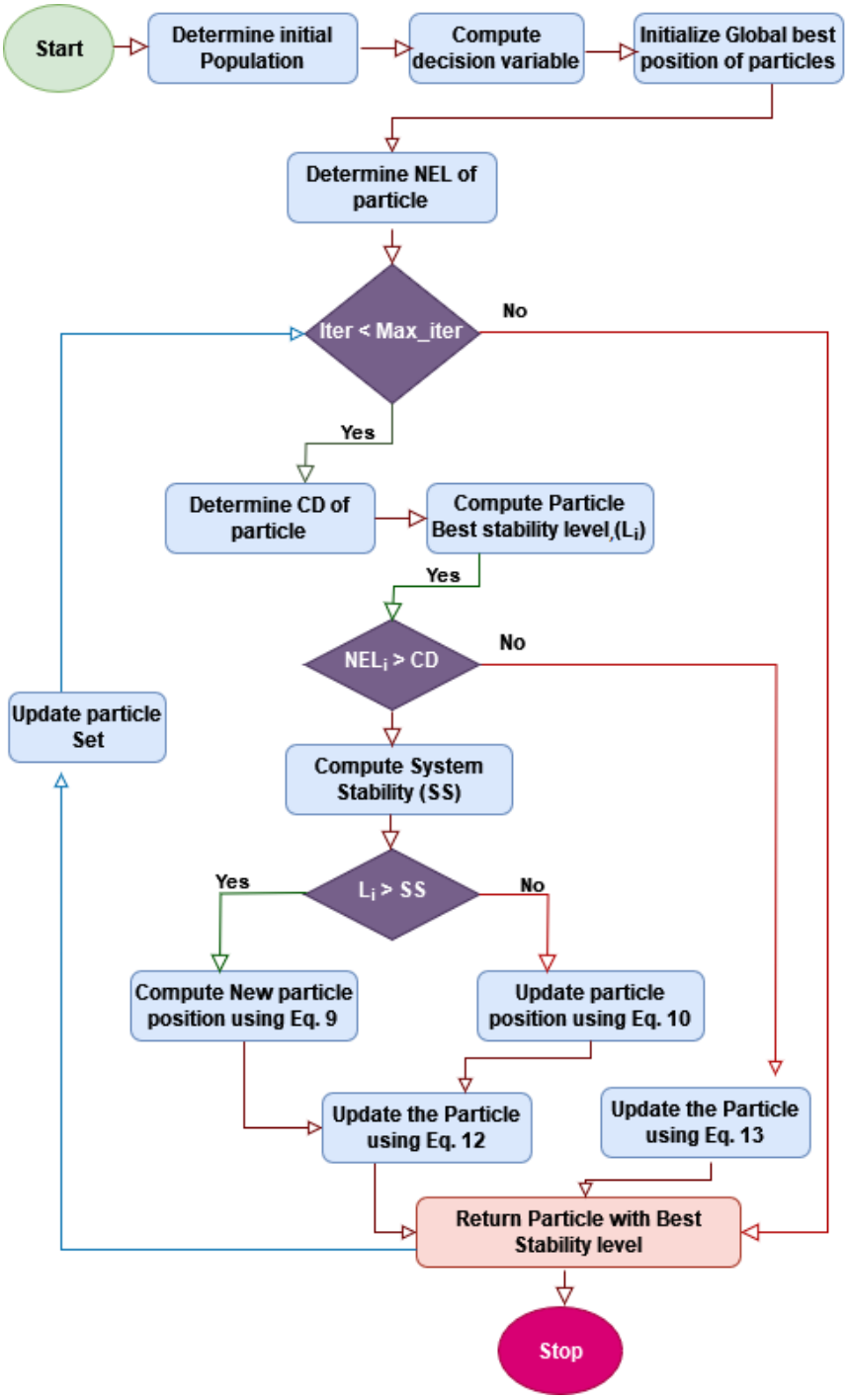


Figure 1: Flowchart of the proposed PPDA algorithm.

	TEST ID	TEST CASE	PRE-CONDITIONS	TEST STEPS	PRECEDENCE	TEST EFFORT	TEST DATA	EXPECTED RESULT	ACTUAL RESULT	PASS/FAIL
0	LOAD_001	Test Page load functionality through URL	None	Enter Invalid URL	H	8	NaN	NaN	NaN	NaN
1	LOAD_002	Test Page Reload without crashes	None	NaN	H	3	NaN	NaN	NaN	NaN
2	LOAD_003	Test 404 Error for Invalid URL in domain	None	NaN	L	7	NaN	NaN	NaN	NaN
3	SELECT_CITY_001	Test City Choice using Icons	LOAD_001	Choose city from name and icons	M	4	NaN	NaN	NaN	NaN
4	SELECT_CITY_002	Test Search City using Search Bar	LOAD_001	Choose city by Search Bar	M	6	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...
95	BOOK_EVENT_003	Test Add & Remove person functionality	BOOK_EVENT_001	NaN	M	6	NaN	NaN	NaN	NaN
96	BOOK_EVENT_004	Test choice of seats	BOOK_EVENT_003	NaN	H	7	NaN	NaN	NaN	NaN
97	EVENT_CONFIRM_001	Test transfer to Payment Init	BOOK_EVENT_004	NaN	H	4	NaN	NaN	NaN	NaN
98	EVENT_CONFIRM_002	Test transfer to Confirmation Page after payment	PAY_PORTAL_003	NaN	H	4	NaN	NaN	NaN	NaN
99	HELP_001	Test help link functionality	HOME_001	NaN	L	5	NaN	NaN	NaN	NaN

Figure 2: Sample description of the dataset used for experimentation.

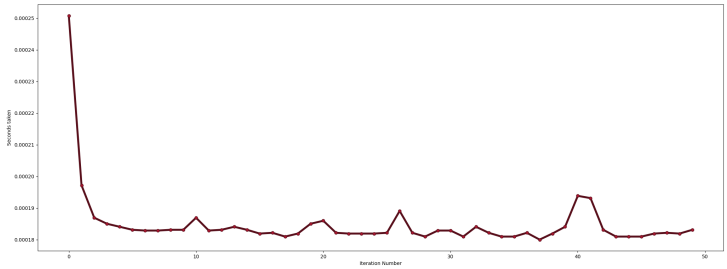
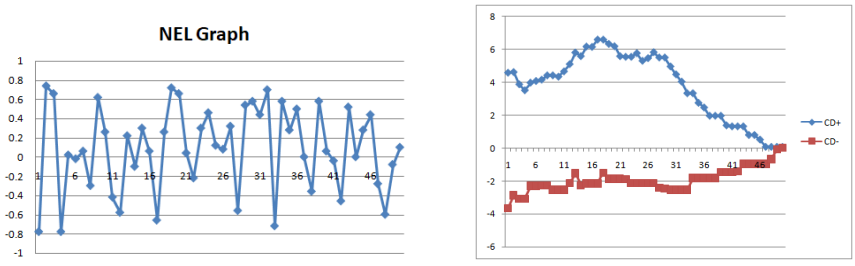


Figure 3: Performance analysis of the proposed algorithm in between the iterations count and time taken by the algorithm.



(a) NEL Curve (b) CD Curve

Figure 4: Convergence of the NEL and CD on the iteration count 50

Table 4: Performance analysis of the proposed algorithm with ten benchmark functions

Benchmark Function	**Parameters**	**Experimented value**
Bukin Function	Mean	0.11958
	Std. Dev	1.15982
	Best	3.24853
Cross-in-Tray Function	Mean	20.8456
	Std. Dev	4.12124
	Best	9.05642
Easom Function	Mean	27.5468
	Std. Dev	11.2619
	Best	8.4653
Goldstein-Price Function	Mean	12.5887
	Std. Dev	3.8911
	Best	4.5997
Himmelblau's Function	Mean	-2.805118
	Std. Dev	3.131312
	Best	3.584428
Levi Function	Mean	13.1087
	Std. Dev	7.7724
	Best	19.4376
Six-Hump Camel Function	Mean	0.0898
	Std. Dev	-0.7126
	Best	0.7126
Schaffer Function	Mean	5.47×10^{-13}
	Std. Dev	5.47×10^{-12}
	Best	0
Weierstrass Function	Mean	1.53×10^{-03}
	Std. Dev	1.41×10^{-03}
	Best	1.18×10^{-06}
Xin-She Yang Function	Mean	1.25×10^{-03}
	Std. Dev	1.12×10^{-03}
	Best	6.96×10^{-06}

5 Comparative analysis

The comparative analysis of the proposed algorithm with the other benchmark metaheuristic algorithms like ACO, GA, and PSO on the ten selected benchmark functions is presented in Table 5. We explored some recent works on the optimization of the problems and selected few named, as gaining-sharing knowledge-based algorithm (GSK) [23], war strategy optimization (WSO) [5], teamwork optimization algorithm (TOA) [13], and Coronavirus herd immunity optimizer (CHIO) [3]. The comparative analysis using some recently proposed metaheuristic algorithms is also presented in Table 5.

Table 5: Comparative analysis of the proposed algorithm with other metaheuristic optimizations on ten benchmark functions

Benchmark Functions	Parameters	Experimented value							
		Proposed Algorithm	ACO	GA	PSO	GSK	WSO	TOA	CHIO
Bukin Function	Mean	0.11958	0.0452	0.0845	0.1045	0.6424	0.7594	0.4645	0.8215
	Std. Dev	1.15982	0.8875	0.8952	0.9973	0.0353	0.5204	0.5329	0.9762
	Best	3.24853	1.2975	2.0128	2.9861	1.5612	1.8345	2.4699	2.9164
Cross-in-Tray Function	Mean	20.8456	12.5465	14.5612	17.7789	11.9246	12.3320	15.4976	18.5564
	Std. Dev	4.12124	2.2498	2.8167	3.4242	1.7783	1.9201	2.1286	2.5546
	Best	9.05642	5.1971	7.1267	8.9462	2.5610	3.8848	5.5585	6.4319
Easom Function	Mean	27.5468	17.5641	20.5731	23.1973	17.2564	21.5645	20.0843	23.4997
	Std. Dev	11.2619	8.8891	10.0245	11.1189	8.1469	10.1213	11.0046	11.1873
	Best	8.4653	5.2307	6.7239	7.8912	4.3469	4.8501	5.4321	6.0197
Goldstein-Price Function	Mean	12.5887	9.1327	10.8123	11.7831	7.8873	8.8213	9.4631	9.9799
	Std. Dev	3.8911	1.0891	1.9921	2.5618	1.2390	2.0049	2.7391	3.7721
	Best	4.5997	2.2080	3.0891	4.1129	1.9951	2.5739	3.0852	3.7894
Himmelblau's Function	Mean	-2.805118	-4.1978	-3.331946	-2.913354	-5.3349	-4.0167	-2.4223	-2.8521
	Std. Dev	3.131312	0.1974	1.079641	2.139734	0.5216	1.0841	1.9898	2.0173
	Best	3.584428	0.1941	1.943137	2.139424	1.0409	1.9744	2.3367	3.0889
Levi Function	Mean	13.1087	9.1346	11.4621	1.7921	7.2252	8.5545	10.1446	9.8878
	Std. Dev	7.7724	4.9134	6.1389	7.5134	4.4464	5.1973	6.5565	5.8552
	Best	19.4376	12.9434	16.4673	18.9124	10.2881	12.4652	17.8446	15.5550
Six-Hump Camel Function	Mean	0.0898	1.1913	0.1843	0.0661	0.9979	0.6732	0.2212	0.1101
	Std. Dev	-0.7126	-1.7349	-0.9943	-0.7112	-1.2333	-1.1249	-0.9999	-0.4156
	Best	0.7126	1.9431	0.9937	0.6574	0.7449	0.7552	0.7449	0.8510
Schaffer's Function	Mean	5.47E-13	9.77E-13	7.87E-13	6.34E-13	9.44E-13	8.14E-13	6.49E-13	5.11E-13
	Std. Dev	5.47E-12	7.71E-12	6.89E-12	6.07E-12	8.11E-12	7.57E-12	6.04E-12	6.01E-12
	Best	0	1.6972	0.9234	0.0569	0.6646	0.7411	1.1012	1.6887
Weierstrass Function	Mean	1.53E-03	3.33E-03	2.89E-03	1.943E-03	2.84E-03	2.41E-03	1.72E-03	1.50E-03
	Std. Dev	1.41E-03	2.61E-03	1.99E-03	1.65E-03	2.43E-03	1.90E-03	1.99E-03	1.57E-03
	Best	1.18E-06	4.81E-06	3.11E-06	2.01E-06	4.71E-06	3.11E-06	1.53E-06	1.96E-06
Xin-She Yang Function	Mean	1.25E-03	2.55E-03	2.012E-03	1.95E-03	1.53E-03	2.84E-03	2.61E-03	1.47E-03
	Std. Dev	1.12E-03	2.56E-03	1.99E-03	1.24E-03	1.55E-03	1.72E-03	1.30E-03	1.24E-03
	Best	6.96E-06	4.75E-06	5.09E-06	6.66E-06	4.21E-06	5.01E-06	5.78E-06	6.11E-06

To test the significance of the proposed model, the statistical analysis is also computed and compared. To do this, the well-known statistical mea-

sure test named Kolmogorov–Smirnov (KS) test is used in this work. The KS test is a nonparametric test that is used to determine if a sample comes from a specific distribution. It is particularly useful for assessing whether a sample follows a particular theoretical distribution, such as a normal distribution. The test is based on the cumulative distribution function of the empirical distribution and the theoretical distribution being tested. The KS test statistic, denoted as D, is the maximum absolute difference between the empirical distribution function of the sample and the cumulative distribution function of the theoretical distribution. If the calculated test statistic is greater than the critical value from the KS distribution table (or the p-value is less than the significance level), then the null hypothesis will be rejected. The experimentation results of the KS test are presented in Table 6. The KS test results are given in Table 6, and since the test's p-value is less than 0.05, the hypothesis regarding the data's normal distribution is satisfied. As a result, nonparametric statistical tests can be used for more research.

Table 6: The KS test values showing the significance of proposed method

Algorithm	**Data type**	**Existing Algorithms**						
		ACO	**GA**	**PSO**	**GSK**	**WSO**	**TOA**	**CHIO**
Proposed Algorithm	Best	8.22E-06	8.22E-06	8.22E-06	6.45E-06	8.22E-06	7.19E-06	7.84E-06
	Mean	1.65E-04	1.93E-04	1.65E-04	1.65E-04	2.01E-04	1.99E-04	1.65E-04
	SD	2.12E-02	2.12E-02	1.96E-02	2.76E-02	2.12E-02	2.12E-02	2.76E-02

6 Conclusion

In conclusion, the proposed metaheuristic algorithm based on the physics informed particle decay principle, specifically for test case minimization, shows promising potential. By leveraging concepts from physics, such as particle stability and different modes of decay, the algorithm offers a unique approach to optimize test case generation and reduce redundancy.

The algorithm's utilization of physics principles allows for the generation of stable particles that represent high-quality test cases, while unstable particles can decay and be replaced by new particles to explore different areas of

the search space. This enables the algorithm to efficiently converge towards optimal or near-optimal solutions.

Through the use of the proposed algorithm, the test case minimization process can benefit from the inherent properties of particles in physics, such as stability and decay. This approach can enhance the efficiency of test case generation by reducing the number of redundant or irrelevant test cases, leading to improved resource utilization and reduced testing time.

Also, the proposed metaheuristic algorithm for test case minimization, validated on ten benchmark functions, has demonstrated its effectiveness and potential in optimizing the generation and selection of test cases. The algorithm has been evaluated and compared against well-known mathematical functions, which serve as standard test cases for assessing the performance of optimization algorithms.

The results of the validation process indicate that the proposed metaheuristic algorithm performs well in terms of minimizing the test case suite while maintaining the required coverage and effectiveness. The algorithm has shown competitive performance in terms of convergence towards optimal or near-optimal solutions, as well as the ability to handle different types of objective functions.

Through the use of metaheuristic techniques, the algorithm has successfully explored the search space and identified high-quality test cases that meet the specified criteria. The algorithm's ability to adapt and evolve over iterations has been instrumental in achieving good performance across the benchmark functions.

The validation of ten benchmark functions has provided insights into the algorithm's strengths and weaknesses. The algorithm's robustness and generalization capabilities have been evaluated by testing its performance on various types of objective functions. The results indicate that the algorithm is capable of effectively minimizing test cases across different problem domains.

Future work will still need to be done in order to evaluate the algorithm on a larger variety of benchmark functions and real-world test situations. The algorithm's performance can also be increased by refining its parameters and taking into account knowledge particular to the situation. The work could

be extended with the design of the nonconvex objective function in place of the convex function. Introducing nonconvexity can make the optimization problem more challenging, as it increases the likelihood of multiple local minima.

Compliance with ethical standards

Conflict of Interest:

The authors of this manuscript declare that there is no conflict of interest.

Data Availability Statement:

The dataset and code generated during and/or analyzed during the current study are available from the corresponding author at a reasonable request.

Funding Information:

The author declares that there is no funding associated with this project.

Ethics Statement:

The author of this manuscript confirms that:

(i) Informed, written consent has been obtained from the relevant sources wherever required;

(ii) All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1964 and its later amendments.

(iii) Approval and/or informed consent were not required for the study as the dataset is collected from an open-source website which is freely available.

References

[1] Ahmed, B.S. *Test case minimization approach using fault detection and combinatorial optimization techniques for configuration-aware structural testing*, Eng. Sci. Technol. an Int. J. 19(2) (2016), 737–753.

[2] Akour, M., Abuwardih, L., Alhindawi, N. and Alshboul, A. *Test case minimization using genetic algorithm: pilot study*, In: 2018 8th International Conference on Computer Science and Information Technology (CSIT), 66 –70. IEEE (2018).

[3] Al-Betar, M.A., Alyasseri, Z.A.A., Awadallah, M.A. and Abu Doush, I. *Coronavirus herd immunity optimizer (chio)*, Neural Comput. Appl. 33 (2021), 5011–5042.

[4] Arasteh, B., Gharehchopogh, F.S., Gunes, P., Kiani, F. and Torkamanian Afshar, M. *A novel metaheuristic based method for software mutation test using the discretized and modified forrest optimization algorithm*, J. Electron. Test. (2023), 1–24.

[5] Ayyarao, T.S., Ramakrishna, N., Elavarasan, R.M., Polumahanthi, N., Rambabu, M., Saini, G., Khan, B. and Alatas, B. *War strategy optimization algorithm: a new effective metaheuristic algorithm for global optimization*, IEEE Access 10, (2022), 25073–25105.

[6] Bajaj, A. and Sangwan, O.P. *Discrete and combinatorial gravitational search algorithms for test case prioritization and minimization*, Int. J. Inf. Technol. 13 (2021), 817–823.

[7] Bajaj, A., Sangwan, O.P. and Abraham, A. *Improved novel bat algorithm for test case prioritization and minimization*, Soft Comput. 26(22) (2022), 12393–12419.

[8] Bajaj, A., Abraham, A., Ratnoo, S. and Gabralla, L.A. *Test case prioritization, selection, and reduction using improved quantum-behaved particle swarm optimization*, Sensors 22(12) (2022), 4374.

[9] Bharathi, M. *Hybrid particle swarm and ranked firefly metaheuristic optimization-based software test case minimization*, Int. J. Appl. Metaheuristic Comput. (IJAMC) 13(1) (2022), 1–20.

[10] Bhatia, P.K. *Test case minimization in cots methodology using genetic algorithm: a modified approach*, In: Proceedings of ICETIT 2019: Emerging Trends in Information Technology, 219 –228. Springer, 2020.

[11] Bian, Y., Li, Z., Zhao, R. and Gong, D. *Epistasis based aco for regression test case prioritization*, IEEE Trans. Emerg. Top. Comput. Intell. 1(3) (2017), 213–223.

[12] Boukhlif, M., Hanine, M. and Kharmoum, N. *A decade of intelligent software testing research: A bibliometric analysis*, Electronics 12(9) (2023), 2109.

[13] Dehghani, M. and Trojovsk'y, P. *Teamwork optimization algorithm: A new optimization approach for function minimization/maximization*, Sensors 21(13) (2021), 4567.

[14] Deneke, A., Assefa, B.G. and Mohapatra, S.K. *Test suite minimization using particle swarm optimization*, Mater. Today: Proc. 60 (2022), 229–233.

[15] Geetha, U. and Sankar, S. *Multi-objective modified particle swarm optimization for test suite reduction (mompso)*, Comput. Syst. Sci. Eng. 42(3) (2022), 899–917.

[16] Guizzo, G., Califano, F., Sarro, F., Ferrucci, F. and Harman, M. *Inferring test models from user bug reports using multi-objective search*, Empir. Softw. Eng. 28(4) (2023), 95.

[17] Habib, A.S., Khan, S.U.R. and Felix, E.A. *A systematic review on searchbased test suite reduction: State-of-the-art, taxonomy, and future directions*, IET Software 17(2) (2023), 93–136.

[18] Hashim, N.L. and Dawood, Y.S. *Test case minimization applying firefly algorithm*, Int. J. Adv. Sci. Eng. Inf. Technol. 8(4-2) (2018), 1777–1783.

[19] Joseph, A. and Radhamani, G. *Hybrid test case optimization approach using genetic algorithm with adaptive neuro fuzzy inference system for regression testing*, J. Test. Eval. 45(6) (2017), 2283–2293.

[20] Khatibsyarbini, M., Isa, M.A. and Abang Jawawu, D.N. *A hybrid weight-based and string distances using particle swarm optimization for prioritizing test cases*, J. Theor. Appl. Inf. Technol. 95(12) (2017).

[21] Khoshniat, N., Jamarani, A., Ahmadzadeh, A., Haghi Kashani, M. and Mahdipour, E. *Nature-inspired metaheuristic methods in software testing*, Soft Comput. (2023), 1–42.

[22] Kocher, D.C. *Radioactive decay data tables*, Tech. rep., Oak Ridge National Lab., TN (USA), 1981.

[23] Mohamed, A.W., Hadi, A.A. and Mohamed, A.K. *Gaining-sharing knowledge based algorithm for solving optimization problems: a novel natureinspired algorithm*, Int. J. Mach. Learn. Cybern. 11(7)(2020), 1501–1529.

[24] Mohanty, S., Mohapatra, S.K. and Meko, S.F. *Ant colony optimization (aco-min) algorithm for test suite minimization*, In: Progress in Computing, Analytics and Networking: Proceedings of ICCAN 2019, 55 –63. Springer, 2020.

[25] Nayak, G. and Ray, M. *Modified condition decision coverage criteria for test suite prioritization using particle swarm optimization*, Int. J. Intell. Comput. Cybern. 12(4) (2019), 425–443.

[26] Pachariya, M.K. *Building ant system for multi-faceted test case prioritization: An empirical study*, Int. J. Softw. Innov. (IJSI) 8(2)(2020), 23–37.

[27] Pan, R., Ghaleb, T.A. and Briand, L. *Atm: Black-box test case minimization based on test code similarity and evolutionary search*, 2023

IEEE/ACM 45th International Conference on Software Engineering (ICSE). IEEE, 2023.

[28] Sahin, O. and Akay, B. *Comparisons of metaheuristic algorithms and fitness functions on software test data generation*, Appl. Soft Comput. 49 (2016), 1202–1214.

[29] Sahoo, R.R. and Ray, M. *Pso based test case generation for critical path using improved combined fitness function*, J. King Saud Univ. - Comput. Inf. Sci. 32(4) (2020), 479–490.

[30] Sheikh, R., Babar, M.I., Butt, R., Abdelmaboud, A. and Eisa, T.A.E. *An optimized test case minimization technique using genetic algorithm for regression testing*, Comput. Mater. Contin. 74(3) (2023), 6789–6806.

[31] Sun, J., Chen, J. and Wang, G. *Multi-objective test case prioritization based on epistatic particle swarm optimization*, Int. J. Performability Eng. 14(10) (2018), 2441.

[32] Suri, B. and Singhal, S. *Analyzing test case selection & prioritization using ACO*, ACM SIGSOFT Software Engineering Notes 36(6) (2011), 1–5.

[33] Tyagi, M. and Malhotra, S. *Test case prioritization using multi objective particle swarm optimizer*, In: 2014 International Conference on Signal Propagation and Computer Technology (ICSPCT 2014), 390–395. IEEE, 2014.

[34] Verma, A.S., Choudhary, A. and Tiwari, S. *Regression test case selection: A comparative analysis of metaheuristic algorithms*, In: Proceedings of Second Doctoral Symposium on Computational Intelligence: DoSCI 2021, 605–615. Springer, 2022.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 2, 2024, pp 449–474
https://doi.org/10.22067/ijnao.2024.82434.1269
https://ijnao.um.ac.ir/

Research Article

Dynamics of Cholera Pathogen Carriers and Effect of Hygiene Consciousness in Cholera Outbreaks

S.F. Abubakar*, M.O. Ibrahim

Abstract

We derive a deterministic mathematical model that scrutinizes the dynamics of cholera pathogen carriers and the hygiene consciousness of individuals, before the illness, during its prevalence, and after the disease's outbreaks. The dynamics can effectively help in curtailing the disease, but its effects had less coverage in the literature. Boundedness of the solution of the model, its existence, and uniqueness are ascertained. Effects of cholera pathogen carriers and hygiene consciousness of individuals in controlling the disease or allowing its further spread are analyzed. The differential transformation method is used to obtain series solutions of the differential equations that make the system of the model. Simulations of

*Corresponding author

Received 24 May 2023; revised 31 January 2024; accepted 14 February 2024

Sani Fakai Abubakar
Department of Mathematics, Faculty of Physical Sciences, Kebbi State University of Science and Technology, Aliero, Nigeria. saniabufakai@hotmail.com

Mohammed Olanrewaju Ibrahim
Department of Mathematics, Faculty of Physical Sciences, University of Ilorin, Ilorin, Nigeria, moibraheem@yahoo.com

How to cite this article
Abubakar, S.F. and Ibrahim, M.O., Dynamics of Cholera Pathogen Carriers and Effect of Hygiene Consciousness in Cholera Outbreaks. *Iran. J. Numer. Anal. Optim.*, 2024; 14(2): 449-474. https://doi.org/10.22067/ijnao.2024.82434.1269

the series solutions of the model are carried out and displayed in graphs. The dynamics of the concerned state variables and parameters in the model are interpreted via the obtained graphs. It is observed that higher hygiene consciousness of individuals can drastically reduce catching cholera disease at onset and further spread of its infections in the population, this in turn, shortens the period of cholera epidemic.

AMS subject classifications (2020): Primary 92D30; Secondary 92B05.

Keywords: Differential transform; carriers; simulation; hygiene consciousness; series.

1 Introduction

Mathematical modeling can be defined as the process of selecting and utilizing relevant mathematical ideas to scrutinize and portray situations in a better way to understand and apply safer decisions about physical phenomena, and the entire real-life problems. Models are believed to aid in understanding various ways phenomena and real-life problems operate. This includes epidemic processes, and it helps in designing effective control strategies [1].

The world has witnessed a remarkable improvement in hygiene, sanitary measures, and effective treatment of infectious diseases. This has drastically reduced the threat of diseases like cholera in many countries. In addition to this, many studies, experiments, and theories that were carried out and developed now and in the past have aided in reducing cholera mobility [8]. A report from "acap" stated that *"there have been 93,932 suspected cholera cases, that caused 3,293 deaths from these cases in Nigeria and that 5,343 confirmed cholera cases with 156 deaths in Niger Republic as of October 2021"*.

Due to poor sanitation systems, the shortage of good drinking water, and the high density of the population, Africa become a victim of cholera outbreaks, leading to a higher risk of outbreaks, and the spread of the disease. This coupled with shifting priority to COVID-19 resulted in the diversion of most resources and attention to it, leaving other health problems. This in-

turn caused lapses in the entire health sector and affected the process of contact tracing and thorough management of cholera [27].

1.1 Cholera

Cholera is an illness caused by gram-negative, nearly comma-shaped bacillus bacteria called *Vibro cholerae*. The sickness's symptoms are mostly, severe gray color diarrhea, with flecks of mucus, extreme thirst, drop in blood pressure, abdominal pain, and kidney failure [29]. It usually comes with vomiting and can dehydrates the patient in quite short period of time if prompt treatment measures are not taken. Its outbreaks usually comes after natural and artificial catastrophes that disrupt the public strong sanitation service systems. The disease is still a threat and problem to global public health [26], especially in developing countries.

Mishandling faeces and vomit of infected individuals easily trigger spread of the disease. This is so because Vibro cholerae bacteria that pass through the human digestive system are more infective and stubborn to control by human body immunity than others. Among its crucial characteristics is that it can survive even with limited oxygen and iron [17]. It has been estimated that (though, depending on the individual's age, stomach acidity and blood group), an individual has to ingest up to a dose range of 10^3 to 10^8 *Vibro cholerae* bacteria cells, before developing cholera disease. The disease's incubation period ranges (that is, from ingestion of *Vibro cholerae* bacteria to cholera infection) from 12 hours to 5 days [11, 20]. The range is pegged at, from 18 hours to 5 days, in [24]'s study (as cited in [12]). *Vibro cholerae*'s mechanism of causing cholera was described by [25] that usually start with surviving the stomach acidity, shedding its flagellum, swim deep into intestine mucus. It colonizes and secretes a toxin that stimulates intestine *mucosal* cells to pump body water due to osmosis. This leads to dehydrating the patient.

The onset outbreak of cholera is usually caused by the consumption of contaminated water. The subsequent spread of the disease can be through contaminated food, contact with cholera patients and unhygienic handling of

faeces, and vomit of cholera patients. Most mathematical models on cholera we visited, have not categorically scrutinized cholera pathogens since the pathogens cannot on their own spread without human contribution. It is hence vital to study the dynamics of *Vibro cholerae* as causative agents or pathogens that cause cholera disease [28].

Cholera is an extremely virulent disease that affects all ages and economic classes of a nation. It has been estimated that about two billion people drink water from sources that may be contaminated with faeces and that 1.3 billion people are at risk of cholera disease in cholera endemic areas and regions of the world [29]. 75% of cholera pathogens carriers, manifest no symptoms but spread the pathogen to their immediate environment through faeces for a period of 7–14 days, of ingesting the pathogens by the carriers [13]. People with low gastric acidity are at high risk of catching cholera. Therefore transmission of the disease in such a class of people is rampant [1]. The study intends to apply mathematical modeling to best portray the transmission thresholds of the infections by *Vibro cholerae* in the host, and the magnitudes it attains so as to get the best ways to eradicate it.

1.2 Related literature

Elhia et al. [10] proposed a mathematical model on the global stability of the SIR cholera epidemic model that dealt with two infectious stages and treatment of cholera. They carried out an analysis of the model's global stability, obtained R_0, and illustrated the analytical results numerically.

A mathematical model on the controls of cholera, including hygiene consciousness as a control strategy and a compartment of the model was formulated in [21].

Mathematical models on the dynamics of individuals who ingested cholera pathogens but have not developed any symptoms of cholera disease are very rare. These categories of individuals unknowingly spread the pathogens of the disease in the environment. This in turn causes the spread of the disease. Therefore, incorporating all possible and necessary compartments in a mathematical model, will surely warrant a chance to determine the essential

dynamics it entails. This will also help in determining the basic mechanisms of transmission of the disease. Its outbreak and ways individuals can be infected by the disease. It can also help in exploring various ways of its spread, methods of controlling it, and what can make it escalate or persist.

These motivate us to formulate a deterministic mathematical model of cholera by marrying [21]'s SICRB cholera model with [10]'s SIUTR cholera model, to come up with SICUTRB deterministic mathematical model of cholera, in order to investigate the dynamics, spread, and control measures of cholera disease, especially through hygiene consciousness. The model has two incidences, linear and nonlinear. The linear incidence represents the human-to-human transmission of cholera, and the nonlinear is the logistic response to the increase in the cholera pathogen B, and ϕ is represented by $v1\frac{B}{K+B}$.

Differential transform method (DTM) is a numerical-analytical (others call it semi-analytical) technique. It is used in solving linear and nonlinear, ordinary and partial differential equations, usually with known initial values. It is carried out by obtaining the approximate series solutions of the equations. It was first used by [30] in solving linear and nonlinear initial value problems in electrical circuit analysis. Pukhov earlier proposed differential transform by computing the image of a transformed function using differential operations, in his work 'Differential Transformations of Functions and Equations' published in 1980. Zhou pioneered using the method to solve the initial value problem [5].

The method is an iterative procedure for getting analytic Taylor series solutions of differential equations by iterations [4, 31]. In other words, DTM solves an induced recursive equation from a given differential equation by determining coefficients of the Taylor series of the concern function [2]. Its analytical solutions can be constructed in the form of polynomials, whose obtained solutions can be presented in terms of convergence series. This in turn makes it easier to compute especially nonlinear equations. Application of the method reduces the bulk computation and helps arrive at more accurate approximate answers while solving linear and nonlinear differential equations. In recent years, the method has been employed in solving one-dimensional planar Bratu's problem, differential-difference equations, delay differential

equations, differential-algebraic equations, and integro-differential systems [32].

Some basic definitions and properties of DTM as obtained in related literature, such as [2, 3, 14, 18, 19, 23], are as follows: The transformation of the bth derivative of a function with one variable is

$$G[b] = \frac{1}{b!}\left[\frac{d^b g(t)}{dt^b}\right]_{t=t_0}. \tag{1}$$

Here $g(t)$ is the original function, while $G[b]$ is the transformed function, and the inverse transformation is given as

$$g(t) = \sum_{b=0}^{\infty} G[b](t-t_0)^b. \tag{2}$$

By Taylor series, the inverse transformation of (1) can be obtained as

$$g(t) = \sum_{b=0}^{\infty} t^b G[b]. \tag{3}$$

Substituting (1) into (3) gives

$$g(b) = \sum_{b=0}^{\infty}\left[\frac{t^b}{b!}\left(\frac{d^b g(t)}{dt^b}\right)\right]_{t=t_0}. \tag{4}$$

Some functions and their respective differential transformations are displayed in Table 1.

2 Model formulation

State variables and parameters used in the model are described in Tables 2 and 3, respectively.

Table 1: Some functions and their respective differential transforms

Original Function	Transformed Function
$g(t) = f(t) \pm h(t)$	$G[b] = F[b] \pm H[b]$
$g(t) = mf(t)$	$G[b] = mF[b]$ m is a constant
$g(t) = 1$	$G[b] = \delta[b]$
$g(t) = t$	$G[b] = \delta[b-1]$
$g(t) = f(t)h(t)$	$G[b] = \sum_{a=0}^{b} F[a]H[b-a]$
$g(t) = e^{\lambda t}$	$G[b] = \frac{\lambda^b}{b!}$
$g(t) = e^t$	$G[b] = \frac{1}{b!}$
$g(t) = \frac{df(t)}{dt}$	$G[b] = (b+1)G[b+1]$
$g(t) = \frac{d^2f(t)}{dt^2}$	$G[b] = (b+1)(b+2)G[b+2]$
$g(t) = t^a$	$G[b] = \delta(b-a) = \begin{cases} 1, & \text{if } b = a \\ 0, & \text{if b} \neq \text{a} \end{cases}$ where δ is the Kronecker delta
$g(t) = (1+t)^a$	$G[b] = \frac{a(a-1)\cdots(a-b+1)}{b!}$
$g(t) = \sin(\theta t + \alpha)$	$G[b] = \frac{\theta^b}{b!} \sin\left(\frac{b\pi}{2} + \alpha\right)$
$g(t) = \cos(\theta t + \alpha)$	$G[b] = \frac{\theta^b}{b!} \cos\left(\frac{b\pi}{2} + \alpha\right)$

Table 2: State variables description

Variable	Description
S	Susceptible Individuals
I	Infected Individuals
C	Hygiene Conscious Individuals
U	Untreated *Vibro cholerae* Carries
T	Infected Individuals Under Treatment
R	Recovered Individuals
B	Concentration of *Vibro cholerae* Bacteria in Food and Water Consume by the Human Population

Table 3: Parameters description

State Variables & Parameter	Description	Value	Reference
S	Susceptible Individuals	900	Estimated
I	Infected Individuals	50	Calculated
C	Hygiene Conscious Individuals	500	Estimated
U	Untreated *Vibro cholerae* Carries	25	Calculated
T	Infected Individuals Under Treatment	22	Estimated
R	Recovered Individuals	3	Calculated
B	Concentration of *Vibro cholerae*Bacteria in Food	600	Estimated
q	Recruitment rate of human population	0.104	Estimated
δ	Per capital natural human death rate	0.0101	[10]
v_1	Rate of ingesting *Vibro Cholerae* bacteria from contaminated food and water consume by the human population	0.5	[21]
v_2	Rate of getting *Vibro cholerae* bacteria through human to human interaction	0.03	[21]
θ	Recovery rate of hygiene individuals	0.38	[21]
h	Hygiene consciousness of individuals	0.6	Estimated
β	Rate of *Vibro cholerae* decrease	0.78	Estimated
η	Rate at which infected individuals increase *Vibro cholerae* bacteria in the environment	10	[6]
f	Rate of freedom from *Vibro cholerae* by the hygiene conscious individuals	0.7	Estimated
k	Concentration of cholera pathogen that yields 50% chance of individual developing cholera disease	10^6	[6]
g_1	Progression of recovery of treated individuals to untreated *Vibro cholerae* carriers	0.08	Estimated
g_2	Rate of interruption of treatment	0.2	[10]
r_1	Recovery rate of untreated *Vibro Cholerae* carriers	0.8	Estimated
r_2	Recovery of treated cholera patients	0.9	Estimated
ϵ	Disease induced death rate for individuals in I class	0.105	Estimated
μ	Disease induced death rate for individuals in U class	0.1	Estimated
α	Disease induced death rate for individuals in T class	0.08	Estimated
m	Rate of movement of the infected class	0.4	[10]

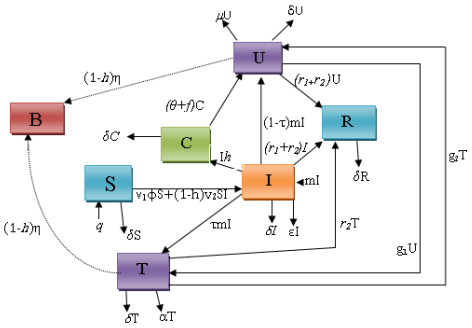


Figure 1: Schematic diagram of the model

2.1 Basic assumptions of the model

The following basic assumptions are used in formulating the model. They serve as conditions under which the model will be meaningful and work efficiently.

1. A homogeneous human population is used.

2. Infected individuals are classified into under-treatment and untreated healthy *Vibro cholerae* carriers.

3. State variables and parameters of the model are positive and remain positive.

4. Only few cholera pathogen carriers develop severe cholera disease (about 75% of cholera infected individuals are healthy cholera pathogen carriers) during a cholera outbreak.

5. Recovery of severe cholera patients is through treatment and asymptomatic cholera pathogen carriers can heal untreated.

6. Recruitment into the population is through birth and immigration while the exit is through migration and natural and disease-related deaths.

   These assumptions were used to draw the flow diagram in Figure 1, then used to derive the model's system of differential equations, as follows:

$$\frac{dS}{dt} = q - v_1\phi S - (1-h)v_2 SI - \delta S, \tag{5}$$
$$\frac{dI}{dt} = v_1\phi S + (1-h)v_2 SI - (\delta + \varepsilon + r_1 + r_2 + h)I, \tag{6}$$
$$\frac{dC}{dt} = hI - (\theta + f + \delta)C, \tag{7}$$
$$\frac{dU}{dt} = (1-\tau)mI + (\theta + f)C + g_2 T - (\delta + r_1 + r_2 + \mu + g_1)U, \tag{8}$$
$$\frac{dT}{dt} = \tau mI + g_1 U - (\delta + \alpha + g_2 + r_2)T, \tag{9}$$
$$\frac{dR}{dt} = (r_1 + r_2)I + r_2 T + (r_1 + r_2)U - \delta R, \tag{10}$$
$$\frac{dB}{dt} = (1-h)\eta I - \beta B. \tag{11}$$

2.2 Existence and uniqueness of solution of the model

Formulated mathematical models need to undergo a surety test to ascertain the existence and uniqueness of their solution. Using Lipschitz criteria as in the definition of basic terms, the existence and uniqueness of a solution of the model are established as follows:

Theorem 1 stated below, is from Derrick and Grossman (1981), also available in [9, 16], together with Theorem 2 are adopted to prove the existence and uniqueness of solution of the model.

Theorem 1. Suppose that D^1 denotes the region

$$| t - t_0 | \leq l, \qquad \| a - a_0 \| \leq m,$$

$$a = (a_1, a_2, \ldots, a_n), \qquad a_0 = (a_{10}, a_{20}, \ldots, a_{n0}),$$

and that

$$\|g(a_1, t) - g(a_2, t)\| \leq p \|a_1 - a_2\|. \tag{12}$$

If the pair (a_1, t) and (a_2, t) belong to D^1, where p is a positive constant, then there is a constant $\delta > 0$ such that there exist a unique continuous vector solution $a(t)$ of the system in the interval $t - t_0 \leq \delta$. This condition is satisfied provided that

$$\frac{\partial g_i}{\partial t_j}, \qquad i = 1, 2, \ldots,$$

is continuous and bounded in D^1, (where i represents equations in the system and j represents the compartments).

Now, considering the region $0 \leq w \leq \mathbb{R}$. Bounded solution of the system of equations (5)–(11) will be looked for in the same region whose partial derivatives satisfy $\delta \leq w \leq \mathbb{R}$, where δ and w are positive constants, as in the theorem that follows.

Theorem 2. Let **D** denotes the region $0 \leq w \leq \mathbb{R}$. Then the system of equations (5)–(11) has a unique solution if it is established that

$$\frac{\partial l_i}{\partial t_j}, \qquad i = 1, 2, 3, 4, 5, 6, 7,$$

are continuous and bounded in **D**.

The proof of the existence and uniqueness of solution of model two system of equations (5)–(11) are as follows:

Proof. Let

$$\begin{aligned}
l_1 &= q - v_1\phi S - (1-h)v_2 SI - \delta S,\\
l_2 &= v_1\phi S + (1-h)v_2 SI - (\delta + \varepsilon + r_1 + r_2 + h)I,\\
l_3 &= hI - (\theta + f + \delta)C,\\
l_4 &= (1-\tau)mI + (\theta + f)C + g_2 T - (\delta + r_1 + r_2 + \mu + g_1)U,\\
l_5 &= \tau mI + g_1 U - (\delta + \alpha + g_2 + r_2)T,\\
l_6 &= (r_1 + r_2)I + r_2 T + (r_1 + r_2)U - \delta R,\\
l_7 &= (1-h)\eta I - \beta B.
\end{aligned}$$

Partial derivatives with respect to each state variables of the system (5)–(11) are obtained as follows:
For l_1,

$$\left|\frac{\partial l_1}{\partial S}\right| = |-v_1\phi - (1-h)v_2 I - \delta| < \infty, \qquad \left|\frac{\partial l_1}{\partial I}\right| = |-(1-h)v_2 S| < \infty,$$
$$\left|\frac{\partial l_1}{\partial C}\right| = \left|\frac{\partial l_1}{\partial U}\right| = \left|\frac{\partial l_1}{\partial T}\right| = \left|\frac{\partial l_1}{\partial R}\right| = \left|\frac{\partial l_1}{\partial B}\right| = 0 < \infty.$$

Similarly, for l_2, l_3, l_4, l_5, l_6, and l_7, all have the absolute values of their partial derivatives less than infinity. □

Solution of the system of equations (5)–(11) exists and is unique in $\mathbb{R}^7$ since all the partial derivatives are bounded and defined.

2.3 Boundedness of solutions of the model

Let P_H be the human population of the model represented by equations (5)–(10). The variables and parameters used are nonnegative.

Theorem 3. All solutions of equations (5)–(10) of the model are bounded in $\mathbb{R}^6_+$, and solution to equation (11) is bounded in $\mathbb{R}_+$.

Proof. Let

$$P_H = S + I + C + U + T + R. \tag{13}$$

Differentiating (13) with respect to t gives

$$\frac{dP_H}{dt} = \frac{dS}{dt} + \frac{dI}{dt} + \frac{dC}{dt} + \frac{dU}{dt} + \frac{dT}{dt} + \frac{dR}{dt}. \tag{14}$$

Simplifying (14) gives

$$\frac{dP_H}{dt} = q - \delta S - (\varepsilon - m + \delta)I - (\delta)C - (\mu + \delta)U - (\alpha + \delta)T - \delta R. \tag{15}$$

Taking and using any $\rho > 0$, multiplying the ρ by (13), then adding the result to respective sides of (15), and simplifying, thus we have

$$\rho(P_H) = \rho(S + I + C + U + T + R),$$

$$\begin{aligned}\frac{dP_H}{dt} + \rho P_H =& q - (\delta - \rho)S - (\varepsilon - m + \delta - \rho)I - (\delta - \rho)C - (\mu + \delta - \rho)U \\ & - (\alpha + \delta - \rho)T - (\delta - \rho)R.\end{aligned}$$

Let

$$A_q = (\delta-\rho)S-(\varepsilon-m+\delta-\rho)I-(\delta-\rho)C-(\mu+\delta-\rho)U-(\alpha+\delta-\rho)T-(\delta-\rho)R$$

with

$$\rho = \min\{\delta, \varepsilon - m + \delta, \delta, \mu + \delta, \alpha + \delta, \delta\},$$

which means that ρ is the minimum value of δ. Hence $\delta - \rho \leq 0$. Similarly, ρ is the minimum values of $\varepsilon - m + \delta$, $\mu + \delta$, and $\alpha + \delta$. This implies $\varepsilon - m + \delta - \rho \leq 0$, $\mu + \delta - \rho \leq 0$, and $\alpha + \delta - \rho \leq 0$.

Therefore

$$\frac{dP_H}{dt} + \rho P_H = q - A_q$$

but $A_q \leq 0$. Taking $A_q = 0$ gives

$$\frac{dP_H}{dt} + \rho P_H \leq q. \tag{16}$$

Now, using the differential inequality theorem from Gronwall (1919) and the Brikhoff and Rota version as used in [7, 15, 21, 22], (16) gives

$$\frac{dP_H}{dt} \leq q \ - \rho P_H. \tag{17}$$

Separating the differential inequality (17) gives

$$\frac{dP_H}{q \ - \rho P_H} \leq dt. \tag{18}$$

Integrating the both sides of (18) gives

$$-\frac{1}{\rho}\ln(q - \rho P_H) \leq t + c. \tag{19}$$

Solving (19) in terms of c gives

$$-t - \frac{1}{\rho}\ln(q - \rho P_H) \leq c. \tag{20}$$

Evaluating (20) at $t = 0$ and $P_H = P_H(0)$ gives

$$-\frac{1}{\rho}\ln(q - \rho P_H(0)) \leq c. \tag{21}$$

Substituting (21) into (19) gives

$$-\frac{1}{\rho}\ln(q - \rho P_H) \leq t + \left(-\frac{1}{\rho}\ln(q - \rho P_H(0))\right). \tag{22}$$

Multiplying (22) by $-\rho$ gives

$$\ln\left(q-\rho P_H\right) \leq -\rho t + \ln\left(q-\rho P_H(0)\right). \tag{23}$$

Taking exponent of (23) gives

$$q-\rho P_H \leq e^{-\rho t}\left(q-\rho P_H(0)\right) \tag{24}$$

as $t \to \infty$,

$$e^{-\rho t} \to 0.$$

Also,

$$e^{-\rho t}\left(q-\rho P_H(0)\right) \to 0,$$

$$q-\rho P_H \leq 0,$$

so $P_H \leq \frac{q}{\rho}$. All state variables and parameters of the model are positive. Hence the sum P_H is greater than zero. Therefore

$$0 < P_H \leq \frac{q}{\rho}.$$

This cumulatively implies the system of equations (5)–(10) is bounded and its solutions enter into the region

$$H_h = \left\{(S,I,C,U,T,R) \in \mathbb{R}_+^6 \ : 0 < P_H \leq \frac{q}{\rho}\right\}. \tag{25}$$

Hence, the solution of equations (5)–(10) is bounded and belongs to feasible region H_h. On the other hand, the nonhuman population is described by equation (11). Its boundedness is sought as follows.

Let P_B denote the number of cholera pathogens, so that equation (11) can be written as

$$\frac{dP_B}{dt} = (1-h)\eta I - \beta P_B.$$

From (13) $\frac{dP_H}{dt} \leq q - \rho P_H$. Since $P_H \leq \frac{q}{\rho}$, then from the same (13), it implies that $I \leq \frac{q}{\rho}$.

Therefore

$$\frac{dP_B}{dt} \leq \frac{q}{\rho}(1-h)\eta - \beta P_B$$

holds and can be written as

$$\frac{dP_B}{\frac{q\eta}{\rho}(1-h) - \beta P_B} \leq dt. \tag{26}$$

Integrating the both sides of (26) gives

$$-\frac{1}{\beta}\ln\left(\frac{q\eta}{\rho}(1-h)-\beta P_B\right)\leq t+c_1. \tag{27}$$

Evaluating (27) at $t=0$ and $P_B=P_B(0)$ gives

$$-\frac{1}{\beta}\ln\left(\frac{q\eta}{\rho}(1-h)-\beta P_B(0)\right)<c_1. \tag{28}$$

Substituting (28) into (27) gives

$$-\frac{1}{\beta}\ln\left(\frac{q\eta}{\rho}(1-h)-\beta P_B\right)\leq t-\frac{1}{\beta}\ln\left(\frac{q\eta}{\rho}(1-h)-\beta P_B(0)\right). \tag{29}$$

Multiplying both sides of (29) by $-\beta$ gives

$$\ln\left(\frac{q\eta}{\rho}(1-h)-\beta P_B\right)\leq -\beta t+\ln\left(\frac{q\eta}{\rho}(1-h)-\beta P_B(0)\right). \tag{30}$$

Taking exponent of both sides of (30) gives

$$\frac{q\eta}{\rho}(1-h)-\beta P_B\leq e^{-\beta t+\ln\left(\frac{q\eta}{\rho}(1-h)-\beta P_B(0)\right)},$$

which can be written as

$$\frac{q\eta}{\rho}(1-h)-\beta P_B\leq e^{-\beta t}\left(\frac{q\eta}{\rho}(1-h)-\beta P_B(0)\right).$$

Then, as $t\to\infty$, $e^{-\beta t}\to 0$, and also, $e^{-\beta t}\left(\frac{q\eta}{\rho}(1-h)-\beta P_B(0)\right)\to 0$.

it is obtained that

$$P_B\leq\frac{q\eta}{\beta\rho}(1-h),$$

$$\Rightarrow P_B\to\frac{q\eta}{\beta\rho}(1-h).$$

Therefore

$$0<P_B\leq\frac{q\eta}{\beta\rho}(1-h).$$

The nonhuman population equation solution is hence bounded and enters the region

$$H_b=\left\{B\in\mathbb{R}_+, 0<P_B\leq\frac{q\eta}{\beta\rho}(1-h)\right\}. \tag{31}$$

Altogether, the feasible solutions of the system of equations (5)–(11) are bounded and enter the region

$$H = \Big\{(S, I, C, U, T, R) \in \mathbb{R}^6_+, B \in \mathbb{R}_+, S, I, C, U, T, R, B \geq 0,$$
$$0 < P_H \leq \frac{q}{\rho}, 0 < P_B \leq \frac{q\eta}{\beta\rho}(1-h)\Big\}. \quad (32)$$

□

3 Results of the model using DTM

Let

$$S = s(t), \quad I = i(t), \quad C = c(t), \quad U = u(t), \quad T = t_1(t), \quad R = r(t), \quad B = b_1(t),$$

$$G_1 = \delta + \varepsilon + r_1 + r_2 + h,$$
$$G_2 = \delta + \mu + r_1 + r_2 + g_1,$$
$$G_3 = \delta + \alpha + g_2 + r_2,$$
$$G_4 = \delta + \theta + f,$$
$$G_5 = \theta + f,$$
$$G_6 = \eta - \eta h,$$
$$G_7 = v_1\phi,$$
$$G_8 = v_2 - v_2 h,$$
$$G_9 = m\tau,$$
$$G_{10} = m - m\tau.$$

The system of ordinary differential equations of the model is now written as

$$\frac{ds(t)}{dt} = q - G_7 s(t) - G_8 s(t) i(t) - \delta s(t), \quad (33)$$

$$\frac{di(t)}{dt} = G_7 s(t) + G_8 s(t) i(t) - G_1 i(t), \quad (34)$$

$$\frac{dc(t)}{dt} = hi(t) - G_4 c(t), \quad (35)$$

$$\frac{du(t)}{dt} = G_{10} i(t) + G_5 c(t) + g_2 t_1, (t) - G_2 u(t) \quad (36)$$

$$\frac{dt_1(t)}{dt} = G_9 i(t) + g_1 u(t) - G_3 t_1(t), \tag{37}$$

$$\frac{dr(t)}{dt} = G_5 c(t) + r_1 u(t) + r_2 t_1(t) - \delta r(t), \tag{38}$$

$$\frac{db_1(t)}{dt} = G_6 i(t) - \beta b_1(t). \tag{39}$$

Let $S(b)$, $I(b)$, $C(b)$, $U(b)$, $T(b)$, $R(b)$, and $B(b)$ denote the differential transformation of $s(t)$, $i(t)$, $c(t)$, $u(t)$, $t_1(t)$, $r(t)$, and $b_1(t)$, respectively. Applying the fundamental operations of DTM provided in Table 1, the following recurrence relations are obtained:

$$S(b+1) = \frac{1}{b+1}\left(qL(b,0) - (G_7+\delta)S(b) - G_8\sum_{a=0}^{b}(S(a)I(b-a))\right), \tag{40}$$

$$I(b+1) = \frac{1}{b+1}\left(G_7 S(b) + G_8\sum_{a=0}^{b}(S(a)I(b-a)) - G_1 I(b)\right), \tag{41}$$

$$C(b+1) = \frac{1}{b+1}\left(hI(b) - G_4 C(b)\right), \tag{42}$$

$$U(b+1) = \frac{1}{b+1}\left(G_{10} I(b) + G_5 C(b) + g_2 T(b) - G_2 U(b)\right), \tag{43}$$

$$T(b+1) = \frac{1}{b+1}\left(G_9 I(b) + g_1 U(b) - G_3 T(b)\right), \tag{44}$$

$$R(b+1) = \frac{1}{b+1}\left(G_5 C(b) + r_1 U(b) + r_2 T(b) - \delta R(b)\right), \tag{45}$$

$$B(b+1) = \frac{1}{b+1}\left(G_6 I(b) - \beta B(b)\right). \tag{46}$$

Substituting the initial conditions
$s(0) = 900$, $i(0) = 50$, $c(0) = 500$, $u(0) = 25$, $t_1(0) = 22$, $r(0) = 3$, $b_1(0) = 600$
and the values of the parameters
$q = 0.1$, $\delta = 0.0101$, $v_1 = 0.5$, $v_2 = 0.03$, $\theta = 0.38$, $h = 0.6$, $\beta = 0.8$, $\eta = 10$, $f = 0.7$, $g_1 = 0.08$, $g_2 = 0.2$, $r_1 = 0.8$, $r_2 = 0.9$, $\epsilon = 0.1$, $\mu = 0.1$, $\alpha = 0.08$, $m = 0.4$, $\tau = 0.5$, $\phi = 0.000599$,
into DTM conditions (40) to (46). Using Maple 18, we obtain

$$
\begin{aligned}
&S(1) = -549.25955, && I(1) = 419.76455,\\
&C(1) = -515.05, && U(1) = 507.1475,\\
&S(2) = -2099.094692, && I(2) = 1596.031182,\\
&C(2) = 406.6573675, && U(2) = -716.84851,\\
&S(3) = -4396.37805, && I(3) = 3121.246753,\\
&C(3) = 171.4405043, && U(3) = 709.150619,\\
&S(4) = -2483.196769, && I(4) = 613.668424,\\
&C(4) = 421.4651895, && U(4) = -129.7781802,\\
&S(5) = 15561.71779, && I(5) = -15852.50219,\\
&C(5) = -18.24762974, && U(5) = 170.74677881,\\
&S(6) = 56847.30013, && I(6) = -50505.80978,\\
&C(6) = -1581.934929, && U(6) = -585.9512738,
\end{aligned}
$$

$$
\begin{aligned}
&T(1) = -14.1822, && R(1) = 579.7697, && B(1) = 320,\\
&T(2) = 70.7014731, && R(2) = -84.577827, && B(2) = 3230.1164,\\
&T(3) = 59.23884417, && R(3) = -23.26776309, && B(3) = 7650.80193,\\
&T(4) = 152.6203129, && R(4) = 201.506551, && B(4) = 1095.82663,\\
&T(5) = -13.8564008, && R(5) = 97.33658518, && B(5) = 210.966696,\\
&T(6) = -523.3916988, && R(6) = 17.2393537, && B(6) = -42301.46807,\\
& && R(7) = -378.3539126, && B(7) = -110607.3974,
\end{aligned}
$$

The series solutions of the transformed equations b ranges from 1 to 6 or 7:

$$
\begin{aligned}
s(t) &= \sum_{a=0}^{b} t^b S(b) = 900 - 549.25955t - 2099.094692t^2 - 4396.37805t^3\\
&\quad - 2483.196769t^4 + 15561.71779t^5 + 56847.30013t^6 + \cdots,\\
i(t) &= \sum_{a=0}^{b} t^b I(b) = 50 + 419.76455t + 1596.031182t^2 + 3121.246753t^3\\
&\quad + 613.668424t^4 - 15852.50219t^5 - 50505.80978t^6 + \cdots,\\
c(t) &= \sum_{a=0}^{b} t^b C(b) = 500 - 515.05t + 406.6573675t^2 + 171.4405043t^3\\
&\quad + 421.4651895t^4 - 18.24762974t^5 - 1581.934929t^6 + \cdots,\\
u(t) &= \sum_{a=0}^{b} t^b U(b) = 25 + 507.1475t - 716.84851t^2 + 709.150619t^3\\
&\quad - 129.7781802t^4 + 170.74677881t^5 - 585.9512738t^6 + \cdots,
\end{aligned}
$$

$$
\begin{aligned}
t_1(t) &= \sum_{a=0}^{b} t^b T(b) = 22 - 14.1822t + 70.7014731t^2 + 59.23884417t^3 \\
&\quad + 152.6203129t^4 - 13.8564008t^5 - 523.3916988t^6 + \cdots, \\
r(t) &= \sum_{a=0}^{b} t^b R(b) = 3 + 579.7697t - 84.577827t^2 - 23.26776309t^3 \\
&\quad + 201.506551t^4 + 97.33658518t^5 + 17.2393537t^6 \\
&\quad - 378.3539126t^7 + \cdots, \\
b_1(t) &= \sum_{b=0}^{7} t^b B(b) = 600 + 330t + 3226.6364t^2 + 7662.06045t^3 \\
&\quad + 10960.11186t^4 + 214.6803636t^5 - 42259.50383t^6 \\
&\quad - 110607.3974t^7 + \cdots.
\end{aligned}
$$

The Maple 18 plots of $s(t)$, $i(t)$, $c(t)$, $u(t)$, $t_1(t)$, $r(t)$, and $b_1(t)$ are displayed in Figures 2–6.

4 Conclusion

The model proposed how hygiene consciousness of individuals influences the susceptible and infected individuals. A high rate of h, which is hygiene consciousness of individuals, increases number of susceptible individuals while the low rate of h decreases their number, for details, see Figures 3, 4 and 5. The untreated *Vibro Cholerae* carriers, are seen physically as healthy, because they do not manifest any symptoms of cholera, but spread the pathogen without knowing. The rate of freedom of individuals from *Vibro Cholerae* was determined by their hygiene consciousness (see Figure 6). In Figure 3, the number of susceptible individuals reduced from 900 to about 100 individuals in less than a week, infection reduced to zero in about half of a week, the peak of the infection is 350 out of 900 susceptible individuals when the rate of hygiene consciousness is 63%. While the number of susceptible individuals reduced from 900 to 90 individuals, in about two weeks' time, infection reduced to zero in two weeks, while the peak of infection is 200 out of 900 susceptible individuals, when the rate of hygiene consciousness is 82%. The number of

susceptible individuals dropped from 900 to 400 individuals, in about a week, infection reduced from 50 to zero individuals in the same period, peak of the infection is 50 out of 900 susceptible, when the rate of hygiene consciousness is 99%, as in Figure 5. Reduction of hygiene consciousness results in the prolongation of the cholera illness period to about two weeks and the maximum number of infected individuals reach 200.

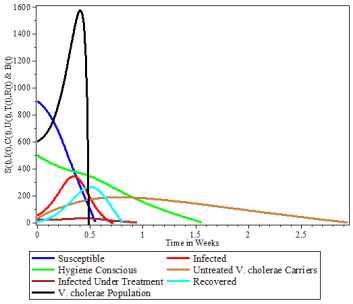


Figure 2: Combined dynamics of the seven compartments of the model plotted against time.

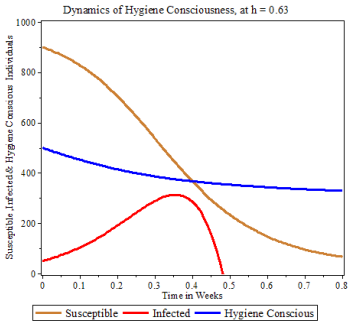


Figure 3: Dynamics of hygiene consciousness, its effect on susceptible and infected individuals at $h = 0.63$. As higher hygiene consciousness is maintained, number of individuals susceptible to cholera drastically reduced, and cholera infection stopped.

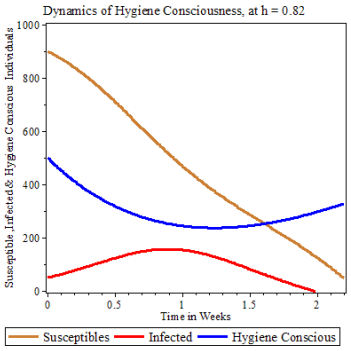


Figure 4: Dynamics of hygiene consciousness, its effect on susceptible and infected individuals at $h = 0.82$. Maintaining higher rate of hygiene consciousness, will make the number of individual susceptible to cholera reduced drastically and cholera spread and infection stopped immediately.

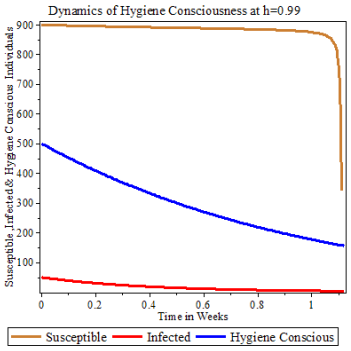


Figure 5: Dynamics of hygiene consciousness, its effect on susceptible and infected individuals at $h = 0.99$. Applying and maintaining about 100% hygiene consciousness, in societies, will within just about 10 days, make the number of susceptible individuals to approach zero as well as the rate of cholera infection.

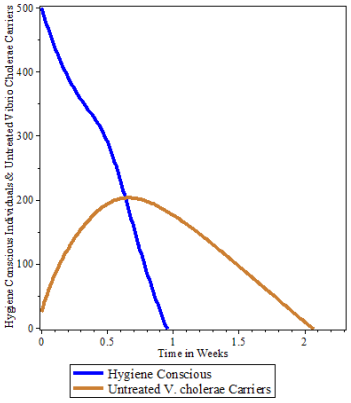


Figure 6: Effect of hygiene consciousness on untreated *Vibrio cholerae* bacteria carriers.

Acknowledgment

I would like to thank Kebbi State University of Science and Technology, Aliero, for sponsoring my Ph.D. at the University of Ilorin. This work is part of my Ph.D. thesis.

References

[1] Adagbada, A.O., Adesida, S.A., Nwaokorie, F.O., Niemogha, M.T. and Coker, A.O. *Cholera epidemiology in Nigeria: an overview.* Pan Afr. Med. J. 12(1) (2012).

[2] Ahmad, M.Z., Alsarayreh, D., Alsarayreh, A. and Qaralleh, I. *Differential transformation method (DTM) for solving SIS and SI epidemic models.* Sains Malays. 46 (10) (2017) 2007–2017.

[3] Akinboro, F.S., Alao, S. and Akinpelu, F.O. *Numerical solution of SIR model using differential transformation method and variational iteration method.* General Mathematics Notes 22(2) (2014) 82–92.

[4] Batiha, K. and Batiha, B. *A new algorithm for solving linear ordinary differential equations.* World Appl. Sci. J. 15 (12) (2011) 1774–1779.

[5] Chakraborty, A.K., Shahrear, P. and Islam, M.A. *Analysis of epidemic model by differential transform method.* J. Multidiscip. Eng. Sci. Technol. 4(2) (2017) 6574–6581.

[6] CCodeço, C.T., *Endemic and epidemic dynamics of cholera: the role of the aquatic reservoir.* BMC Infect. Dis. 1 (2001) 1–14.

[7] Edwards, C.H. and Penney, D.E. *Differential equations and boundary value problems: computing and modeling.* Pearson Educación. 2000.

[8] Edward, S. and Nyerere, N. *A mathematical model for the dynamics of cholera with control measures.* Comput. Appl. Math. 4(2) (2015) 53–63.

[9] Egbetade, S.A. and Ibrahim, M.O. *Modelling the Impact of BCG vaccines on tuberculosis epidemics.* Journal of Mathematical Modelling and Application, 9(1) (2014) 49–55.

[10] Elhia, M., Laaroussi, A., Rachik, M., Rachik, Z. and Labriji, E., *Global stability of a susceptible-infected-recovered (SIR) epidemic model with two infectious stages and treatment.* Int. J. Sci. Res. 3(5) (2014) 114–121.

[11] Harris JB, LaRoque RC, Qadri F, Ryan ET, Calderwood SB. Seminar, The Lancet 2012.

[12] Jahan, S., *Cholera–epidemiology, prevention and control.* Significance, Prevention and Control of Food Related Diseases. Croatia: InTech, (2016) 145–157.

[13] Javidi, M. and Ahmad, B., *A study of a fractional-order cholera model.* Appl. Math. Inf. Sci. 8(5) (2014) p.2195.

[14] Kenmogne, F. *Generalizing of differential transform method for solving nonlinear differential equations.* J. Appl. Comp. Math. 4 (2015) p.196.

[15] Mafuta, P., Mushanyu, J. and Nhawu, G. *Invariant region, endemic equilibria and stability analysis.* IOSR J. Math. 10(2) (2014) 118–120.

[16] Madubueze, C.E., Kimbir, A.R., Onah, E.S. and Aboiyar, T., *Existence and uniqueness of solution of ebola virus disease model with contact tracing and quarantine as controls.* Nigerian J. Math. Appl. 25 (2016) 111–121.

[17] Merrell, D.S., Butler, S.M., Qadri, F., Dolganov, N.A., Alam, A., Cohen, M.B., Calderwood, S.B., Schoolnik, G.K. and Camilli, A., *Host-induced epidemic spread of the cholera bacterium.* Nature, 417(6889) (2002) 642–645.

[18] Mirzaee, F., *Differential transform method for solving linear and non-linear systems of ordinary differential equations.* Appl. Math. Sci. 5(70) (2011) 3465–3472.

[19] Najafgholipour, M. and Soodbakhsh, N., *Modified differential transform method for solving vibration equations of MDOF systems.* Civil Eng. J. 2(4) (2016) 123–139.

[20] Nelson, E.J., Harris, J.B., Glenn Morris Jr, J., Calderwood, S.B. and Camilli, A., *Cholera transmission: the host, pathogen and bacteriophage dynamic.* Nat. Rev. Microbiol. 7(10) (2009) 693–702.

[21] Ochoche, M.J., Madubueze, E.C. and Akaabo, T.B., *A mathematical model on the control of cholera: hygiene consciousness as a strategy.* J. Math. Comput. Sci. 5(2) (2015) 172–187.

[22] Panja, P. and Mondal, S.K., *A mathematical study on the spread of Cholera.* South Asian J. Math. 4(2) (2014) 69–84.

[23] Patil, N. and Khambayat, A., *Differential transform method for ordinary differential equations.* Res. J. Math. Stat. Sci. 3 (2014) 330–337.

[24] Penrose, K., Castro, M.C.D., Werema, J. and Ryan, E.T., *Informal urban settlements and cholera risk in Dar es Salaam, Tanzania.* PLoS Negl. Trop. Dis. 4(3) (2010) p.e631.

[25] Todar K. Online Textbook of Bacteriology. www.onlinetextbookofbacteriology.net 2008.

[26] Tuite, A.R., Chan, C.H. and Fisman, D.N., *Cholera, canals, and contagion: Rediscovering Dr Beck's report.* J. Public Health Policy, 32 (2011) 320–333.

[27] Uwishema, O., Okereke, M., Onyeaka, H., Hasan, M.M., Donatus, D., Martin, Z., Oluwatomisin, L.A., Mhanna, M., Olumide, A.O., Sun, J. and Adanur, I., *Threats and outbreaks of cholera in Africa amidst COVID-19 pandemic: a double burden on Africa's health systems.* Tropical medicine and health, 49(1) (2021) p.93.

[28] Wang, X. and Wang, J., *Analysis of cholera epidemics with bacterial growth and spatial movement.* J. Biol. Dyn. 9(sup1) (2015) 233–261.

[29] Wang, J., *Mathematical models for cholera dynamics—A review.* Microorganisms, 10(12) (2022) p.2358.

[30] Soltanalizadeh, B. and Branch, S., *Application of differential transformation method for solving a fourth-order parabolic partial differential equations.* Int. J. Pure Appl. Math. 78(3) (2012) 299–308.

[31] Munganga, J.M.W., Mwambakana, J.N., Maritz, R., Batubenge, T.A. and Moremedi, G.M., *Introduction of the differential transform method*

to solve differential equations at undergraduate level. Int. J. Math. Educ. Sci. Technol. 45(5) (2014) 781–794.

[32] Saeed, R.K. and Rahman, B.M., *Differential transform method for solving system of delay differential equation.* Aust. J. Basic Appl. Sci. 5(4) (2011) 201–206.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 2, 2024, pp 475–499
https://doi.org/10.22067/ijnao.2024.85218.1335
https://ijnao.um.ac.ir/

Research Article

Drug release from a two-layer stent coating considering the viscoelastic property of the arterial wall: A mathematical and numerical study

E. Azhdari* and A. Emami

Abstract

Atherosclerosis is one of the most common diseases in the world. Medication with metal stents plays an important role in treating this disease. There are many models for delivering drugs from stents to the arterial wall. This paper presents a model that describes drug delivery from the stent coating layers to the arterial wall tissue. This model complements the previous models by considering the mechanical properties of the arterial wall tissue, which changes due to atherosclerosis and improves results

*Corresponding author

Received 6 November 2023; revised 5 February 2024; accepted 28 February 2024

Ebrahim Azhdari
Department of Mathematics, Salman Farsi University of Kazerun, Iran. e-mail: e.azhdari@kazerunsfu.ac.ir

Aram Emami
Department of Mathematics, Salman Farsi University of Kazerun, Iran. e-mail: emami@kazerunsfu.ac.ir

How to cite this article
Azhdari, E. and Emami, A., Drug release from a two-layer stent coating considering the viscoelastic property of the arterial wall: A mathematical and numerical study. *Iran. J. Numer. Anal. Optim.*, 2024; 14(2): 475-499. https://doi.org/10.22067/ijnao.2024.85218.1335

for designing stents. The stability behavior of the model is analyzed, and a number of numerical results are provided with explanations. A comparison between numerical and experimental results, which examine a more accurate match between the in vivo and in vitro, is shown.

AMS subject classifications (2020): 65M06.

Keywords: Stent coating; Viscoelastic; Mathematical model; Numerical simulation.

1 Introduction

Atherosclerosis is a condition in which plaques made up of cholesterol, fat, cellular waste, calcium, and fibrin (a type of insoluble coagulation protein) form inside the arteries and block them [21] (see Figure 1). Methods used to reduce the severity of atherosclerosis include bypass grafting (BG), coronary skin intervention, and less invasive balloon angioplasty [12].

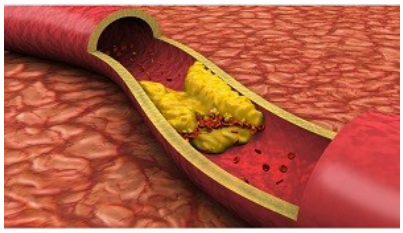


Figure 1: Atherosclerosis. Source: Intermountain Medical Center https://healthcare-in-europe.com/en/news/potential-breakthrough-in-determining-who-s-at-risk.html

Recently, though, a new procedure called stent-based delivery has been introduced that claims to help open blocked vessels and prevent restenosis [3, 28]. The bare metal stent is the first type of stent that is used as a scaffold to prevent restenosis of blood vessels (see Figure 2). Another kind of stent that is used to open blocked vessels and prevent restenosis is a drug-eluting stent [20]. In drug-eluting stents, the drug is gradually removed from the

stent, which makes the drug enter the vessels with sufficient concentration in a timely manner and a biologically active state. This helps keep the arteries smooth and open and allows blood to flow well. Clinical evidence shows that drug effectiveness depends on release rate, drug diffusion rate, drug distribution in tissue, stent design, and so on. Therefore, it is necessary to study the relationship between stent design and the release and retention of eluted drugs.

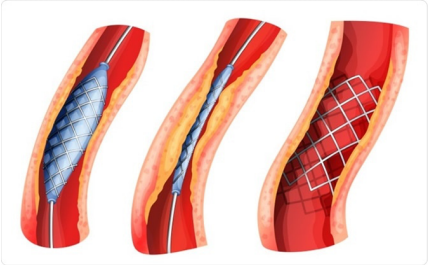


Figure 2: Stent. Source: Image by brgfx on Freepik.com
https://bjbas.springeropen.com/articles/10.1186/s43088-023-00382-9/figures/1

Recently, researchers have studied different types of drug delivery based on stents, including spreading the drug from a metal stent surface as a Dirichlet condition [15, 22], single-layer coating of the drug on the metal wire [11, 16], and also multi-layer coating [14, 24]. For example, Sarifuddin and Mandal [24] have studied the benefits of replacing a single-layer coating with a multi-layer coating along with transfer, which includes better control over drug delivery. Multi-layer coating with different drug forms or release kinetics can provide better control over the drug transfer and treatment process.

Atherosclerosis can affect the thickest arterial layer in advanced stages, and this effect can cause a significant change in the mechanical properties of the arterial wall tissue [10]. All of the above studies have examined drug transfer from stents to arterial wall tissue, regardless of the viscoelastic properties of the arterial wall tissue. The model we present is based on a model presented in [24] and has the potential to help stent designers reduce adverse patient outcomes. The present study aims to investigate more precisely the drug delivery through a two-layer stent coating to the arterial wall tissue taking into account the mechanical properties of the arterial wall tissue. More-

over, the model presented in this paper better agrees with the experimental results than the previous models.

The paper is organized as follows: Section 2 presents a brief clinical perspective of the work. Section 3 describes the model, including the model geometry, a mathematical description of the equations governing drug delivery from the stent coating to the arterial wall tissue, and initial, boundary, and interface boundary conditions. The stability behavior of the model is analyzed in Section 4. Numerical simulations for drug delivery from stent layers to the arterial wall tissue are given in Section 5. Finally, some discussions and conclusions are mentioned in Sections 6 and 7.

2 Clinical perspective

Atherosclerosis is one of the most serious and common cardiovascular diseases in the artery wall. This disease causes blockage of the lumen and reduces blood flow. Surgical techniques, such as coronary artery BG and percutaneous coronary intervention, are techniques used to reduce the severity of advanced atherosclerotic stenosis in the coronary artery. Balloon angioplasty was considered a less invasive intervention to open the blocked lumen. Clinical observations show that balloon angioplasty is associated with restenosis. The advent of bare metal stents has revolutionized balloon angioplasty by reducing the risk of restenosis. However, after several decades of using these types of stents, internal restenosis, that is, re-narrowing of the lumen, was seen again, which was a major defect. Therefore, the drug-eluting stent was invented to overcome the restenosis within the stent after implantation. Atherosclerosis mainly affects the intima, but the periphery, which is the thickest arterial layer, can be affected in advanced stages. This causes a significant change in the mechanical properties of the arterial wall. Several authors have studied mathematical models for drug delivery in cardiovascular tissues. Most of these studies focus on drug diffusion while ignoring the mechanical properties of the arterial wall. In this paper, we study the effect of the mechanical properties of the arterial wall, which has not been considered in the models that are used to deliver drugs from two-layer stents to the arterial wall.

3 Mathematical model

This section describes the mathematical model of the problem, which consists of four parts: geometry of the model (Section 3.1), description of the drug in the stent coating layers (Section 3.2), description of the drug in the arterial wall tissue (Section 3.3), and interface boundary conditions (Section 3.4). The governing equations describe the release of the drug from stent coating layers S_1 (inner layer) and S_2 (outer layer) into the arterial wall tissue V. The parameters used in the model are demonstrated in Table 1 with the relevant descriptions and units.

Table 1: The values used in the model for the parameters.

Variable/Parameter	Description	Value	Reference
$U_{\max}$	maximum velocity (cm/s)	14	[24, 29]
ρ_t	tissue density (g/cm^3)	1.06	[5]
μ_t	viscosity (g/cms)	0.0072	[5]
V_{filt}	filtration velocity (cm/s)	5.8×10^{-6}	[25]
P_k	permeability (cm^2)	2×10^{-14}	[5]
D_t	tissue diffusivity (cm^2/s)	3.65×10^{-8}	[6]
D_1	coating (S_1) diffusivity (cm^2/s)	1×10^{-9}	[6]
D_2	coating (S_2) diffusivity (cm^2/s)	1×10^{-9}	[6]
D_v	viscoelastic diffusion coefficient $(g/(cmsPa))$	5×10^{-11}	Ref. model
B_M	tissue-binding capacity (mM)	1.3	[13]
K_D	constant (mM)	0.136	[13]
D_a	Damkohler number	2700	[27]
r_0	lumen radius (cm)	0.15	[2, 30]
r_1	thickness of coating (S_1) (cm)	0.0025	[24]
r_2	thickness of coating (S_2) (cm)	0.0025	[24]
T	thickness of wall (cm)	0.05	[2, 30]
τ	relaxation time (s)	0.5	Ref. model
τ_σ	relaxation time (s)	0.1	Ref. model
κ	Young modulus (MPa)	0.0001	Ref. model

3.1 Geometry of the model

This section describes the model's geometry using a special stent shape with a half-embedded metal piece and a two-layer coating. To simplify the calculations, a two-dimensional figure has been used instead of a three-dimensional one (see Figure 3). Also, S_1 and S_2 show the inner and outer layers of the stent coating, and V shows the arterial wall tissue. The boundary between lumen and arterial wall tissue is called Γ_l. Γ_w is a boundary between arterial wall tissue and other arterial wall tissue. The boundary between arterial wall tissue and perivascular wall is called Γ_a. Moreover, $\Gamma_{S_1,V}$ is an interface boundary between arterial wall tissue and the inner layer of the stent coating. The interface boundary between the inner and outer layers of the stent coating is called Γ_{S_1,S_2}, and finally, the interface boundary between the outer layer of the stent coating and the metal part of the stent is called Γ_S.

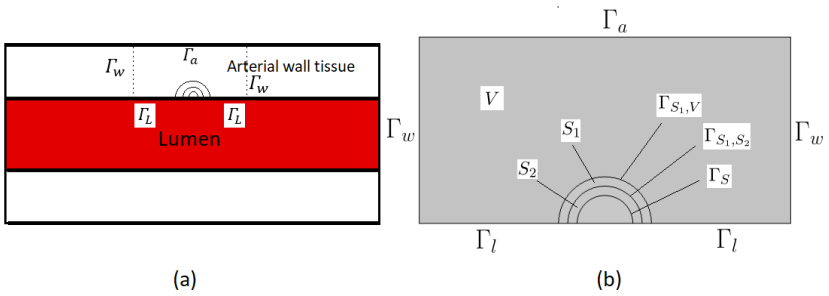


Figure 3: A schematic of the geometric shape of the problem

3.2 Description of the drug in the stent coating layers

In this section, we will study the release of the drug from the stent coating layers. The drug is first loaded solidly into the stent coating. As the stent coating comes into contact with the fluid and the fluid enters the layers through their boundaries, the layer's pores open. The fluid contacts the drug, and the drug begins to release. The drug released along the stent coating layers is described by the following non-dimensionalized equations:

$$\frac{\partial c_m}{\partial t} = -\nabla . J_{c_m} \text{ in } S_m \times (0, T_f], \quad m = 1, 2, \tag{1}$$

where $c_m, m = 1, 2$ are the concentrations of the drug in the stent coating $(S_m, m = 1, 2)$, respectively, $J_{c_m} = -\frac{1}{Pe_m}\nabla c_m$ is the flux of the drug concentration in the stent coating $(S_m, m = 1, 2)$, and T_f is the final time. The parameters $Pe_m, m = 1, 2$ are the Peclet numbers for the coating layers defined by $Pe_m = \frac{R_0 U_0}{D_m}, m = 1, 2$, where $D_m, m = 1, 2$ are the diffusion coefficient of the drug in coating layers $S_m, m = 1, 2$, respectively, and $R_0 = \frac{r_0}{6}$ and $U_0 = \frac{U_{\max}}{6}$ in which r_0 and $U_{\max}$ denote the luminal radius and the maximum velocity at the inlet, respectively.

At time $t = 0$, we assume that the amount of drug released after contacting the loaded drug with the fluid in the stent coating is as follows:

$$c_m(0) = 1, \qquad m = 1, 2. \tag{2}$$

We assume a non-flux boundary condition on the Γ_S, that $J_{c_2}.\eta = 0$ because the drug is impermeable to the metal part of the stent, where η is the exterior unit normal. The drug enters the inner stent coating via boundary conditions from the outer stent coating and then enters the arterial wall tissue through the inner stent coating. Section 3.4 describes the interface boundary conditions.

3.3 Description of the drug in the arterial wall tissue

The equations used to describe the distribution of the drug in the wall are as follows:

Convection is caused by the flow of interstitial fluid within the porous arterial wall, and the continuity equation can be written in dimensionless form in two-dimensional form. The Brinkman equation provides more accurate results because it combines the linear momentum and mass conservation of

fluids in large pores and channels with the Darcy equation for regions with unresolved pores [7]. Therefore, the Brinkman equation is used as follows:

$$\begin{cases} \dfrac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}.\nabla)\mathbf{v} = -\nabla P + \dfrac{1}{Re}\nabla.(\nabla \mathbf{v} + (\nabla \mathbf{v})^T) - \dfrac{1}{ReK_p}\mathbf{v}, \\ \nabla.\mathbf{v} = 0 \quad \text{in} \quad V \times (0, T_f], \end{cases} \tag{3}$$

where $Re = \frac{\rho_t U_0 R_0}{\mu_t}$ and $K_p = \frac{P_k}{R_0^2}$ in which P_k is the Darcy permeability. Here, $\mathbf{v}$ is the velocity vector for the tissue domain, ρ_t is the density, μ_t is the viscosity, and P is the pressure.

To complete the equation (3), the following boundary and initial conditions are given: We assume the initial condition $\mathbf{v}(0) = 0$ and boundary conditions on the boundaries Γ_l, Γ_a and Γ_w (see Figure 3), as follows:

$$\begin{cases} \mathbf{v} = 0 \quad \texttt{on } \Gamma_w, \\ \mathbf{v} = (v_x, v_y) = (0, V_{filt}) \quad \texttt{on } \Gamma_l, \\ P = 0 \quad \texttt{on } \Gamma_a, \end{cases} \tag{4}$$

where V_{filt} is the filtration velocity of plasma.

A drug that does not bind to tissue receptors is known as a free drug, the concentration which is indicated by c_t. Its equation in the arterial wall tissue is as follows:

$$\frac{\partial c_t}{\partial t} = -\nabla.J_{c_t} - \frac{\partial b_t}{\partial t} \quad \text{in} \quad V \times (0, T_f], \tag{5}$$

where $J_{c_t} = -(\frac{1}{Pe_t}\nabla c_t + \frac{1}{D_\sigma}\nabla \sigma - \mathbf{v}c_t)$ is the free drug mass flux, and σ is the response of arterial wall tissue to strain caused by drug molecules. Moreover, Pe_t is the Peclet number for tissue defined by $Pe_t = \frac{R_0 U_0}{D_t}$, where D_t is the diffusion coefficient of the drug in the arterial wall tissue. Also, $D_\sigma = \frac{R_0 U_0}{D_v}$, where D_v is the viscoelastic diffusion coefficient. Finally, b_t is the concentration of a drug that attaches to tissue receptors in the arterial wall and is known as a bound drug, which is described by the following equation:

$$\frac{\partial b_t}{\partial t} = k_a c_t (B_M - b_t) - k_d b_t \quad \text{in} \quad V \times (0, T_f], \tag{6}$$

where $k_a = \frac{K_a R_0}{U_0}$ and $k_d = \frac{K_d R_0}{U_0}$ in which K_a and K_d are the aggregation and disintegration rate constants, respectively, where $K_a = \frac{D_t D_a}{B_M T^2}$ and $K_d = K_a K_D$.

The second sentence of J_{c_t} in equation (5), which is $\frac{1}{D_\sigma}\nabla\sigma$, describes the mechanical behavior of the arterial wall tissue. In other words, it describes the arterial wall as a barrier that prevents the drug from entering the wall. Using the same symbols used in [8] for stress and viscoelastic diffusion coefficient, the model from [4]

$$\sigma + \tau\frac{\partial\sigma}{\partial t} = -\kappa(c_t + \tau_\sigma\frac{\partial c_t}{\partial t}) \quad \text{in} \quad V \times (0, T_f], \tag{7}$$

is similar to the equation used by [19]. Also, κ is the Young modulus, and τ and τ_σ are the relaxation time.

The initial and boundary conditions are taken into account to complete the equations (5) and (7). As in the initial time, there is no free drug and bound drug in the arterial wall, so we have

$$c_t(0) = b_t(0) = 0. \tag{8}$$

Due to the fact that the free drug does not cross the arterial walls ($\Gamma_w \cup \Gamma_a$) as well as the boundary between the lumen and the tissue of the arterial wall (Γ_l), there is a boundary condition $J_{c_t}.\eta = 0$ on the boundaries $\Gamma_l \cup \Gamma_w \cup \Gamma_a$.

3.4 Interface boundary conditions

Aside from the initial and boundary conditions, Systems (1), (3), (5), and (7) are completed by the interface boundary conditions between the first and second layers of the stent coating (Γ_{S_1,S_2}) and the second layer of the stent coating and the arterial wall ($\Gamma_{S_1,V}$) as follows:

$$\begin{cases} J_{c_1}.\eta = -A_2 c_2 \ \texttt{on}\ \Gamma_{S_1,S_2}, \\ J_{c_1}.\eta = -J_{c_2}.\nu \ \texttt{on}\ \Gamma_{S_1,S_2}, \\ J_{c_1}.\eta = A_1 c_1 \ \texttt{on}\ \Gamma_{S_1,V}, \\ J_{c_1}.\eta = -J_{c_t}.\nu \ \texttt{on}\ \Gamma_{S_1,V}, \end{cases} \tag{9}$$

where A_1 and A_2 are partition coefficients.

4 Energy estimates

In this section, we obtain an upper bound for the energy estimate, which is defined as follows:

$$\mathcal{E}(c(t)) = \|c(t)\|^2 + \int_0^t \|\nabla c(s)\|^2 ds + \int_0^t \|c(s)\|^2_{(\Gamma_{S_1,S_2}\cup\Gamma_{S_1,V})} ds,$$

where $c(t) = (c_1(t), c_2(t), c_t(t))$ and $\|\cdot\|$ represent the usual norm in $L^2(\Omega)$ $\big(\Omega = S_1 \cup S_2 \cup V\big)$, which is induced by the corresponding inner product $(\cdot,\cdot)$ [1]. After solving (7), we get the following equation:

$$\sigma = \frac{\kappa}{\tau}(\frac{\tau_\sigma}{\tau} - 1)\int_0^t e^{\frac{1}{\tau}(s-t)} c_t(s) ds + \frac{\kappa\tau_\sigma}{\tau} e^{\frac{-t}{\tau}} c_t(0) - \frac{\kappa\tau_\sigma}{\tau} c_t + e^{\frac{-t}{\tau}}\sigma(0). \quad (10)$$

Replacing (6) in (5) to get

$$\frac{\partial c_t}{\partial t} = \nabla.(\frac{1}{Pe_t}\nabla c_t + \frac{1}{D_\sigma}\nabla\sigma - \mathbf{v}c_t) - k_a c_t(B_M - b_t) + k_d b_t. \quad (11)$$

The weak formulation of the differential problem (1) and (11) is defined as follows: Find $c(t) \in H^1(S_1) \times H^1(S_2) \times H^1(V)$ such that $\frac{\partial c}{\partial t}(t) \in L^2(S_1) \times L^2(S_2) \times L^2(V)$, $c(0) = (1, 1, 0)$, and

$$\begin{aligned}(\frac{\partial c}{\partial t}(t), u) = \sum_{i=1,2,t} (-\frac{1}{Pe_i}\nabla c_i, \nabla u_i) + (A_2c_2, u_1)_{\Gamma_{S_1,S_2}} - (A_1c_1, u_1)_{\Gamma_{S_1,V}} \\ -(A_2c_2, u_2)_{\Gamma_{S_1,S_2}} - ((\frac{1}{D_\sigma}\nabla\sigma - \mathbf{v}c_t), \nabla u_t) + (A_1c_1, u_t)_{\Gamma_{S_1,V}} \\ -(k_a c_t(B_M - b_t), u_t) + (k_d b_t, u_t), \end{aligned} \quad (12)$$

where $(u_1, u_2, u_t) \in (H^1(S_1) \times H^1(S_2) \times H^1(V))$.

Replacing (10) in (12) by considering $\sigma(0)$ and $c_t(0)$ as a constant to get

$$\begin{aligned}(\frac{\partial c}{\partial t}(t), u) = \sum_{i=1,2,t} (-\frac{1}{Pe_i}\nabla c_i, \nabla u_i) + (A_2c_2, u_1)_{\Gamma_{S_1,S_2}} - (A_1c_1, u_1)_{\Gamma_{S_1,V}} \\ -(A_2c_2, u_2)_{\Gamma_{S_1,S_2}} + \frac{\kappa}{D_\sigma\tau}(1 - \frac{\tau_\sigma}{\tau})\int_0^t e^{\frac{1}{\tau}(s-t)}(\nabla c_t(s), \nabla u_t(s)) ds \\ +\frac{\kappa\tau_\sigma}{D_\sigma\tau}(\nabla c_t, \nabla u_t) + (\mathbf{v}c_t, \nabla u_t) + (A_1c_1, u_t)_{\Gamma_{S_1,V}} \\ -(k_a c_t(B_M - b_t), u_t) + (k_d b_t, u_t). \end{aligned} \quad (13)$$

Theorem 1. If $c(t) = (c_1(t), c_2(t), c_t(t))$ is a solution to the variational problem (13), then

$$\mathcal{E}(c(t)) \leq \frac{1}{\phi_{\min}} \mathcal{E}(c(0)) e^{\phi t}, \tag{14}$$

where $\phi = \frac{\phi_{\max}}{\phi_{\min}}$ is a time independent positive constant, with $\phi_{\min} = \min\{1, \frac{2}{Pe_1}, \frac{2}{Pe_2}, 2(\frac{1}{Pe_t} - \frac{\kappa\tau_\sigma}{D_\sigma\tau} - \epsilon_2^2), +2A_1, 2A_2\}$ and

$$\begin{aligned}\phi_{\max} = \max\{&(\frac{1}{2\epsilon_2^2}\|\mathbf{v}\|^2_{L^\infty(L^\infty)} - 2k_aB_M + 2k_a\delta\|b_t\|_{L^\infty(L^\infty)}), 2\epsilon_1^2A_2, \frac{A_2}{2\epsilon_1^2},\\ &2\frac{\kappa}{D_\sigma\tau}(1 - \frac{\tau_\sigma}{\tau}), 2\epsilon_3^2A_1, \frac{A_1}{2\epsilon_3^2}\},\end{aligned}$$

where $\epsilon_1, \epsilon_2, \epsilon_3 \neq 0$ are positive constants and $\left(\frac{1}{Pe_t} - \frac{\kappa\tau_\sigma}{D_\sigma\tau} - \epsilon_2^2\right) > 0$.

Proof. In (13), replace u_1, u_2, and u_t with c_1, c_2 and c_t, respectively, to get

$$\begin{aligned}\frac{1}{2}\frac{\partial}{\partial t}\|c(t)\|^2 = &-\frac{1}{Pe_i}\sum_{i=1,2,t}\|\nabla c_i\|^2 + (A_2c_2, c_1)_{\Gamma_{S_1,S_2}} - A_1\|c_1\|^2_{\Gamma_{S_1,V}}\\ &-A_2\|c_2\|^2_{\Gamma_{S_1,S_2}} + \frac{\kappa}{D_\sigma\tau}(1 - \frac{\tau_\sigma}{\tau})\int_0^t e^{\frac{1}{\tau}(s-t)}\|\nabla c_t(s)\|^2 ds\\ &+\frac{\kappa\tau_\sigma}{D_\sigma\tau}\|\nabla c_t\|^2 + (\mathbf{v}c_t, \nabla c_t) + (A_1c_1, c_t)_{\Gamma_{S_1,V}}\\ &-(k_ac_t(B_M - b_t), c_t) + (k_db_t, c_t).\end{aligned} \tag{15}$$

Using the Cauchy–Schwarz inequality and since $k_d = K_Dk_a$ assuming $\delta > 1$ such that $c_t \geq \frac{K_D}{\delta - 1}$

$$\begin{aligned}\frac{1}{2}\frac{\partial}{\partial t}\|c(t)\|^2 \leq &\sum_{i=1,2,t} -\frac{1}{Pe_i}\|\nabla c_i\|^2 + \epsilon_1^2A_2\|c_2\|^2_{\Gamma_{S_1,S_2}} + \frac{A_2}{4\epsilon_1^2}\|c_1\|^2_{\Gamma_{S_1,S_2}}\\ &-A_1\|c_1\|^2_{\Gamma_{S_1,V}} - A_2\|c_2\|^2_{\Gamma_{S_1,S_2}}\\ &+\frac{\kappa}{D_\sigma\tau}(1 - \frac{\tau_\sigma}{\tau})\int_0^t e^{\frac{1}{\tau}(s-t)}\|\nabla c_t(s)\|^2 ds + \frac{\kappa\tau_\sigma}{D_\sigma\tau}\|\nabla c_t\|^2\\ &+\epsilon_2^2\|\nabla c_t\|^2 + \frac{1}{4\epsilon_2^2}\|\mathbf{v}\|^2_{L^\infty(L^\infty)}\|c_t\|^2 + \epsilon_3^2A_1\|c_1\|^2_{\Gamma_{S_1,V}}\\ &+\frac{A_1}{4\epsilon_3^2}\|c_t\|^2_{\Gamma_{S_1,V}} - k_aB_M\|c_t\|^2 + k_a\delta\|b_t\|_{L^\infty(L^\infty)}\|c_t\|^2.\end{aligned}$$

Let $\epsilon_2 > 0$ such that $\left(\frac{1}{Pe_t} - \frac{\kappa\tau_\sigma}{D_\sigma\tau} - \epsilon_2^2\right) > 0$ and $\left(\frac{1}{4\epsilon_2^2}\|\mathbf{v}\|^2_{L^\infty(L^\infty)} - k_aB_M + k_a\delta\|b_t\|_{L^\infty(L^\infty)}\right) > 0$. Then

$$\frac{1}{2}\frac{\partial}{\partial t}\|c(t)\|^2+\sum_{i=1,2}\frac{1}{Pe_i}\quad\|\nabla c_i\|^2+\big(\frac{1}{Pe_t}-\frac{\kappa\tau_\sigma}{D_\sigma\tau}-\epsilon_2^2\big)\|\nabla c_t\|^2+A_1\|c_1\|^2_{\Gamma_{S_1,V}}$$
$$+A_2\|c_2\|^2_{\Gamma_{S_1,S_2}}\le\big(\frac{1}{4\epsilon_2^2}\|\mathbf{v}\|^2_{L^\infty(L^\infty)}-k_aB_M+k_a\delta\|b_t\|_{L^\infty(L^\infty)}\big)\|c_t\|^2$$
$$+\epsilon_1^2A_2\|c_2\|^2_{\Gamma_{S_1,S_2}}+\frac{A_2}{4\epsilon_1^2}\|c_1\|^2_{\Gamma_{S_1,S_2}}$$
$$+\frac{\kappa}{D_\sigma\tau}(1-\frac{\tau_\sigma}{\tau})\int_0^t e^{\frac{1}{\tau}(s-t)}\|\nabla c_t(s)\|^2ds$$
$$+\epsilon_3^2A_1\|c_1\|^2_{\Gamma_{S_1,V}}+\frac{A_1}{4\epsilon_3^2}\|c_t\|^2_{\Gamma_{S_1,V}}.\qquad(16)$$

Multiplying (16) by two and taking integration to get

$$\mathcal{E}(c(t))\le\frac{1}{\phi_{\min}}\mathcal{E}(c(0))+\phi\int_0^t\mathcal{E}(c(s))ds,$$

using the Gronwall lemma [9] to obtain (14).

□

The estimation we obtained in Theorem 1 allows us to determine the uniqueness of the solution (13) and its stability for a limited time.

5 Numerical simulations

COMSOL 5.3 Multiphysics is utilized for meshing and numerical results. The arterial wall and coating layers are meshed, as shown in Figure 4. The following modules in the COMSOL are used:

- Brinkman equations in order to calculate the velocity and pressure of the arterial wall;
- Coefficient form partial differential equation (c) in order to calculate the drug concentration in the coating layers and arterial wall tissue.

To make sure that the numerical results do not depend on the mesh size, we consider the following meshes:

- Mesh 1: 1642 elements;

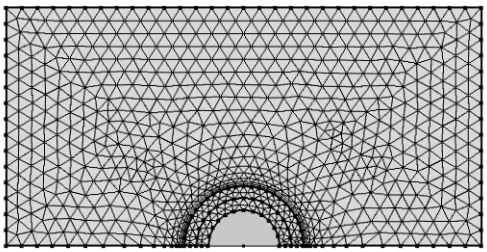


Figure 4: Computational meshes in the domain

Table 2: Different errors for free drug concentration in arterial wall.

h	L^∞	L^2	H^1
0.001	1.3×10^{-12}	1.54×10^{-11}	3.81×10^{-10}
0.01	2.2×10^{-12}	2.14×10^{-11}	5.88×10^{-10}
0.1	3.6×10^{-12}	3.56×10^{-11}	7.43×10^{-10}

- Mesh 2: 3132 elements;
- Mesh 3: 6162 elements;
- Mesh 4: 12340 elements.

The adequacy of the mesh density has been investigated using convergence tests. By considering two different meshes, the solutions' convergence and the mesh's independence have been investigated. To illustrate this claim more clearly, Figure 5 and Table 2 show the comparison among different meshes for the average drug concentration in arterial wall tissue. Of note, the reference solution is defined by $h_{\max} = 1 \times 10^{-4}$. According to this figure, the changes among the meshes are almost unrecognizable, suggesting that the numerical results do not depend on the mesh size.

Based on Figure 6, the velocity vectors affect both the distribution and retention of the drug. This result indicates the boundary condition for the

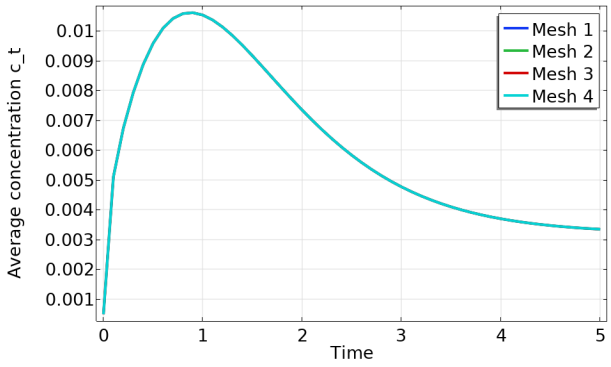


Figure 5: Average drug concentration (c_t) in the arterial wall.

interface boundary between the lumen and the arterial wall tissue. The velocity vector is indicated by an arrow. These results are consistent with the results in [28].

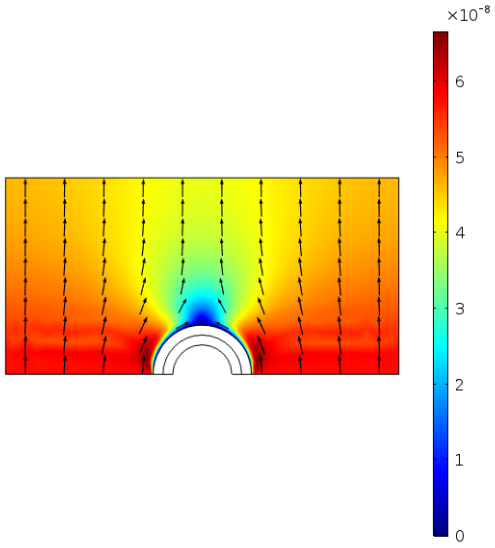
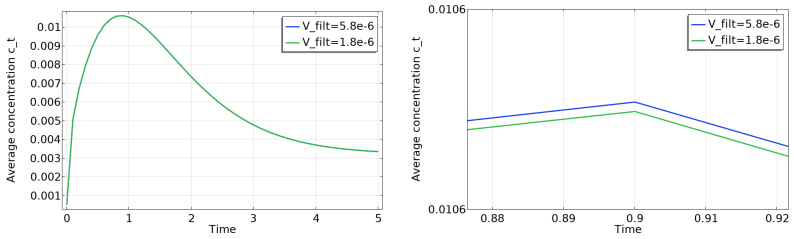


Figure 6: Dissemination of the velocity vector

The filtration velocity used in [24], which is derived from experimental results performed on animals [25], is similar to the amounts obtained in [17, 23, 30]. Figure 7 shows the low effect of filtration velocity on the average

drug concentration in arterial wall tissue for $1.8 \times 10^{-6}(cm/s)$ and $5.8 \times 10^{-6}(cm/s)$.

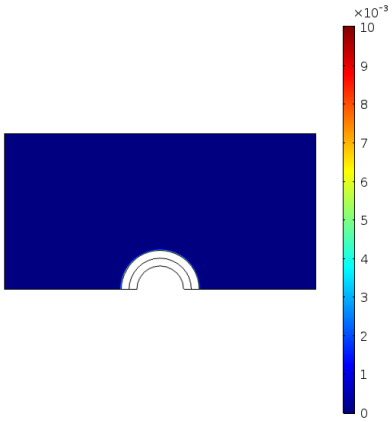


(a) (b) zoom of (a) around t=0.9

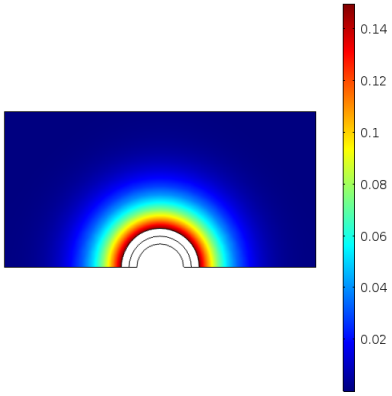
Figure 7: The effect of filtration velocity (V_{filt}) on the average drug concentration c_t in the arterial wall tissue.

The release of drug concentration from the stent coating layers into the arterial wall tissue during 300 seconds has been demonstrated in Figure 8. As shown in Figure 8, the drug concentration enters from the inner layer of the stent coating due to the interface boundary condition on $\Gamma_{S_1,V}$. Over time, the drug concentration enters from the inner layer of the stent coating and increases in the arterial wall tissue.

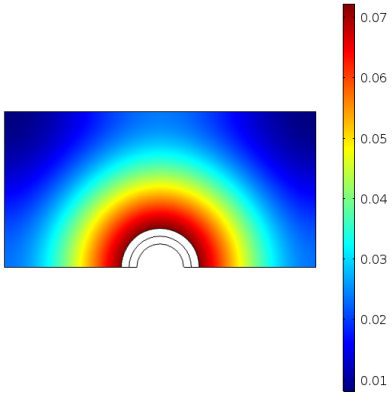
The local concentration of the free and bound drug, which is indicated by a red point (see Figure 9 (f)), is illustrated in Figure 9. According to these figures, the concentration of free drug increases for a certain period of time and then diminishes (Figure 9 (a) and (c)), although the bound drug goes up to reach a steady state (Figure 9 (b) and (d)). By comparing the rate of drug release from the inner and outer layers of the stent, it can be seen that a faster release leads to a higher concentration of the free and bounded drug in the arterial wall. For further explanation, we need to point out that the rapid diffusion in the inner layer compared to the outer layer has a higher concentration in the arterial wall tissue, which is around 0.16 in the inner layer (Figure 9 (a)) and around 0.12 in the outer layer (Figure 9 (c)). This is because the inner layer is closer to the arterial wall. Moreover, Figure 9 (e) shows the role of convective flow in drug delivery to (at the reference point) the arterial wall tissue. It is clear that convection contributes significantly to



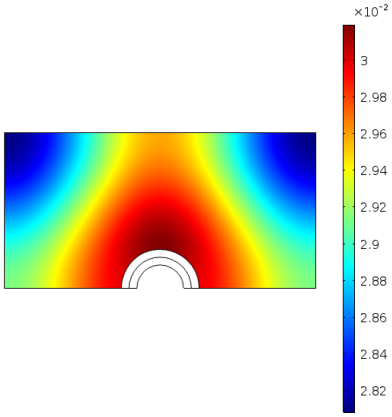
(a) Drug release, 0 second



(b) Drug release, 25 second



(c) Drug release, 100 second



(d) Drug release, 300 second

Figure 8: Dissemination for different times

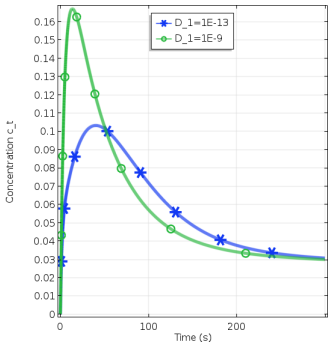
drug delivery, which must be taken into account in order to achieve better results.

The behavior of mean drug concentration in the inner and outer layers of the stent for different diffusion coefficients is shown in Figure 10. This figure shows that the reduction of the diffusion coefficient in the inner and outer layers of the stent makes the drug enter the arterial wall tissue with a delay from the stent. This enables the stent designers to deliver the required amount of drug to the arterial wall tissue.

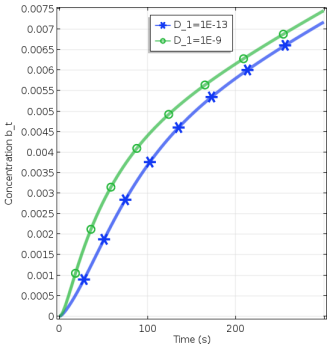
The average free drug in arterial wall tissue is depicted in Figure 11. This figure shows that the drug concentration tends to increase from zero to maximum and diminishes afterward. It is interesting to note that in the case of fast release of the inner layer ($D_1 = 1 \times 10^{-9} cm^2/s$) and slow release of the outer layer ($D_2 = 1 \times 10^{-11} cm^2/s$), this maximum is higher than in the case of slow release of the inner layer ($D_1 = 1 \times 10^{-11} cm^2/s$) and slow release of the outer layer ($D_2 = 1 \times 10^{-11} cm^2/s$).

The arterial wall tissue has a viscoelastic property, which significantly affects drug delivery. As a result, we presented this effect in Figure 12. It can be seen that the higher the viscoelastic diffusion coefficient, D_v, the less drug is observed in the arterial wall tissue. In other words, the tissue of the arterial wall has a barrier property against drug delivery.

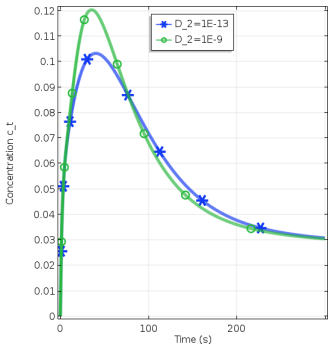
In order to validate the application of the model presented in this paper, in Figure 13, we compare the numerical results obtained for the free drug concentration in the arterial wall with the in-vivo experimental data (Sarifuddin and Mandal [24], Tzafriri et al. [26], and Mongrain et al. [18]). It is clear that considering the viscoelastic properties of the arterial wall can affect the transfer of free drugs from the stent coating layers to the arterial wall. These results indicate that the proposed model is congruent with the models proposed in Sarifuddin,Tzafriri,Mongrain, although they did not take into account the viscoelastic properties of the arterial wall. In other words, the model presented in this paper is more consistent with the experimental results than other models since the viscoelastic property, which is one of the mechanical properties of the arterial wall tissue that acts as a barrier, has been considered. Figure 13 illustrates the advantages of this article compared to the previous ones.



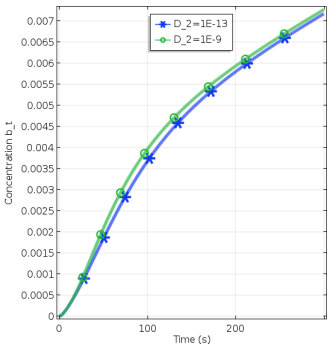
(a) The effect of D_1 on the free drug



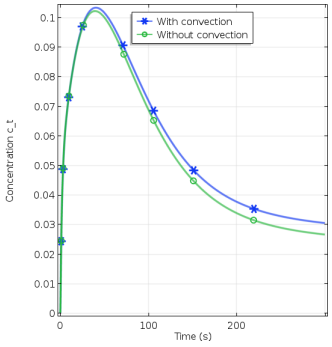
(b) The effect of D_1 on the bound drug



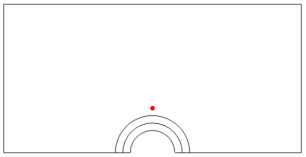
(c) The effect of D_2 on the free drug



(d) The effect of D_2 on the bound drug

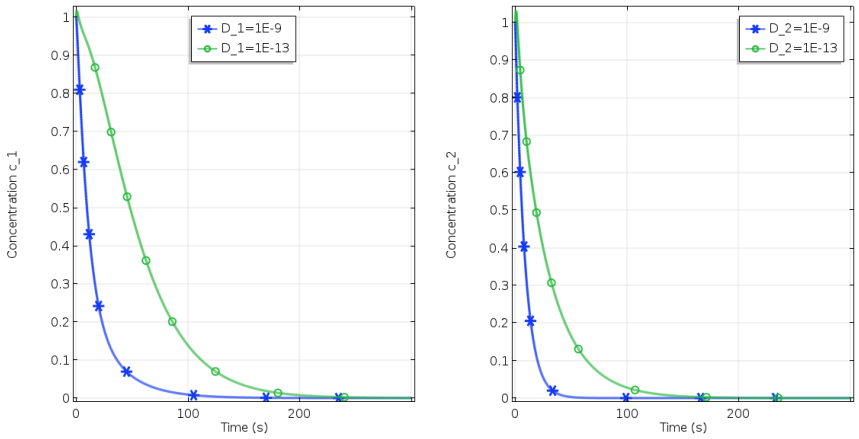


(e) The role of convection on the free drug



(f) reference point

Figure 9: The effect of the diffusion coefficient of the layers of stent coating on free and bound drug concentration and the effect of convection on free drug concentration in the arterial wall tissue.



(a) inner layer, S_1 (b) outer layer, S_2

Figure 10: The effect of diffusion coefficients on the drug in coating layers

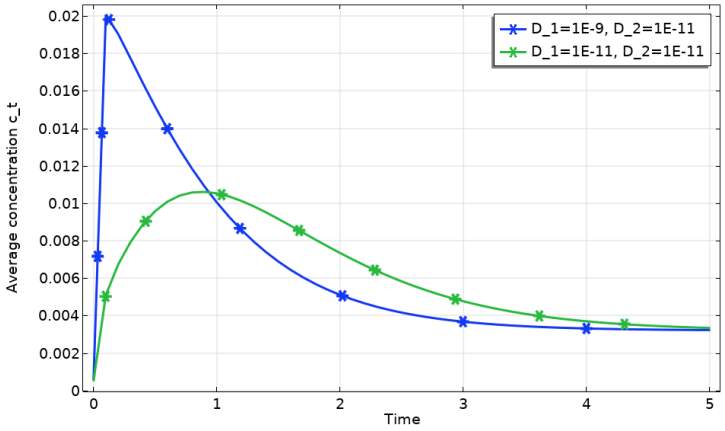


Figure 11: The average concentration of free drug in arterial wall tissue (the effect of diffusion coefficients on drug in coating layers)

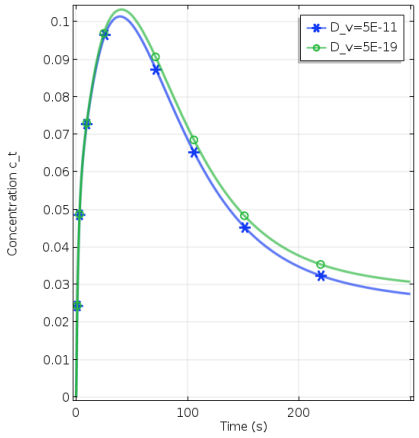


Figure 12: The influence of D_v on the concentration of free drug in arterial wall tissue

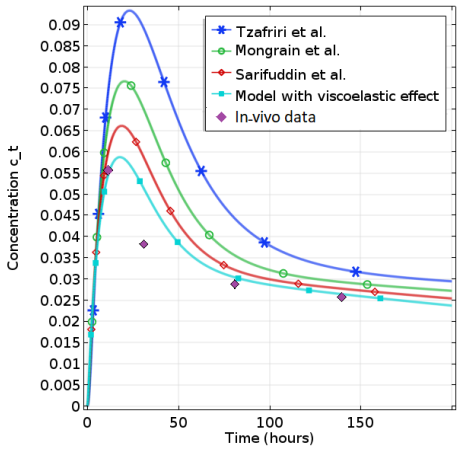


Figure 13: Comparison between experimental data and the model presented for drug delivery from stent coating layers into the arterial wall

6 Discussion

In order to discuss the transfer of drugs from a half-embedded two-layer coating stent into arterial walls with viscoelastic properties, the following points can be specified. In Figure 5, the independence of numerical results from the mesh size is discussed. The effect of the velocity vectors on the distribution and retention of the drug is shown in Figure 6, which corroborates the results in [28]. The effect of filtration velocity on the average drug concentration in arterial wall tissue is shown in Figure 7. An increase in time indicates a rise in the concentration of the drug in the tissue of the arterial wall, which can be seen in Figure 8. As Figure 9 demonstrates, the concentration of the bound drug increased and reached a steady state, while the concentration of the free drug went up for some time but then decreased. In addition, faster release results in higher concentrations of free and bound drugs in the arterial wall. Because the inner layer is closer to the arterial wall, the rapid diffusion of the inner layer is greater than the concentration of the outer layer in the arterial wall tissue. Additionally, convection significantly contributes to drug delivery. Figure 10 shows that the reduction of the diffusion coefficient in the inner and outer layers of the stent makes the drug enter the arterial wall tissue with a delay, which allows the stent designers to determine the amount of drugs required for the arterial wall tissue. When the inner layer is faster than the outer layer, compared with slow release in both layers, the maximum drug concentration is higher, which is observed in Figure 11. Figure 12 illustrates that the viscoelastic property of arterial wall tissue acts as a barrier against drug transmission, which is not discussed in similar works. Finally, a comparison between the experimental data and the proposed model for drug delivery from the stent coating layers to the arterial wall, which exhibits greater agreement than other models, is presented in Figure 13.

7 Conclusion

Considering the mechanical properties of the arterial wall due to atherosclerotic changes provided a deeper understanding of the stent's performance. It

enabled a better agreement between the mathematical model and the experimental results. A mathematical model is presented to consider the transport of drugs from the stent layers to the arterial wall and describes the convective flow due to the flow of interstitial fluid in the arterial wall tissue as well as the viscoelastic property of the arterial wall. Presenting an upper bound for the energy estimate determines the uniqueness of the solution and its stability for a limited time. Finally, the numerical simulation verifies the agreement between the mathematical model, numerical results, and experimental results.

Acknowledgments

The authors would like to express their profound gratitude to the editor and anonymous reviewers for their remarkable dedication and invaluable feedback, which greatly enhanced the quality of the paper.

References

[1] Azhdari, E. and Naghipoor, J. *The effect of viscoelasticity of the tissue on the magneto-responsive drug delivery system*, J. Math. Biol. 84 (13) (2022) 1–26.

[2] Baroldi, G., Bigi, R. and Cortigiani, L. *Ultrasound imaging versus morphopathology in cardiovascular diseases. Coronary atherosclerotic plaque*, Cardiovasc Ultrasound 2(1) (2004) 1–10.

[3] Bozsak, F., Chomaz, J-M. and Barakat, A.I. *Modeling the transport of drugs eluted from stents: physical phenomena driving drug distribution in the arterial wall*, Biomech. Model. Mechanobiol.13 (2014) 327–347.

[4] Brinson, H.F. and Brinson, L.C. '*Mechanical bioeffects of ultrasound-Polymer Engineering Science and Viscoelasticity: An Introduction*, New York: Springer .2010

[5] Chung, S. and Vafai, K. *Low-density lipoprotein transport within a multilayered arterial wall-effect of the atherosclerotic plaque/stenosis*, J. Biomech. 46 (3) (2013) 574–585.

[6] Creel, C.J., Lovich, M.A. and Edelman, E.R. *Arterial paclitaxel distribution and deposition*, Circ. Res.86 (8) (2000) 879–884.

[7] de Paulo Ferreira, L., de Oliveira, T.D.S., Surmas, R., da Silva, M.A.P. and Peçanha, R.P. *Brinkman equation in reactive flow: Contribution of each term in carbonate acidification simulations*, Adv. Water Resour. 144 (2020) 103696.

[8] Ferreira, J.A., Grassi, M., Gudino, E. and De Oliveira, P. *A new look to non-Fickian diffusion*, Appl. Math. Model. 39 (1) (2015) 194–204.

[9] Gronwall, T.H. *Note on the derivative with respect to a parameter of the solutions of a system of differential equations*, Ann. Math. 20 (1919) 292–296.

[10] Holzapfel, G.A., Gasser, T.C. and Ogden, R.W. *A new constitutive framework for arterial wall mechanics and a comparative study of material models*, J. Elast. 61 (2000) 1–48.

[11] Hossainy, S. and Prabhu, S. *A mathematical model for predicting drug release from a biodurable drug-eluting stent coating,* J. Biomed. Mater. Res A. 87 (2) (2008) 487–493.

[12] Huang, Y., Ng, H.C.A., Ng, X.W. and Subbu, V., *Drug-eluting biostable and erodible stents,* J. Control. Release.193 (2014) 188–201.

[13] Kolachalama, V.B., Pacetti, S.D., Franses, J.W., Stankus, J.J., Zhao, H.Q., Shazly, T., Nikanorov, A., Schwartz, L.B., Tzafriri, A.R. and Edelman, E.R. *Mechanisms of tissue uptake and retention in zotarolimus-coated balloon therapy*, Circulation. 127 (20) (2013) 2047–2055.

[14] McGinty, S., King, D. and Pontrelli, G. *Mathematical modelling of variable porosity coatings for controlled drug release*, Med. Eng. Phys. 45 (2017) 51–60.

[15] McGinty, S., McKee, S., McCormick, C. and Wheel, M. *Release mechanism and parameter estimation in drugeluting stent systems: analytical solutions of drug release and tissue transport,* Math. Med. Biol. 32 (2) (2015) 163–186.

[16] McKittrick, C.M., McKee, S., Kennedy, S., Oldroyd, K., Wheel, M., Pontrelli, G., Dixon, S., McGinty, S. and McCormick, C. *Combining mathematical modelling with in vitro experiments to predict in vivo drug-eluting stent performance*, J. Control. Release.303 (2019) 151–161.

[17] Meyer, G., Merval, R.G. and Tedgui, A. *Effects of pressure-induced stretch and convection on low-density lipoprotein and albumin uptake in the rabbit aortic wall*, Circ. Res. 79 (1996) 532–540.

[18] Mongrain, R., Faik, I., Leask, R.L., Rodés-Cabau, J., Larose, E. and Bertrand, O.F. *Effects of diffusion coefficients and struts apposition using numerical simulations for drug eluting coronary stents*, J. Biomech. Eng. 129 (5) (2007) 733–742.

[19] Naghipoor, J., Ferreira, J.A., de Oliveira, P. and Rabczuk, T. *Tuning polymeric and drug properties in a drug eluting stent: A numerical study*, Appl. Math. Model. 40 (17-18) (2016) 8067–8086.

[20] Nishimoto, Y., Ueda, Y., Sugihara, R., Murakami, A., Ueno, K., Takeda, Y., Hirata, A., Kashiwase, K., Higuchi, Y. and Yasumura, Y., *Comparison of angioscopic findings among secondgeneration drug-eluting stents,* J. Cardiol. 70 (2017) 297–302.

[21] Pericevic, I., Lally, C., Toner, D. and Kelly, D.J. *The influence of plaque composition on underlying arterial wall stress during stent expansion: The case for lesion-specific stents*, Med. Eng. Phys. 31(4), (2009) 428–433.

[22] Pontrelli, G. and De Monte, F. *A multi-layer porous wall model for coronary drug-eluting stents,* Int. J. Heat. Mass. Trans. 53 (2010) 3629–3637.

[23] Prosi, M., Zunino, P., Perktold, K. and Quarteroni, A. *Mathematical and numerical models for transfer of low-density lipoproteins through the arterial walls: a new methodology for the model set up with applications to the study of disturbed lumenal flow,* J. Biomech. 38 (4) (2005) 903–917.

[24] Sarifuddin, Roy, S. and Mandal, P.K. *Computational model of stent-based delivery from a half-embedded two-layered coating*, Comput. Methods. Biomech. Biomed. Engin. 23 (12) (2020) 815–831.

[25] Tedgui, A. and Lever, M. *Filtration through damaged and undamaged rabbit thoracic aorta*, Am. J. Physiol. 247 (5) (1984) H784–H791.

[26] Tzafriri, A.R., Groothuis, A., Price, G.S. and Edelman, E.R. *Stent elution rate determines drug deposition and receptor-mediated effects*, J. Control. Release. 161 (2012) 918–926.

[27] Tzafriri, A.R., Levin, A.D. and Edelman, E.R. *Diffusion limited binding explains binary dose response for local arterial and tumour drug delivery*, Cell Prolif.42(3) (2009) 348–363.

[28] Vairo, G., Cioffi, M., Cottone, R., Dubini, G. and Migliavacca, F. *Drug release from coronary eluting stents: a multidomain approach,* J. Biomech. 43(8) (2010) 1580–1589.

[29] Welkowitz, W. *Engineering hemodynamics: application to cardiac assist devices*, J. Clin. Eng. 13, no. 2 (1988) 79.

[30] Yang, N. and Vafai, K. *Modeling of low-density lipoprotein (LDL) transport in the artery effects of hypertension*, Int. J. Heat. Mass. Transf. 49 (5-6) (2006) 850–867.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 2, 2024, pp 500–521
https://doi.org/10.22067/ijnao.2024.85581.1349
https://ijnao.um.ac.ir/

Research Article

Mathematical analysis and forecasting of controlled Spatio-temporal dynamics of the EG.5 Virus

E.M. Moumine*, O. Balatif and M. Rachik

Abstract

In this article, we propose a mathematical approach that connects an innovative spatio-temporal model to the problem of the **EG.5** variant of COVID-19 in a human population. We demonstrate the existence and

*Corresponding author

Received 27 November 2023; revised 19 February 2024; accepted 19 February 2024

El Mehdi Moumine
Laboratory of Analysis, Modeling and Simulation, Department of Mathematics and Computer Science, Faculty of science Ben M'sik, University Hassan II, Casablanca, Morocco. e-mail: moumine.maths@gmail.com

Omar Balatif
Laboratory of Fundamental Mathematics and Their Applications , Department of Mathematics Faculty of Sciences El Jadida, Chouaib Doukkali University, El Jadida, Morocco. e-mail: balatif.maths@gmail.com

Mostafa Rachik
Laboratory of Analysis, Modeling and Simulation, Department of Mathematics and Computer Science, Faculty of science Ben M'sik, University Hassan II, Casablanca, Morocco. e-mail: m_rachik@yahoo.fr

How to cite this article
Moumine, E.M., Balatif, O. and Rachik, M., Mathematical analysis and forecasting of controlled Spatio-temporal dynamics of the EG.5 Virus. *Iran. J. Numer. Anal. Optim.*, 2024; 14(2): 500-521. https://doi.org/10.22067/ijnao.2024.85581.1349

uniqueness of the global positive solution for our suggested system. The implementation and analysis of an applicable optimal control issue are as follows. The methods of optimal control theory are applied in this work to demonstrate the existence of optimal control, and with necessary optimality conditions, we discover the explicit expression of optimal control that minimizes the negative impacts of this infectious disease on countries. We provide numerical simulations at the conclusion to demonstrate the efficacy of our chosen strategy.

AMS subject classifications (2020): 49J15, 92B05, 93A30, 93C10

Keywords: Spatio-temporal Model; EG.5 Virus; Optimal Control; Mathematical Model.

1 Introduction

The COVID-19 pandemic swept into our world suddenly, causing global disruption and emphasizing the need for cutting-edge scientific methods to understand and stop the spread of the disease. Among these methods, quantitative analysis and optimal control are essential for understanding the spatio-temporal spread of the virus and formulating powerful strategies to fight the epidemic.

Since the discovery of the first cases in December 2019, the epidemic has wreaked havoc on the economy, society, and public health worldwide. The disease quickly crossed international borders, confounding health agencies and efficiencies around the world. Scientists and researchers have used mathematical techniques to study the dynamics of the spread of COVID-19 and formulate customized public health policies in this fight against an implacable foe (see [15, 9, 2, 12, 1, 14, 16, 23, 11, 10, 7, 8, 21, 18, 3]). Although the pandemic might appear to be a distant recollection for some, the coronavirus persists in its spread, with a surge in new cases attributed to emerging mutant strains. One such mutation, known as **EG.5**, has gained prominence in Europe since its initial detection earlier this year (2023). Recently, the world health organization (WHO) has classified it as a "variant of interest" in response to the escalating global case numbers.

EG.5 virus represents a subcategory of the Omicron variant of COVID-19 and exhibits a close genetic relationship with other globally circulating variants. It signifies a modified form of the virus. Its prevalence has surged on a global scale, rising from 7.6% of COVID-19 cases in late June to 17.4% by the close of July (2023), prompting the WHO to designate it as a matter of particular concern. Health authorities worldwide are currently conducting investigations into this novel virus. Similar to its predecessors, this fresh variant has undergone multiple mutations. Scientists are tirelessly working to grasp its attributes, encompassing its transmissibility, severity, and potential implications for vaccine efficacy. Presently, the public is strongly encouraged to persist in adhering to safety protocols, including mask-wearing, practicing social distancing, and receiving vaccinations, in order to mitigate the virus's propagation.

In addition, it is necessary to model and predict the spread of the virus and develop effective control strategies by offering theoretical models that take into account the complex interactions between people, the environment, and the properties of the virus itself. This mathematical study makes it possible to understand the spatio-temporal transmission of viruses. These mathematical models, like epidemiological models, provide a framework to analyze and predict disease progression in a specific population. Additionally, they enable quantification of the impact of various control methods such as social isolation, mask use, and vaccination on virus spread.

However, to maximize the effectiveness of control strategies, an optimal control-based strategy must be used. The goal of the latter is to select the best preventive measures considering limitations and available resources to reduce the spread of disease. This method combines mathematical analysis with optimization techniques to pinpoint actions that will have the greatest impact on reducing case incidence, mortality and the strain on the healthcare system.

In this article, the optimal management of the spatio-temporal transmission of EG.5 virus is reviewed together with current developments in the field of mathematical analysis. We look at the spread of the disease as described by the epidemiological model. We also move from optimal control strategies that we can implement in our spatio-temporal model to create effective pub-

lic health regulations that take into account the specific characteristics of the disease, available resources, and ethical considerations. In order to evaluate the outcomes of treatments and to support intelligent decision making.

2 Background

However, merely simply simulating how the virus spreads will not be enough to stop a pandemic. This is where it is important to have good control. The goal of optimal control is to consider the limitations and resources at hand and identify proper management methods to reduce virus spread [4, 13, 6, 17, 20]. Determining the timing and level of containment measures, vaccination campaigns, managing medical resources, or even adapting screening and research strategies are some possible solutions.

The researchers conducted an evaluation of various interventions, employing a combination of mathematical analysis and optimal control techniques to identify the most effective approaches for reducing viral transmission. In one study [15], the authors introduced a model featuring two distinct controls. The first control demonstrated an increase in the proportion of individuals adopting virus-preventive measures, coupled with a decrease in those not doing so. These changes were attributed to media campaigns and direct community engagement efforts. The second control illustrated that limiting social interactions effectively lowered disease transmission risk, subsequently reducing the number of infected individuals. In another study [9], the authors aimed to develop a multi-regional model by incorporating four alternative strategies as controls. The first control represented awareness campaigns in the region "j," with a focus on disseminating health precautions to protect individuals from viral infections. The second control addressed health and security measures to restrict movement between regions "j" and "r." The third control involved promoting quarantine measures for exposed individuals within the same regions, while the fourth aimed to eliminate infected animals.

We will examine a spatio-temporal mathematical model, which is based on the one we used in [15]. To gain deeper insights into the virus's temporal and spatial propagation dynamics, our analysis will incorporate a spatial variable.

This integration involves introducing a diffusion term utilizing the Laplace operator (well explained in [24]). Furthermore, we will conduct an optimal control investigation to enhance our understanding. The management of the spatio-temporal transmission of EG.5 virus will then be examined, with a focus on the role of mathematical analysis and optimal control in evaluating the efficacy of current interventions, exploring novel avenues, and offering recommendations based on reliable quantitative data. By understanding the potential impact of this study, we will be better prepared to meet the future challenges linked to this world pandemic.

3 Model formulation

The progression of the EG.5 virus in a human population can be explained by the SEIHR model that we suggest, which is shown in Figure 1. It is clear

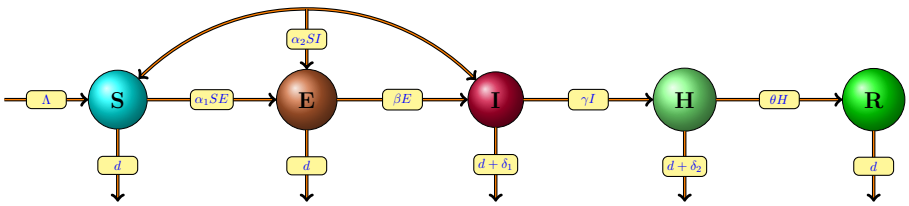


Figure 1: Diagram depicting the evolution of the EG.5 virus.

that the EG.5 virus is described in the following temporal dynamic system:

$$\begin{cases} \dfrac{dS(t)}{dt} = \Lambda - [\alpha_1 E(t) + \alpha_2 I(t)]\, S(t) - dS(t), \\ \dfrac{dE(t)}{dt} = [\alpha_1 E(t) + \alpha_2 I(t)]\, S(t) - (d+\beta)\, E(t), \\ \dfrac{dI(t)}{dt} = \beta E(t) - (d+\delta_1+\eta)\, I(t), \\ \dfrac{dH(t)}{dt} = \eta I(t) - (d+\delta_2+\theta)\, H(t), \\ \dfrac{dR(t)}{dt} = \theta H(t) - dR(t). \end{cases} \tag{1}$$

The different compartments of our spatio-temporal model are listed below, Table 1, indicating their descriptions (see [15]). It is clear that the temporal

Table 1: The description of the different compartments of our model.

Compartment	Description
S	The number of people likely to be infected by the virus.
E	The number of infected people who are symptomless.
I	The number of infected people showing symptoms
H	The number of hospitalized people
R	The number of the recovered people

dynamic system (1) cannot adequately describe how the virus spreads over the area. We advocate using the Laplacian operator to close this gap. To be specific, we add the followings to the deterministic epidemic EG.5 mode:

$$\begin{cases} \dfrac{dS(t,z)}{dt} = d_S \Delta S(t,z) + \Lambda - (1-u)\left[\alpha_1 E(t,z) + \alpha_2 I(t,z)\right] S(t,z) - dS(t,z), \\ \dfrac{dE(t,z)}{dt} = d_E \Delta E(t,z) + (1-u)\left[\alpha_1 E(t,z) + \alpha_2 I(t,z)\right] S(t,z) - (d+\beta) E(t,z), \\ \dfrac{dI(t,z)}{dt} = d_I \Delta I(t,z) + \beta E(t,z) - (d+\delta_1+\eta) I(t,z), \\ \dfrac{dH(t,z)}{dt} = d_H \Delta H(t,z) + \eta I(t,z) - (d+\delta_2+\theta) H(t,z), \\ \dfrac{dR(t,z)}{dt} = d_R \Delta R(t,z) + \theta H(t,z) - dR(t,z). \end{cases} \tag{2}$$

The initial conditions and no-flux boundary conditions are expressed as follows:

$$S(0,z) = S_0,\ E(0,z) = E_0,\ I(0,z) = I_0,\ H(0,z) = H_0,\ R(0,z) = R_0, \quad z \in \Omega. \tag{3}$$

Here, Ω represents a bounded domain within $\mathbb{R}^2$ with smooth boundary denoted as $\partial\Omega$. One needs to add the equation

$$\frac{\partial S}{\partial \upsilon} = \frac{\partial E}{\partial \upsilon} = \frac{\partial I}{\partial \upsilon} = \frac{\partial H}{\partial \upsilon} = \frac{\partial R}{\partial \upsilon} = 0, \quad (t,z) \in [0; t_f] \times \partial\Omega, \tag{4}$$

where υ represents the outward unit normal vector on the boundary $\partial\Omega$, and $\dfrac{\partial}{\partial \upsilon}$ corresponds to the outward normal derivative on $\partial\Omega$.

Additionally, we have a control function $u : [0; t_f] \times \Omega \longrightarrow [0; 1[$, which is designed to decrease the count of infected individuals while concurrently increasing the count of susceptible individuals. This is achieved through the implementation of measures aimed at reducing social interactions, conducting

security campaigns to inhibit migration, and promoting health precautions to minimize virus transmission.

Our objective is to optimize both the count of individuals affected by the illness and the cost of the treatment program. To achieve this, we formulate the problem as follows: Minimize the cost functional represented by

$$C(S,E,I,H,R,u)=\int_{\Omega}\int_{0}^{t_f} aI(t,z)dtdz+\frac{b}{2}\|u(t,z)\|^2_{L^2([0,t_f])} \tag{5}$$

subject to the controlled system (2).

Additionally, we define U_{ad} as the set of admissible controls, characterized by

$$U_{ad}=\{u\in L^{\infty}(Q), 0\leq u\leq 1 \quad \text{a.e. on } Q\}, \tag{6}$$

where $Q=[0,t_f]\times\Omega$, with Ω representing a bounded region within $\mathbb{R}^2$ featuring a smooth boundary $\partial\Omega$. Further details about the parameters of our spatio-temporal model can be found in Table 2.

Table 2: The description of the parameters of our spatio-temporal model (2).

Symbol	Description
Λ	Recruitment rate
d	Natural mortality rate
α_1	Virus transmission rate between an infected person without symptoms and a person users
α_2	Virus transmission rate between an infected person with symptoms and a person users
β	Rate of infected person without symptoms
δ_1	Infected viral mortality rate
δ_2	Viral mortality rate of an infected person during hospitalization
η	Hospitalized people rate
θ	Recovered rate
d_S	Diffusion of susceptible individuals
d_E	Diffusion of infected people without symptoms
d_I	Diffusion of infected people with symptoms
d_H	Diffusion of hospitalized people
d_R	Diffusion of recovered people

4 Existence and uniqueness of global solution

In this part, we demonstrate that our model (2) has a strong global solution.

We consider $M(\Omega) = (L_2(\Omega))^5$,
$H^1(\Omega) = \left\{ u \in L_2(\Omega) : \dfrac{\partial u}{\partial x} \in L_2(\Omega) \text{ and } \dfrac{\partial u}{\partial y} \in L_2(\Omega) \right\}$,
and $H^2(\Omega) = \left\{ u \in H^1(\Omega) : \dfrac{\partial^2 u}{\partial y^2}, \dfrac{\partial^2 u}{\partial y^2}, \dfrac{\partial^2 u}{\partial x \partial y}, \dfrac{\partial^2 u}{\partial y \partial x} \in L_2(\Omega) \right\}$ the Hilbert spaces.

Let $L^2\left(0, t_f; H^2(\Omega)\right)$ be the space of all strongly measurable functions $v : [0, t_f] \longmapsto H^2(\Omega)$ such that

$$\int_0^{t_f} \|v(t,z)\|_{H^2(\Omega)} \mathrm{d}t < \infty.$$

Additionally, we define $L^\infty\left(0, t_f; H^1(\Omega)\right)$ as the set of all functions $v : [0, t_f] \longmapsto H^1(\Omega)$ that satisfy

$$\sup_{t \in [0, t_f]} \left(\|v(t,z)\|_{H^1(\Omega)} \right) < \infty.$$

The norm in $L^\infty\left(0, t_f; H^1(\Omega)\right)$ is given by

$$\|v\|_{L^\infty(0,t_f;H^1(\Omega))} := \inf \left\{ m \in \mathbb{R}_+ : \|v(t,z)\|_{H^1(\Omega)} < m \right\}.$$

Our spatio-temporal model (2) can be expressed in the following form

$$\frac{\partial y(t,z)}{\partial t} = Ay(t,z) + f(t, y(t,z)), \tag{7}$$

where $y = (y_1, y_2, y_3, y_4, y_5) = (S, E, I, H, R)$ and $f = (f_1, f_2, f_3, f_4, f_5)$ is defined by

$$\begin{cases} f_1 = \Lambda - (1-u)\left[\alpha_1 y_2 + \alpha_2 y_3\right] y_1 - d y_1, \\ f_2 = (1-u)\left[\alpha_1 y_2 + \alpha_2 y_3\right] y_1 - (d+\beta)\, y_2, \\ f_3 = \beta y_2 - (d + \delta_1 + \eta)\, y_3, \\ f_4 = \eta y_3 - (d + \delta_2 + \theta)\, y_4, \\ f_5 = \theta y_4 - d y_5. \end{cases}$$

For all $k \in \{1, 2, 3, 4, 5\}$

$$\frac{\partial y_k}{\partial t} = d_k \Delta y_k + f_k(y(t,z)).$$

Let A denote the linear operator, mapping from $D(A) \subset M(\Omega)$, defined as follows:

$$Ay = (d_S \Delta y_1, d_E \Delta y_2, d_I \Delta y_3, d_H \Delta y_4, d_R \Delta y_5)$$

with

$$y \in D(A) = \Big\{ y = (y_1, y_2, y_3, y_4, y_5) \in \left(H^2(\Omega)\right)^5 : \\ \frac{\partial y_1}{\partial \upsilon} = \frac{\partial y_2}{\partial \upsilon} = \frac{\partial y_3}{\partial \upsilon} = \frac{\partial y_4}{\partial \upsilon} = \frac{\partial y_5}{\partial \upsilon} = 0 \quad \text{on} \quad \partial\Omega \Big\}.$$

Theorem 1. Consider a bounded region Ω in $\mathbb{R}^2$ with a sufficiently smooth boundary. Assuming nonnegativity constraints for y_k^0 on Ω (for $k \in \{1,2,3,4,5\}$), as well as nonnegative parameters $\Lambda, d, \alpha_1, \alpha_2, \beta, \delta_1, \delta_2, \eta$, and θ, and an admissible control $u \in U_{ad}$ with an initial condition
$y^0 \in D(A)$, the system (2) possesses a unique strong nonnegative solution $y \in W^{1,2}([0,t_f]; M(\Omega))$, satisfying the following properties:

$$y_1, y_2, y_3, y_4, y_5 \in L^2\left(0,t_f; H^2(\Omega)\right) \cap L^\infty\left(0,t_f; H^1(\Omega)\right) \cap L^\infty(Q).$$

Moreover, there exists a constant $M_0 > 0$, independent of the control u, such that for all $t \in [0,t_f]$ and all $k \in \{1,2,3,4,5\}$

$$\left\| \frac{\partial y_k}{\partial t} \right\|_{L^2(Q)} + \|y_k\|_{L^2(0,t_f,H^2(\Omega))} + \|y_k\|_{H^1(\Omega)} + \|y_k\|_{L^\infty(Q)} \leq M_0. \tag{8}$$

Proof. Since the operator Δ possesses characteristics of dissipation, self-adjointness, and the capacity to generate a C_0 semigroup of contractions on $M(\Omega)$ [22], it is readily apparent that the function f represented as $(f_1, f_2, f_3, f_4, f_5)$ maintains Lipschitz continuity concerning the variable y denoted by $(y_1, y_2, y_3, y_4, y_5)$, and this continuity remains uniform with respect to t within the interval $[0,t_f]$. Consequently, the system possesses a unique strong solution $y \in W^{1,2}([0,t_f]; M(\Omega))$ with

$$y_k \in L^2\left(0,t_f; H^2(\Omega)\right) \qquad \text{for all} \quad k \in \{1,2,3,4,5\}. \tag{9}$$

We prove now that for all $k \in \{1,2,3,4,5\}, y_k \in L^\infty(Q)$.

Set $c = \max\left\{ \|f_k\|_{L^\infty(Q)}\right), \left\|y_k^0\right\|_{L^\infty(\Omega)} : k \in \{1,2,3,4,5\}\Big\}$, and let

$$U_k(t,z) = y_k(t,z) - ct - \left\|y_k^0\right\|_{L^\infty(\Omega)}.$$

It is clear that the function U_k satisfies the system

$$\begin{cases} \dfrac{\partial U_k(t,z)}{\partial t} = d_k \Delta U_k(t,z) + f_k(t, y(t,z)) - c, \quad t \in [0, t_f], \\ U_k(0,z,z') = y_k^0 - \left\| y_k^0 \right\|_{L^\infty(\Omega)}. \end{cases}$$

This system admits a unique strong solution defined by

$$U_k(t,z) = \Gamma(t) \left(y_k^0 - \left\| y_k^0 \right\|_{L^\infty(\Omega)} \right) + \int_0^t \Gamma(t-x) \left(f_k(y(x)) - c \right) dx,$$

where $\Gamma(t)$ is an infinitesimal semigroup related to the operator $d_k \Delta$.

It comes that $U_k(t,z) \leq 0$, so $y_k \leq ct + \left\| y_k^0 \right\|_{L^\infty(\Omega)}$.

Using the similar manner, we can prove that function $V_k(t,z) = y_k(t,z) + ct + \left\| y_k^0 \right\|_{L^\infty(\Omega)}$ is nonnegative. So, $y_k \geq -ct - \left\| y_k^0 \right\|_{L^\infty(\Omega)}$, then $|y_k(t,z),| \leq ct + \left\| y_k^0 \right\|_{L^\infty(\Omega)}$, which implies that

$$y_k \in L^\infty(Q) \ \ \forall (t,z) \in [0, t_f] \times \Omega, \qquad \text{for } k \in \{1,2,3,4,5\}. \tag{10}$$

Now, we establish that $y_k \in L^\infty \left(0, t_f; H^1(\Omega)\right)$ for all $k \in \{1,2,3,4,5\}$. Consider $k \in \{1,2,3,4,5\}$, and start from the equation

$$\frac{\partial y_k(t,z)}{\partial t} - d_k \Delta y_k(t,z) = f_k(t, y(t,z)), \quad (t,z) \in [0, t_f] \times \Omega.$$

We have

$$\int_0^t \int_\Omega \left(\frac{\partial y_k(t,z)}{\partial t} - d_k \Delta y_k(t,z) \right)^2 dz ds = \int_0^t \int_\Omega \left(f_k(t, y(t,z)) \right)^2 dz ds.$$

By employing Green's formula, we arrive at

$$\begin{aligned} &\int_0^t \int_\Omega \left(\frac{\partial y_k}{\partial t} \right)^2 dz ds + d_k^2 \int_0^t \int_\Omega \left(\Delta y_k \right)^2 dz ds \\ &\quad = 2 d_k \int_0^t \int_\Omega \frac{\partial y_k}{\partial t} \times \Delta y_k dz ds + \int_0^t \int_\Omega \left(f_k\left(t, y_k\right) \right)^2 dz ds \\ &\quad = d_k \int_\Omega |\nabla y_k^0|^2 dz - d_k \int_\Omega |\nabla y_k|^2 dz + \int_0^t \int_\Omega \left(f_k\left(t, y_k\right) \right)^2 dz ds. \end{aligned}$$

Since $y_k^0 \in H^2(\Omega)$ and $\|y_k\|_{L^\infty(Q)}$ are bounded independently of u, it follows that

$$y_k \in L^\infty \left(0; t_f; H^1(\Omega)\right), \qquad \text{for } k \in \{1,2,3,4,5\}. \tag{11}$$

Based on (9), (10), and(11), it is determined that the inequality presented in (8) holds.

Furthermore, we conclude that the solution $(y_1, y_2, y_3, y_4, y_5)$ remains non-negative using arguments similar to those employed for the Field-Noyes equations in [19]. Consider the set

$$\Pi = \{(y_1, y_2, y_3, y_4, y_5) : 0 \leq y_k \leq D \text{ for } k \in \{1, 2, 3, 4, 5\}\},$$

and define the convex functions F_k over Π as $F_k(y_1, y_2, y_3, y_4, y_5) = -y_k$. This leads to the following relationships:

$$\begin{aligned}
\nabla(F_1) \cdot f|_{y_1=0} &= \nabla(-y_1) \cdot f|_{y_1=0} = -\Lambda \leq 0, \\
\nabla(F_2) \cdot f|_{y_2=0} &= \nabla(-y_2) \cdot f|_{y_2=0} = -(1-u)\alpha_2 y_1 y_3 \leq 0, \\
\nabla(F_3) \cdot f|_{y_3=0} &= \nabla(-y_3) \cdot f|_{y_3=0} = -\beta y_2 \leq 0, \\
\nabla(F_4) \cdot f|_{y_4=0} &= \nabla(-y_4) \cdot f|_{y_4=0} = -\eta y_3 \leq 0, \\
\nabla(F_5) \cdot f|_{y_5=0} &= \nabla(-y_5) \cdot f|_{y_5=0} = -\theta y_4 \leq 0.
\end{aligned}$$

As shown in [19], the set Π is positively invariant. □

5 Existence of an optimal control

In this part, we aim to establish the presence of an optimal control for minimizing cost function (5) subject to the reaction diffusion system outlined in (2)–(4) with $u \in U_{ad}$. The primary outcome we seek to demonstrate in this section is as follows.

Theorem 2. Given the assumptions outlined in Theorem 1, the optimal control problem (2)–(5) possesses an optimal solution denoted as (y^*, u^*).

Proof. Since u and y_k are uniformly bounded in $L^\infty(Q)$ for $k \in \{1, 2, 3, 4, 5\}$, the infimum of the cost function is ensured. Let $C^* = \inf_{u \in U_{ad}} C(y, u)$. Consider a minimizing sequence $\{u_n\} \subset U_{ad}$ such that $\lim_{n \to +\infty} C(y^n, u_n) = C^*$, where $(y_1^n, y_2^n, y_3^n, y_4^n, y_5^n)$ is the solution of the system (2)–(3) corresponding to the control u_n. This leads to the following system:

$$\begin{cases} \dfrac{dy_1^n}{dt} = d_S \Delta y_1^n + \Lambda - (1-u_n)\left[\alpha_1 y_2^n + \alpha_2 y_3^n\right] y_1^n - d y_1^n, \\ \dfrac{dy_2^n}{dt} = d_E \Delta y_2^n + (1-u_n)\left[\alpha_1 y_2^n + \alpha_2 y_3^n\right] y_1^n - (d+\beta)\, y_2^n, \\ \dfrac{dy_3^n}{dt} = d_I \Delta y_3^n + \beta y_2^n - (d+\delta_1+\eta)\, y_3^n, \\ \dfrac{dy_4^n}{dt} = d_H \Delta y_4^n + \eta y_3^n - (d+\delta_2+\theta)\, y_4^n, \\ \dfrac{dy_5^n}{dt} = d_R \Delta y_5^n + \theta y_4^n - d y_5^n, \end{cases} \tag{12}$$

where $\dfrac{\partial y_1^n}{\partial \upsilon} = \dfrac{\partial y_2^n}{\partial \upsilon} = \dfrac{\partial y_3^n}{\partial \upsilon} = \dfrac{\partial y_4^n}{\partial \upsilon} = \dfrac{\partial y_5^n}{\partial \upsilon} = 0$ on Q.

Given that $H^1(\Omega))$ is compactly embedded in $L^2(\Omega)$, we can conclude that $y_k^n(t,z)$ is compact in $L^2(\Omega)$ for $k \in \{1,2,3,4,5\}$.

Additionally, we need to establish that $\{y_k^n(t,z), n \geq 1\}$ is equicontinuous in $C\left([0,t_f], L^2(\Omega)\right)$.

The boundedness of $\dfrac{\partial y_k^n}{\partial t}$ in $L^2(Q)$ implies the existence of a positive constant h such that

$$\left| \int_\Omega \left(y_k^n\right)^2 (t,z) dz - \int_\Omega \left(y_k^n\right)^2 (x,z) dz \right| \leq h|t-x|.$$

Hence, by the Ascoli–Arzela Theorem (see [5]), we can confirm that y_k^n is compact in $C\left([0,t_f], L^2(\Omega)\right)$. Consequently, $y_k^n \to y_k^*$ uniformly in $L^2(\Omega)$ with respect to t.

Considering the boundedness of Δy_k^n in $L^2(Q)$, there exists a subsequence, denoted again as Δy_k^n, that converges weakly in $L^2(Q)$. Therefore, for all distributions φ,

$$\int_Q \varphi \Delta y_k^n = \int_Q y_k^n \Delta \varphi \to \int_Q y_k^* \Delta \varphi = \int_Q \varphi \Delta y_k^*.$$

This implies that $\Delta y_k^n \to \Delta y_k^*$ weakly in $L^2(Q)$. Furthermore, $\dfrac{\partial y_k^n}{\partial t} \to \dfrac{\partial y_k^*}{\partial t}$ weakly in $L^2(Q)$, and $y_k^n \to y_k^*$ weakly in $L^2\left(0,t_f; H^2(\Omega)\right)$ and $y_k^n \to y_k^*$ weakly in $L^\infty\left(0,t_f; H^1(\Omega)\right)$.

Since $y_1^n y_2^n - y_1^* y_2^* = \left(y_1^n - y_1^*\right) y_2^n + y_1^* \left(y_2^n - y_2^*\right)$, we deduce that $y_1^n y_2^n \to y_1^* y_2^*$ in $L^2(Q)$. As a result, $u_n \to u^*$ in $L^2(Q)$. Furthermore, since U_{ad} is a closed set, it follows that $u^* \in U_{ad}$.

By taking the limit as $n \to \infty$ in (12), we can establish that y^* is a solution to (7) associated with u^*. Thus,

$$
\begin{aligned}
C(y^*, u^*) &= \int_0^{t_f} a y_3^*(t,z)dtdz + \frac{b}{2}\|u^*(t,z)\|^2_{L^2(Q])} \\
&\leq \liminf \int_0^{t_f} a y_3^n(t,z)dtdz + \frac{b}{2}\|u^n(t,z)\|^2_{L^2(Q)} \\
&\leq \lim \int_0^{t_f} a y_3^n(t,z)dtdz + \frac{b}{2}\|u^n(t,z)\|^2_{L^2(Q)} = C^*.
\end{aligned}
$$

This confirms that C reaches its minimum at (y^*, u^*). □

6 Necessary optimality conditions

In this section, we will delve into the optimality conditions for problem (2)–(5) with $u \in U_{ad}$ and provide insights into the characterization of the optimal control. Consider an optimal pair denoted as (y^*, u^*), and let $u^\epsilon = u^* + \epsilon u$, where $\epsilon > 0$, represents a control function satisfying $u \in L^2(Q)$ and $u \in U_{ad}$. We denote the corresponding solutions of (2) associated with the controls u^ϵ and u^* as $y^\epsilon = (y_1^\epsilon, y_2^\epsilon, y_3^\epsilon, y_4^\epsilon, y_5^\epsilon)$ and $y = (y_1^*, y_2^*, y_3^*, y_4^*, y_5^*)$, respectively. We put

$$
G = \begin{pmatrix}
-d - (1-u^*)\left(\alpha_1 y_2^* + \alpha_2 y_3^*\right) & -(1-u^*)\alpha_1 y_1^* & -(1-u^*)\alpha_2 y_1^* & 0 & 0 \\
(1-u^*)\left(\alpha_1 y_2^* + \alpha_2 y_3^*\right) & (1-u^*)\alpha_1 y_1^* - (d+\beta) & (1-u^*)\alpha_2 y_1^* & 0 & 0 \\
0 & \beta & -(d+\delta_1+\eta) & 0 & 0 \\
0 & 0 & \eta & -(d+\delta_2+\eta) & 0 \\
0 & 0 & 0 & \theta & -d
\end{pmatrix}
$$

and

$$
P = \begin{pmatrix}
(\alpha_1 y_2^* + \alpha_2 y_3^*)\, y_1^* \\
-(\alpha_1 y_2^* + \alpha_2 y_3^*)\, y_1^* \\
0 \\
0 \\
0
\end{pmatrix}.
$$

We have the following theorem.

Theorem 3. The mapping $y : U_{ad} \to W^{1,2}([0,t_f], M(\Omega))$ for $k \in \{1,2,3,4,5\}$ is Gateaux differentiable with respect to u^*. For any $u \in U_{ad}$, $y_k'(u^*)\,u = Y_k$, consequently, $Y = (Y_1, Y_2, Y_3, Y_4, Y_5)$ represents the unique solution to the problem given by

$$
\frac{\partial Y}{\partial t} = AY + GY + uP \quad \text{subject to } Y(0,z) = 0.
$$

Proof. We define $Y_k^\varepsilon = \dfrac{y_k^\varepsilon - y_k^*}{\varepsilon}$ for $k \in \{1,2,3,4,5\}$. By subtracting the systems corresponding to y_k^ε and y_k^*, we derive the following equation:

$$\frac{\partial Y^{\varepsilon}}{\partial t} = AY^{\varepsilon} + G^{\varepsilon}Y^{\varepsilon} + uP \text{ subject to } Y^{\varepsilon}(0,z) = 0, \quad \text{for all } z \in \Omega$$

with

$$G^{\varepsilon} = \begin{pmatrix} -d-(1-u^{\varepsilon})\left(\alpha_1 y_2^* + \alpha_2 y_3^*\right) & -(1-u^{\varepsilon})\alpha_1 y_1^{\varepsilon} & -(1-u^{\varepsilon})\alpha_2 y_1^{\varepsilon} & 0 & 0 \\ (1-u^{\varepsilon})\left(\alpha_1 y_2^* + \alpha_2 y_3^*\right) & (1-u^{\varepsilon})\alpha_1 y_1^{\varepsilon} - (d+\beta) & (1-u^{\varepsilon})\alpha_2 y_1^{\varepsilon} & 0 & 0 \\ 0 & \beta & -(d+\delta_1+\eta) & 0 & 0 \\ 0 & 0 & \eta & -(d+\delta_2+\eta) & 0 \\ 0 & 0 & 0 & \theta & -d \end{pmatrix}.$$

Let $(\Gamma(t), t \geq 0)$ be the semi-group generated by A. The solution to this system can be expressed as

$$Y^{\varepsilon}(t,z) = \int_0^t \Gamma(t-x)G^{\varepsilon}Y^{\varepsilon}(x,z)dx + \int_0^t \Gamma(t-x)uPdx.$$

Since the coefficients of the matrix G^{ε} are uniformly bounded with respect to ε, we can apply Grönwall's inequality to establish that Y_k^{ε} is bounded in $L^2(Q)$. Consequently, $y_k^{\varepsilon} \to y_k^*$ in $L^2(Q)$. Taking the limit as $\varepsilon \to 0$, we obtain

$$\frac{\partial Y}{\partial t} = AY + GY + uP \quad \text{subject to} \quad Y(0,z) = 0, \quad \text{for all } z \in \Omega.$$

Applying a similar approach, we conclude that $Y_k^{\varepsilon} \to Y_k^*$ as $\varepsilon \to 0$. □

Consider the adjoint variable $q = (q_1, q_2, q_3, q_4, q_5)$, and let G^* represent the adjoint of the Jacobian matrix G. The dual system associated with our problem can be defined as follows:

$$-\frac{\partial q}{\partial t} - Aq - G^*q = D^*D\psi \quad \text{subject to } q(t_f, z) = 0, \tag{13}$$

where

$$D = \begin{pmatrix} 0&0&0&0&0 \\ 0&0&0&0&0 \\ 0&0&1&0&0 \\ 0&0&0&0&0 \\ 0&0&0&0&0 \end{pmatrix} \quad \text{and} \quad \psi = \begin{pmatrix} 0 \\ 0 \\ a \\ 0 \\ 0 \end{pmatrix}.$$

Lemma 1. Under the conditions of Theorem 1, if (y^*, u^*) represents an optimal pair, then there exists a unique strong solution denoted as $q \in W^{1,2}([0,t_f], M(\Omega))$ for the dual system (13). Specifically, the components $q_k \in L^2\left(0,t_f; H^2(\Omega)\right) \cap L^{\infty}\left(0,t_f; H^1(\Omega)\right)$, where $k = \{1,2,3,4,5\}$.

Proof. This lemma can be established by introducing a change of variable $t' = t_f - t$ and employing a similar manner as demonstrated in the proof of Theorem 3. □

By employing linkages between the state and adjoint equations, objective functional, and standard optimality approaches, we are able to characterize the control optimal and get the necessary conditions for the optimal control problem.

Theorem 4. If there exists an optimal control u^* and its corresponding solution $y^* \in W^{1,2}([0,t_f];M(\Omega))$, then the optimal control u^* can be expressed as

$$u^* = \min\left(u_{\max}, \max\left(0, \frac{(q_2-q_1)\left(\alpha_1 y_2^* + \alpha_2 y_3^*\right) y_1^*}{b}\right)\right). \tag{14}$$

Proof. Let (y^*,u^*) be an optimal pair. Let $u^\varepsilon = u^* + \varepsilon h \in U_{ad}$ with $h \in L^2(\Omega)$ and y^ε be the corresponding state solution. We get

$$\begin{aligned} C'(u^*)(h) &= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left(C(u^\varepsilon) - C(u^*)\right) \\ &= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left(a\int_0^{t_f}\int_\Omega (y_3^\varepsilon - y_3^*)\,dzdt + \frac{b}{2}\int_0^1\int_\Omega \left((u^\varepsilon)^2 - (u^*)^2\right)dzdt\right) \\ &= \lim_{\varepsilon\to 0}\left(a\int_0^{t_f}\int_\Omega \left(\frac{y_3^\varepsilon - y_3^*}{\varepsilon}\right)dzdt + \frac{b}{2}\int_0^1\int_\Omega \left(2hu^* + \varepsilon h^2\right)dzdt\right). \end{aligned}$$

Since $\lim\limits_{\varepsilon\to 0}\dfrac{y_3^\varepsilon - y_3^*}{\varepsilon} = \lim\limits_{\varepsilon\to 0}\dfrac{y_3(u^*+\varepsilon h) - y_3^*}{\varepsilon} = Y_3$, $\lim\limits_{\varepsilon\to 0} y_3^\varepsilon = y_3^*$ and $y_3^\varepsilon, y_3^* \in L^\infty(Q)$, it follows that C is Gateaux differentiable with respect to u^*, and we obtain

$$\begin{aligned} C'(u^*)(h) &= \int_0^{t_f}\int_\Omega aY_3 dzdt + b\int_0^{t_f}\int_\Omega hu^* dzdt \\ &= \int_0^{t_f}\langle D\psi, DY\rangle dt + \int_0^1 \langle bu^*, h\rangle_{L^2(\Omega)}\,dt. \end{aligned}$$

If we put $h = v - u^*$, then

$$C'(u^*)(v-u^*) = \int_0^{t_f}\langle D\psi, DY\rangle dt + \int_0^1 \langle bu^*, v-u^*\rangle_{L^2(\Omega)}\,dt.$$

Since

$$\begin{aligned}\int_0^{t_f} \langle D\psi, DY\rangle dt &= \int_0^{t_f} \langle D^* D\psi, Y\rangle \, dt \\ &= \int_0^{t_f} \left\langle -\frac{\partial q}{\partial t} - Aq - G^* q, Y \right\rangle dt \\ &= \int_0^{t_f} \left\langle q, \frac{\partial Y}{\partial t} - AY - GY \right\rangle dt \\ &= \int_0^{t_f} \langle q, P\,(v - u^*)\rangle \, dt \\ &= \int_0^{t_f} \langle P^* q, v - u^*\rangle_{L^2(\Omega)} \, dt,\end{aligned}$$

and U_{ad} is convex. Therefore, $C'(u)(v-u) \geq 0$ for every $v \in U_{ad}$, which is equivalent to

$\int_0^{t_f} \langle P^* q + bu^*, v - u^*\rangle_{L^2(\Omega)} \, dt \geq 0$ for every $v \in U_{ad}$.

So, $bu^* = -P^* q$, thus $u^* = \dfrac{(q_2 - q_1)(\alpha_1 y_2^* + \alpha_2 y_3^*) y_1^*}{b}$. Since $u^* \in U_{ad}$, we have that (14) holds. □

7 Numerical simulation and presentation of the results

To achieve the interest of optimal control and evaluate the precision of our spatio-temporal model with and without control, we carried out a numerical simulation of our optimal control problem in the context of modeling the spread of disease. We adopt a rigorous methodological approach. Our optimality system is formulated by state equations with initial and boundary conditions (2)–(4), adjoint equations with transversality conditions (13), and a characterization of optimal control (14). To solve this system, we use the forward-backward scanning method in an iterative process. The temporal discretization of the state equations is carried out using the explicit Euler method, while the adjoint equations are solved retroactively in time. The space considered, representing a city $\Omega = 50Km^2$, is initially ($t = 1$ day) populated randomly in the city, and it is assumed that the epidemic began to spread randomly in different areas of the city. The results will be presented as follows.

The values of the parameters connected to our spatio-temporal model with and without control are presented in Table 3.

Table 3: Parameters for the model without and with control.

Parameters	Values
$(\Lambda, d, \alpha_1, \alpha_2, \beta, \delta_1, \delta_2, \eta, \theta)$	(0.06,0.125,0.65,0.65,0.5,0.0001,0.001,0.05,0.02)
$(d_S, d_E, d_I, d_H, d_R)$	(0.2,0.2,0.2,0.2,0.2)

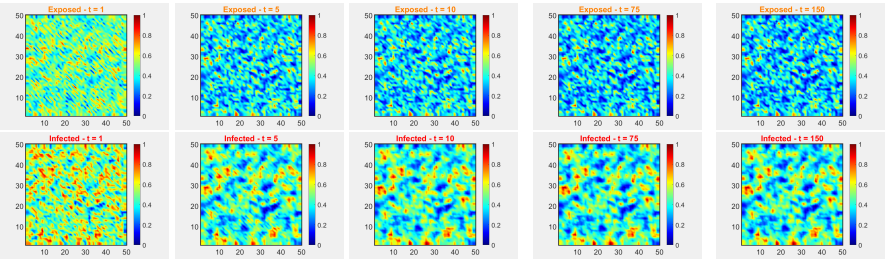


Figure 2: The behavior of compartments E and I without control

Before the enforcement of control measures, a pervasive presence of asymptomatic infections was observed in multiple locations throughout the region, suggesting a widespread dissemination of the epidemic within the city. Figure 4 illustrates a persistent high percentage of asymptomatic cases, consistently standing at 11% of the total population over time. This phenomenon played a role in the formation of a substantial cluster of symptomatic individuals across the city, comprising 31% of the total population. The visual representation underscored the extensive transmission of the infection in various areas of the city.

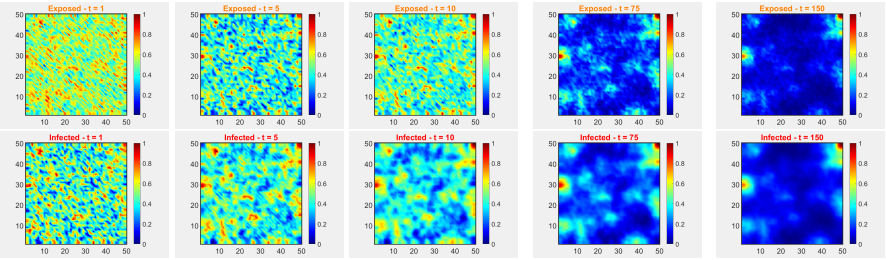


Figure 3: The behavior of compartments E and I with control

After the enforcement of stringent control measures, a noticeable transformation has taken place in the landscape of the epidemic. These measures, which strategically aimed to curtail the widespread prevalence of asymptomatic infections in various locations across the region, including crucial components such as reducing social interactions, conducting safety campaigns to discourage migration, and promoting health precautions to minimize virus transmission. The proactive interventions have resulted in a significant decrease in the overall percentage of asymptomatic cases, plummeting from the initial 11% to a mere 0.2%. Consequently, there has been a substantial reduction in the number of symptomatic individuals, with the affected population now comprising only 0.3% of the total. This positive development indicates the effectiveness of the implemented controls in containing the spread of the disease and alleviating its impact on the population. The newly established situation reflects a more controlled and manageable scenario, underscoring the success of the measures in restricting the transmission of the infection within the city.

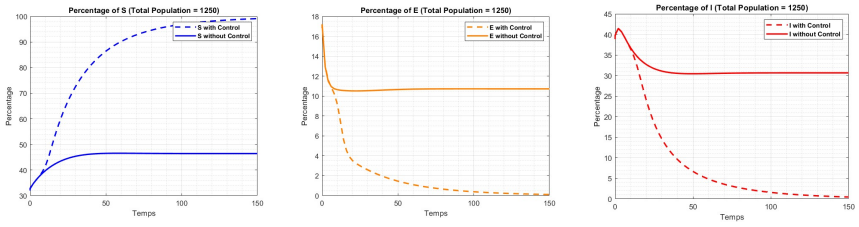


Figure 4: The percentage distribution of compartments S, E, and I relative to the total population

8 Conclusion

The spatio-temporal model used in this work captures population size, movement models, and social relationships that are important in the evolution of EG.5 virus transmission, and as our study suggests, we now understand the role and potential of controls using intervention strategies, such as advocacy campaigns, creating social barriers, and focused testing on, to prevent the spread of the virus while reducing associated costs and societal disorders.

In conclusion, our analysis of this spatio-temporal epidemiological model contributes to well-known decision analyses and important mechanisms of epidemic development. This approach has the potential to combine rigorous mathematical analysis with real data to determine effective targeted therapies, ultimately supporting global efforts to combat the virus and protect humanity wall. Finally, to be more accurate and effective models, we can improve these models by combining other factors such as age, socio-economic characteristics and behavioral models. Additionally, efforts should be made to improve the data collection system and to develop reports that will provide modern, reliable data for accurate sampling, so that we can develop a reduction strategy epidemic and protect human health.

Acknowledgements

Authors are grateful to there anonymous referees and editor for their constructive comments.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Abbasi, Z., Zamani, I., Mehra, A.H.A., Shafieirad, M. and Ibeas, A. *Optimal control design of impulsive SQEIAR epidemic models with ap-*

plication to COVID-19. Chaos Solit. Fractals. 139 (2020) 110054.

[2] El Alami Laaroussi, A. and Rachik, M. *On the regional control of a reaction-diffusion system SIR*. Bull. Math. Biol. 82 (2020) 1–25.

[3] Aldila, D., Padma, H., Khotimah, K., Desjwiandra, B. and Tasman, H. *Analyzing the MERS disease control strategy through an optimal control problem*. Int. J. Appl. Math. Comput. Sci. 28(1) (2018) 169–184.

[4] Athans, M. and Falb, P.L. *Optimal control: An introduction to the theory and its applications.*, Dover Publications, 2006.

[5] Brezis, H., Ciarlet, P.G. and Lions, J.L. *Analyse fonctionnelle: théorie et applications*, vol 91. Dunod, Paris, 1999.

[6] Chinchuluun, A., Pardalos, P.M., Enkhbat, R., and Tseveendori. I, Eds., *Optimization and optimal control: Theory and applications*, vol. 39, Springer, 2010.

[7] Drosten, C., Seilmaier, M., Corman, V.M., Hartmann, W., Scheible, G., Sack, S., Guggemos, W., Kallies, R., Muth, D., Junglen, S. and Müller, M.A. *Clinical features and virological analysis of a case of Middle East respiratory syndrome coronavirus infection*. Lancet. Infect. Dis. 13 (2013) 745–751.

[8] Guery, B., Poissy, J., El Mansouf, L., Séjourné, C., Ettahar, N., Lemaire, X., Vuotto, F., Goffard, A., Behillil, S., Enouf, V. and Caro, V. *Clinical features and viral diagnosis of two cases of infection with Middle East respiratory syndrome coronavirus: A report of nosocomial transmission*. Lancet 381(9885) (2013) 2265–2272.

[9] Khajji, B., Kada, D., Balatif, O. and Rachik, M. *A multi-region discrete time mathematical modelling of the dynamics of Covid-19 virus propagation using optimal control*. J. Appl. Math. Comput. 64(1-2) (2020) 255–281.

[10] Kim, Y., Lee, S., Chu, C., Choe, S., Hong, S. and Shin, Y. *The characteristics of Middle Eastern respiratory syndrome coronavirus transmission dynamics in South Korea*. Osong Public Health Res. Perspect. 7(1) (2016) 49–55.

[11] Kouidere, A., Khajji, B., El Bhih, A., Balatif, O. and Rachik, M. *A mathematical modeling with optimal control strategy of transmission of covid-19 pandemic virus*. Commun. Math. Biol. Neurosci., 2020 (2020), Article ID 24.

[12] Laaroussi, A.E.A., Ghazzali, R., Rachik, M. and Benrhila, S. *Modeling the spatiotemporal transmission of Ebola disease and optimal control: a regional approach*. Int. J. Dynam. Control, 7 (2019) 1110–1124.

[13] Liberzon, D. *Calculus of Variations and Optimal Control Theory*. A Concise Introduction, Princeton University Press, Princeton, NJ, USA, 2012.

[14] Madubueze, C.E., Dachollom, S. and Onwubuya, I.O. *Controlling the spread of covid-19: Optimal control analysis*. Comput. Math. Methods Med. (2020) Article ID 6862516.

[15] El Mehdi, M., Said, M., Bouchaib, K., Omar, B. and Mostafa, R. *Mathematical Study Aiming at Adopting an Effective Strategy to Coexist with Coronavirus Pandemic*. J. Math. Comput. Sci. 11(1) (2021) 44–60.

[16] Perkins, T.A. and España, G. *Optimal control of the COVID-19 pandemic with non-pharmaceutical interventions*. Bull. Math. Biol. 82(9) (2020) 118.

[17] Pontryagin, L.S. *Mathematical theory of optimal processes*, John Wiley and Sons, London, UK, 1962.

[18] Sari, R.A., Habibah, U. and Widodo, A., *Optimal control on model of SARS disease spread with vaccination and treatment*. J. Exp. Life Sci. 7(2) (2017) 61–68

[19] Smoller, J. *Shock waves and reaction-diffusion equations,* Grundlehren der mathematischen Wissenschaften, 258, (GL, volume 258) Springer-Verlag, New York, 1994.

[20] Stengel, R.F, *Optimal control and estimation*, Dover, New York, NY, USA, 1994.

[21] Tahir, M., IS, A.S., Zaman, G. and Khan, T. *Prevention strategies for mathematical model MERS-Corona virus with stability analysis and optimal control*. J. Nanosci. Nanotechnol. Appl. 1(1) (2019) 1.

[22] Vrabie, I.I. *C0-semigroups and applications,* North-Holland Mathematics Studies, 191, North-Holland Publishing Co., Amsterdam, Volume 191 (2003) 1–373.

[23] Yang, C. and Wang, J. *A mathematical model for the novel coronavirus epidemic in Wuhan, China.* Math. Biosci. Eng. 17(3) (2020) 2708–2724.

[24] Zine, H., Adraoui, A.E. and Torres, D.F. *Mathematical analysis, forecasting and optimal control of HIV/AIDS spatiotemporal transmission with a reaction diffusion SICA model*. AIMS Mathematics 7(9) (2022) 16519–16535.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 2, 2024, pp 522–544
https://doi.org/10.22067/ijnao.2024.83955.1303
https://ijnao.um.ac.ir/

Research Article

Optimizing natural gas liquids (NGL) production process: A multi-objective approach for energy-efficient operations using genetic algorithm and artificial neural networks

N. Sabeghi*, B. Kiani Talaei, R. Ghanbari and K. Ghorbani-Moghadam

*Corresponding author

Received 10 August 2023; revised 8 March 2024; accepted 12 March 2024

Narjes Sabeghi
Department of Mathematics, Faculty of Basic Sciences, Velayat University, Iranshahr, Iran. e-mail: n.sabeghi@velayat.ac.ir

Behrouz Kiani Talaei
University of Sistan and Baluchestan, Zahedan, Iran.e-mail: Behrouzkiani@pgs.usb.ac.ir

Reza Ghanbari
Ferdowsi university of Mashhad, Mashhad, Iran. e-mail: rghanbari@um.ac.ir

Khatere Ghorbani-Moghadam
Mosaheb Institute of Mathematics, Kharazmi University, Tehran, Iran. e-mail: k.ghorbani@khu.ac.ir

How to cite this article
Sabeghi, N., Kiani Talaei, B., Ghanbari, R. and Ghorbani-Moghadam, K., Optimizing natural gas liquids (NGL) production process: A multi-objective approach for energy-efficient operations using genetic algorithm and artificial neural networks. *Iran. J. Numer. Anal. Optim.*, 2024; 14(2): 522-544. https://doi.org/10.22067/ijnao.2024.83955.1303

Abstract

There are various techniques for separating natural gas liquid (NGL) from natural gas, one of which is refrigeration. In this method, the temperature is reduced in the dew point adjustment stage to condense the NGLs. The purpose of this paper is to introduce a methodology for optimizing the NGLs production process by determining the optimal values for specific set-points such as temperature and pressure in various vessels and equipment. The methodology also focuses on minimizing energy consumption during the NGL production process. To do this, this research defines a multi-objective problem and presents a hybrid algorithm, including a genetic algorithm (NSGA II) and artificial neural network (ANN) system. We solve the defined multi-objective problem using NSGA II. In order to design a tool that is a decision-helper for selecting the appropriate set-points, the ability of the ANN algorithm along with multi-objective optimization is evaluated. We implement our proposed algorithm in an Iranian chemical factory, specifically the NGL plant, which separates NGL from natural gas, as a case study for this article. Finally, we demonstrate the effectiveness of our proposed algorithm using the nonparametric statistical Kruskal–Wallis test.

AMS subject classifications (2020):

Keywords: Natural gas liquid (NGL); Multi-objective optimization; Artificial neural network; NSGA II.

1 Introduction

Natural gas extracted from underground sources can contain varying amounts of heavier hydrocarbon liquids, including ethane, propane, butanes, and natural gasoline, collectively known as Natural Gas Liquids (NGL). Due to their added value and transportation requirements, NGLs are consistently separated from the treated feed gas. The utilization of NGLs has extended across different sectors, with ethane serving as a feedstock for petrochemical plants and propane and butanes being marketed as liquefied petroleum gas (LPG). Additionally, the heavier fraction of raw gas ($C5^{+}$) can be used as a blending stock for gasoline. NGLs are characterized as the liquid-state

heavier hydrocarbons of natural gas under atmospheric conditions. In a typical natural gas refining plant, various components of NGLs are separated one by one from the natural gas stream by applying a series of fractionation columns, namely demethanizer, deethanizer, depropanizer, and debutanizer. However, process industries face significant operational challenges related to monitoring and controlling these processes. Factors such as outlet chiller temperatures, column pressures, by-pass flows, and control schemes affect the recovery rate, energy consumption, and economic benefits of different NGL extraction methods.

The majority of investigations on problems in gas processing plant focused on different kinds of controlling systems (see [3, 5]), methods of preventing product loss with special attention to the environmental aspects, and energy consumption (see [7, 20, 2]). Some studies have examined the dynamic conditions of the process (see [6, 15]), but determining optimal set-points in terms of energy consumption, and NGL quality has received less attention. It should be noted that the set-point of the the controlled variable is not a single point and can change in a certain range. Designing a tool to determine the optimal set-point is important for making operational decisions because there are numerous set-points in such processes, and even small changes in them can result in significant variations in energy consumption and product quality. On the other word, the main objective of such tools is to define the optimal set-points for the multiple control loops in the process, aiming to maximize the profits while saving energy.

Addressing this complex optimization problem, data-driven model approaches offer the ability to simplify large-scale process models and perform optimization (see [27, 14]). Among the various approaches that utilize multivariate statistics and machine learning, partial least squares [23], artificial neural network (ANN) [21], support vector regression [10], and Gaussian process regression [1] have been widely implemented in constructing surrogate models in numerous research fields. Most of these methods build predictive models using the correlations among different variables (spatial correlations).

By leveraging the computational power of ANNs, industries can achieve efficient and effective optimization, leading to improved process performance, reduced costs, and increased productivity. The application of ANNs in the

industrial sector provides a powerful tool for addressing complex optimization problems. Through their ability to simplify large-scale process models and capture relationships among variables, ANNs enable efficient optimization and predictive modeling in various research fields, ultimately contributing to improved industrial processes and outcomes. Controlling production in industrial factories to achieve optimal conditions in terms of product quality and energy consumption poses a significant challenge, given the longevity of factories. The complexity of chemical plants, characterized by numerous set-points, further exacerbates this challenge. This prompts the exploration of the possibility of designing a tool that can determine the best set-points based on the process condition, aiding operators in their decision-making. Additionally, the feasibility of specifying set-points solely using recorded data, such as process data sheets and laboratory results, is investigated. The primary objective of this paper is to assess the capability of ANNs in determining optimal set-points, considering both the quantity and quality of products as well as energy consumption. So far, there are some research based on ANN in chemical engineering (e.g., see [4, 17, 18, 19, 24, 9, 22, 25]). Also, some research works consider the optimization of NGL production based on a genetic algorithm (see [16]). In this work, for an NGL optimization process, a multi-objective model is introduced. Then, an ANN is combined with a genetic algorithm, named NSGA II, to solve it.

In the rest of the paper, first, we describe the process under investigation in section 1. Then we explain the simulation process briefly. In section 2, we will discuss the optimization methodology, and finally, the numerical results for a real case study, NGL production plant in Iran, will be presented in section 3.

1.1 Process description

The extraction of natural gas from wells yields a mixture of various hydrocarbons, including heavier components that tend to liquefy. Referred to as NGL, these liquids can be categorized based on their vapor pressures, such as low vapor pressure (condensate), intermediate vapor pressure (natural gasoline),

and high vapor pressure (LPG). Commonly found NGL components include ethane, propane, butane, isobutene, and natural gasoline. Several methods are employed to separate NGL from natural gas, including absorption, refrigeration, and cryogenic turbo-expansion. In this study, the focus is on a specific Iranian NGL plant designed using the refrigeration method. According to this approach, the natural gas stream undergoes cooling to approximately -32°C, causing the NGL within the mixture to condense into a liquid phase and separate from the natural gas. Figure 1 illustrates the process block diagram. To prevent ice crystallization in the dew point adjustment section, the feed streams, saturated with water, require the separation of the aqueous phase from the organic phase. The dew point adjustment section facilitates temperature reduction to facilitate NGL condensation; however, it may also result in the entrapment of lean hydrocarbons in the NGL product. The stabilization section is responsible for separating these hydrocarbons from the NGL products, guided by an increase in NGL temperature.

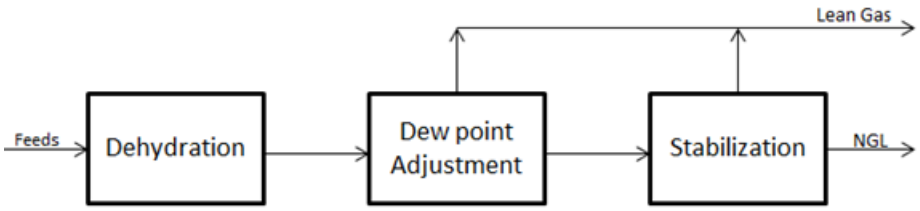


Figure 1: NGL production block diagram

1.2 Process simulation

In this section, we briefly explain the simulation process of NGL production. The process simulation will be explained based on design and operational conditions.

1.2.1 Process simulation based on design condition

This study employs an ANN algorithm, a data-driven approach, to analyze process variables; however, certain data from the plant's distribution control system (DCS) were unavailable. To overcome this limitation and avoid time-consuming data recording, a process simulation was utilized as a virtual representation of the actual process to generate the required data. The simulation incorporated design condition data such as feed and production specifications, equipment parameters, and mechanical information. To evaluate the simulation accuracy, a comparison was made between the simulated product flow values and the design values, which are presented in Tables 1 and 2. These simulation results were then utilized for both training and testing the ANN model. Subsequently, a detailed explanation of the NGL process simulation is provided. The Aspen Plus simulator was used for simulating the NGL production process (Figure 2), which includes the specification of binary parameters of various thermodynamic models. The Peng–Robinson model was specifically employed as the thermodynamic model. In Figure 2, stream (1) represents the high-pressure gas feed, while stream (2) represents the liquid feed, both of which combine in the V-1101 separator. Additionally, glycol is injected into the inlet stream of the E–1101 exchanger (stream 4).

Table 1: Design conditions for NGL plant feed and production

Parameter	Units	Inlet Gas	Inlet Liquid	Lean Gas Outlet	Lean Gas Outlet
Phase	–	Vapor	Liquid	Vapor	Liquid
Temperature	$°C$	41	36	−40.9	22
Pressure	bara	42	42	39.5	21
Average MW	–	23	50	19	45
Mole Flows	kmol/hr	14360	409	1604	2722

Table 2: Simulation precision

Parameter	Lean Gas Outlet (kmol/hr)	NGL Outlet (kmol/hr)
Design condition	1604	2722
Simulation result	1596	2734
precision%	99.5	99.6

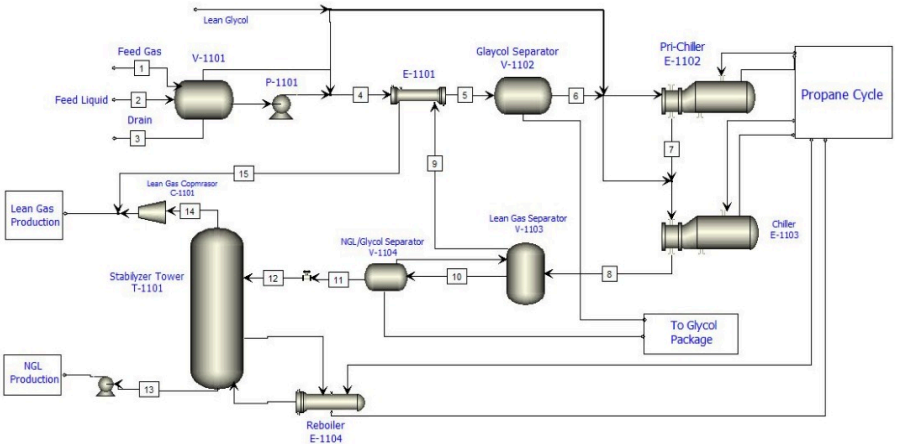


Figure 2: NGL production block diagram

This operation causes that water phase changing in E-1101 from gas to liquid occurs more quickly and prevents hydration. The V-1102 separator is employed to separate glycol solution (referred to as rich glycol) from hydrocarbons, a process commonly known as dehydration. As the hydrocarbons pass through the pre-chiller and chiller, their temperature undergoes a significant decrease of approximately -32°C. These heat exchangers enable a phase change by facilitating heat transfer between the hydrocarbons and liquid propane (used as a refrigerant). Specifically, heavy hydrocarbons transition into the liquid phase, while the liquid propane absorbs heat from the hydrocarbons and transitions into the gas phase. To prevent hydration, glycol is injected into these exchangers since the complete removal of water in the preceding stage is impractical. The outlet stream of the chiller now comprises three phases, including a gas phase. The effective operation of the pre-chiller and chiller is vital for maintaining product quality and minimizing energy consumption. The propane compression cycle incorporates three large electro-compressors, emphasizing the importance of energy efficiency through precise equipment set point tuning and operational strategies. This aspect is known as dew point adjustment. A V-1103 separator is used to separate the liquid phases (oily and aqueous) from the gas phase. Since the gaseous stream from V-1103, representing the light hydrocarbon fraction, is at a low temperature, it is directed to the E-1502 exchanger to pre-cool the

hydrocarbon stream. The liquid phases are directed to V-1104, where the NGL and the aqueous solution, which includes water and glycol, are separated. To achieve the stabilization of the NGL product, it is necessary to remove the trapped light hydrocarbons in the liquid NGL at a lower pressure than the feed pressure. For this purpose, the NGL stream passes through a valve, where the pressure is reduced to half of the feed pressure. The T-1101 distillation column is responsible for the stabilization of the NGL. Distillation, a process involving selective boiling and condensation, is employed to separate the components or substances within a liquid mixture. The reboiler supplies the necessary energy for the evaporation of light hydrocarbons, such as methane and ethane. The NGL enters the top of the column and then flows down to the lowest tray, while the heat from the reboiler causes the evaporation of light hydrocarbons, ensuring the desired stabilization of the NGL product.

1.2.2 Process simulation based on operation condition

Considering the stable composition of the feeds in the plant, it is noted that while the temperature, pressure, and flow rates may undergo changes, the feed flow remains constant. In order to observe the influence of temperature and pressure variations on energy consumption patterns, it is preferable to assume a constant feed flow rate. Figures 3 and 4 illustrate the impact of pressure fluctuations on the vapor stream outlet of the V-1103 separator, specifically for methane and ethane. These figures are based on fixed values of pre-chiller and chiller set-points, as well as a feed temperature of -5°C, a pre-chiller temperature of -32°C, and a chiller temperature of 44°C. The results depicted in the figures demonstrate that an increase in pressure prompts the transition of methane into the liquid phase, consequently escalating the reboiler duty during the stabilization process. On the other hand, the desired effect of pressure variation is the transfer of other hydrocarbons such as ethane into the liquid phase, enriching the production of NGL with heavy hydrocarbons without incurring additional energy consumption. As a result, this leads to a reduction in reboiler duty.

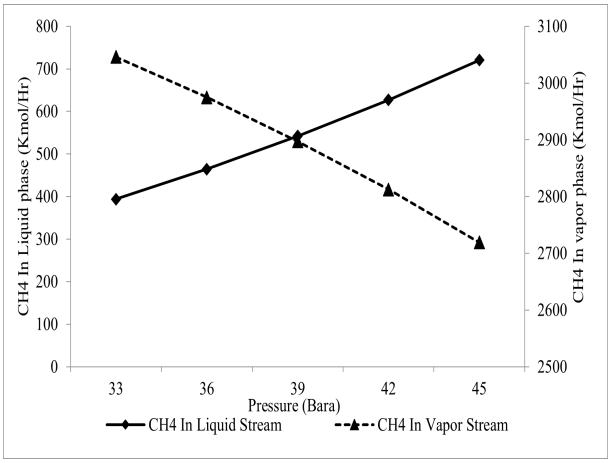


Figure 3: Methane flow rate via pressure

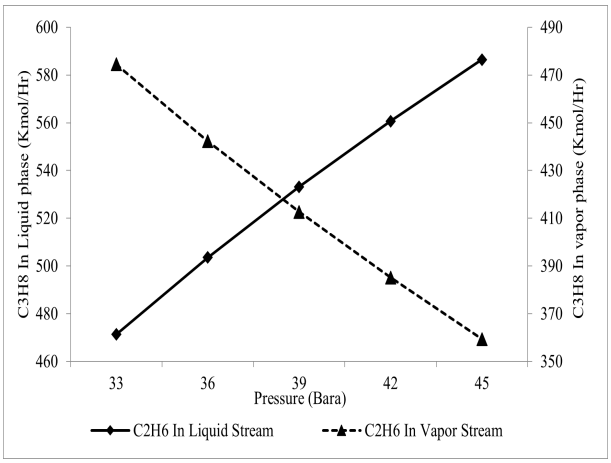


Figure 4: Ethane flow rate via pressure

As mentioned, we try to apply the ANN's ability to determine the set-points, based on which the high-quality and quantity of production with low energy consumption can be achieved. Due to this goal, according to the principles of the Monte Carlo method (see [13]) in different feed pressures and temperatures, the set-point of the pre-chiller, chiller, and the distillation column (the set-point of the tenth tray of distillation column adjust the reboiler duty), as the inputs (Table 3), are randomly changed and the variable shown in Table 4, as the output, are recorded.

Table 3: Set-point criteria ($\circ C$)

pre-chiller	$-9 \leq Temperature \leq -2$
chiller	$-36 \leq Temperature \leq -28$
tenth tray of distillation column	$-8 \leq Temperature \leq -2$

Table 4: Output variable

Output	Type	unit
NGL flow	Mass flow	kg/hr
lean gas flow	Mass flow	kg/hr
pre-chiller duty	Power	w
chiller duty	Power	w
reboiler duty	Power	w
CH4 in NGL stream	Mass fraction	-
CH4 in lean gas stream	Mass flow	-
C2H6 in NGL stream	Mass flow	-
C2H6 in lean gas stream	Mass flow	-

2 Algorithm framework

With the advancements in computer science, a wide range of software has been developed for simulation and optimization across various fields, including chemical engineering. However, most of these software programs rely on basic or "white box" models as their foundation. White box models are derived from basic knowledge and physical insight into the process. These models are primarily developed for the planning and designing of processing units and therefore describe the ideal state and steady-state of the process. While white box modeling or analytical modeling can effectively extrapolate the properties of a process, developing these models requires significant technical expertise and a considerable investment of time and effort. One of the main challenges in these types of modeling is determining the appropriate model parameters. In practice, the mechanisms underlying the system being

modeled are often incomplete or imprecise, making it difficult to accurately capture the true nature of the system. Additionally, because systems can change unpredictably over time, even well-designed models may not perfectly reflect the real world.

In addition, such direct modeling is often not possible. For example, industrial processes are typically extremely complex, and accurate analytical modeling can be challenging or even infeasible. In such cases, black box modeling is often a practical solution. Black box models rely on data from the process itself to establish an input-output relationship and describe the behavior of the system. Data-driven models are particularly effective in accurately capturing the actual conditions of a process unit, as they are built using measured data. It is worth noting that with the increasing use of DCS in factories, a wealth of data is now available for potential use in developing data-driven optimization methods.

An ANN is a simplified model of the natural nervous system and has the ability to learn, like the brain, by processing experimental data [11]. In fact, by performing calculations on numerical data or examples, the grids learn general rules, and therefore, they are referred to as smart systems. The advantage of the neural network in comparison with simulation software is direct learning of data without the need to estimate their statistical characteristics. A neural network, regardless of any initial hypothesis and previous knowledge of the relationships between the parameters studied, is able to find the relationship between the set of inputs and outputs to predict any output corresponding to the desired input.

In this paper, we consider a multi-objective optimization problem and use a genetic algorithm as an optimization method.

A distinguishing feature of our approach, in comparison to similar works, is the use of an ANN to evaluate the target values.

2.1 Objective functions

Clearly, the amount of produced NGL and its quality will vary by changing the set-point of pre-chiller, chiller and tenth tray of distillation column.

Therefore, determining the appropriate set-points in a variety of feed conditions can be an optimization problem in this process. The objective functions that we consider in this paper are the NGL flow, lean gas flow, the pre-chiller duty, chiller duty, reboiler duty and the amount of methane and ethane in NGL and lean gas, which should be maximized or minimized according to Table 5.

Table 5: Problem objective functions

Maximize	NGL flow
	lean gas flow
	mass fraction of ethane in NGL flow
	mass fraction of methane in lean gas flow
Minimize	mass fraction of methane in NGL flow
	mass fraction of ethane in lean gas flow
	pre-chiller duty
	chiller duty
	reboiler duty

To evaluate the objective functions value, we use an ANN model.

2.2 Variables and constraints

The variables and their upper and lower bounds are considered in Table 3.

2.3 Genetic algorithm

Here, our optimization problem is a multi-objective problem, and we use a version of the genetic algorithm for multi-objective problems, namely NSGA II (see [8]), to obtain its solution. The specifications of the used genetic algorithm are shown in Table 6.
It should be noted that during the various implementations of the algorithm,

the values reported for the parameters in Table 6 are the most appropriate values.

Table 6: the parameters of NSGA-II

Parameter	-
Population Size	50
Selection Method	Randomly
Crossover Method	Arithmetic
Probability of Crossover	0.7
Number of Crossover Points	2
Mutation Method	Gaussian
Probability of Mutation	0.4
Mutation Rate	0.02
Number of Iteration	100

2.4 Artificial neural network

In this research, in order to evaluate the objective functions (cost functions) during the NSGAII, we use an ANN [11]. The specifications of the used ANN are shown in Table 7. For convenience, the hybrid algorithm is shown in the

Table 7: The parameters of ANN

Data division	Random
Training function	Levenberg-Marquardt
Performance	Mean squared error
The percentage of data used for training	80%
The number of hidden layers	5
Calculations	MATLAB

flowchart, Figure 5.

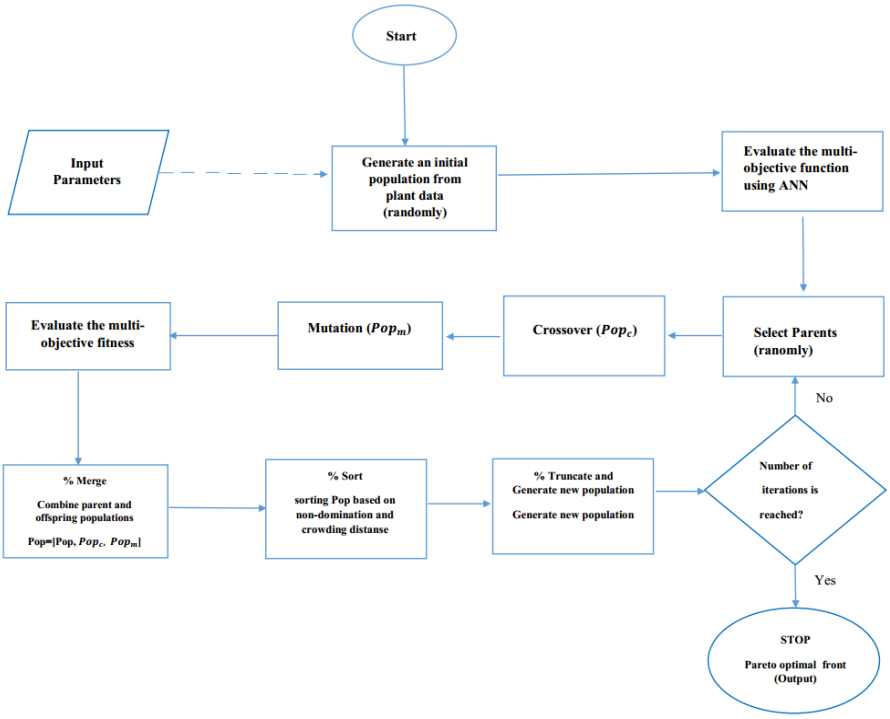


Figure 5: Diagram flow of the hybrid ANN and GA approches

3 Computational results and discussion

Our goal in this article is to evaluate the ability of ANN in determining optimal set-points for the NGL factory. No process knowledge or complex modeling is required to implement the ANN. This is a black box modeling specification. ANN has the ability to solve complex optimization problems based on a set of data. As described in Section 1.2.2, in each of the different conditions of feed, the amount of pre-chiller, chiller, and tenth tray distillation column set-points were randomly changed 100 times in Aspen Plus software, and the changes in product characteristics and energy consumption were recorded. It should be noted that in operating conditions, the temperature and pressure of the feed are always changing, but in most of the time, the set-points are constant. Based on this recorded information, the ANN was able to identify the optimal set-points for the pre-chiller, chiller, and tenth

tray distillation column under various feed conditions, which are presented in Tables 8–13. Each of these tables represents a pressure change from 30 to 45 Bar at a constant temperature, consistent with the factory's design conditions.

Table 8: Results of applying hybrid algorithm when the temperature of input feed is 20 with six different values for its pressure

feed Temperature, Pressure	20, 30	20, 33	20, 36	20, 39	20, 42	20, 45
Tpch	−2.4558	−8.1039	−2.8469	−3.4545	−8.1039	−2.8469
Tch	−34.739	−28.007	−28.089	−28.324	−28.007	−28.089
Tr	−2.176	−6.973	−7.997	−4.066	−6.973	−7.997
NGL Flow(kg/hr)	51068.32	48476.67	49697.35	50486.30	51320.65	52172.75
Lean Gas Flow(kg/hr)	75581.7377	78279.2254	77061.5338	76277.0166	75440.0092	74596.1124
Pch-duty(w)	−2992828.8	−2569350.7	−1789626.3	−1940885.5	−2803750.5	−1924607.6
Ch-duty(w)	−5279878.9	−3233329	−4092233.5	−4033362	−3235192.5	−4104263.6
R-duty(w)	1628711.5	1127823.2	1198274.1	1474674.4	1440698.7	1506065.4
NGL (Methane)	0.02981	0.03417	0.03486	0.03128	0.03366	0.03443
NGL (ethane)	0.2981	0.2306	0.239	0.2466	0.2518	0.2576
Lean (Methane)	0.7099	0.6837	0.6936	0.7027	0.7085	0.7156
Lean (Methane)	0.2104	0.2205	0.2149	0.2097	0.2057	0.2011

Table 9: Results of applying hybrid algorithm when the temperature of input feed is 25 with six different values for its pressure

feed Temperature, Pressure	25, 30	25, 33	25, 36	25, 39	25, 42	25, 45
Tpch	−6.18338	−8.7039	−3.4545	−8.1039	−2.8469	−6.1833
Tch	−28.137	−28.216	−28.324	−28.007	−28.089	−28.137
Tr	−5.587	−6.865	−4.066	−6.973	−7.973	−5.587
NGL Flow(kg/hr)	47212.595	48637.763	49587.753	50533.892	51478.857	52037.833
Lean Gas Flow(kg/hr)	79556.8129	78132.8698	77188.3086	76241.8668	75300.9841	74751.2638
Pch-duty(w)	−2539579.1	−3106388.4	−2362577.7	−3202990.2	−2423875.9	−2946120.6
Ch-duty(w)	−3568959.3	−3169936.6	−4032455.1	−3231098.3	−4096165.8	−3573937.9
R-duty(w)	1080486.7	1145095.2	1367966.3	1337918.5	1404682.4	1609338.3
NGL (Methane)	0.03323	0.03406	0.03148	0.03380	0.03455	0.03226
NGL (ethane)	0.2216	0.2313	0.2396	0.2456	0.2521	0.2576
Lean (Methane)	0.6739	0.6850	0.6946	0.7013	0.7092	0.7157
Lean (ethane)	0.22596	0.21999	0.2146	0.2102	0.2053	0.2011

Table 10: Results of applying hybrid algorithm when the temperature of input feed is 30 with six different values for its pressure

feed Temperature, Pressure	30, 30	30, 33	30, 36	30, 39	30, 42	30, 45
Tpch	−8.7039	−7.4553	−7.7454	−2.6906	−2.8408	−8.5214
Tch	−28.216	−35.99	−28.249	−28.037	−35.648	−35.320
Tr	−6.865	−6.865	−5.555	−4.081	−3.88	−7.59
NGL Flow(kg/hr)	47379.997	53543.9022	49673.5941	50363.0825	55608.3122	56407.8474
Lean Gas Flow(kg/hr)	79400.8129	73240.2306	77111.0292	76431.8586	71190.8946	70395.6043
Pch-duty(w)	−3201496.2	−2970380.75	−3363020.30	−2684177.7	−2642137.9	−3601514.4
Ch-duty(w)	−3174206.3	−4637885.7	−3327603.1	−4108695.4	−5345604.1	−4393631.2
R-duty(w)	1037903.7	1656311.4	1305177.2	1455658.1	2139106	2072962.8
NGL (Methane)	0.0343	0.0334	0.0327	0.0313	0.0302	0.0335
NGL (ethane)	0.2220	0.2586	0.2394	0.2455	0.2766	0.2807
Lean (Methane)	0.6745	0.7289	0.6945	0.7013	0.7515	0.7570
Lean (ethane)	0.2257	0.1993	0.2146	0.2103	0.1834	0.1791

Table 11: Results of applying hybrid algorithm when the temperature of input feed is 35 with six different values for its pressure

feed Temperature, Pressure	35, 30	35, 33	35, 36	35, 39	35, 42	35, 45
Tpch	−8.1564	−2.3367	−2.7848	−4.375	−3.9652	−4.4152
Tch	−28.147	−35.39	−33.948	−28.470	−35.809	−28.974
Tr	−5.019	−3.748	−2.199	−2.146	−4.550	−2.586
NGL Flow(kg/hr)	47237.1015	52964.270	52951.511	50512.7140	55776.8986	52361.6583
Lean Gas Flow(kg/hr)	79560.0112	73835.9236	73847.042	76293.1329	71037.8673	70395.6043
Pch-duty(w)	−3370471	−2424978.4	−2689324	−3219069.8	−3104113.9	−3369385.6
Ch-duty(w)	−3251258.3	−5364446.5	−5056929.11	−3907831.5	−5190935.7	−3996550.7
R-duty(w)	1103174	1745392.4	1832591.2	1562914.2	2122881.6	1794971
NGL (Methane)	0.0327	0.0307	0.0293	0.0296	0.0308	0.0296
NGL (ethane)	0.2216	0.2563	0.2588	0.2468	0.2771	0.2620
Lean (Methane)	0.6741	0.7253	0.7262	0.7036	0.7526	0.7203
Lean (ethane)	0.2259	0.2013	0.1995	0.2093	0.1828	0.1989

For our comparisons, in different temperatures, we used the Kruskal–Wallis nonparametric statistical hypothesis test [26]. We considered the null and alternative hypotheses as follows:

$$\begin{cases} H_0 : \mu_{20} = \mu_{25} = \mu_{30} = \mu_{35} = \mu_{40} = \mu_{45}, \\ H_1 : \mu_{20} \neq \mu_{25} \neq \mu_{30} \neq \mu_{35} \neq \mu_{40} \neq \mu_{45}, \end{cases} \tag{1}$$

where μ_i $(i = 20, 25, 30, 35, 40, 45)$ is the mean value of the results due to NSGA II in different temperature of the input feed. According to the Kruskal–Wallis, since the p-value (p-value is defined as probability under the assumption of the null hypothesis; see the detail on p-value in [12]) is bigger

Table 12: Results of applying hybrid algorithm when the temperature of input feed is 40 with six different values for its pressure

feed Temperature, Pressure	40, 30	40, 33	40, 36	40, 39	40, 42	40, 45
Tpch	−7.4888	−3.9254	−2.3948	−2.0129	−8.6673	−8.9663
Tch	−35.992	−34.847	−28.417	−34.945	−28.043	−31.298
Tr	−2.560	−4.184	−5.282	−2.272	−6.759	−3.335
NGL Flow(kg/hr)	52157.4104	52706.1927	49822.8637	54399.9901	51437.3716	53811.6187
Lean Gas Flow(kg/hr)	74661.0344	74113.3967	77001.8408	72426.5602	75392.4430	73018.6954
Pch-duty(w)	−3318042.2	−2955805.4	−3029352.5	−2956612.1	−4321120	−4337468.1
Ch-duty(w)	−4636169.8	−5019948.5	−4216945.8	−5349044.9	−3148542.7	−3646772.6
R-duty(w)	1706350.2	1689083.7	1327406.6	2029210.2	1452874.3	193927.7
NGL (Methane)	0.0299	0.0312	0.0325	0.0290	0.0334	0.03
NGL (ethane)	0.2498	0.2545	0.2399	0.2686	0.2518	0.2678
Lean (Methane)	0.7182	0.7224	0.6956	0.7401	0.7091	0.7336
Lean (ethane)	0.2064	0.2028	0.2141	0.191	0.2054	0.1922

Table 13: Results of applying hybrid algorithm when the temperature of input feed is 45 with six different values for its pressure

feed Temperature, Pressure	45, 30	45, 33	45, 36	45, 39	45, 42	45, 45
Tpch	−2.4418	−3.4545	−8.1039	−3.4545	−8.1039	−2.2458
Tch	−35.745	−28.324	−28.007	−28.324	−28.007	−34.739
Tr	−4.438	−4.066	−6.973	−4.066	−6.973	−2.176
NGL Flow(kg/hr)	52171.1324	48612.0988	49693.756	50603.0594	51429.8852	55661.4816
Lean Gas Flow(kg/hr)	74662.7311	78212.3323	77134.2149	76226.8468	75394.6536	71170.3579
Pch-duty(w)	−2740606.2	−3288061.8	−4223594.7	−3614717.7	−4506176.1	−3466013.3
Ch-duty(w)	−5408302.9	−4033571.1	−3229535.1	−4031765.8	−3233724.1	−5306934
R-duty(w)	1612112.6	1260970.7	1235050.4	1476449.4	1441966.1	2265692.4
NGL (Methane)	0.0315	0.0317	0.0339	0.0312	0.0336	0.0286
NGL (ethane)	0.2490	0.2314	0.2386	0.2465	0.2517	0.2778
Lean (Methane)	0.7170	0.6858	0.6935	0.7031	0.7089	0.753
Lean (ethane)	0.2069	0.2198	0.2150	0.2095	0.2055	0.1816

than 0.05, H_0 is accepted. Thus, at the $\alpha = 0.05$ level of significance, there is evidence to conclude that the mean value of our proposed algorithm is equal in different temperatures.

NGL and lean gas are the factory productions that sales to the Petrochemical and Gas Company. Each of these customers has quality restrictions when purchasing products. In this article, considering that the selected set-points are in the operational range of the factory, the quality restrictions of the products are satisfied.

Table 14 shows the comparison between the factories fixed set-points ($-5°C$ for pre-chiller and reboiler and $-30°C$ for chiller) and the optimal

Table 14: The comparison between the current operating set-points and the optimal set-points

Production Stream	Variables	plant setpoint	Optimal Setpoint	Difference (%)
NGL	NGL flow (kg/hr)	49767	47380	4.80
	Methane Mass Fraction	0.03	0.03	-5.86
	Ethane Mass Fraction	0.24	0.22	5.77
Lean Gas	lean flow (kg/hr)	77014	79401	-3.10
	Methane Mass Fraction	0.70	0.67	3.03
	ethane Mass Fraction	0.22	0.23	-3.99
Total Energy (mw)		8.23	7.41	9.92

set-points in terms of production rate, product quality, and energy consumption in feed conditions with temperature 30 and pressure 30. The amount of difference in this table is calculated relative to the plant set-point values. In this comparison, about 9.9% less energy is used for optimal set-points. This is equivalent to 0.817 MW and also equivalent to 13.8 tons per day of carbon dioxide emissions. However, under optimal conditions, the amount of NGL production is lower (about 4.8%) compared to the current operating set-points, which is the opposite for light gas (about -3.1%). As mentioned above, the purpose of this paper is to evaluate the ability of ANN to determine the optimal set-points in terms of quantity and quality of products and energy consumption so this study can be the basis for the next study at a higher level (managerial perspective) to determine the set-points in terms of profitability and reduce production costs.

In the used neural network, we considered five neurons in the hidden layer and based on that, the results of Tables 8–13 were obtained. The performance of the designed network for all cases is desirable. The neural network performance and regression diagrams for the data, when the input feed temperature is 44, are shown in Figures 6 and 7, respectively. It should be noted that by changing the number of neurons in the hidden layer and even changing the number of layers, the performance of the network did not change much. The same was true for the other cases.

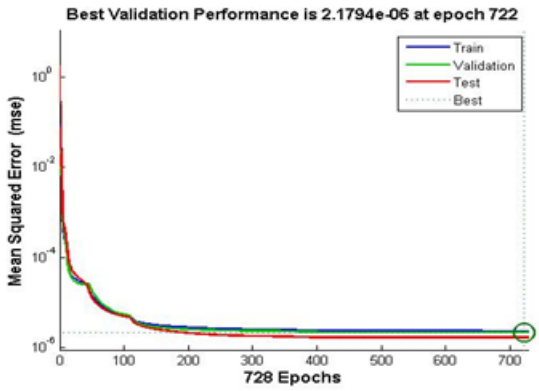


Figure 6: ANN Performance diagram

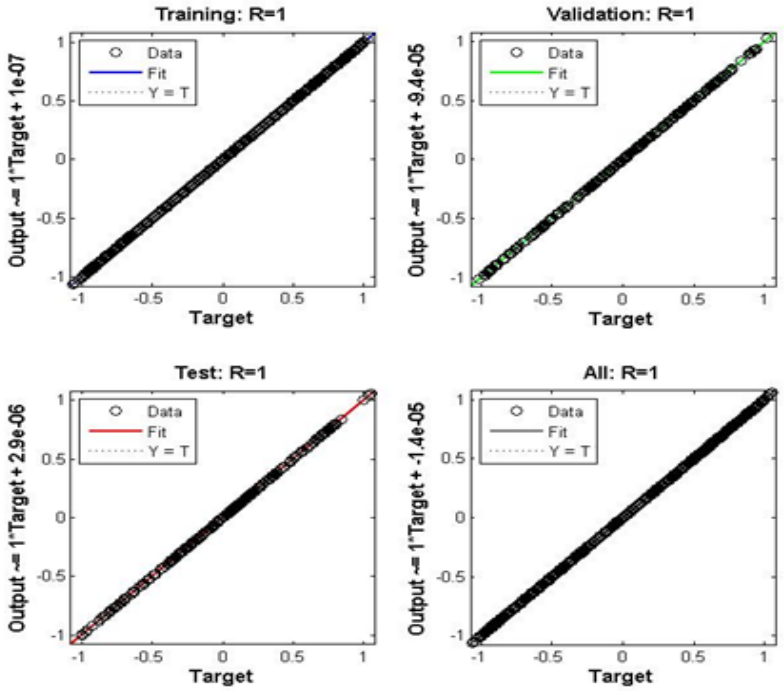


Figure 7: ANN Regression diagram

4 Conclusions and future suggestions

In this paper, we have tried to determine the best amount of set-points in the NGL production process by considering several objective functions, including the amount of NGL, lean gas, the energy consumption of pre-chiller, chiller, and reboiler, and the amount of ethane and methane in NGL and lean gas. This is done using the NSGA-II algorithm and an ANN. The contribution of the presented method is its ability to adapt to any process, its capacity for upgrades, and its utilization of available plant data for machine learning. If there are no sufficient data in the plant, then we can use a simulation just for generating data to train and test the ANN. We used the non-parametric statistical Kruskal–Wallis test to analyze the performance of our algorithm. The obtained results showed that the mean value of our proposed algorithm is equal at different temperatures.

The purpose of this paper is to assess the capability of ANN in identifying the optimal set-points that result in high-quality products, reduced energy consumption, and increased efficiency. This study lays the foundation for further research at a managerial level, which will determine the set-points based on profitability and cost-effectiveness. Additionally, there are two crucial materials used in the process, namely propane and ethylene glycol. Propane is utilized to cool the gas in the pre-chiller and chiller, while ethylene glycol is used to prevent the hydrate phenomenon. Future studies should investigate the factors that contribute to propane wastage and optimize its usage, as well as explore ways to minimize the use of ethylene glycol, thereby providing further opportunities for optimization and can be considered in future research.

Acknowledgements

The authors would like to thank Khorasan Science and Technology Park, Behpoyan company for their support and for giving authors the opportunity to do this research. The first author would like to thank Faculty of Mathe-

matical Sciences, Ferdowsi University of Mashhad, for their assistance during this research.

References

[1] Alipanahi, E., Moradkhani, M.A., Zolfaghari, A., and Bayati, B. *Robust intelligent approaches to predict the CO2 frosting temperature in natural gas mixtures under cryogenic conditions.* Int. J. Refrig. 154 (2023) 281–289.

[2] Amidpour, M., Hamedi, M.H., Mafi, M., Ghorbani, B., Shirmohammadi, R., and Salimi, M. *Sensitivity analysis, economic optimization, and configuration design of mixed refrigerant cycles by NLP techniques.* J. Nat. Gas Sci. Eng. 24, (2015) 144–155.

[3] Ansari, R.M., and Tade, M.O. *Nonlinear model based multivariable control of a debutanizer.* J. Process Control, 8(4) (1998) 279–286.

[4] Arora, S., Shen, W., and Kapoor, A. *Neural network based computational model for estimation of heat generation in LiFePO4 pouch cells of different nominal capacities.* Comput. Chem. Eng. 101 (2017) 81–94.

[5] Bahmani, M., Shariati, J., and Rouzbahani, A.N. *Simulation and optimization of an industrial gas condensate stabilization unit to modify LPG and NGL production with minimizing CO2 emission to the environment.* Chin. J. Chem. Eng. 25(3) (2017) 338–346.

[6] Chebeir, J., Salas, S.D., and Romagnoli, J.A. *Operability assessment on alternative natural gas liquids recovery schemes.* J. Nat. Gas Sci. Eng. 71, (2019) 102974.

[7] Davoudi, M., Aleghafouri, A., and Safadoost, A. *Flaring networks assessment in South Pars Gas processing plant.* J. Nat. Gas Sci. Eng. 21, (2014) 221–229.

[8] Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T.A.M.T. *A fast and elitist multiobjective genetic algorithm: NSGA-II.* IEEE Trans. Evol. Comput. 6(2) (2002) 182–197.

[9] Dos Santos, R. M., Szklo, A., Lucena, A., and Poggio, M. *Evaluating strategies for monetizing natural gas liquids from processing plants–Liquid fuels versus petrochemicals.* J. Nat. Gas Sci. Eng. 99 (2022) 104413.

[10] Ebrahimi, A., and Khamehchi, E. *Developing a novel workflow for natural gas lift optimization using advanced support vector machine.* J. Nat. Gas Sci. Eng. 28 (2016) 626–638.

[11] Fausett, L. V. *Fundamentals of neural networks: architectures, algorithms and applications.* Pearson Education India, 2006.

[12] Freund, J. E. *Mathematical statistics.* Second edition. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1971.

[13] Kroese, D. P., Brereton, T., Taimre, T., and Botev, Z. I. *Why the Monte Carlo method is so important today.* Wiley Interdiscip. Rev.: Comput. Stat. 6(6) (2014) 386–392.

[14] Leperi, K.T., Yancy-Caballero, D., Snurr, R.Q., and You, F. *110th anniversary: surrogate models based on artificial neural networks to simulate and optimize pressure swing adsorption cycles for CO2 capture.* Ind. Eng. Chem. Res., 58(39), (2019) 18241–18252.

[15] Luyben, W. L. *NGL demethanizer control.* Ind. Eng. Chem. Res., 52(33), (2013) 11626–11638.

[16] Mehrpooya, M., Gharagheizi, F., and Vatani, A. *An optimization of capital and operating alternatives in a NGL recovery unit.* Chem. Eng. Technol.: Industrial Chemistry-Plant Equipment-Process Engineering-Biotechnology, 29(12), (2006) 1469–1480.

[17] Nascimento, C.A.O., Giudici, R., and Guardani, R. *Neural network based approach for optimization of industrial chemical processes.* Comput. Chem. Eng. 24(9-10), (2000) 2303–2314.

[18] Park, Y.I., and Kim, J.H. *Artificial neural network based prediction of ultimate buckling strength of liquid natural gas cargo containment system under sloshing loads considering onboard boundary conditions.* Ocean Eng. 249, (2022) 110981.

[19] Roudari, S., Sadeghi, A., Gholami, S., Mensi, W., and Al-Yahyaee, K.H. *Dynamic spillovers among natural gas, liquid natural gas, trade policy uncertainty, and stock market.* Resour. Policy, 83, (2023) 103688.

[20] Rouzbahani, A.N., Bahmani, M., Shariati, J., Tohidian, T., and Rahimpour, M.R. *Simulation, optimization, and sensitivity analysis of a natural gas dehydration unit.* J. Nat. Gas Sci. Eng. 21, (2014) 159–169.

[21] Safarvand, D., Aliazdeh, M., Samipour Giri, M., and Jafarnejad, M. *Exergy analysis of NGL recovery plant using a hybrid ACOR-BP neural network modeling: a case study.* Asia-Pacific J. Chem. Eng. 10(1) (2015) 133–153.

[22] Schweidtmann, A.M., Huster, W.R., Lüthje, J.T., and Mitsos, A. *Deterministic global process optimization: Accurate (single-species) properties via artificial neural networks.* Comput. Chem. Eng. 121, (2019) 67–74.

[23] Shang, C., Huang, X., Suykens, J.A., and Huang, D. *Enhancing dynamic soft sensors based on DPLS: A temporal smoothness regularization approach.* J. Process Control, 28 (2015) 17–26.

[24] Shoghl, S.N., Naderifar, A., Farhadi, F., and Pazuki, G. *Thermodynamic analysis and process optimization of a natural gas liquid recovery unit based on the Joule–Thomson process.* J. Nat. Gas Sci. Eng. 96,(2021). 104265.

[25] Wang, Y., Zhang, Y., Wu, Z., Li, H., and Christofides, P.D. *Operational trend prediction and classification for chemical processes: A novel convolutional neural network method based on symbolic hierarchical clustering.* Chem. Eng. Sci. 225 (2020) 115796.

[26] Wasserman, L. *All of nonparametric statistics.* Springer Texts in Statistics. Springer, New York, 2006.

[27] Yunfei, Ch. and Fengqi, Y. *Integrated planning, scheduling, and dynamic optimization for batch processes: MINLP model formulation and efficient solution methods via surrogate modeling.* Ind. Eng. Chem. Res. 52(47) (2014) 16851–16869.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 2, 2024, pp 545–579
https://doi.org/10.22067/ijnao.2024.82197.1252
https://ijnao.um.ac.ir/

Research Article

A new bi-level data envelopment analysis model to evaluate the Human Development Index

E. Hajinezhad *, H. Hajinezhad and M.R. Alirezaee

Abstract

In the 1990s, the united nations development programme (UNDP) introduced the human development index (HDI) to determine the development degrees of countries. One deficiency in the HDI calculation is the use of equal weights for its sub-indicators. Many scholars have tried to solve this problem using a data envelopment analysis (DEA) method, particularly the one enhanced by weight restrictions. Indeed no specific method

*Corresponding author

Received 04 May 2023; revised 23 February 2024; accepted 24 February 2024

Ensie Hajinezhad
School of Mathematics, Iran University of Science and Technology, Tehran, Iran. e-mail: e.hajinezhad@gmail.com

Haniye Hajinezhad
Department of Mathematics, Payame Noor University, Tehran, Iran. e-mail: H.Hajinezhad@pnu.ac.ir

Mohammadreza Alirezaee
School of Mathematics, Iran University of Science and Technology, Tehran, Iran. e-mail: mralirez@iust.ac.ir

How to cite this article
Hajinezhad, E., Hajinezhad, H. and Alirezaee, M.R., A new bi-level data envelopment analysis model to evaluate the Human Development Index. *Iran. J. Numer. Anal. Optim.*, 2024; 14(2): 545-579. https://doi.org/10.22067/ijnao.2024.82197.1252

has been yet suggested to determine the parameters of the weight restrictions. In this paper, we use four DEA/benefit of the doubt (BoD) models enriched by the assurance regions type I (AR-I) constraints to assess human development; we aim to objectively determine the AR-I bounds. Therefore, we consider a basis as the accepted human development values and propose a bi-level optimization problem to extract the AR-I bounds in such a way that the efficiency scores are almost the same as the basic values. On the other hand, the HDI is a globally accepted index that shows small changes year by year. So, if the UNDP decides to apply a BoD model for calculating the HDI instead of the traditional method, then it is better than the scores obtained by the BoD model, showing small changes in comparison with the HDI, at least in the first few years. Therefore, the HDI values are considered as the basis. Moreover, the objectively achieved AR-I bounds provide us with an insight into the way the sub-indicators affect the development scores. The bounds can be modified by the experts opinions, in the future.

AMS subject classifications (2020): 91B06; 90C29

Keywords: Human development index (HDI); Data envelopment analysis (DEA), Benefit of the doubt (BoD); Assurance regions type I (AR-I); Bi-level optimization problem (BOP).

1 Introduction

In the 1990s, the united nations development programme (UNDP) introduced the human development index (HDI) [87] to show that people's capabilities should be measured not only by economic advances but also by human well-being. The HDI is an incorporation of three important dimensions of human development: a long and healthy life, education, and standards of living. The sub-indicators used for the assessment of these dimensions, respectively, are, expected years of life at birth, mean of years of schooling for adults and expected years of schooling for children, and gross national income per capita.

Since the introduction of the HDI, it has been criticized from different aspects [47, 46, 42]. One main criticism is the arbitrariness that is caused

by subjective choices in the HDI construction. To solve this problem, many authors proposed the use of the data envelopment analysis (DEA) technique [58, 27, 21, 7, 61, 60]. The method used in most papers is to construct an index by applying the benefit of the doubt (BoD) approach, which is a DEA model considering only desirable attributes.

DEA is a nonparametric method for evaluating the performance of decision-making units (DMUs), where each DMU consumes multiple inputs to produce multiple outputs. Since the appearance of the first DEA model, it has been extended and developed vastly. One of the most important developments is the introduction of weight restrictions, which limit the total flexibility in choosing the weights by the DMUs. These weight restrictions also increase the discrimination power of the DEA models. So, a DEA model with weight restrictions seems to be a suitable model for evaluating the HDI (see Section 2.2). However, the parameters associated with the weight restrictions are arbitrarily determined. Blancard and Hoarau [9] suggested that the arbitrariness of choosing these parameters should be eliminated. So, in this paper, we focus on a procedure that objectively determines the parameters of the weight restriction constraints.

Here, we use four BoD approaches (like Mariano, Ferraz, and Oliveira Gobbo [60]), namely, the traditional, the multiplicative, the slacks-based measure (SBM), and the range-adjusted measure (RAM) approaches. We enrich these BoD models with the constraints of the assurance regions type I (AR-I) [83]. Our proposed procedure is to consider a set of values as the basis and then objectively determine the bounds of the AR-I constraints in such a way that the enriched BoD model produces the basis. This procedure is somewhat inspired by the work of Edirisinghe and Zhang [31], where input/output selection is endogenously performed by an iterative procedure that seeks the maximum correlation between the DEA-based performance scores and the stock returns. Also, it is somehow similar to the approach used in Alrezaee, Hajinezhad, and Paradi [2], in which the returns to scale are extracted from the data by using a method based on data mining. The key question is how to select the basis. To answer this question, we have reviewed those research papers that used DEA to solve the problem of equal weights for evaluating the HDI (see Section 2.2). In some papers, the scores

obtained by the DEA models were compared with the HDI [58, 27, 28, 56]; in cases where there were large differences, the reasons were explained. Now, two ambiguities arise: 1) why did the researchers refuse to justify small differences? 2) why does it seem that a small difference between the two rankings is desirable?

On the other hand, if the UNDP decides to shift the HDI evaluation from geometric average to a BoD model, then the HDI values and rankings may change dramatically, which is not usual; by the current evaluation method, the ranking will change a little between two consecutive years, because of the nature of the HDI sub-indicators; except for gross national income per capita, the changes in other sub-indicators are very slow from year to year. So, it would be wise to manipulate the BoD approach in such a way that its results make the lowest changes in the human development values or the rankings in comparison with the last year. Since the UNDP's HDI has such a property, we consider it as the basis and try to extract the required information for setting up the BoD model to obtain efficiency scores that are closest to the basis. By using the proposed method, the effects of the sub-indicators on the calculation of human development values are determined and can be manipulated by the experts' opinions.

To objectively determine the bounds of the AR-I constraints, we propose a bi-level optimization problem (BOP) in which the upper-level is the minimization of the differences between the bases, namely, the UNDP's HDI values and the enriched BoD scores; the lower-level is a linear multi-objective optimization problem derived from the integration of the enriched BoD model for all units. The proposed model is NP-hard. To search for a globally optimal solution for the proposed BOP, we apply a nested approach; the upper-level is optimized by a genetic algorithm (GA) that is enhanced by two local search methods. Also, the lower-level linear multi-objective optimization problem is transformed into N linear optimization problems, which are solved via an exact method. However, the final solution of the BOP is the one which produces the scores that are closest to the UNDP's HDI values, solutions with the objective function (of the upper-level) less than a determined value are saved in *BestSet* during the process of solving the BOP. If the final solution is not desirable, then a solution from *BestSet* can be applied. Of course, the

bounds can be manipulated according to the experts' opinions, and they can be gradually modified each year.

It is worthwhile to note that shifting from a traditional performance assessment system that uses fixed equal weights to a more complicated one, namely, a DEA model, could be difficult from the approval aspect. For example, suppose that a DEA model is used instead of the arithmetic average to evaluate the performance of bank branches. The DEA model may produce efficiency scores that are substantially different from the results of the traditional model. Moreover, it may show big changes in comparison with the last assessment. As a result, it is difficult for the managers to accept the new rankings. For this reason, it is advisable to apply the proposed BOP to extract the AR-I bounds in such a way that the DEA results do not differ dramatically from the results of the traditional method. However, the bounds could be manipulated by the experts' opinions and also be changed in the future. Moreover, considering a set of rankings as the base and applying the proposed BOP for extracting the parameters of weight restrictions are helpful in cases where the experts cannot define the parameters explicitly but are able to determine a ranking of the units.

The contributions of this research can be described as follows:

1) We propose a multi-objective BOP to objectively determine the AR-I bounds in the enriched BoD models, and we apply their results for evaluating human development values.

2) The proposed BOP represents a road to shift slowly from the traditional evaluation of the HDI (or any index) to a BoD approach (or any DEA model) with the least changes in the efficiency values or rankings.

3) This structure, which is used for initializing the parameters of a BoD model, can be applied to objectively determine all parameters that appeared in the DEA models.

This paper is organized as follows. Section 2 reviews the literature related to DEA, the HDI, and BOP. In Section 3, the BoD approaches for evaluating the human development are reviewed. In Section 4, the BOP is proposed for tuning the AR-I bounds. We propose an approach to solve the BOP model

in Section 5. In Section 6, the numerical results of the proposed method are presented. The last section is devoted to a brief conclusion.

2 Literature review

This section reviews the literature related to DEA method.

2.1 Data envelopment analysis

The origins of DEA date back to Farrell's seminal paper [34]. Charnes, Cooper, and Rhodes [17] developed Farrell's idea and presented a model that is able to measure the efficiencies of units having multiple inputs and multiple outputs. This first model of DEA is called CCR. It has inspired several extensions, most notably the BCC model of Banker, Charnes, and Cooper [5] and the additive model of Charnes et al. [16].

The most well-known advantage of the DEA models is the flexibility in choosing weights. As a result of this, a DMU that is determined to be inefficient cannot argue that its inefficiency is due to the weights selected for its inputs and outputs. However, this full flexibility has been criticized from different aspects (see [3, 65]). Although the DMUs may have their own circumstances and objectives in general, they are assumed to be homogeneous. So, it is not sensible that the weights significantly differ from one DMU to another [71]. To resolve this problem, weight restriction is introduced to ensure that all inputs and outputs are considered and that higher weights are assigned to the more important ones.

Now, let us mention the most prominent weight restriction constraints in the DEA literature. The most straightforward weight restriction is to set simple numerical limits on each of the weights [30, 69]. This may lead to infeasible linear programming.

Thompson et al. [83] proposed the concept of assurance regions type I (AR-I), which incorporates into the model the relative importance of the

inputs/outputs. Moreover, they introduced assurance regions type II (AR-II), which is the relative value of inputs to outputs.

It should be noted that applying the AR-II constraints may cause infeasibility. The Cone-Ratio (CR) technique, developed by Charnes et al. [20] in 1989, is the most well-known weight restriction that acts on the data and transforms them [82]. AR-I and CR are related in the sense that an AR-I constraint can be represented as a CR one. However, CR is more general than AR-I.

Wong and Beasley [91] presented a weight restriction based on virtual inputs/outputs which limited the importance of input i to DMUj .

The DEA literature is vast, leading to numerous conducted reviews. Liu, Lu, and Lu [52] delved into the evolving landscape of DEA research, employing a network clustering method to identify four distinct research fronts, shedding light on key areas such as bootstrapping, undesirable factors, cross-efficiency, and network DEA. Building on this, Zhou et al. [96] offered a comprehensive overview of DEA's origins, popular models, applications, and its evolving trends through a bibliometric analysis. Complementing these works, Camanho and D'Inverno [13] provided a historical overview of DEA, discussing its main models, recent developments, and successful applications, emphasizing its relevance in organizational management and public policy. Furthermore, Krmac and Mansouri Kaleibar [49] conducted a systematic review of DEA's applications in port efficiency evaluation, revealing its potential as a valuable tool for assessing future port performance, while also identifying research gaps. Emrouznejad et al. [33] contributed to the discourse by outlining recent theoretical developments and novel applications of DEA, showcasing its versatility across diverse fields. Liu et al. [53] reviewed the areas in which DEA has been applied. They concluded that the top five industries addressed by researchers were banking, healthcare, agriculture, transportation, and education. The applications of DEA are surveyed on the field of banking [64], healthcare [44], agriculture [50], transportation [59], and education [81].

2.2 The human development index and DEA

DEA was originally proposed in the microeconomic environment to measure the performance of schools, hospitals, and so on. It is also well-suited for the evaluation of macroeconomic performance [55]. When DEA is extended to analyze social performance, the concepts of "input" and "output" are not considered in the same sense as their traditional sense of DEA; instead, they are considered desirable and undesirable attributes, respectively, [67].

The first applications of DEA to the evaluation of macroeconomic performance can be found in [63, 39]. In this area, researchers have used DEA to construct a composite index and to measure the efficiencies of countries or regions in generating welfare [61]. The most important advantage of using DEA is that the arbitrariness in weighting the sub-indicators is eliminated, and they are determined endogenously [45]. In fact, the DEA model is used to retrieve some information on appropriate weighting hidden in the performance of a country; the relatively strong (poor) performance of the country in one indicator shows that its policies consider that dimension as more (less) important than the others [21].

For constructing an index, the DEA approach with only the desirable attributes is alternatively called the BoD (benefit of the doubt). This was originally proposed by Melyn and Moesen in 1991 [63]. In the BoD approach, all the indicators or sub-indicators are considered as outputs, along with a dummy input equal to 1 for all the units. Unlike the original HDI calculated by the equal weights, the HDI obtained using the BoD approach applies the most advantageous weights for each country or region [10]. A good review of the BoD approach and the different weight restrictions, which can be used in it, is provided by Cherchye et al. [21].

The first application of the BoD approach for recalculating the HDI dates back to the work of Mahlberg and Obersteiner [58]. They stated two reasons for using the DEA method.

(a) The HDI should be evaluated against the best performance of the other countries.

(b) The weights of the sub-indicators should be determined directly from the data and independently for each country.

Initially, they used an output-oriented CCR model. Then, to improve the discrimination of the efficient countries, a BoD approach with AR-I constraints was applied; they examined three different arbitrary intervals for the AR-I constraints. The ranking of countries resulting from this method and the ranking obtained by the UNDP had a correlation of 0.83. Although, in some cases, there were significant differences. They believed that the main reason for such differences was the weights expertly assigned to the sub-indicators of the HDI.

Despotis [27] presented a method for evaluating the HDI based on a modified DEA model. In this method, a BoD approach is used, and then, by using a goal linear programming model, common weights for all the countries are obtained. Thus, a new estimate of the HDI is achieved, which improves comparability. According to Despotis [28], when the HDI is used to evaluate the human development of countries, the sub-indicators of life expectancy, education, and standards of living are considered as means for achieving social welfare. Although this approach is accepted, human development could also be considered differently; standards of living improve the sub-indicators of life expectancy and education. In such cases, the competency of countries in converting income into a higher quality of life can be modeled and evaluated. Thus, to evaluate the human development of countries, Despotis used a DEA model with variable returns to scale in which the sub-indicator of standards of living was the input and the sub-indicators of life expectancy and education were the outputs.

Lee, Lin, and Fang [51] presented a fuzzy multi-objective DEA model to evaluate the performance of countries based on the human development perspective. In their approach, like Despotis' model [28], the efficiencies of all countries were maximized using a common set of weights.

In 2010, the UNDP changed the method of measuring the HDI [88]. One of these changes was the use of geometric averages instead of arithmetic averages, which eliminated compensation among the HDI sub-indicators. However, the problem of equal weights has remained unresolved. To solve this problem, Zhou et al. [95, 93] used two BoD approaches in the multiplicative

form [18, 19] to calculate the best and the worst efficiency scores. Then, the convex combination of these two efficiency scores was regarded as the efficiency of the country under study.

In some DEA results, the optimal weights of some sub-indicators may be equal to 0. To resolve this problem, Zhou et al. applied the weight restriction proposed by Wong and Beasley [91] and set the parameters arbitrarily. It should be noted that by using the logarithmic transformation, the multiplicative model is changed into the additive model. Also, Blancard and Hoarau [8] used a similar BoD approach with the same weight restriction to evaluate the HDI. One of the drawbacks of their approach is that the weight parameter in the convex combination of the best and worst efficiency scores significantly affects the value and rank of the HDI [40]. In addition, the multiplicative models in [95, 93, 8] are not scale-invariant, which is problematic for linearizing the model because the values of sub-indicators lie between 0 and 1 (see Section 3).

Hatefi and Torabi [40, 41] used a set of common weights to recalculate the HDI of Asian and Pacific countries. To do so, they used a one-stage optimization model to calculate the common weights in such a way that the maximum deviation from the full efficiency score was minimized. The drawback of this model is that the weights of the sub-indicators are substantially influenced by the countries that have poor performance [84]. By using the dual of the proposed model, Hatefi and Torabi [41] performed an analysis to improve the performance of countries with low HDI values.

Tofallis [85] applied a multiplicative DEA model. The model involves an extra factor that absorbs the effect of multiplying any sub-indicator by a constant. The method proposed in [85] is somewhat similar to Despotis' model [27, 28]. Firstly, the DEA model is applied to all countries. Subsequently, a set of common weights is estimated by using the least squares regression.

Puyenbroeck and Rogge [89] compared groups instead of individuals to cope with the problems caused by intermediate steps in the construction of the HDI. They used a BoD approach with weight restrictions to analyze the HDI levels in seven regions. The method they utilized to limit the weights was based on the work of Wong and Beasley [91], with the difference that only the unit under study was considered.

Mariano, Sobreiro, and Rebelatto [61] provided a valuable review of the research works that used DEA to recalculate the HDI. They highlighted twenty gaps in this field. To address the lack of systematic studies considering the advantages and challenges of applying different DEA approaches to the evaluation of the HDI, Mariano, Ferraz, and Oliveira Gobbo [60] compared multiple BoD approaches, including the traditional, the multiplicative, the SBM, the RAM, weight restrictions, common weights, and tiebreaker methods. For the weight restriction, they used the method proposed by Puyenbroeck and Rogge [89]. The analysis was done for raw and normalized data by using social network analysis (SNA) and the information derived from the model itself. Finally, they made the following useful suggestions:

- The normalized data are preferable because of avoiding the outliers.
- The nonradial models are more appropriate because they do not generate false efficiencies.
- To improve the results, it is recommended that weight restrictions be considered in the DEA model.

A group of researchers has investigated the efficiency of countries or regions in transforming the minimum possible units of resources into the maximum possible levels of the sub-indicator of quality of life [4, 66]; it can be seen as the sustainable HDI [9, 15].

2.3 Bi-level optimization problems

Multi-level and bi-level optimization problems have emerged as important research areas in mathematical programming [80]. These optimization problems are widely used in the modeling of real-world problems in which there are many decision-makers at different levels. A BOP is a special case of multi-level optimization problems (MOPs) with two levels, where the upper-level is optimized subject to the best solution of the lower-level. The first formulation of BOPs goes back to the work of Bracken and McGill [12]. Later, Candler and Norton [14] proposed the term "bi-level programming" in a technical report. Since then, many researchers have focused on this area of optimization.

MOPs and BOPs have been applied to various fields, including supply chain management [68], energy-efficient scheduling [92], and robot motion planning [97]. Recently, BOPs have been applied to solve parameter tuning problems, where the procedure of algorithm configuration is designed by using a BOP [78, 62].

BOPs are strongly NP-hard [76]. Many studies have been dedicated to solving BOPs using classical methods such as the Karush–Kuhn–Tucker (KKT) approach [6, 74], descent methods [90], and the penalty function methods [57]. However, in the case of complex BOPs, evolutionary techniques are more suitable. For a review of solution methodologies, we refer the reader to [76, 22]. One of the most popular evolutionary algorithms for solving BOPs is the nested methods, where the lower-level optimization problem is solved completely corresponding to any given upper-level decision [77]. The upper-level optimization is handled using an evolutionary algorithm, and the lower-level is handled by using a classical method or an evolutionary one. As generalizations of BOPs, multi-objective BOPs have been considered by a number of researchers [75, 70, 26].

3 Some BoD approaches used to evaluate the HDI

Here, the different BoD approaches addressed in this work are represented (for more details; see Mariano, Ferraz, and Oliveira Gobbo [60]). We utilize the multiplier form of the BoD models in the subsequent sections. Additionally, for a clearer comprehension of the models, we present the envelopment forms. Suppose that E_j, H_j, and G_j are the sub-indicators of education, health, and standards of living, and represent the outputs for country j ($j = 1, 2, \ldots, N$), respectively, and assume that W_E, W_H, and W_G are the weights of the sub-indicators of education, health, and standards of living, respectively. Also, E_o, H_o, and G_o are the respective sub-indicators related to the DMU under evaluation. Moreover, v is a scalar value being free in sign. A dummy input equal to 1 is considered for all units, and N is the number of countries under evaluation. For the envelopment form, suppose that S_E, S_H, and S_G are nonnegative slacks of the education, health, and standards of living, respectively, and that λ_j is the importance value of benchmark j for the

Table 1: Traditional BoD

Traditional BoD in multiplier form	Traditional BoD in envelopment form
min $\frac{1}{E_o} = v$ Subject to: $W_E \cdot E_o + W_H \cdot H_o + W_G \cdot G_o = 1,$ $-v + W_E \cdot E_j + W_H \cdot H_j + W_G \cdot G_j \leq 0,$ for all $j,$ $W_E, W_H, W_G \geq 0.$	max $\frac{1}{E_o} = \theta$ Subject to: $\sum \lambda_j E_j \geq E_o,$ $\sum \lambda_j H_j \geq H_o,$ $\sum \lambda_j G_j \geq G_o,$ $\sum \lambda_j = 1,$ $\lambda_j \geq 0.$

country under analysis. The traditional BoD approach [63] is the output-oriented BCC model. The traditional BoD approach 57 is the output-oriented BCC model. The multiplier and envelopment forms are as follows:

The multiplier and envelopment forms of the multiplicative BoD approach, which are similar to the model presented by Tofallis [85], are as follows:

Table 2: Multiplicative BoD

Multiplicative BoD in multiplier form	Multiplicative BoD in envelopment form
Max $Ef_o = e^v \times E_o^{W_E} \times H_o^{W_H} \times G_o^{W_G}$ Subject to: $W_E + W_H + W_G = 1,$ $e^v \times E_j^{W_E} \times H_j^{W_H} \times G_j^{W_G} \leq 1,$ for all j , $W_E, W_H, W_G \geq 0.$	Min $Ef_o = \theta$ Subject to: $\frac{\theta \times \prod_j \lambda_j}{S_E} = E_o,$ $\frac{\theta \times \prod_j \lambda_j}{S_H} = H_o,$ $\frac{\theta \times \prod_j \lambda_j}{S_G} = G_o,$ $\sum \lambda_j = 1,$ $\lambda_j, S_E, S_H, S_G \geq 0.$

Applying the natural logarithm to the objective function and the constraints, the multiplicative BoD in Table 2 are linearized as follows:

Table 3: Linearized Multiplicative BoD

Linearized Multiplicative BoD in multiplier form	Linearized Multiplicative BoD in envelopment form
Max $\ln(E_f) = v + W_E \ln(E_o)$ $+W_H \ln(H_o) + W_G \ln(G_o)$ Subject to: $W_E + W_H + W_G = 1,$ $v + W_E \ln(E_j) + W_H \ln(H_j) + W_G \ln(G_j) \leq 0,$ for all $j,$ $W_E, W_H, W_G \geq 0.$	Min $\ln(E_f) = \ln(\theta)$ Subject to: $\ln(\theta) + \sum_j \ln(E_j)\lambda_j - \ln(S_E) = \ln(E_o),$ $\ln(\theta) + \sum_j \ln(H_j)\lambda_j - \ln(S_H) = \ln(H_o),$ $\ln(\theta) + \sum_j \ln(G_j)\lambda_j - \ln(S_G) = \ln(G_o),$ $\sum_j \lambda_j = 1,$ $\lambda_j, \ln(S_E), \ln(S_H), \ln(S_G) \geq 0.$

The multiplier and envelopment forms of the SBM-BoD model, which are based on the output-oriented SBM model (Tone [86]), are as follows. Note that, as the slack related to the input is zero, the input term has been removed, and the inverse of the objective function is maximized.

Table 4: SBM-BoD

SBM-BoD in multiplier form	SBM-BoD in envelopment form
$Min \frac{1}{Ef_o} = v - W_E E_o - W_H H_o - W_G G_o$	$Max \frac{1}{Ef_o} = 1 + \frac{1}{3}(\frac{S_E}{E_o} + \frac{S_H}{H_o} + \frac{S_G}{G_o}$
Subject to:	Subject to:
$v - W_E E_j - W_H H_j - W_G G_j \geq 1,$ for all j,	$\sum_j \lambda_j E_j - S_E = E_o,$
$W_E \geq \frac{1}{3E_o}$,	$\sum_j \lambda_j H_j - S_H = H_o,$
$W_H \geq \frac{1}{3H_o}$,	$\sum_j \lambda_j G_j - S_G = G_o,$
$W_G \geq \frac{1}{3G_o}$.	$\sum_j \lambda_j = 1,$
	$\lambda_j, S_E, S_H, S_G \geq 0.$

Finally, the RAM-BoD model, derived from the output-oriented RAM model (Aida et al. [1]), is scale and translation invariant. It is represented in the multiplier and envelopment forms as follows:

Table 5: RAM-BoD

RAM-BoD in multiplier form	RAM-BoD in envelopment form
$MaxEf_o = v + W_E E_o + W_H H_o + W_G G_o$	$MinEf_o = 1 - \frac{1}{3}(\frac{S_E}{R_E} + \frac{S_H}{R_H} + \frac{S_G}{R_G})$
Subject to:	Subject to:
$v + W_E \cdot E_j + W_H \cdot H_j + W_G \cdot G_j \leq 1,$ for all j,	$\sum_j \lambda_j E_j - S_E = E_o,$
$W_E \geq \frac{1}{3R_E}$,	$\sum_j \lambda_j H_j - S_H = H_o,$
$W_H \geq \frac{1}{3R_H}$,	$\sum_j \lambda_j G_j - S_G = G_o,$
$W_G \geq \frac{1}{3R_G}$.	$\sum_j \lambda_j = 1,$
	$\lambda_j, S_E, S_H, S_G \geq 0.$

Here, R_E, R_H, and R_G are the ranges of education, health, and standards of living in the N countries, respectively. It is worth noting that $\frac{1}{3}(\frac{S_E}{R_E} + \frac{S_H}{R_H} + \frac{S_G}{R_G})$ represents the range-adjusted inefficiency of the country under evaluation.

We enrich the aforementioned BoD models by using the AR-I restrictions

$$L_1 \leq \frac{W_E}{W_H} \leq U_1, \tag{1}$$

$$L_2 \leq \frac{W_E}{W_G} \leq U_2, \tag{2}$$

where L_1, U_1, L_2, and U_2 are constant and nonnegative values. From here on, we refer to RAM-BoD mode (Table 5) with the AR-I constraints (1)–(2) as the enriched RAM-BoD model or enriched model (Table 5). The parameters of the AR-I constraints could be determined by experts' opinions. However, to avoid problems associated with the application of the BoD model instead of the traditional assessment system, we objectively determine the unknown AR-I parameters in such a way that the ranking of the UNDP is reproduced by the enriched BoD models. To this end, we propose a BOP. This provides a useful tool to better understand the relative effects of sub-indicators in calculating the HDI values and manipulating them later. The BOP is explained in the next section.

4 The proposed BOP

In this section, the HDI values presented by the UNDP are considered an acceptable base. It should be noted that the HDI classifications are based on fixed cutoff values. We set up the BoD model (each of the models in Tables 1, 3, 4, and 5) with the AR-I parameters (constraints (1)–(2)) in such a way that it results in the base values. To do so, we determine the unknown bounds of the AR-I constraints by using a BOP, where the upper-level is to minimize the total difference between the enriched BoD scores and the HDI values; the lower-level is a multi-objective optimization problem for evaluating the BoD efficiency scores corresponding to each country subject to that BoD constraints and also the AR-I constraints with variable bounds. In other words, the upper-level optimizes the values of unknown AR-I parameters in the lower-level based on the upper-level objective function, while the lower-level maximizes the efficiency scores under the parameters assigned by the upper-level, subject to the BoD and AR-I constraints.

We want to reproduce the HDI values by the enriched BoD model (each of the models in Tables 1, 3, 4, and 5); so, in the best case, we obtain $Ef_t = HDI_t$ for $t = 1, \ldots, N$. Thus, the upper-level objective function is defined as the minimization of the sum of absolute differences between the

efficiency scores (Ef_t) and the HDI values (HDI_t) for countries $t = 1, \ldots, N$:

$$\min \ z = \sum_{t=1}^{N} |EF_t - HDI_t|. \tag{3}$$

Here, the decision variables of the upper-level are constrained as follows:

$$0 \leq L_1^v \leq U_1^v, \tag{4}$$

$$0 \leq L_2^v \leq U_2^v. \tag{5}$$

It should be noted that if the ranking is considered as the base, then the upper-level objective function is changed to the absolute difference between the two rankings.

The efficiency scores Ef_t, for $(t = 1, 2, \ldots, N)$, are the responses of the lower-level multi-objective optimization problem to the upper-level decision on the AR-I variable bounds. The lower-level problem can be one of the BoD approaches in Tables 1, 3, 4, or 5. We consider the RAM-BoD model in Table 5 to explain the BOP. The objectives of the lower-level, similar to the model in Table 5, are as follows:

$$\max \ Ef_t = v_t + W_{E_t} E_t + W_{H_t} H_t + W_{G_t} G_t, \quad t = 1, 2, \ldots, N. \tag{6}$$

In equations (6), the index t is considered for the aforementioned weights, because the weights of all N countries are included in a single model. The variables v_t, W_{E_t}, W_{H_t}, and W_{G_t} are the decision variables of the lower-level optimization problem, where, similar to the model in Table 5, the weights for a country must be assigned in such a way that the following constraints are satisfied:

$$v_t + W_{E_t} E_j + W_{H_t} H_j + W_{G_t} G_j \leq 1, \quad j = 1, 2, \ldots, N, \ t = 1, \ldots, N, \tag{7}$$

$$W_{E_t} \geq \frac{1}{3R_E}, \qquad t = 1, \ldots, N, \tag{8}$$

$$W_{H_t} \geq \frac{1}{3R_H}, \qquad t = 1, \ldots, N, \tag{9}$$

$$W_{G_t} \geq \frac{1}{3R_G}, \qquad t = 1, \ldots, N. \tag{10}$$

Next, we have to consider the AR-I constraints. The relationships between the decision variables of the two levels of the BOP are expressed by the following inequalities:

$$L_1^v \leq \frac{W_{E_t}}{W_{H_t}} \leq U_1^v, \quad t = 1, \ldots, N, \tag{11}$$

$$L_2^v \leq \frac{W_{E_t}}{W_{G_t}} \leq U_2^v, \quad t = 1, \ldots, N. \tag{12}$$

The constraints (11)–(12) are similar to the AR-I constraints (1)–(2) except that the bounds (L_1^v, U_1^v, L_2^v, and U_2^v) are considered as variables. The combination of equations (3)–(12) is our proposed BOP model for evaluating the bounds of the AR-I constraints in the enriched model in Table 5, to achieve the maximum alignment with the UNDP's HDI. The BOP for the enriched BoD models in Tables 1, 3, and 4 is the same.

By solving the proposed BOP, we obtain the bounds L_1^v, U_1^v, L_2^v, and U_2^v, and also the weights W_{G_t}, W_{H_t}, W_{E_t}, and v_t for $t = 1, \ldots, N$. We replace the bounds of the AR-I constraints in the enriched model in Table 5, namely, L_1, U_1, L_2, and U_2 in constraints (1)–(2), with the optimized variable bounds L_1^v, U_1^v, L_2^v, and U_2^v, respectively. Thus, we obtain the enriched RAM-BoD model with objectively determined bounds. These bounds enable us to analyze the effects of sub-indicators on the evaluation of the HDI values.

Let us note that it is necessary to consider both objective functions (equations (3) and (6)) in a bi-level structure. If our BOP is considered with (3) as the only objective function or the objective functions of the two levels are merged in some way, then applying the optimized AR-I bound variables to the enriched model in Table 5 could produce efficiency scores that are different from the ones obtained from the BOP model and are also so far from the UNDP's HDI values.

The objective functions and the constraints of the BOP are linear or at least convex. So, the proposed BOP model is a convex bi-level multi-objective optimization problem. Therefore, it seems effective to change the lower-level optimization problem into the constraints of the upper-level optimization problem by using the KKT optimality conditions. However, since the lower-level is a multi-objective optimization problem, the number of constraints regarding the KKT conditions is so large that it is even too difficult to find a

feasible solution. Hence, we use a nested approach in which the upper-level is managed by a GA, and the lower-level optimization problems are solved exactly. The methods we apply to solve the proposed BOP are described in the following section.

5 Solving the proposed BOP

In this section, we propose a nested approach to solve our proposed BOP; the upper-level is handled by a GA, which is described in the following subsection; by replacing the optimized upper-level decision variables (the optimized variable bounds of the AR-I constraints), the lower-level problem changes to a linear multi-objective optimization problem, which in turn can be transformed into N linear single-objective optimization problems. In fact, these linear optimization problems are the enriched model in Table 5 for $o = 1, 2, \ldots, N$, and they are solved exactly.

5.1 The GA heuristic

GAs are heuristic search algorithms that can be applied to an extensive range of optimization problems. For the first time, the basic theoretical concepts of GAs were developed by Holland [43] in 1975. These probabilistic search techniques imitate the natural evolution process to find the best or the nearly best solution. A candidate solution is represented as a chromosome, and the optimization problem is formulated as the fitness function. A GA starts with a random population of chromosomes that evolves using nature-inspired operators: selection, crossover, mutation, and replacement. A more exhaustive overview of GAs can be found in [48].

In our GA, we seek the best solution to the upper-level optimization problem (3)–(5). So, we consider a chromosome as an array of length 4 whose numbers show the lower and upper bounds for equations (11)–(12). The population is of the fixed size $P = 20$. By considering the maximum value 5 for the upper bound, each chromosome is randomly initialized in

the interval $[0, 5]$. After producing a chromosome (bounds), the *Evaluate* algorithm is run; the chromosome is checked to be feasible regarding the constraints (4)–(5). If any bound is negative, then its value is fixed to 0, and if the lower bound is greater than its respective upper bound, then it is fixed to be one-half of the value of the upper bound.

Once the feasibility of the solution is ensured, its values are replaced in equations (11)–(12). We have already mentioned that solving the lower-level multi-objective optimization problem (6)–(12) with the fixed AR-I bounds is equivalent to solving the enriched model in Table 5 with those fixed AR-I bounds, for $o = 1, 2, \ldots, N$. So, by using the current solution, we set the AR-I bounds in the enriched model in Table 5 and evaluate Ef_o for $o = 1, 2, \ldots, N$. If it is feasible, then the fitness function (equation (3)) is calculated. However, the determined bound variables may cause infeasibility. We use the GAMS (with CPLEX solver) to solve the enriched model in Table 5. In case of infeasibility, we modify the bound variables in a way that provides wider intervals for the corresponding proportion of weights, as follows:

$$L_1^v = \min_t \frac{W_{E_t}}{W_{H_t}}, \quad U_1^v = \max_t \frac{W_{E_t}}{W_{H_t}},$$
$$L_2^v = \min_t \frac{W_{E_t}}{W_{G_t}}, \quad U_2^v = \max_t \frac{W_{E_t}}{W_{G_t}}.$$

Subsequently, the enriched model in Table 5 with the modified bounds for $o = 1, \ldots, N$ is solved again. If the feasibility is achieved, then the GA proceeds; otherwise, the value of the first cell of the chromosome that is not considered yet is decreased (or increased) by 10% if it shows the lower (or upper) bound. Then, the enriched model in Table 5 with the newly determined bounds is solved. This proportional increase or decrease in the bounds is repeated until the feasibility is achieved.

For any produced solution, the *Evaluate* algorithm is run. Parents are chosen by using a binary tournament; two pools of members are formed, where each one consists of two randomly selected members. The best member of each pool is a parent. The crossover operator randomly takes the bounds of equations (11) and (12) from the two parents to produce a single child. By the probability of 0.3, a child is mutated; a cell in the chromosome is randomly

selected, and its value is changed to its respective upper or lower bound. Also, we propose two methods to locally seek solutions. If the mutation is not done, then the child is subject to the local search (LS); one of the methods is chosen with equal probability. The worst member of the population (considering the objective function (3)) is replaced by the new child. This process is repeated until the best fitness function is not improved within M iterations. Here, we set $M = 10$.

During the running of the GA, the best solutions are saved; we consider the solution with the lowest fitness function during the random initialization of the population, then the produced solutions whose fitness functions are lower than that, enter *BestSet*. After the termination of the GA, *BestSet* is sorted and the first, middle, and last members are selected to be improved by using local search methods. Ultimately, the solution with the best fitness function is selected as the final solution. However, the other members of *BestSet* are possible solutions, one of which may be selected by the expert if it seems to be more sensible.

5.2 The LS methods

To improve a solution to the upper-level optimization problem, we use two LS methods. During the first LS method (LS1), each bound is increased (or decreased) by α percent while the fitness function (calculated by the *Evaluate* algorithm) improves. In the second LS method (LS2), each lower bound is increased (or decreased) together with its respective upper bound by α percent while the fitness function (calculated by the *Evaluate* algorithm) improves. In LS2, the distance between the upper bound and its respective lower bound remains fixed. During the running of the GA, we use the LS methods with $\alpha = 10\%$. After the termination of the GA, LS1, LS2, and then LS1 are run with $\alpha = 10\%$. Next, LS1 and then LS2 are run with $\alpha = 1\%$ for the selected solution of *BestSet*.

6 Numerical results and their analysis

To examine the proposed BOP model, we use the UNDP data of 2019. A summary of the HDI and its sub-indicators (from the UNDP data) for 189 countries is presented in Table 6.

Table 6: Summary of the HDI and its sub-indicators for 189 countries (the UNDP data of 2019)

	Education	Life expectancy	Standards of living	HDI
Mean	0.659	0.811	0.714	0.722
Median	0.682	0.832	0.732	0.740

We solve the proposed BOP model by applying the heuristic GA; the optimized variable bounds of the AR-I constraints for the different enriched BoD models are obtained in Table 7.

Table 7: The optimized bounds of the AR-I constraints for the different enriched BoD models

	L_1^v	U_1^v	L_2^v	U_2^v
Traditional BoD in Table 1	2.9737	2.9934	1.0271	1.0285
Multiplicative BoD in Table 3	2.2988	2.3123	0.9091	0.9146
SBM-BoD in Table 4	1.7715	2.6488	0.8487	2.2152
RAM-BoD in Table 5	0.6985	0.7023	0.9911	1.0019

It is worthwhile to note that all three solutions selected from *BestSet* almost converge to the best solution, presented in Table 7, by using the local search methods. For the traditional enriched BoD model, the weight of the education sub-indicator is 2.9 times that of the health sub-indicator. For the enriched BoD models other than the RAM, the weight of the education sub-indicator is at least 1.8 times that of the health sub-indicator. Except for the enriched SBM-BoD model, the bounds for the other enriched BoD models show that the weights of the education and income sub-indicators are almost equal.

We use these AR-I bounds to enrich the BoD models. The correlation between the rankings obtained from the HDI values and the efficiency scores resulting from the BoD models with and without AR-I constraints are represented in Table 8.

Table 8: The correlation between the rankings from the HDI values and the BoD models

		$+AR-I$
Traditional	0.538	0.997
Multiplicative	0.997	0.997
SBM	0.999	0.999
RAM	0.998	0.998

As evident in Table 9, the strong correlations between BoD models and HDI rankings—excluding the traditional BoD model without AR-I constraints—indicate that BoD models are generally suitable for assessing the human development of countries. Nevertheless, Tables 8 and 9 reveal specific differences among these models in detail.

Table 9 presents a summary of the errors: the absolute values of differences between the efficiency scores (E) and the HDI values are ($|HDI-E|$), and the absolute values of differences between the rankings based on the HDI values (R_{HDI}) and the rankings based on the efficiency scores (R_E) are ($|R_{HDI}-R_E|$). Moreover, the errors of the BoD models without the AR-I constraints are represented.

As mentioned before, the classification of countries is based on the fixed cutoff points of the HDI: low (less than 0.550), medium (0.550–0.699), high (0.700–0.799), and very high (0.800 or greater). So, the objective function of the proposed BOP is to minimize the changes in human development values. To further investigate the results, we represent the number of changes in the classification of countries in comparison with the UNDP's HDI in Table 10.

Regarding Tables 7–10, we discuss some points.

- The fitness function of the BOP is the sum of the values of |HDI-E|. So, the best enriched BoD model to reproduce the basis is the SBM, and the worst one is the RAM. Also, the best and worst errors for

Table 9: A summary of the errors of the BoD models without and with the AR-I constraints

BoD model	Weight restriction		Average	Median	Min	Max	Sum
Traditional		$\lvert HDI - E\rvert$	0.171	0.141	0.041	0.606	32.285
		$\lvert R_E - R_{HDI}\rvert$	39.196	29	0	184	7408
	+AR-I	$\lvert HDI - E\rvert$	0.019	0.016	0	0.050	3.512
		$\lvert R_E - R_{HDI}\rvert$	2.900	2	0	17	548
Multiplicative		$\lvert HDI - E\rvert$	0.105	0.093	0.037	0.260	19.824
		$\lvert R_E - R_{HDI}\rvert$	12.847	10	0	45	2428
	+AR-I	$\lvert HDI - E\rvert$	0.019	0.018	0	0.049	3.683
		$\lvert R_E - R_{HDI}\rvert$	2.847	2	0	17	538
SBM		$\lvert HDI - E\rvert$	0.028	0.0131	0	0.062	5.346
		$\lvert R_E - R_{HDI}\rvert$	1.249	1	0	8	236
	+AR-I	$\lvert HDI - E\rvert$	0.012	0.007	0	0.048	2.186
		$\lvert R_E - R_{HDI}\rvert$	1.661	1	0	7	314
RAM		$\lvert HDI - E\rvert$	0.096	0.079	0	0.289	18.134
		$\lvert R_E - R_{HDI}\rvert$	2.730	2	0	10	516
	+AR-I	$\lvert HDI - E\rvert$	0.095	0.079	0	0.289	17.968
		$\lvert R_E - R_{HDI}\rvert$	2.529	2	0	9	478

misclassification are obtained by the enriched SBM and the RAM with or without AR-I constraints, respectively (Table 10).

- It is evident from Tables 9 and 10 that the enriched traditional and multiplicative BoD models show the most improvements in becoming aligned with the UNDP's HDI in comparison with their original models. This is because of the wide feasibility region in the original BoD model that is limited by the AR-I constraints in an optimal way.

- After considering the AR-I constraints, the errors of misclassification (Table 10) obtained for the SBM-BoD model show notable improvement. However, the changes in the RAM-BoD model are not considerable.

- It seems that applying the AR-I constraints to the SBM-BoD model makes the errors of the human development values slightly better; however, it makes the errors of the rankings rather worse. Although, the changes in the new rankings do decrease misclassifications (Table 10).

- Tables 9–10 show as the results of the BoD model without the AR-I constraints get farther away from the UDNP's HDI, the optimized

upper and lower bounds of the AR-I constraints get closer to each other (in the traditional BoD), and vice versa (in the SBM-BoD).

Table 10: The number of changes in the human development classification by using the BoD models with or without AR-I constraints

Traditional		Multiplicative		SBM		RAM	
	+AR-I		+AR-I		+AR-I		+AR-I
112	12	102	14	27	6	79	79

7 Discussion

The proposed bi-level optimization framework for the DEA model serves as a versatile structure for initializing parameters within the DEA context. While parameters can often be determined through expert opinions, real-world scenarios may demand a more gradual transition from traditional ranking methods to DEA-based approaches. For instance, organizations such as banks might initially possess baseline efficiencies derived from conventional methods and seek a seamless transition to DEA methods. This gradual transition allows for the calculation of subsequent efficiencies with bearable and acceptable changes, ensuring a smooth adoption process without drastic disruptions to existing operations. One important aspect to consider is the potential challenge of translating certain constraints that can be introduced to DEA models into meaningful relationships among inputs, outputs, or DMUs. In such cases, establishing a baseline efficiency scores, if attainable, proves valuable. The computational time required to solve the bi-level DEA model grows as the number of DMUs, inputs, and outputs increases. Nonetheless, as the lower-level optimization is linear and the upper level is managed by a heuristic, the computational complexity remains linear. Furthermore, the bi-level optimization is solved once, and subsequently, using the obtained parameters, the DEA model is employed to assess DMUs. It is important to note that solving the bi-level DEA model is challenging because it falls under NP-hard problems, leading to difficulties with local optima. The effectiveness of the heuristic method in navigating these challenges hinges on thoughtful de-

sign and proper initialization of parameters. Employing a greedy search for parameter initialization is a common strategy that can enhance the model's efficiency and effectiveness.

8 Conclusion

In the 1990s, the UNDP introduced the HDI. This is currently measured as the geometric average of three sub-indicators. One of the drawbacks of this method is the use of equal weights for the three sub-indicators. To resolve this problem, it is suggested that a DEA model be employed, particularly one that is enriched by weight restrictions. In this paper, we considered multiple BoD models enriched by the AR-I constraints. We suggested a novel objective method for determining the parameters associated with the AR-I constraints. The proposed method was based on a BOP, and by considering a basis, it sought the best bounds in such a way that the results of the enriched BoD model were able to reproduce the basis. It should be noted that because of the nature of its sub-indicators, the UNDP's HDI usually shows small changes in comparison with last year's assessment. To maintain this quality for the new method of assessing human development and to make it more acceptable, the UNDP's HDI was fixed as the basis. According to our numerical results, the scores obtained by the enriched SBM-BoD model were similar to those obtained by the UNDP's HDI. Except for the enriched RAM-BoD, which could not be aligned with the UNDP's HDI, the effect of the education sub-indicator on the evaluation of human development was determined to be much more than that of the health sub-indicator.

References

[1] Aida, K., Cooper, W.W., Pastor, J.T. and Sueyoshi, T. *Evaluating water supply services in Japan with RAM: a range-adjusted measure of inefficiency*, Omega, 26(2) (1998), 207–232.

[2] Alirezaee, M., Hajinezhad, E. and Paradi, J.C. *Objective identification of technological returns to scale for data envelopment analysis models*,

Eur. J. Oper. Res. 266(2) (2018), 678–688.

[3] Allen, R., Athanassopoulos, A., Dyson, R.G. and Thanassoulis, E. *Weights restrictions and value judgements in data envelopment analysis: evolution, development and future directions*, Ann. Oper. Res. 73 (1997), 13–34.

[4] Arcelus, F.J., Sharma, B. and Srinivasan, G. *The Human Development Index Adjusted for Efficient Resource Utilization*, in Inequality, Poverty and Well-being, M. McGillivray, Editor., Palgrave Macmillan UK: London, 2006, 177–193.

[5] Banker, R.D., Charnes, A. and Cooper, W.W. *Some models for estimating technical and scale ineffciencies in data envelopment analysis*, Manag. Sci. 30(9) (1984), 1078–1092.

[6] Bard, J.F. and Falk, J.E. *An explicit solution to the multi-level programming problem*, Comput. Oper. Res. 9(1) (1982) 77–100.

[7] Bilbao-Ubillos, J. *Another Approach to Measuring Human Development: The Composite Dynamic Human Development Index*, Soc. Indic. Res. 111(2) (2013), 473–484.

[8] Blancard, S. and Hoarau, J.F. *Optimizing the formulation of the united nations' human development index: an empirical view from data envelopment analysis*, Econ. Bull. 31(1) (2011) 989–1003.

[9] Blancard, S. and Hoarau, J.F. *A new sustainable human development indicator for small island developing states: A reappraisal from data envelopment analysis*, Econ. Model. 30 (2013) 623–635.

[10] Bougnol, M.L., Dulá, J.H., Lins, M.E. and Da Silva, A.M. *Enhancing standard performance practices with DEA*, Omega, 38(1-2) (2010) 33–45.

[11] Boussofiane, A., Dyson, R.G. and Thanassoulis, E. *Applied data envelopment analysis*, Eur. J. Oper. Res. 52(1) (1991), 1–15.

[12] Bracken, J. and McGill, J. *Mathematical programs with optimization problems in the constraints*, Oper. Res. 21 (1973), 37–44.

[13] Camanho, A.S. and D'Inverno, G. *Data Envelopment Analysis: A Review and Synthesis*, In: Macedo, P., Moutinho, V., Madaleno, M. (eds), Advanced Mathematical Methods for Economic Efficiency Analysis, Lecture Notes in Economics and Mathematical Systems, vol 692, Springer, Cham, 2023.

[14] Candler, W. and Norton, R.D. *Multi-level programmin*, World Bank, 20, 1977.

[15] Chansarn, S. *The evaluation of the sustainable human development: A cross-country analysis employing slack-based DEA*, Procedia. Environ. Sci. 20 (2014) 3–11.

[16] Charnes, A., Cooper, W.W., Golany, B., Seiford, L. and Stutz, J. *Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions*, J. Econom. 30(1-2) (1985) 91–107.

[17] Charnes, A., Cooper, W.W. and Rhodes, E. *Measuring the efficiency of decision making unit*, Eur. J. Oper. Res. 2 (1978) 429–444.

[18] Charnes, A., Cooper, W.W., Seiford, L. and Stutz, J. *A multiplicative model for efficiency analysis*, Socio-Econ. Plan. Sci. 16(5) (1982) 223–224.

[19] Charnes, A., Cooper, W.W., Seiford, L. and Stutz, J. *Invariant multiplicative efficiency and piecewise Cobb-Douglas envelopments*, Oper. Res. Lett. 2(3) (1983) 101–103.

[20] Charnes, A., Cooper, W.W., Wei, Q.L. and Huang, Z.M. *Cone ratio data envelopment analysis and multi-objective programming*, Int. J. Syst. Sci. 20(7) (1989) 1099–1118.

[21] Cherchye, L., Moesen, W., Rogge, N. and Puyenbroeck, T.V. *An introduction to 'benefit of the doubt'composite indicators*, Soc. Indic. Res. 82(1) (2007) 111–145.

[22] Colson, B., Marcotte, P. and Savard, G. *Bilevel programming: A survey*, 4or, 3(2) (2005), 87–107.

[23] Cook, W.D. and Seiford, L.M. *Data envelopment analysis (DEA) – Thirty years on*, Eur. J. Oper. Res. 192(1) (2009) 1–17.

[24] Cook, W.D., Tone, K. and Zhu, J. *Data envelopment analysis: Prior to choosing a model*, Omega, 44 (2014), 1–4.

[25] Cooper, W.W., Seiford, L.M., Tone, K. and Zhu, J. *Some models and measures for evaluating performances with DEA: past accomplishments and future prospects*, J. Product. Anal. 28(3) (2007)151–163.

[26] Deb, K. and Sinha, A. *Solving bilevel multi-objective optimization problems using evolutionary algorithms*, in International conference on evolutionary multi-criterion optimization, Springer, 2009.

[27] Despotis, D.K. *A Reassessment of the human development index via data envelopment analysis*, J. Oper. Res. Soc. 56(8) (2005), 969–980.

[28] Despotis, D.K. *Measuring human development via data envelopment analysis: the case of Asia and the Pacific*, Omega, 33(5) (2005), 385–390.

[29] Dyson, R.G., Allen, R., Camanho, A.S., Podinovski, V.V., Sarrico, C.S. and Shale, E.A. *Pitfalls and protocols in DEA*, Eur. J. Oper. Res. 132(2) (2001) 245–259.

[30] Dyson, R.G. and Thanassoulis, E. *Reducing weight flexibility in data envelopment analysis*, J. Oper. Res. Soc. 39(6) (1988) 563–576.

[31] Edirisinghe, N.C.P. and Zhang, X. *Generalized DEA model of fundamental analysis and its application to portfolio optimization*, J. Bank. Finance. 31(11) (2007), 3311–3335.

[32] Emrouznejad, A., Parker, B.R. and Tavares, G. *Evaluation of research in efficiency and productivity: A survey and analysis of the first 30 years of scholarly literature in DEA*, Socio-Econ. Plan. Sci. 42(3) (2008), 151–157.

[33] Emrouznejad, A., Yang, Gl., Khoveyni, M. and Michali, M. *Data Envelopment Analysis: Recent Developments and Challenges*, In: Salhi, S., Boylan, J. (eds), The Palgrave Handbook of Operations Research (2022) 307–350.

[34] Farrell, M.J. *The measurement of productive efficiency*, J. R. Stat. Soc. 120(3) (1957) 253–281.

[35] Fethi, M.D. and Pasiouras, F. *Assessing bank efficiency and performance with operational research and artificial intelligence techniques: A survey*, Eur. J. Oper. Res. 204(2) (2010) 189–198.

[36] Gattoufi, S., Oral, M., Kumar, A. and Reisman, A. *Epistemology of data envelopment analysis and comparison with other fields of OR/MS for relevance to applications*, Socio-Econ. Plan. Sci. 38(2-3) (2004) 123–140.

[37] Gattoufi, S., Oral, M. and Reisman, A. *Data envelopment analysis literature: a bibliography update (1951–2001)*, Socio-Econ. Plan. Sci. 38 (2004), 159–229.

[38] Golany, B. and Roll, Y. *An application procedure for DEA*, Omega, 17(3) (1989), 237–250.

[39] Hashimoto, A. and Ishikawa, H. *Using DEA to evaluate the state of society as measured by multiple social indicators*, Socio-Econ. Plan. Sci. 27(4) (1993) 257–268.

[40] Hatefi, S.M. and Torabi, S.A. *A common weight MCDA–DEA approach to construct composite indicators*, Ecol. Econ. 70(1) (2010) 114–120.

[41] Hatefi, S.M. and Torabi, S.A. *A slack analysis framework for improving composite indicators with applications to human development and sustainable energy indices*, Econom. Rev. 37(3) (2018) 247–259.

[42] Hirai, T. *The human development index and its evolution,* In The Creation of the Human Development Approach, T. Hirai, Editor. Springer International Publishing, Cham., (2017) 73–121.

[43] Holland, J. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control and AI,* The University of Michigan. Ann Arbor, MI, 1975.

[44] Jung, S., Son, J., Kim, C. and Chung, K. *Efficiency Measurement Using Data Envelopment Analysis (DEA) in Public Healthcare: Research Trends from 2017 to 2022.* Processes, 11(3) (2023), 811.

[45] Junior, P.N.A., Mariano, E.B. and Nascimento Rebelatto, D.A. do. *Using data envelopment analysis to construct human development index*, in Emerging Trends in the Development and Application of Composite Indicators. IGI Global, 2017, 298–323.

[46] Kelley, A.C. *The Human Development Index: "Handle with Care"*, Popul. Dev. Rev. 17(2) (1991) 315–324.

[47] Kovacevic, M. *Review of HDI critiques and potential improvements*, Hum. Dev. Res. paper, 33 (2010) 1–44.

[48] Kramer, O. *Genetic algorithms, in Genetic algorithm essentials*, Springer, 2017, 11–19.

[49] Krmac, E. and Mansouri Kaleibar, M. *A comprehensive review of data envelopment analysis (DEA) methodology in port efficiency evaluation*, Marit. Econ. Logist. 25(4) (2023) 817–881.

[50] Kyrgiakos, L.S., Kleftodimos, G., Vlontzos, G. and Pardalos, P.M. *A systematic literature review of data envelopment analysis implementation in agriculture under the prism of sustainability,* Oper. Res. 23(7) (2023) 7.

[51] Lee, H.S., Lin, K. and Fang, H.H. *A Fuzzy Multiple Objective DEA for the Human Development Index*, in Knowledge-Based Intelligent Information and Engineering Systems: 10th International Conference, KES 2006, Bournemouth, UK, October 9-11, 2006. Proceedings, Part II, B. Gabrys, R.J. Howlett, and L.C. Jain, Editors., Springer Berlin Heidelberg: Berlin, Heidelberg. 2006, 922–928.

[52] Liu, J.S., Lu, L.Y. and Lu, W.M. *Research fronts in data envelopment analysis*, Omega, 58 (2016) 33–45.

[53] Liu, J.S., Lu, L.Y., Lu, W.M. and Lin, B.J. *A survey of DEA applications,* Omega, 41(5) (2013) 893–902.

[54] Liu, J.S., Lu, L.Y., Lu, W.M. and Lin, B.J. *Data envelopment analysis 1978–2010: A citation-based literature survey*, Omega, 41(1) (2013) 3–15.

[55] Lovell, C.K., Pastor, J.T. and Turner, J.A. *Measuring macroeconomic performance in the OECD: A comparison of European and non-European countries*, Eur. J. Oper. Res. 87(3) (1995) 507–518.

[56] Lozano, S. and Gutiérrez, E. *Data envelopment analysis of the human development index*, Int. J. Soc. Syst. Sci. 1(2) (2008) 132–150.

[57] Lv, Y., Hu, T., Wang, G. and Wan, Z. *A penalty function method based on Kuhn–Tucker condition for solving linear bilevel programming*, Comput. Appl. Math. 188(1) (2007) 808–813.

[58] Mahlberg, B. and Obersteiner, M. *Remeasuring the HDI by Data Envelopement Analysis*, Available at SSRN 1999372. 2001.

[59] Mahmoudi, R., Emrouznejad, A., Shetab-Boushehri, S.N. and Hejazi, S.R. *The origins, development and future directions of data envelopment analysis approach in transportation systems,* Socio-Econ. Plan. Sci. 69 (2020) 100672.

[60] Mariano, E.B., Ferraz, D. and Oliveira Gobbo, S.C. de. *The Human Development Index with Multiple Data Envelopment Analysis Approaches: A Comparative Evaluation Using Social Network Analysis*, Soc. Indic. Res. (2021) 1–58.

[61] Mariano, E.B., Sobreiro, V.A. and Rebelatto, D.A.d.N. *Human development and data envelopment analysis: A structured literature review*, Omega, 54 (2015), 33–49.

[62] Mejía-de-Dios, J.A., Mezura-Montes, E. and Quiroz-Castellanos, M. *Automated parameter tuning as a bilevel optimization problem solved by a surrogate-assisted population-based approach*, Appl. Intell. 2021, 1–23.

[63] Melyn, W. and Moesen, W. *Towards a synthetic indicator of macroeconomic performance: unequal weighting when limited information is available*, J. Public Econ. 1991, 1–24.

[64] Paradi, J.C. and Zhu, H. *A survey on bank branch efficiency and performance research with data envelopment analysis*, Omega, 41(1) (2013), 61–79.

[65] Pedraja-Chaparro, F., Salinas-Jimenez, J. and Smith, P. *On the role of weight restrictions in data envelopment analysis*, J. Product. Anal. 8(2) (1997), 215–230.

[66] Prasetyo, A.D. and Zuhdi, U. *The government expenditure efficiency towards the human development*, Procedia Econ. Financ. 5 (2013), 615–622.

[67] Ramanathan, R. *Evaluating the comparative performance of countries of the Middle East and North Africa: A DEA application*, Socio-Econ. Plan. Sci. 40(2) (2006) 156–167.

[68] Roghanian, E., Sadjadi, S.J. and Aryanezhad, M.B. *A probabilistic bi-level linear multi-objective programming problem to supply chain planning*, Comput. Appl. Math. 188(1) (2007) 786–800.

[69] Roll, Y., Cook, W.D. and Golany, B. *Controlling factor weights in data envelopment analysis*, IIE Trans. 23(1) (1991), 2–9.

[70] Ruuska, S. and Miettinen, K. *Constructing evolutionary algorithms for bilevel multiobjective optimization*, in 2012 IEEE Congress on Evolutionary Computation, IEEE, 2012.

[71] Schmidt, P. *Frontier production functions*, Econom. Rev. 4(2) (1985), 289–328.

[72] Seiford, L.M. *Data envelopment analysis: The evolution of the state of the art (1978–1995)*, J. Product. Anal. 7 (1996), 99–137.

[73] Seiford, L.M. *A bibliography for data envelopment analysis (1978–1996).*, Ann. Oper. Res. 73 (1997), 393–438.

[74] Shi, C., Lu, J. and Zhang, G. *An extended Kuhn–Tucker approach for linear bilevel programming*, Comput. Appl. Math. 162(1) (2005), 51–63.

[75] Shi, X. and Xia, H.S. *Model and interactive algorithm of bi-level multi-objective decision-making with multiple interconnected decision makers*, Journal of Multi-Criteria Decision Analysis, 10(1) (2001), 27–34.

[76] Sinha, A., Malo, P. and Deb, K. *A review on bilevel optimization: from classical to evolutionary approaches and applications*, IEEE Trans. Evol. Comput. 22(2) (2017) 276–295.

[77] Sinha, A., Malo, P., Frantsev, A. and Deb, K. *Finding optimal strategies in a multi-period multi-leader–follower Stackelberg game using an evolutionary algorithm*, Comput. Oper. Res. 41 (2014), 374–385.

[78] Sinha, A., Malo, P., Xu, P. and Deb, K. *A bilevel optimization approach to automated parameter tuning*, in Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation, 2014.

[79] Song, M., An, Q., Zhang, W., Wang, Z. and Wu, J. *Environmental efficiency evaluation based on data envelopment analysis: A review*, Renew. Sustain. Energy Rev. 16(7) (2012) 4465–4469.

[80] Talbi, E.G. *Metaheuristics for Bi-level Optimization.* Springer Berlin Heidelberg, 2013.

[81] Thanassoulis, E., De Witte, K., Johnes, J., Johnes, G., Karagiannis, G. and Portela, C.S. *Applications of Data Envelopment Analysis in Education. In: Zhu, J. (eds)* Data envelopment analysis: A handbook of empirical studies and applications. International Series in Operations Research & Management Science, (2016), 367–438.

[82] Thanassoulis, E., Portela, M.C. and Allen, R. *Incorporating value judgments in DEA*, Handbook on data envelopment analysis, 2004, 99–138.

[83] Thompson, R.G., Langemeier, L.N., Lee, C.T., Lee, E. and Thrall, R.M. *The role of multiplier bounds in efficiency analysis with application to Kansas farming*, J. Econom. 46(1-2) (1990), 93–108.

[84] Tofallis, C. *Multicriteria ranking using weights which minimize the score range*, In New developments in multiple objective and goal programming 2010 (pp. 133–140). Springer Berlin Heidelberg.

[85] Tofallis, C. *An automatic-democratic approach to weight setting for the new human development index*, J. Popul. Econ. 26(4) (2013), 1325–1345.

[86] Tone, K. *A slacks-based measure of efficiency in data envelopment analysis*, Eur. J. Oper. Res. 130(3) (2001), 498–509.

[87] UNDP, *Human Development Report (1990)*, New York: United Nations Development Programme, Oxford University Press, 1990.

[88] UNDP, *Human Development Report 2010: 20th Anniversary Edition*, Palgrave Macmillan, 2010.

[89] Van Puyenbroeck, T. and Rogge, N. *Comparing regional human development using global frontier difference indices*, Socio-Econ. Plan. Sci. 70 (2020) 100663.

[90] Vicente, L., Savard, G. and Júdice, J. *Descent approaches for quadratic bilevel programming*, J. Optim. Theory Appl. 81(2) (1994), 379–399.

[91] Wong, Y.H. and Beasley, J. *Restricting weight flexibility in data envelopment analysis*, J. Oper. Res. Soc. 1990, 829–835.

[92] Yan, J., Li, L., Zhao, F., Zhang, F. and Zhao, Q. *A multi-level optimization approach for energy-efficient flexible flow shop scheduling*, J. Clean. Prod. 137 (2016), 1543–1552.

[93] Zhou, P., Ang, B.W. and Poh, K.L. *A mathematical programming approach to constructing composite indicators*, Ecol. Econ. 62(2) (2007), 291–297.

[94] Zhou, P., Ang, B.W. and Poh, K.L. *A survey of data envelopment analysis in energy and environmental studies*, Eur. J. Oper. Res. 189(1) (2008), 1–18.

[95] Zhou, P., Ang, B.W. and Zhou, D.Q. *Weighting and Aggregation in Composite Indicator Construction: a Multiplicative Optimization Approach*, Soc. Indic. Res. 96(1) (2010), 169–181.

[96] Zhou, H., Yang, Y., Chen, Y. and Zhu, J. *Data envelopment analysis application in sustainability: The origins, development and future directions*, Eur. J. Oper. Res. 264(1) (2018) 1–16.

[97] Zimmermann, S., Hakimifard, G., Zamora, M., Poranne, R. and Coros, S. *A multi-level optimization framework for simultaneous grasping and motion planning*, IEEE Robot. Autom. Lett. 5(2) (2020) 2966–2972.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 2, 2024, pp 580–612
https://doi.org/10.22067/ijnao.2024.85987.1365
https://ijnao.um.ac.ir/

Research Article

Modeling individual mobility's impact on COVID-19 transmission: Insights from a two-patch SEIR-V approach

M. Bouziane*, M.A. Boubekeur, M.E.B. Keddar and O. Belhamiti

*Corresponding author

Received 20 December 2023; revised 22 February 2024; accepted 11 March 2024

Mehdi Bouziane
Department of Mathematics, Higher Normal School of Mostaganem, Mostaganem 27000, Algeria. e-mail: mehdi.bouziane@univ-mosta.dz

Maroua Amel Boubekeur
Department of Mathematics and Computer Science,
Abdelhamid Ben Badis University, Mostaganem 27000, Algeria. e-mail: maroua.boubekeur.etu@univ-mosta.dz

Meriem El Batoul Keddar
Department of Mathematics and Computer Science,
Abdelhamid Ben Badis University, Mostaganem 27000, Algeria. e-mail: batoulkeddar@gmail.com

Omar Belhamiti
Department of Mathematics and Computer Science,
Abdelhamid Ben Badis University, Mostaganem 27000, Algeria. e-mail: omar.belhamiti@univ-mosta.dz

How to cite this article
Bouziane, M., Boubekeur, M.A., Keddar, M.E.B. and Belhamiti, O., Modeling individual mobility's impact on COVID-19 transmission: Insights from a two-patch SEIR-V approach. *Iran. J. Numer. Anal. Optim.*, 2024; 14(2): 580-612. https://doi.org/10.22067/ijnao.2024.85987.1365

Abstract

This research explores the influence of individual mobility on COVID-19 transmission, utilizing a temporal mathematical model to clarify disease spread and vaccination dynamics across diverse regions. Employing a computationally efficient two-patch configuration that emphasizes regional interactions, our study aims to guide optimal disease control strategies. The introduced SEIR-V model with a two-patch setup estimates the vaccination reproduction number, R_v, while equilibrium points and system stability are identified. Visualizations from numerical simulations and sensitivity analyses illustrate key parameters affecting the vaccination reproduction number and COVID-19 control measures. Our findings underscore system responsiveness, emphasizing the intricate relationship between R_v, migration rates, and disease prevalence.

AMS subject classifications (2020): Primary 39A12; Secondary 92C60, 92D30.

Keywords: COVID-19; Metapopulation; Global stability; Local stability; Vaccination reproduction number.

1 Introduction

In December 2019, a coronavirus emerged in Wuhan, China, transmitted from animals to humans [22]. The outbreak escalated rapidly, with 44 reported cases in China by January 3, 2020 [29]. The virus swiftly spread nationwide, prompting the declaration of a state of epidemic in March 2020, exacerbated by global human migration and travel between cities [18, 20]. The facilitation of worldwide travel and commerce contributed to the rapid global dissemination, resulting in over 585 million reported cases and 6.4 million reported deaths as of August 2022 [28]. Consequently, a multitude of researchers have been captivated by the dynamics of COVID-19 dissemination [16, 6, 7, 9, 12].

The investigation focused on the movement patterns of individuals across a network during the global spread of COVID-19 in 203 countries [13]. Researchers observed that human mobility, particularly through cities and tourism, significantly contributed to the virus's dissemination, supported by Arino and others who affirmed that epidemic spread increases during trans-

portation [2]. Numerous epidemiologists, including Baister et al. [4], Ahmad et al. [1], McCarthy et al. [21], and Iyaniwura et al. [15] employed metapopulation models to explore COVID-19 transmission dynamics. These models considered diverse scenarios, such as studying individual mobility and social interactions in major cities, investigating vaccination strategies in a prison setting, and using hybrid gravity-metapopulation modeling to analyze the spread between different regions. Other studies delved into specific regions like Ireland and formulated generalized n-patch SEIR epidemiological models to devise effective control strategies against the disease [14, 19].

This paper introduces metapopulation modeling to investigate the impact of individual mobility on the spread of COVID-19, building upon our previous model [7]. Our focus is on developing a temporal mathematical model to describe COVID-19 transmission with vaccination, aiming to optimize disease control strategies within and between different regions. Using a two-patch configuration ($n = 2$), we address the complexities of calculations, ensuring the boundedness and positivity of solutions. We calculate the vaccination reproduction number (R_v) for the two-patch metapopulation and analyze the local and global stability of the disease-free equilibrium (DFE), establishing that if $R_v > 1$, then the disease becomes endemic. Sensitivity analysis and numerical simulations further explore the impact of key parameters on Rv, confirming the existence and stability of the endemic equilibrium. Additionally, we examine the influence of migration rates on COVID-19 transmission between the two patches.

The paper's structure unfolds as follows: Section 2 revisits the previously introduced and analyzed model. In Section 3, we introduce the metapopulation model. Section 4 proposes the SEIR-V model with a two-patch configuration for COVID-19, estimating the vaccination reproduction number (R_v). We identify equilibrium points and conduct stability analysis for the dynamic system of COVID-19 propagation. Section 5 presents numerical simulations and sensitivity analyses to illustrate theoretical results. Finally, Section 6 encapsulates our conclusions.

2 Baseline model

We first recall the model introduced and analyzed in [7], which describes the transmission dynamics of COVID-19. In this model, the population is subdivided into seven distinct groups: susceptible individuals (S), exposed but not yet infectious individuals (E), vaccinated individuals (V), asymptomatic infectious individuals (I_a), symptomatic and hospitalized infectious individuals (I_s), individuals who have succumbed to the disease (D), and those who have recovered (R).

It is assumed that there is no vertical transmission of the disease so that all births occur into the susceptible human class, at the rate $\mu > 0$. Moreover, we assume that the total human population N remains constant; thus, birth and death rates are equal.

Hence, the normalized reduced system is given by

$$\begin{cases} S'(t) = \mu - (\alpha_1\eta + \alpha_2(1-\eta))\, S(t)\,(I_a(t) + I_s(t)) - \mu S(t) \\ \qquad\qquad +\lambda_2 V(t) + \gamma R(t) - \lambda_1 S(t), \\ E'(t) = (\alpha_1\eta + \alpha_2(1-\eta))\, S(t)\,(I_a(t) + I_s(t)) - (\beta_1 + \beta_2 + \mu + \lambda_3)\, E(t), \\ V'(t) = \lambda_1 S(t) + \lambda_3 E(t) + \lambda_4 I_a(t) - (\lambda_2 + \lambda_5 + \mu)\, V(t), \\ I_a'(t) = \beta_1 E(t) - (\gamma_1 + \lambda_4 + \mu)\, I_a(t), \\ I_s'(t) = \beta_2 E(t) - (\gamma_2 + \delta + \mu)\, I_s(t), \\ R'(t) = \gamma_1 I_a(t) + \gamma_2 I_s(t) + \lambda_5 V(t) - \gamma R(t) - \mu R(t), \end{cases} \tag{1}$$

where

α_1 is the infection rate of confined susceptible and α_2 is the infection rate of unconfined susceptible such as $\alpha_1 < \alpha_2$.

η is the confinement rate within the population.

$(\beta_i)_{i=1,2}$ are infected-infectious rates.

γ is the reinfection rate after recovery from a first infection.

$(\gamma_i)_{i=1,2}$ are the recovery rates.

$(\lambda_i)_{i=1,3,4}$ are the vaccination rates.

λ_2 is the rate of vaccine ineffectiveness.

λ_5 is the vaccine effectiveness rate.

δ is the COVID-19 mortality rate.

μ is the natality rate.

The study of the model is detailed in [7]. For this model (1), the vaccination reproduction number is given by

$$R_v = \frac{(\alpha_1\eta + \alpha_2(1-\eta))}{(\gamma_1+\lambda_4+\mu)(\gamma_2+\delta+\mu)} \times \frac{(\beta_1(\gamma_2+\delta+\mu)+\beta_2(\gamma_1+\lambda_4+\mu))(\lambda_2+\lambda_5+\mu)(\gamma+\mu)}{(\beta_1+\beta_2+\mu+\lambda_3)((\lambda_2+\lambda_5+\mu)(\gamma+\mu)+\lambda_1(\gamma+\lambda_5+\mu))}. \tag{2}$$

3 Metapopulation model

In works [4, 1, 21, 15, 14, 2, 19, 3], a comprehensive system of differential equations is developed to characterize human mobility. Within this model, subpopulations are identified based on their origin and their current location. Building upon the model detailed in the preceding section, we expand it by introducing the concept of a neighborhood wherein interactions occur between infected and non-infected individuals.

Consider a network comprising p nodes. Within this network, individuals are characterized by two key attributes: the node of origin, signifying their residence, and the node they currently occupy at time t. We assume that the total human population N_i in each node remains constant and strictly positive, and the transition rates from one pathological state to another also remain the same for all patches. Furthermore, we suppose that birth occurs in the resident node while deaths take place in any node where the human is present.

The population in the patch i, is divided into compartments of susceptible (S_i), exposed (E_i), vaccinated(V_i), asymptomatic infectious $(I_{a,i})$, symptomatic and hospitalized infectious $(I_{s,i})$, death (D_i), and the recovered as (R_i).

The total number of individuals in the patch i size at time t is defined by

$$N_i = S_i(t) + E_i(t) + V_i(t) + I_{a,i}(t) + I_{s,i}(t) + R_i(t) + D_i(t),$$

and the total population is given by

$$N = \sum_{i=1}^{p} N_i.$$

We adopt a methodology akin to that employed by Arino and van Den Driessche [3]. Migration is observed between any pair of patches, transpiring at the specified rates:

- φ_{ij}^{S} migration rate of susceptible individuals from the patch i to the patch j,
- φ_{ij}^{E} migration rate of infected individuals (noninfectious) from the patch i to the patch j,
- φ_{ij}^{V} migration rate of vaccination individuals from the patch i to the patch j,
- $\varphi_{ij}^{I_a}$ migration rate of asymptomatic infectious individuals from the patch i to the patch j,
- $\varphi_{ij}^{I_s}$ migration rate of symptomatic infectious individuals from the patch i to the patch j,
- φ_{ij}^{R} migration rate of recovered and immunized infectious individuals from the patch i to the patch j,

where

$$\varphi_{ii}^{S} = \varphi_{ii}^{E} = \varphi_{ii}^{V} = \varphi_{ii}^{I_a} = \varphi_{ii}^{I_s} = \varphi_{ii}^{R} = 0.$$

Hence, the metapopulation system of differential equations that will model the dynamics of coronavirus spread is the system given by

$$
\begin{cases}
\dfrac{dS_i}{dt} = \mu \frac{N_i}{N} - (\alpha_1 \eta_i + \alpha_2 (1-\eta_i)) (I_{a,i} + I_{s,i}) S_i + \sum\limits_{j=1(j\neq i)}^{p} \varphi_{ji}^{S} S_j \\
\qquad\qquad - \left(\sum\limits_{j=1(j\neq i)}^{p} \varphi_{ij}^{S} \right) S_i - (\mu + \lambda_1) S_i + \lambda_2 V_i + \gamma R_i, \\
\dfrac{dE_i}{dt} = (\alpha_1 \eta_i + \alpha_2 (1-\eta_i)) (I_{a,i} + I_{s,i}) S_i + \sum\limits_{j=1(j\neq i)}^{p} \varphi_{ji}^{E} E_j \\
\qquad\qquad - \left(\sum\limits_{j=1(j\neq i)}^{p} \varphi_{ij}^{E} \right) E_i - (\beta_1 + \beta_2 + \mu + \lambda_3) E_i, \\
\dfrac{dV_i}{dt} = \lambda_1 S_i + \lambda_3 E_i + \lambda_4 I_{a,i} + \sum\limits_{j=1(j\neq i)}^{p} \varphi_{ji}^{V} V_j \\
\qquad\qquad - \left(\sum\limits_{j=1(j\neq i)}^{p} \varphi_{ij}^{V} \right) V_i - (\lambda_2 + \lambda_5 + \mu) V_i, \\
\dfrac{dI_{a,i}}{dt} = \beta_1 E_i + \sum\limits_{j=1(j\neq i)}^{p} \varphi_{ji}^{I_a} I_{a,j} - \left(\sum\limits_{j=1(j\neq i)}^{p} \varphi_{ij}^{I_a} \right) I_{a,i} - (\gamma_1 + \lambda_4 + \mu) I_{a,i}, \\
\dfrac{dI_{s,i}}{dt} = \beta_2 E_i - (\gamma_2 + \delta + \mu) I_{s,i}, \\
\dfrac{dR_i}{dt} = \gamma_1 I_{a,i} + \gamma_2 I_{s,i} + \lambda_5 V_i + \sum\limits_{j=1(j\neq i)}^{p} \varphi_{ji}^{R} R_j \\
\qquad\qquad - \left(\sum\limits_{j=1(j\neq i)}^{p} \varphi_{ij}^{R} \right) R_i - (\gamma + \mu) R_i, \\
\dfrac{dD_i}{dt} = \delta I_{s,i},
\end{cases}
$$

with initial conditions

$$
\begin{cases}
S_i(0) > 0, & E_i(0) \geqslant 0, & \sum_{i=1}^{p} E_i(0) > 0, \\
V_i(0) > 0, & I_{a,i}(0) \geqslant 0, & \sum_{i=1}^{p} I_{a,i}(0) > 0, \\
I_{s,i}(0) \geqslant 0, & \sum_{i=1}^{p} I_{s,i}(0) > 0, & R_i(0) > 0,
\end{cases}
$$

for $i = \overline{1,p}$.

4 Two-patch SEIR-V model for COVID-19

To enhance the realism of our study, we confine our investigation to a two-patch model, undertaking a rigorous mathematical analysis of its dynamics.

The diagram of the transmission of COVID-19 in two-patch is shown in Figure 1.

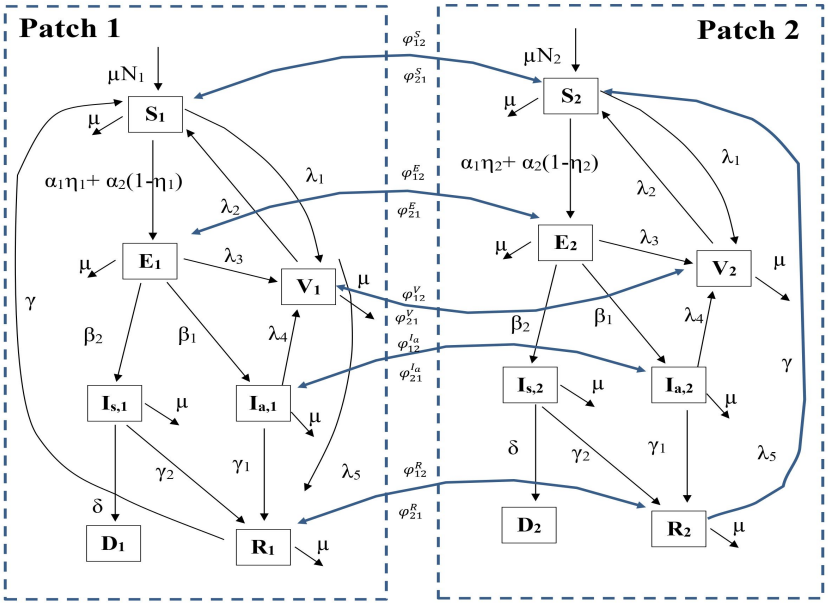


Figure 1: Diagram of the proposed two-Patch SEIR-V Model.

The SEIR-V two-Patch model is formulated as follows:

$$\begin{cases} \dfrac{dS_1}{dt} = \mu n_1 - c_1 \left(I_{a,1} + I_{s,1}\right) S_1 + \varphi_{21}^S S_2 - \left(\varphi_{12}^S + c_0\right) S_1 + \lambda_2 V_1 + \gamma R_1, \\ \dfrac{dS_2}{dt} = \mu n_2 - c_2 \left(I_{a,2} + I_{s,2}\right) S_2 + \varphi_{12}^S S_1 - \left(\varphi_{21}^S + c_0\right) S_2 + \lambda_2 V_2 + \gamma R_2, \\ \dfrac{dE_1}{dt} = c_1 \left(I_{a,1} + I_{s,1}\right) S_1 + \varphi_{21}^E E_2 - \left(\varphi_{12}^E + c_3\right) E_1, \\ \dfrac{dE_2}{dt} = c_2 \left(I_{a,2} + I_{s,2}\right) S_2 + \varphi_{12}^E E_1 - \left(\varphi_{21}^E + c_3\right) E_2, \\ \dfrac{dV_1}{dt} = \lambda_1 S_1 + \lambda_3 E_1 + \lambda_4 I_{a,1} + \varphi_{21}^V V_2 - \left(\varphi_{12}^V + c_4\right) V_1, \\ \dfrac{dV_2}{dt} = \lambda_1 S_2 + \lambda_3 E_2 + \lambda_4 I_{a,2} + \varphi_{12}^V V_1 - \left(\varphi_{21}^V + c_4\right) V_2, \\ \dfrac{dI_{a,1}}{dt} = \beta_1 E_1 + \varphi_{21}^{I_a} I_{a,2} - \left(\varphi_{12}^{I_a} + c_5\right) I_{a,1}, \\ \dfrac{dI_{a,2}}{dt} = \beta_1 E_2 + \varphi_{12}^{I_a} I_{a,1} - \left(\varphi_{21}^{I_a} + c_5\right) I_{a,2}, \\ \dfrac{dI_{s,1}}{dt} = \beta_2 E_1 - c_6 I_{s,1}, \\ \dfrac{dI_{s,2}}{dt} = \beta_2 E_2 - c_6 I_{s,2}, \\ \dfrac{dR_1}{dt} = \gamma_1 I_{a,1} + \gamma_2 I_{s,1} + \lambda_5 V_1 + \varphi_{21}^R R_2 - \left(\varphi_{12}^R + c_7\right) R_1, \\ \dfrac{dR_2}{dt} = \gamma_1 I_{a,2} + \gamma_2 I_{s,2} + \lambda_5 V_2 + \varphi_{12}^R R_1 - \left(\varphi_{21}^R + c_7\right) R_2, \end{cases} \tag{3}$$

with initial conditions

$$\begin{cases} S_i\left(0\right) > 0, \\ E_i\left(0\right) \geqslant 0, \ E_1\left(0\right) + E_2\left(0\right) > 0, \\ V_i\left(0\right) > 0, \\ I_{a,i}\left(0\right) \geqslant 0, \ I_{a,1}\left(0\right) + I_{a,2}\left(0\right) > 0, \\ I_{s,i}\left(0\right) \geqslant 0, \ I_{s,1}\left(0\right) + I_{s,2}\left(0\right) > 0, \\ R_i\left(0\right) > 0, \end{cases} \tag{4}$$

for $i = \overline{1,2}$. Here

$$\begin{cases} n_1 = \frac{N_1}{N}, n_2 = \frac{N_2}{N}, \\ c_0 = (\mu + \lambda_1), \\ c_1 = (\alpha_1\eta_1 + \alpha_2(1-\eta_1)), \\ c_2 = (\alpha_1\eta_2 + \alpha_2(1-\eta_2)), \\ c_3 = \beta_1 + \beta_2 + \mu + \lambda_3, \end{cases} \quad \begin{cases} c_4 = (\lambda_2 + \lambda_5 + \mu), \\ c_5 = \gamma_1 + \lambda_4 + \mu, \\ c_6 = \gamma_2 + \delta + \mu, \\ c_7 = (\gamma + \mu). \end{cases} \tag{5}$$

4.1 Boundedness and positivity of solutions

We consider the following zone of biological interest:

$$\Omega = \Big\{(S_1, S_2, E_1, E_2, V_1, V_2, I_{a,1}, I_{a,2}, I_{s,1}, I_{s,2}, R_1, R_2) \in R_+^{12} : \tag{6}$$
$$0 \leqslant \sum_i S_i + \sum_i E_i + \sum_i V_i + \sum_i I_{a,i} + \sum_i I_{s,i} + \sum_i R_i \leqslant 1, i = \overline{1,2}\Big\}.$$

Theorem 1. Consider system (3) with initial conditions (4).

The set Ω is an attracting and positively invariant with respect to model (3), with S_i, V_i, R_i $\left(i = \overline{1,2}\right)$; remaining positive. The total population in each patch is assumed to be constant, and all their pathological states are also bounded.

Proof. Under initial conditions (4):

- In the event that E_1 becomes zero at t_1 before E_2 becomes zero, then at t_1, we have

$$\frac{dE_1}{dt} = c_1(I_{a,1} + I_{s,1})S_1 + \varphi_{21}^E E_2 \geqslant 0.$$

  This demonstrates that E_1 is a nondecreasing function of t at t_1. Hence, E_1 stays nonnegative. Analogously, the same holds for E_2.

- In the event that $I_{a,1}$ becomes zero at some time t_2 before $I_{a,2}$ becomes zero, then at t_2, we have

$$\frac{dI_{a,1}}{dt} = \beta_1 E_1 + \varphi_{21}^{I_a} I_{a,2} \geqslant 0.$$

  This demonstrates that $I_{a,1}$ is a nondecreasing function of t at t_2. Hence, $I_{a,1}$ stays nonnegative. Analogously, the same holds for $I_{a,2}$.

- In the event that $I_{s,1}$ becomes zero at some time t_3 before $I_{s,2}$ becomes zero, then at t_3, we have

$$\frac{dI_{s,1}}{dt} = \beta_2 E_1 \geqslant 0.$$

This demonstrates that $I_{s,1}$ is a nondecreasing function of t at t_3. Hence, $I_{s,1}$ stays nonnegative. Analogously, the same holds for $I_{s,2}$.

- Suppose now that at some time t_4, $S_1(t_4) = 0$ before S_2 goes to zero. Then at $t = t_4$, from system (3), we have

$$\frac{dS_1}{dt} = \mu\frac{N_1}{N} + \varphi_{21}^S S_2 + \lambda_2 V_1 + \gamma R_1 > 0,$$

which implies that $\dfrac{dS_1}{dt} > 0$ when $\dfrac{N_1}{N}$ is strictly positive.

Thus, there is no time t_4 such that $S_1(t_4) = 0$. Therefore, S_1 stays positive for $t > 0$ when the initial condition $S_1(0) > 0$. Using comparable reasoning, we deduce the positivity of S_2.

- Likewise at some time t_5, $V_1(t_5) = 0$ before V_2 goes to zero. Then at $t = t_5$, from system (3), we have

$$\frac{dV_1}{dt} = \lambda_1 S_1 + \lambda_3 E_1 + \lambda_4 I_{a,1} + \varphi_{21}^V V_2 > 0,$$

which implies that $\dfrac{dV_1}{dt} > 0$ because $S_1(t) > 0$ for $t > 0$.

Thus, there is no time t_5 such that $V_1(t_5) = 0$. Therefore, V_1 stays positive. Using comparable reasoning, we deduce the positivity of V_2.

- Likewise at some time t_6; $R_1(t_6)$ before R_2 goes to zero. Then at $t = t_6$, from system (3), we have

$$\frac{dR_1}{dt} = \gamma_1 I_{a,1} + \gamma_2 I_{s,1} + \lambda_5 V_1 + \varphi_{21}^R R_2 > 0,$$

which implies that $\dfrac{dR_1}{dt} > 0$ because $V_1(t) > 0$ for $t > 0$.

Thus, there is no time t_6 such that $R_1(t_6) = 0$. Therefore, R_1 stays positive. Using comparable reasoning, we deduce the positivity of R_2.

Since the positive set Ω is invariant under system (3) and the total population in each patch $(N_i)_{i=1,2}$ is assumed to be constant, and all their pathological states are also bounded. □

4.2 Disease-Free Equilibrium

The DFE (disease-free equilibrium) of system (3) is expressed as

$$\mathcal{E}_0^* = \left(S_1^* \ S_2^* \ 0 \ 0 \ V_1^* \ V_2^* \ 0 \ 0 \ 0 \ 0 \ R_1^* \ R_2^* \right)^T, \tag{7}$$

where

$$\begin{cases} S_1^* = \dfrac{\mu k_2 k_3 \left(n_1 \left(d_1 + d_3\right) + d_1 n_2\right)}{d_3 \left(d_3 + d_1 + d_2\right)}, \\ S_2^* = \dfrac{\mu k_2 k_3 \left(d_2 n_1 + \left(d_2 + d_3\right) n_2\right)}{d_3 \left(d_3 + d_1 + d_2\right)}, \\ V_1^* = \dfrac{\lambda_1 \left(\varphi_{21}^V + c_4\right) \mu k_2 k_3 \left(n_1 \left(d_1 + d_3\right) + d_1 n_2\right)}{k_2 d_3 \left(d_3 + d_1 + d_2\right)} \\ \qquad + \dfrac{\lambda_1 \varphi_{21k_2}^V \mu k_2 k_3 \left(d_2 n_1 + \left(d_2 + d_3\right) n_2\right)}{d_3 \left(d_3 + d_1 + d_2\right)}, \\ V_2^* = \dfrac{\lambda_1 \varphi_{12}^V \mu k_2 k_3 \left(n_1 \left(d_1 + d_3\right) + d_1 n_2\right)}{k_2 d_3 \left(d_3 + d_1 + d_2\right)} \\ \qquad + \dfrac{\lambda_1 \left(\varphi_{12}^V + c_4\right) \mu k_2 k_3 \left(d_2 n_1 + \left(d_2 + d_3\right) n_2\right)}{k_2 d_3 \left(d_3 + d_1 + d_2\right)}, \\ R_1^* = \dfrac{\lambda_5 \lambda_1 k_4 \mu k_2 k_3 \left(n_1 \left(d_1 + d_3\right) + d_1 n_2\right)}{k_2 k_3 d_3 \left(d_3 + d_1 + d_2\right)} \\ \qquad + \dfrac{\mu k_2 k_3 \lambda_5 \lambda_1 \left(\varphi_{21}^V k_3 + \varphi_{21}^R k_5\right) \left(d_2 n_1 + \left(d_2 + d_3\right) n_2\right)}{k_2 k_3 d_3 \left(\varphi_{12}^R + c_7\right) \left(d_3 + d_1 + d_2\right)}, \\ R_2^* = \dfrac{\mu k_2 k_3 \lambda_1 \lambda_5 \left(\varphi_{12}^V k_3 + \varphi_{12}^R k_4\right) \left(n_1 \left(d_1 + d_3\right) + d_1 n_2\right)}{k_2 k_3 d_3 \left(\varphi_{21}^R + c_7\right) \left(d_3 + d_1 + d_2\right)} \\ \qquad + \dfrac{\lambda_5 \lambda_1 k_5 \mu k_2 k_3 \left(d_2 n_1 + \left(d_2 + d_3\right) n_2\right)}{k_2 k_3 d_3 \left(d_3 + d_1 + d_2\right)}, \end{cases}$$

and

$$\begin{cases} d_1 = \gamma \lambda_1 \lambda_5 k_6 + \lambda_1 \lambda_2 k_3 \varphi_{21}^V + k_3 k_2 \varphi_{21}^S, \\ d_2 = \gamma \lambda_1 \lambda_5 k_7 + \lambda_1 \lambda_2 k_3 \varphi_{12}^V + k_2 k_3 \varphi_{12}^S, \\ d_3 = k_3 k_2 \mu + \lambda_1 \mu \left(\lambda_5 + c_7\right) \left(k_6 + k_7 + c_4 c_7\right), \end{cases} \tag{8}$$

and

$$\begin{cases} k_2 = c_4^2 + \varphi_{12}^V c_4 + \varphi_{21}^V c_4, \\ k_3 = c_7^2 + \varphi_{12}^R c_7 + \varphi_{21}^R c_7, \\ k_4 = \left(\varphi_{21}^V + c_4\right)\left(\varphi_{21}^R + c_7\right) + \varphi_{12}^V \varphi_{21}^R, \\ k_5 = \left(\varphi_{12}^V + c_4\right)\left(\varphi_{12}^R + c_7\right) + \varphi_{12}^R \varphi_{21}^V, \\ k_6 = \varphi_{21}^V\left(\varphi_{21}^R + c_7\right) + \varphi_{21}^R\left(\varphi_{12}^V + c_4\right), \\ k_7 = \varphi_{12}^V\left(\varphi_{12}^R + c_7\right) + \varphi_{12}^R\left(\varphi_{21}^V + c_4\right). \end{cases} \tag{9}$$

The existence of the DFE $\mathcal{E}_0^*$ remains unaffected by any prerequisites regarding positive operational data.

4.3 Vaccination reproduction number R_v of two-patch metapopulation

In this subsection, we present the method used for our SEIR-V model to estimate the vaccination reproduction number R_v. This method is proposed by Diekmann, Heesterbeek, and Metz [26] and elaborated by Van Den Driessche and Watmough [27, 25], which gives a way of determining R_v for an ordinary differential equation compartmental model by using the next-generation matrix.

Let $X = (E_1, E_2, I_{a,1}, I_{a,2}, I_{s,1}, I_{s,2})^T$. Then the system can be written as

$$\frac{dX}{dt} = \mathcal{F}(X) - \mathcal{V}(X),$$

where

$$\mathcal{F}(X) = \begin{pmatrix} c_1\left(I_{a,1} + I_{s,1}\right) S_1 \\ c_2\left(I_{a,2} + I_{s,2}\right) S_2 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

and

$$\mathcal{V}(X) = \begin{pmatrix} -\varphi_{21}^{E} E_2 + \left(\varphi_{12}^{E} + c_3\right) E_1 \\ -\varphi_{12}^{E} E_1 + \left(\varphi_{21}^{E} + c_3\right) E_2 \\ -\beta_1 E_1 - \varphi_{21}^{I_a} I_{a,2} + \left(\varphi_{12}^{I_a} + c_5\right) I_{a,1} \\ -\beta_1 E_2 - \varphi_{12}^{I_a} I_{a,1} + \left(\varphi_{21}^{I_a} + c_5\right) I_{a,2} \\ -\beta_2 E_1 + c_6 I_{s,1} \\ -\beta_2 E_2 + c_6 I_{s,2} \end{pmatrix}.$$

The basic reproduction number, R_v, is calculated by next generation technique. The F and V matrices at the DFE $\mathcal{E}_0^*$ is given as follows:

$$F = \left.\frac{\partial \mathcal{F}(X)}{\partial X}\right|_{\mathcal{E}_0^*} = \begin{pmatrix} 0 & 0 & c_1 S_1^* & 0 & c_1 S_1^* & 0 \\ 0 & 0 & 0 & c_2 S_2^* & 0 & c_2 S_2^* \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

and

$$V = \left.\frac{\partial \mathcal{V}(X)}{\partial X}\right|_{\mathcal{E}_0^*} = \begin{pmatrix} \varphi_{12}^{E} + c_3 & -\varphi_{21}^{E} & 0 & 0 & 0 & 0 \\ -\varphi_{12}^{E} & \varphi_{21}^{E} + c_3 & 0 & 0 & 0 & 0 \\ -\beta_1 & 0 & \varphi_{12}^{I_a} + c_5 & -\varphi_{21}^{I_a} & 0 & 0 \\ 0 & -\beta_1 & -\varphi_{12}^{I_a} & \varphi_{21}^{I_a} + c_5 & 0 & 0 \\ -\beta_2 & 0 & 0 & 0 & c_6 & 0 \\ 0 & -\beta_2 & 0 & 0 & 0 & c_6 \end{pmatrix}.$$

Therefore, the next generation matrix is

$$FV^{-1} = \begin{pmatrix} m_1 S_1^* & m_3 S_1^* & \frac{c_1\left(c_5+\varphi_{21}^{I_a}\right)}{k_1} S_1^* & \frac{c_1 \varphi_{21}^{I_a}}{k_1} S_1^* & \frac{c_1}{c_6} S_1^* & 0 \\ m_2 S_2^* & m_4 S_2^* & \frac{c_2 \varphi_{12}^{I_a}}{k_1} S_2^* & \frac{c_2\left(c_5+\varphi_{12}^{I_a}\right)}{k_1} S_2^* & 0 & \frac{c_2}{c_6} S_2^* \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

where

$$
\begin{aligned}
m_1 &= c_1 \left(\frac{\left(c_3 + \varphi_{21}^{E}\right)}{k_0} \left(\frac{\beta_2}{c_6} + \frac{c_5 \beta_1}{k_1} \right) + \frac{\beta_1 \varphi_{21}^{I_a}}{c_3 k_1} \right), \\
m_2 &= c_2 \left(\frac{\varphi_{12}^{E}}{k_0} \left(\frac{\beta_2}{c_6} + \frac{c_5 \beta_1}{k_1} \right) + \frac{\beta_1 \varphi_{12}^{I_a}}{c_3 k_1} \right), \\
m_3 &= c_1 \left(\frac{\varphi_{21}^{E}}{k_0} \left(\frac{\beta_2}{c_6} + \frac{c_5 \beta_1}{k_1} \right) + \frac{\beta_1 \varphi_{21}^{I_a}}{c_3 k_1} \right), \\
m_4 &= c_2 \left(\frac{\left(c_3 + \varphi_{12}^{E}\right)}{k_0} \left(\frac{\beta_2}{c_6} + \frac{c_5 \beta_1}{k_1} \right) + \frac{\beta_1 \varphi_{12}^{I_a}}{c_3 k_1} \right),
\end{aligned}
$$

and

$$
\begin{aligned}
k_0 &= c_3 \varphi_{12}^{E} + c_3 \varphi_{21}^{E} + c_3^2, \\
k_1 &= c_5 \varphi_{12}^{I_a} + c_5 \varphi_{21}^{I_a} + c_5^2.
\end{aligned}
$$

The reproduction number, R_v, is the spectral radius of the next generation matrix, which is given by

$$
\begin{aligned}
R_v = \frac{1}{2} \frac{\mu k_2 k_3}{d_3 \left(d_3 + d_1 + d_2\right)} \Big(& \left(d_1 + n_1 d_3\right) m_1 + \left(d_2 + n_2 d_3\right) m_4 \\
& + \sqrt{\left(\left(d_1 + n_1 d_3\right) m_1 - \left(d_2 + n_2 d_3\right) m_4\right)^2 + 4 m_2 m_3 \left(d_1 + n_1 d_3\right) \left(d_2 + n_2 d_3\right)} \Big),
\end{aligned}
$$

where (c_i) and (d_i) are defined in (5) and (8), respectively.

4.4 Model analysis

Theorem 2. If $R_v < 1$, then the DFE $\mathcal{E}_0^*$ is locally asymptotically stable and unstable otherwise.

Proof. System (3) has a unique DFE point $\mathcal{E}_0^*$ in the set Ω, given in (7).

The eigenvalues of the Jacobian matrix at the DFE point $\mathcal{E}_0^*$ are given by

$$
\begin{cases}
\rho_1 = \rho_2 = -c_6, \\
\rho_3 = -\left(\varphi_{12}^{E} + \varphi_{21}^{E} + c_3\right), \\
\rho_4 = -c_3, \\
\rho_5 = -\left(c_5 + \varphi_{12}^{I_a} + \varphi_{21}^{I_a}\right), \\
\rho_6 = -c_5,
\end{cases}
$$

where $(\rho_i)_{i=7,12}$ are roots of

$$P(Z) = Z^6 + C_5Z^5 + C_4Z^4 + C_3Z^3 + C_2Z^2 + C_1Z + C_0.$$

By using Lienard–Chipart criteria, the DFE point $\mathcal{E}_0^*$ is locally asymptotically stable if the coefficients C_5, C_4, C_3, C_2, C_1, and C_0 are positive and their Hurwitz's determinants are positive. It shows that the positivity of Hurwitz determinants is very complicated to achieve. However, the computational cost can be reduced by using the simplified version mentioned in [5, 17]. That is, the coefficients that must be computed are C_5, C_4, C_3, C_2, C_1, and C_0 only. Note that C_5, C_4, and C_3 are needed because the simplified version of Lienard–Chipart criteria still require $|H_2| = C_5C_4 - C_3 > 0$ [11].

Indeed, we employed formal calculation software (Maxima Version 5.42.1) to establish that if $R_v < 1$, then all $(C_i)_{i=\overline{0,5}}$ and $|H_2|$ are positive. □

Theorem 3. If $R_v < 1$, then the the DFE $\mathcal{E}_0^*$ is globally asymptotically stable in Ω.

Proof. Since for $i = 1,2$, we have

$$S_1 \leqslant S_1^*, S_2 \leqslant S_2^*,$$

$$\begin{cases} \dfrac{dE_1}{dt} \leqslant -\left(\varphi_{12}^E + c_3\right) E_1 + \varphi_{21}^E E_2 + c_1 S_1^* I_{a,1} + c_1 S_1^* I_{s,1}, \\ \dfrac{dE_2}{dt} \leqslant \varphi_{12}^E E_1 - \left(\varphi_{21}^E + c_3\right) E_2 + c_2 S_2^* I_{a,2} + c_2 S_2^* I_{s,2}, \\ \dfrac{dI_{a,1}}{dt} \leqslant \beta_1 E_1 - \left(\varphi_{12}^{I_a} + c_5\right) I_{a,1} + \varphi_{21}^{I_a} I_{a,2}, \\ \dfrac{dI_{a,2}}{dt} \leqslant \beta_1 E_2 + \varphi_{12}^{I_a} I_{a,1} - \left(\varphi_{21}^{I_a} + c_5\right) I_{a,2}, \\ \dfrac{dI_{s,1}}{dt} \leqslant \beta_2 E_1 - c_6 I_{s,1}, \\ \dfrac{dI_{s,2}}{dt} \leqslant \beta_2 E_2 - c_6 I_{s,2}. \end{cases} \tag{10}$$

Defining an auxiliary linear system using the right-hand side of system (10), we have

$$\begin{cases} \dfrac{d\widetilde{E}_1}{dt} = -\left(\varphi_{12}^{E} + c_3\right)\widetilde{E}_1 + \varphi_{21}^{E}\widetilde{E}_2 + c_1 S_1^{*}\widetilde{I}_{a,1} + c_1 S_1^{*}\widetilde{I}_{s,1}, \\ \dfrac{d\widetilde{E}_2}{dt} = \varphi_{12}^{E}\widetilde{E}_1 - \left(\varphi_{21}^{E} + c_3\right)\widetilde{E}_2 + c_2 S_2^{*}\widetilde{I}_{a,2} + c_2 S_2^{*}\widetilde{I}_{s,2}, \\ \dfrac{d\widetilde{I}_{a,1}}{dt} = \beta_1\widetilde{E}_1 - \left(\varphi_{12}^{I_a} + c_5\right)\widetilde{I}_{a,1} + \varphi_{21}^{I_a}\widetilde{I}_{a,2}, \\ \dfrac{d\widetilde{I}_{a,2}}{dt} = \beta_1\widetilde{E}_2 + \varphi_{12}^{I_a} I_{a,1} - \left(\varphi_{21}^{I_a} + c_5\right)\widetilde{I}_{a,2}, \\ \dfrac{d\widetilde{I}_{s,1}}{dt} = \beta_2\widetilde{E}_1 - c_6\widetilde{I}_{s,1}, \\ \dfrac{d\widetilde{I}_{s,2}}{dt} = \beta_2\widetilde{E}_2 - c_6\widetilde{I}_{s,2}, \end{cases}$$

or, in other words,

$$\frac{d}{dt}\left(\widetilde{E}_1\ \widetilde{E}_2\ \widetilde{I}_{a,1}\ \widetilde{I}_{a,2}\ \widetilde{I}_{s,1}\ \widetilde{I}_{s,2}\right)^{T} = \begin{pmatrix} -\left(\varphi_{12}^{E} + c_3\right) & \varphi_{21}^{E} & c_1 S_1^{*} & 0 & c_1 S_1^{*} & 0 \\ \varphi_{12}^{E} & -\left(\varphi_{21}^{E} + c_3\right) & 0 & c_2 S_2^{*} & 0 & c_2 S_2^{*} \\ \beta_1 & 0 & -\left(\varphi_{12}^{I_a} + c_5\right) & 0\varphi_{21}^{I_a} & 0 & 0 \\ 0 & \beta_1 & \varphi_{12}^{I_a} & -\left(\varphi_{21}^{I_a} + c_5\right) & 0 & 0 \\ \beta_2 & 0 & 0 & 0 & -c_6 & 0 \\ 0 & \beta_2 & 0 & 0 & 0 & -c_6 \end{pmatrix} \begin{pmatrix} \widetilde{E}_1 \\ \widetilde{E}_2 \\ \widetilde{I}_{a,1} \\ \widetilde{I}_{a,2} \\ \widetilde{I}_{s,1} \\ \widetilde{I}_{s,2} \end{pmatrix}.$$

So,

$$\frac{d}{dt}\begin{pmatrix} d\widetilde{E}_1 \\ d\widetilde{E}_2 \\ d\widetilde{I}_{a,1} \\ d\widetilde{I}_{a,2} \\ d\widetilde{I}_{s,1} \\ d\widetilde{I}_{s,2} \end{pmatrix} = (F - V)\begin{pmatrix} d\widetilde{E}_1 \\ d\widetilde{E}_2 \\ d\widetilde{I}_{a,1} \\ d\widetilde{I}_{a,2} \\ d\widetilde{I}_{s,1} \\ d\widetilde{I}_{s,2}. \end{pmatrix}. \tag{11}$$

We have $R_v < 1 \Longleftrightarrow \sigma\left(F - V\right) < 0$, where $\sigma\left(F - V\right)$ is the spectral abscissa of the matrix $F - V$ (see the proof of Theorem 2 [26]). So, when $R_v < 1$, the eigenvalues of $(F - V)$ are with negative real parts. Thus all nonnegative solutions of (11) are such that

$$\lim_{t \longrightarrow +\infty}\left(\widetilde{E}_1, \widetilde{E}_2, \widetilde{I}_{a,1}, \widetilde{I}_{a,2}, \widetilde{I}_{s,1}, \widetilde{I}_{s,2}\right) = (0, 0, 0, 0, 0, 0).$$

By a standard comparison principle (see [24, Theorem B.1]) and the nonnegativity of $(E_1, E_2, I_{a,1}, I_{a,2}, I_{s,1}, I_{s,2})$, we conclude that when $R_v < 1$, all nonnegative solutions of (3) satisfy

$$\lim_{t \longrightarrow +\infty} (E_1, E_2, I_{a,1}, I_{a,2}, I_{s,1}, I_{s,2}) = (0, 0, 0, 0, 0, 0)\,. \tag{12}$$

Since (12) is satisfied, (3) is an asymptotically autonomous system [see [8, Theorem 2.5]] with limit affine system

$$\begin{cases} \dfrac{dS_1}{dt} = \mu + \varphi_{21}^S S_2 - \left(\varphi_{12}^S + c_0\right) S_1 + \lambda_2 V_1 + \gamma R_1, \\ \dfrac{dS_2}{dt} = \mu + \varphi_{12}^S S_1 - \left(\varphi_{21}^S + c_0\right) S_2 + \lambda_2 V_2 + \gamma R_2, \\ \dfrac{dV_1}{dt} = \lambda_1 S_1 + \varphi_{21}^V V_2 - \left(\varphi_{12}^V + c_4\right) V_1, \\ \dfrac{dV_2}{dt} = \lambda_1 S_2 + \varphi_{12}^V V_1 - \left(\varphi_{21}^V + c_4\right) V_2, \\ \dfrac{dR_1}{dt} = \lambda_5 V_1 + \varphi_{21}^R R_2 - \left(\varphi_{12}^R + c_7\right) R_1, \\ \dfrac{dR_2}{dt} = \lambda_5 V_2 + \varphi_{12}^R R_1 - \left(\varphi_{21}^R + c_7\right) R_2. \end{cases} \tag{13}$$

It is easy to verify from (13) that

$$\lim_{t \longrightarrow +\infty} (S_1, S_2, V_1, V_2, R_1, R_2) = (S_1^*, S_2^*, V_1^*, V_2^*, R_1^*, R_2^*)\,.$$

The globally asymptotic stability of the DFE $\mathcal{E}_0^*$ is then proved if $R_v < 1$. □

The presence of endemic equilibrium, characterized by positive infective numbers, has barely been discussed here. Although their existence has not been formally proven, numerical simulations suggest a globally asymptotically stable equilibrium when $R_v > 1$. Investigating the system's persistence properties under the condition $R_v > 1$ would be a valuable avenue for further exploration.

Lemma 1. [30] Let $\phi_t : X \to X$ be a semiflow and let $X_0 \subset X$ be an open set. Define $\partial X_0 = X \setminus X_0$ and $M_\partial = \{x \in \partial X_0 : \phi_t x \in \partial X_0, t \geqslant 0\}$. Assume the following conditions:

1. $\phi_t X_0 \subset X_0$ and ϕ_t has a global attractor A.

2. There exists a finite sequence $\mathcal{M} = \{M_1, \ldots, M_k\}$ of disjoint, compact, and isolated invariant sets in ∂X_0 such that
   (a) $\Omega(M_\partial) := \cup_{x \in M_\partial} w(x) \subset \cup_{i=1}^{k} M_i$.
   (b) No subset of $\mathcal{M}$ forms a cycle in ∂X_0.
   (c) M_i is isolated in X.
   (d) $W^s(M_i) \cap X_0 = \emptyset$, where $W^s(M_i) = \{x \in X_0 : w(x) \subset M_i\}$ for each $1 \leqslant i \leqslant k$.

Then ϕ_t is uniformly persistent with respect to $(X_0, \partial X_0)$, that is, there exists $\theta > 0$ such that

$$\lim_{t \longrightarrow +\infty} \inf d(\phi_t x, \partial X_0) \geqslant \theta \quad \text{for} \quad x \in X_0.$$

Therefore, we have the following result.

Theorem 4. If $R_v > 1$, then system (3) is uniformly persistent, namely, there exists a constant $\theta > 0$ such that

$$\lim_{t \longrightarrow +\infty} \inf S_i(t) > \theta, \lim_{t \longrightarrow +\infty} \inf E_i(t) > \theta, \lim_{t \longrightarrow +\infty} \inf V_i(t) > \theta,$$
$$\lim_{t \longrightarrow +\infty} \inf I_{a,i}(t) > \theta, \lim_{t \longrightarrow +\infty} \inf I_{s,i}(t) > \theta, \lim_{t \longrightarrow +\infty} \inf R_i(t) > \theta,$$

for any initial conditions satisfying (4).

Proof. Choose

$$\begin{cases} X = \Omega \\ X_0 = \{(S_1, S_2, E_1, E_2, V_1, V_2, I_{a,1}, I_{a,2}, I_{s,1}, I_{s,2}, R_1, R_2) \in X : \\ \quad E_1 + E_2 > 0, I_{a,1} + I_{a,2} > 0, I_{s,1} + I_{s,2} > 0\}, \end{cases}$$

and

$$\begin{aligned} \partial X_0 &= X \setminus X_0 \\ &= \Big\{(S_1, S_2, E_1, E_2, V_1, V_2, I_{a,1}, I_{a,2}, I_{s,1}, I_{s,2}, R_1, R_2) \in X : \\ &\qquad E_1 = E_2 = I_{a,1} = I_{a,2} = I_{s,1} = I_{s,2} = 0\Big\}. \end{aligned}$$

Let ϕ_t be the semi flow induced by the solutions of system (3).

We have proved in Theorem *1* that $\phi_t X_0 \subset X_0$ and ϕ_t is ultimately bounded in X_0; so there always exists a global attractor for ϕ_t. It is obvious

that $\mathcal{E}_0^*$ is the unique boundary equilibrium on ∂X_0, which implies that $\mathcal{E}_0^*$ is globally stable on ∂X_0. Moreover, $(S_1, S_2, , V_1, V_2, R_1, R_2)$ converges to $(S_1^*, S_2^*, , V_1^*, V_2^*, R_1^*, R_2^*)$ on ∂X_0.

Let $M_1 = \{\mathcal{E}_0^*\}$ and $\mathcal{M} = \{M_1\}$, then $\cup_{x \in M_\partial} w(x) = M_1$ and no subset of $\mathcal{M}$ forms a cycle in ∂X_0.

If $R_v > 1$, then $\mathcal{E}_0^*$ is unstable in X_0. Therefore, conditions $(2-c)$ and $(2-d)$ of Lemma 1 are satisfied. Then ϕ_t is uniformly persistent with respect to $(X_0, \partial X_0)$, that is, there exists $\theta > 0$ such that

$$\lim_{t \longrightarrow +\infty} \inf d(\phi_t x, \partial X_0) \geqslant \theta \quad \text{for} \quad x \in X_0.$$

We have thus shown that if $R_v > 1$, then the disease is endemic. □

5 Numerical simulation

In this section, we extend our exploration into the numerical simulation of COVID-19 transmission within a two-patch framework, leveraging well-established methodologies commonly employed in epidemiological studies. Building upon the model detailed and examined in Section 4, our objective is to unravel the intricate dynamics of COVID-19 transmission influenced by migration, vaccination, and medical unavailability.

In a prior study [7], we investigated disease transmission dynamics in the absence of migration in a single patch. In the current work, which builds upon our earlier research, we now consider the pivotal role of migration—representing the movement of individuals between patches. Migration rates, denoted by φ_{12} and φ_{21}, exert a substantial influence on the overall transmission dynamics.

To ground our simulations in real-world scenarios, we integrate data from Table 1 into our analysis. Additionally, we set migration rates from the patch 1 to the patch 2 (φ_{12}^S, φ_{12}^E, φ_{12}^V, $\varphi_{12}^{I_a}$, φ_{12}^R) at 0.001 and migration rates from the patch 2 to the patch 1 (φ_{21}^S, φ_{21}^E, φ_{21}^V, $\varphi_{21}^{I_a}$, φ_{21}^R) at 0.005. These rates underscore the interconnectedness of patches and contribute to the comprehensive understanding of disease spread across different regions.

Table 1: The operational parameters of the simulation consist of two categories: those characterizing the population and the disease remain the same across all patches, whereas parameters reflecting the behavior of individuals vary from one patch to another

Parameters	Description	Values (per day)
N	Population size	100000
N_1	Population size in patch 1 (Most populated patch).	75000
N_2	Population size in patch 2 (Least populated patch).	25000
μ	Recruitment rate.	3.38e-05
α_1	Infection rate for unconfined persons.	0.01-0.3
α_2	Infection rate for confined persons.	0.01-0.3
η_1	The confinement rate in patch 1.	0.2-0.5
η_2	The confinement rate in patch 2.	0.2-0.5
β_1	Rate of an exposed becoming asymptomatic infected.	0.092
β_2	Rate of an exposed becoming symptomatic infected.	0.05
λ_1	Vaccination rate of susceptible.	0.01-0.03
λ_2	The ineffectiveness rate of the vaccine	0.001765
λ_3	Vaccination rate of exposed persons.	0.00167
λ_4	Vaccination rate of asymptomatic infected.	0.0167
λ_5	The effectiveness rate of the vaccine.	0.008
γ	Rate of a recovered becoming susceptible.	0.0099
γ_1	The recovery rate of asymptomatic people	0.01
γ_2	the recovery rate of symptomatic people	0.0005
δ	COVID-19 death rate.	0.0055

Our simulation parameters encompass scenarios without migration restrictions, incorporating measures such as containment and vaccination strategies. The outcomes of these simulations gauged through R_v and other pertinent metrics, offer valuable insights into the potential efficacy of interventions aimed at curtailing the virus's propagation.

Subsequently, we delve into an analysis of the sensitivity of key parameters to the vaccine reproduction number R_v. Following this, we meticulously scrutinize the theoretical results previously obtained, aiming to provide empirical substantiation and validate the accuracy of our theoretical framework.

This multifaceted approach enhances our understanding of the practical implications of proposed interventions and reinforces the robustness of our theoretical foundation.

5.1 Sensitivity analysis of the main parameters on R_v

To effectively mitigate the impact of COVID-19 and reduce associated fatalities, understanding the relative importance of various parameters influencing its progression is crucial. The vaccination reproduction number, R_v [10], stands out as a key determinant in the transmission dynamics of COVID-19.

Our investigation employs sensitivity analysis on R_v within the framework of our model (3). This analysis aims to discern the individual importance of each parameter, shedding light on influential factors in disease transmission. The derived sensitivity indices offer insights into how a state variable changes relative to variations in specific parameters.

Sensitivity analysis, a widely used technique, involves computing the first-order partial derivatives of model output concerning input parameters. This process can be conceptualized as calculating gradients in a multidimensional reference parameter space [23], providing valuable information on the model's responsiveness to slight changes in input parameters.

The normalized forward sensitivity index of R_v, denoted as $\mathcal{I}_p^{R_v}$, quantifies the sensitivity with respect to a given parameter p and is defined by the expression:

$$\mathcal{I}_p^{R_v} = \frac{\partial R_v}{\partial p}\frac{p}{R_v}.$$

This index facilitates a quantitative assessment of how alterations in individual parameters influence the value of R_v, a pivotal metric for understanding disease transmission. By leveraging explicit formulas for each parameter affecting R_v, as detailed in reference [31], we gain precise insights into the impact of specific parameters on R_v. Such sensitivity analysis proves invaluable in prioritizing interventions and control measures to effectively curtail the spread of the disease, enabling a more targeted and comprehensive approach to managing COVID-19.

Table 2: Local sensitivity indices for R_v concerning parameters in the proposed model are computed at the baseline parameter values provided in Table 1. Parameters exhibiting the highest sensitivity are highlighted in bold

Parameters	α_1	$\boldsymbol{\alpha_2}$	η_1	η_2	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\lambda_1}$
Sensitivity Index	0.11	**0.89**	-0.11	-0.00006	**-0.34**	**0.37**	**-0.18**

Parameters	λ_2	λ_3	λ_4	λ_5	γ	γ_1	γ_2
Sensitivity Index	0.032	-0.016	-0.018	0.066	0.079	-0.11	-0.059

Parameters	φ_{21}^S	φ_{21}^E	φ_{21}^V	$\varphi_{21}^{I_a}$	φ_{21}^R
Sensitivity Index	0.15	0.0003	0.008	0.002	0.009

Parameters	φ_{12}^S	φ_{12}^E	φ_{12}^V	$\varphi_{12}^{I_a}$	φ_{12}^R
Sensitivity Index	-0.15	-0.006	-0.008	-0.009	-0.009

Table 2 highlights the parameters most responsive to changes in the vaccination reproduction number (R_v), with the transmission probability of unconfined individuals (α_2) and the transition rate from exposed to symptomatic infected (β_2) exhibiting the highest sensitivity. The vaccination rate (λ_1) and the transition rate from exposed to asymptomatic infected (β_1) closely follow in significance. A 10% increase (or decrease) in α_2 or β_2 leads to corresponding 8.9% and 3.7% changes in R_v, respectively, as denoted by $\mathcal{I}_{\alpha_1}^{R_v} = +0.89$ and $\mathcal{I}_{\eta_1}^{R_v} = +0.37$. Conversely, a 10% increase in λ_1 or β_1 corresponds to a 1.8% and 3.4% decrease in R_v, respectively, indicated by $\mathcal{I}_{\lambda_1}^{R_v} = -0.18$ and $\mathcal{I}_{\beta_1}^{R_v} = -0.34$.

The local analysis emphasizes the importance of implementing effective containment measures for managing the transmission probability (α_2) and the transition rate from exposed to symptomatic infected. This involves stringent containment measures, the use of physical barriers, and adherence to social distancing guidelines. Additionally, vaccination efforts, represented by the rate (λ_1), play a crucial role in reducing the spread of the disease. By strategically controlling these parameters, a significant reduction in COVID-19-related deaths can be achieved.

5.2 The stability of the DFE

In this subsection, we will confirm the theoretical results obtained earlier through numerical simulations for the following parameter values (per day):

$$\alpha_1 = 0.010 \ ; \ \alpha_2 = 0.0655 \ ; \ \eta_1 = 0.20 \ ; \ \eta_2 = 0.25 \ ; \ \lambda_1 = 0.01817.$$

The existence and stability conditions in Theorem 2 are satisfied

$$R_v = 0.8375 < 1.$$

Hence, system (3) has a DFE,

$$\begin{aligned}\mathcal{E}_0^* &= \left[S_1^*,\ S_2^*,\ E_1^*,\ E_2^*,\ V_1^*,\ V_2^*,\ I_{a,1}^*,\ I_{a,2}^*,\ I_{s,1}^*,\ I_{s,2}^*\ ,\ R_1^*,\ R_2^*\right]^T \\ &= [0.1915,\ 0.0385,\ 0,\ 0,\ 0.3326,\ 0.0713,\ 0,\ 0,\ 0,\ 0,\ 0.2861,\ 0.0573]^T .\end{aligned}$$

Therefore, the DFE $\mathcal{E}_0^*$ is stable. This also implies that the disease is eradicated, as depicted in Figures 2 and 3.

Furthermore, if we increase the values of the most sensitive parameters such as $\alpha_1 = 0.0653$, $\alpha_2 = 0.1295$, $\eta_1 = 0.20$, $\eta_2 = 0.30$ and $\lambda_1 = 0.00117$, through direct calculation, the vaccination reproduction number $R_v = 6.4251 > 1$. Consequently, according to Theorem 2, the DFE is unstable, and the model (3) has an endemic equilibrium.

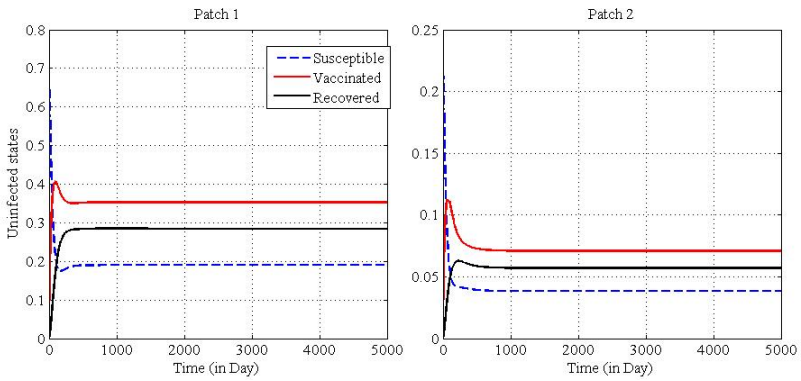


Figure 2: The proposed model has only one DFE $\mathcal{E}_0^*$, which is asymptotically stable when $R_v < 1$.

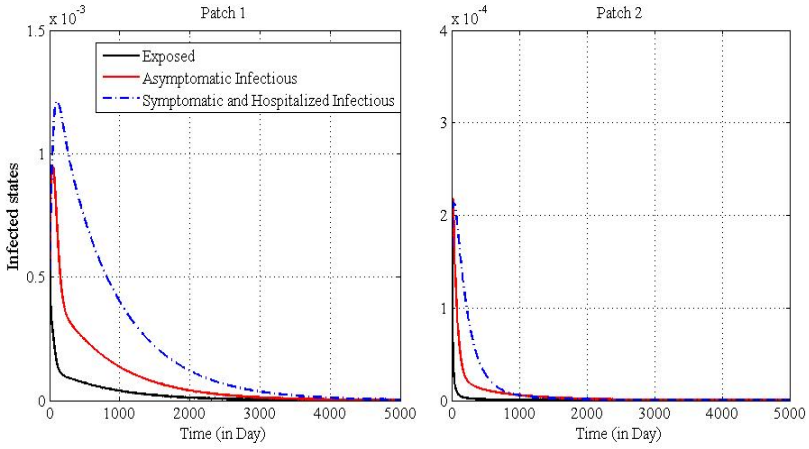


Figure 3: The proposed model has only one DFE $\mathcal{E}_0^*$, which is asymptotically stable when $R_v < 1$.

5.3 The existence and stability of the endemic equilibrium of the model

Now, we deal with the existence of endemic equilibrium of model (3). Consider the following parameter values (per day):

$$\alpha_1 = 0.0653,\ \alpha_2 = 0.1295,\ \eta_1 = 0.20,\ \eta_2 = 0.30,\ \lambda_1 = 0.00117.$$

Hence, we obtain

$$R_v = 6.4251 > 1.$$

System (3) has an endemic equilibrium,

$$\begin{aligned}\mathcal{E}_1^{**} &= \left[S_1^{**},\ S_2^{**},\ E_1^{**},\ E_2^{**},\ V_1^{**}\ ,\ V_2^{**},\ I_{a,1}^{**},\ I_{a,2}^{**},\ I_{s,1}^{**},\ I_{s,2}^{**},\ R_1^{**},\ R_2^{**}\right]^T \\ &= \Big[1.0663e-01,\ 2.3693e-02,\ 6.0569e-04,\ 6.4518e-06, \\ &\quad 1.6194e-02,\ 3.0613e-03,\ 2.0238e-03,\ 8.2520e-05, \\ &\quad 5.0185e-03,\ 5.3556e-05,\ 1.5170e-02,\ 2.7128e-03\Big]^T.\end{aligned}$$

Figures 4 and 5 illustrate the stability of $\mathcal{E}_1^{**}$.

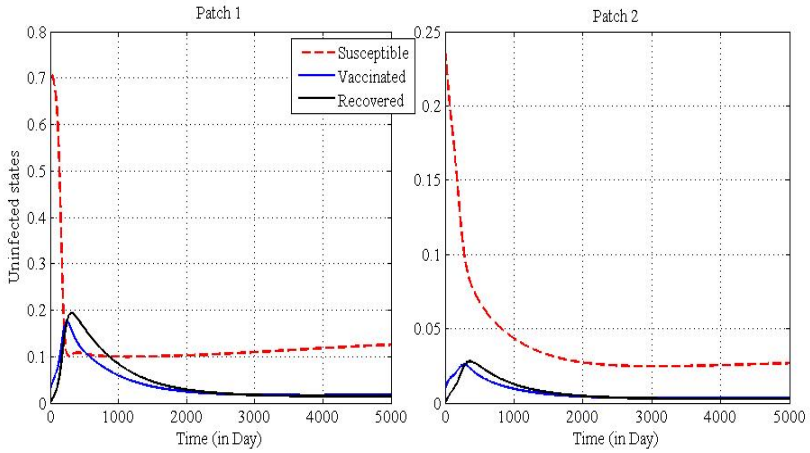


Figure 4: Illustration of the stability of $\mathcal{E}_1^{**}$.

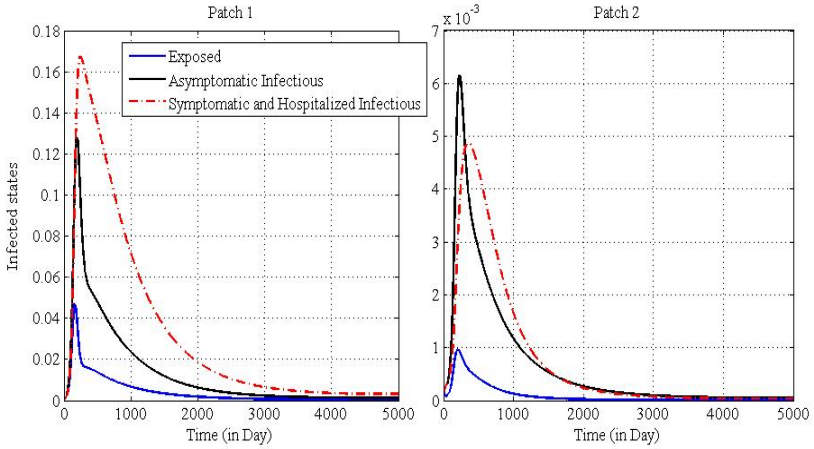


Figure 5: Illustrates the stability of $\mathcal{E}_1^{**}$.

By adjusting the key operational parameters to yield varying values of R_v greater than 1, we have

$$R_v = [3.7816,\ 4.6628,\ 5.5440,\ 6.4251,\ 7.3063,\ 8.1875]^T.$$

We observe in Figure 6, the endemic equilibrium point along with standard deviations for each disease state,

$$SD = \Big[3.6400e-02,\ 7.1634e-03,\ 3.8114e-05,\ 4.1860e-07,\\ 4.1299e-03,\ 8.3947e-04,\ 1.2731e-04,\ 5.2295e-06,\\ 3.1569e-04,\ 3.4790e-06,\ 3.1924e-03,\ 6.5984e-04\Big]^T.$$

These standard deviation values for each disease state indicate that the data points in a set are closely clustered around the mean, suggesting low variability.

Additionally, numerical simulations in Figure 6 show that the solutions with different values of R_v converge to the same endemic equilibrium, where the initial values are

$$S_1(0) = 7.1970e-01,\ S_2(0) = 2.3990e-01,\ E_1(0) = 8.2500e-04,$$
$$E_2(0) = 2.7500e-04,\ V_1(0) = 2.7975e-02,\ V_2(0) = 9.3250e-03,$$
$$I_{a,1}(0) = 3.7500e-04,\ I_{a,2}(0) = 1.2500e-04,\ I_{s,1}(0) = 3.7500e-04,$$
$$I_{s,2}(0) = 1.2500e-04,\ R_1(0) = 4.5000e-04,\ R_2(0) = 1.5000e-04.$$

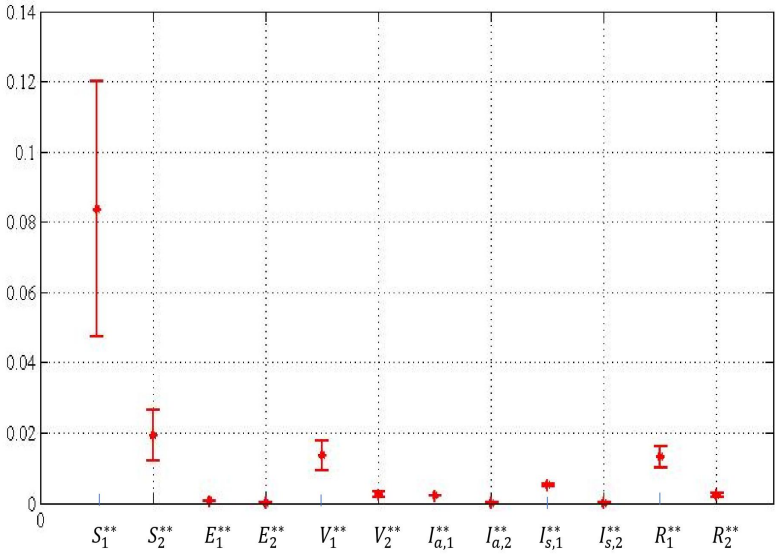


Figure 6: Standard deviation values for each disease state for $\mathcal{E}_1^{**}$.

Certainly, there is a possibility that the endemic equilibrium remains stable given the specified parameters. This leads us to an intriguing question: if the vaccination reproduction number R_v exceeds 1, then the model (3) with migration allows for a locally asymptotically stable endemic equilibrium.

5.4 The effect of migration rate on the transmission of COVID-19 in the two patches

For migration rate values ranging from 0 to 0.01, we observe the results in Figure 7.

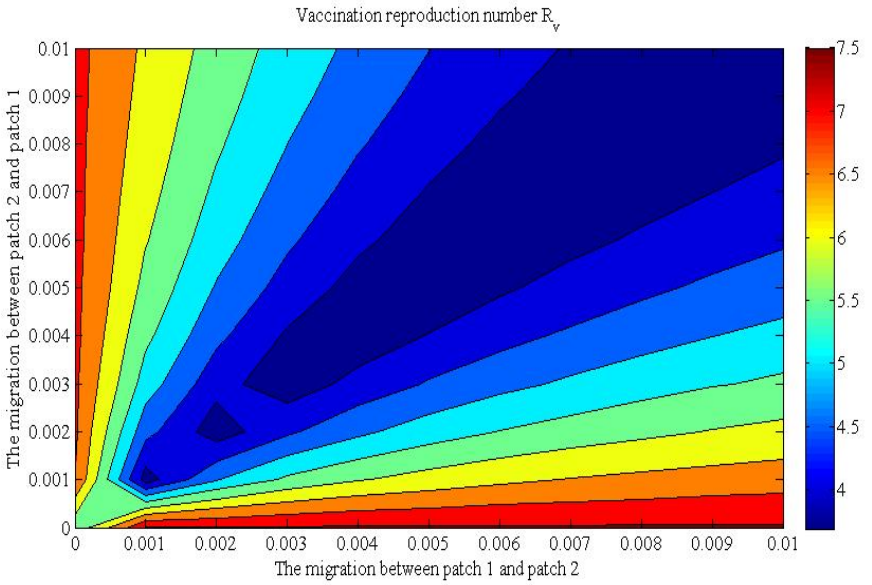


Figure 7: The migration rate effect on R_v.

The numerical simulations presented in Figure 7 underscore the endemic nature of the disease. As the migration rate diminishes from the more populated patch 1 to the less populated patch 2, and concurrently rises from the patch 2 to the patch 1, the vaccine reproduction number Rv demonstrates a proportional increase. This observation validates the outcomes of the sensitivity analysis performed on Rv. Theoretical insights, coupled with these

numerical simulations, emphasize the intricate relationship between COVID-19 prevalence across diverse patches and the corresponding migration rates. These findings contribute to a comprehensive understanding of the dynamics, suggesting that the prevalence is intricately tied to the patterns of migration between distinct patches.

6 Conclusion

In conclusion, the findings from sensitivity analysis shed light on the nuanced interplay between key parameters and the vaccination reproduction number (R_v) in the context of COVID-19 dynamics. Notably, the transmission rate of unconfined individuals (α_2) and the transition rate from exposed to symptomatic infected individuals (β_2) emerge as pivotal factors influencing the increase in R_v. Robust containment measures targeting these aspects, such as stringent protocols, physical barriers, and adherence to social distancing guidelines, are imperative for effective disease control. Furthermore, the vaccination rate (λ_1) and the transition rate from exposed to asymptomatic infected individuals (β_1) closely follow in significance. A 10% increase or decrease in α_2 or β_2 corresponds to substantial changes in R_v, emphasizing the sensitivity of the system. Conversely, a 10% increase in λ_1 or β_1 leads to a notable decrease in R_v. These results provide actionable insights into the impact of parameter variations on the vaccination reproduction number.

For adequate operational data, the stability of DFE $\mathcal{E}_0^*$ suggests the potential for disease eradication. However, elevating the values of sensitive parameters leads to an increase in R_v to 6.4251, surpassing the critical threshold of 1. This instability of $\mathcal{E}_0^*$ indicates the presence of an endemic equilibrium, as per Theorem 2. The theoretical framework aligns with numerical simulations, as depicted in Figures 2 and 3, underscoring the importance of both analytical and computational approaches in understanding the system behavior. The low standard deviation values across different disease states suggest that the data points are closely clustered around the mean, indicating limited variability. Additionally, numerical simulations in Figures 4 and 5 further affirm the endemic nature, demonstrating that varying values of $R_v > 1$ converge to the same endemic equilibrium.

We conclude the simulation by highlighting the complex relationship between COVID-19 prevalence across distinct patches and the corresponding migration rates, offering valuable insights into the dynamic nature of the disease and guiding strategies for effective disease management.

References

[1] Ahmad, R.A., Imron, M.A., Ramadona, A.L., Lathifah, N., Azzahra, F., Widyastuti, K. and Fuad, A. *Modeling social interaction and metapopulation mobility of the COVID-19 pandemic in main cities of highly populated Java Island, Indonesia: An agent-based modeling approach*, Front. Ecol. Evol. 10 (2023), 958651.

[2] Arino, J., Sun, C. and Yang, W. *Revisiting a two-patch SIS model with infection during transport*, Math. Med. Biol. 33(1) (2015), 29–55.

[3] Arino, J. and Van Den Driessche, P. *Disease spread in metapopulations*, Fields Inst. Commun. 48 (2006), 1–12.

[4] Baister, M., McTaggart, E., McMenemy, P., Megiddo, I. and Kleczkowski, A. *COVID-19 in Scottish care homes: A metapopulation model of spread among residents and staff*, medRxiv 2021.08.24.21262524.

[5] Bortolatto, R. *A note on the Lienard-Chipart criterion and roots of some families of polynomials*, arXiv preprint arXiv:1407.4852 (2014).

[6] Boubekeur, M.A. and Belhamiti, O. *Modeling the Impact of Obesity on COVID-19: Evidence from Sensitivity Analysis*, (2023), Submitted.

[7] Bouziane, M., Mezouaghi, A. and Belhamiti, O. *analysis of the vaccination reproduction number and endemic equilibrium to control the Covid-19 spread*, Adv. Math. Sci. App. 32(2) (2023), 399–430.

[8] Castillo-Chavez, C. and Thieme, H. *Autonomous epidemic models*, O. Arino, D. Axelrod, M. Kimmel, M. Langlais (Eds.), Mathematical Population Dynamics: Analysis of Heterogeneity, BU-1248-M (1994) 1–23.

[9] Chen, T.M., Rui, J., Wang, Q.P., Zhao, Z.Y., Cui, J.A. and Yin, L.A. *Model for simulating the phase-based transmissibility of a novel coronavirus*, Infect. Dis. Poverty 9(1) (2020), 1–8.

[10] Chitnis, N., Hyman, J.M. and Cushing, J.M. *Determining important parameters in the spread of malaria through the sensitivity analysis of a mathematical model*, Bull. Math. Biol. 70 (2008), 1272–1296.

[11] Daud, A.A.M. *A note on Lienard-Chipart criteria and its application to epidemic models*, Mathematics and Statistics 9 (2021), 41–45.

[12] Faiçal, N., Ivan, A., Juan, N. and Delfim, T. *Modeling of COVID-19 transmission dynamics with a case study of Wuhan*, Chaos Solit. Fractals. 135 (2020), 109846.

[13] Hancean, M.G., Slavinec, M. and Perc, M. *impact of human mobility networks on the global spread of COVID-19*, J. Complex Netw. 8(6) (2020), 1–14.

[14] Humphries, R., Spillane, M., Mulchrone, K., Wieczorek, S., Riordain, M. and Havel, P. *A metapopulation network model for the spreading of SARS CoV-2: Case study for Ireland*, Infect. Dis. Model. 6 (2021), 420–437.

[15] Iyaniwura, S.A., Ringa, N., Adu, P.A., Mak, S., Janjua, N.Z., Irvine, M.A. and Otterstatter, M. *Understanding the impact of mobility on COVID-19 spread: A hybrid gravity-metapopulation model of COVID-19*, PLoS Comput. Biol. 19(5) (2023), e1011123.

[16] Keddar, M.E.B. and Belhamiti O. *A study of global dynamics and sensitivity analysis of a discrete-time model of the COVID-19 epidemic*, Iranian Journal of Numerical Analysis and Optimization (2023), Accepted. 10.22067/IJNAO.2023.82954.1281

[17] Kim, J.H., SU, W. and Song, Y.J. *On stability of a polynomial*, J. Appl. Math. Inform. 36 (2018), 231–236.

[18] Knobler, S., Mahmoud, A., Lemon, S. and Pray, L. *The impact of globalization on infectious disease emergence and control: Exploring the con-*

sequences and opportunities, Workshop Summary - Forum on Microbial Threats, The National Academies Press, 2006.

[19] Lan Meng, L. and Zhu, W. *SEIR epidemic model for COVID-19 in a multipatch environment, discrete dynamics in nature society*, Discrete Dyn. Nature Soc. 2021 (2021), Article ID 5401253, 1–12.

[20] Martens, P. and Hall, L. *Malaria on the move: human population movement and malaria transmission*, Emerg. Infect. Dis. 6 (2) (2000), 103–109.

[21] McCarthy, C.V., O'Mara, O. and Van Leeuwen, E. *The impact of COVID-19 vaccination in prisons in England and Wales: a metapopulation model*, BMC Public Health, 22 (1003) (2022), 1–17.

[22] Perlman, S., *Another decade, another Coronavirus*, N. Engl. J. Med. 382(8) (2020), 760–762.

[23] Rabitz, H., Kramer, M. and Dacol, D. *Sensitivity analysis in chemical kinetics*, Annu. Rev. Phys. Chem. 34 (1983), 419–461.

[24] Smith, H.L. and Waltman, P. *The theory of the Chemostat*, Cambridge University, 1995.

[25] Van Den Driessche, P. *Reproduction numbers of infectious disease models*, Infect. Dis. Model. 2 (2017), 288–303.

[26] Van Den Driessche, P. and Watmough, J. *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission*, Math. Biosci. 180 (2002), 29–48.

[27] Van Den Driessche, P. and Watmough, J. *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission*, Math. Biosci 180 (2002), 29–48.

[28] World Health Organization (WHO). *Coronavirus (COVID-19) Dashboard*; (accessed April 17, 2022).

[29] World Health Organization (WHO). *Novel Coronavirus* 2019-nCoV, situation report-1. (accessed May 10, 2021).

[30] Zhao, X. *Systems in population biology*, Springer, 2003.

[31] Zi, Z. *Sensitivity analysis approaches applied to systems biology models*, IET Syst. Biol. 5 (2011), 336–346.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 2, 2024, pp 613–637
https://doi.org/10.22067/ijnao.2024.85895.1364
https://ijnao.um.ac.ir/

Research Article

Goursat problem in Hyperbolic partial differential equations with variable coefficients solved by Taylor collocation method

F. Birem, A. Boulmerka, H. Laib*, and C. Hennous

*Corresponding author

Received 17 December 2023; revised 14 March 2024; accepted 30 March 2024

Fouzia Birem
Laboratory of Mathematics and their interactions, University Center Abdelhafid Boussouf, Mila, Algeria. e-mail: fouzia.birem@centre-univ-mila.dz

Aissa Boulmerka
Laboratory of Mathematics and their interactions, University Center Abdelhafid Boussouf, Mila, Algeria. e-mail: aissa.boulmerka@gmail.com

Hafida Laib
Laboratory of Mathematics and their interactions, University Center Abdelhafid Boussouf, Mila, Algeria. e-mail: hafida.laib@gmail.com

Chahinaz Hennous
Laboratory of Mathematics and their interactions, University Center Abdelhafid Boussouf, Mila, Algeria. e-mail: hennouschahinaz@gmail.com

How to cite this article
Birem, F., Boulmerka, A., Laib, H. and Hennous, C., Goursat problem in Hyperbolic partial differential equations with variable coefficients solved by Taylor collocation method. *Iran. J. Numer. Anal. Optim.*, 2024; 14(2): 613-637. https://doi.org/10.22067/ijnao.2024.85895.1364

Abstract

The hyperbolic partial differential equation (PDE) has important practical uses in science and engineering. This article provides an estimate for solving the Goursat problem in hyperbolic linear PDEs with variable coefficients. The Goursat PDE is transformed into a second kind of linear Volterra integral equation. A convergent algorithm that employs Taylor polynomials is created to generate a collocation solution, and the error using the maximum norm is estimated. The paper includes numerical examples to prove the method's effectiveness and precision.

AMS subject classifications (2020): 45D05, 34K28, 45L05, 30K05.

Keywords: Hyperbolic partial differential equations; Goursat problem; Volterra integral equation; Collocation method; Taylor polynomials.

1 Introduction

Partial differential equations (PDEs) are commonly used to address problems in various fields, such as engineering, physics, and finance. These equations are crucial in the study of various phenomena such as electric currents, gravity, heat transfer, water wave movement, fluid mechanics, electromagnetism, elasticity, quantum mechanics, population dynamics, stock and option pricing, chemical reaction-diffusion, as well as in the modeling of Schrödinger's equation.

A second-order linear hyperbolic PDE with two variables for $\omega(\epsilon, \eta)$ has the form

$$A\frac{\partial^2\omega}{\partial\epsilon^2} + B\frac{\partial^2\omega}{\partial\epsilon\partial\eta} + C\frac{\partial^2\omega}{\partial\eta^2} + D\frac{\partial\omega}{\partial\epsilon} + E\frac{\partial\omega}{\partial\eta} + F\omega + G = 0, \tag{1}$$

where $B^2 - 4AC > 0$, and A, B, C, D, E, F, and G are functions of the variables ϵ and η. By a suitable change of the independent variables, we shall show that any equation of the form (1) can be reduced to the canonical form or normal form of the hyperbolic equation called the Goursat problem

$$\frac{\partial^2\omega(t,s)}{\partial t\partial s} = \phi\left(t, s, \omega, \frac{\partial\omega(t,s)}{\partial t}, \frac{\partial\omega(t,s)}{\partial s}\right). \tag{2}$$

Equations (1) and (2) are hyperbolic PDEs and appear frequently in the study of ecological and cosmological phenomena [26]. To accurately simulate these events, it is important to develop efficient and reliable methods for solving these equations. Several numerical techniques have been proposed for this purpose, including finite difference methods [14, 4], finite element methods [2], Taylor matrix method [7, 6], Legendre multi-wavelet Galerkin method [28, 27], rational Chebyshev method [10], modified variational iteration method [1], and Chebyshev wavelet scheme [13].

The numerical solution of Goursat problem (2) has been investigated by many authors. For instance, Scott [25] considered the homogeneous Goursat problem (3) with the coefficients depending on the same variable. Evans and Sanugi [12] proposed a nonlinear trapezoidal formula based on geometric means. Day [9] used the Runge–Kutta method to approximate the solution of (2). Pandey [24] presented a novel exponential finite difference to obtain a numerical solution of (2). Drignei [11] developed an algorithm to find the quadruple solution to a Goursat problem in a triangular domain.

In this paper, we consider a collocation solution of the linear Goursat problem of the second-order with variable coefficients

$$\begin{aligned}&\frac{\partial^2\omega(t,s)}{\partial t\partial s}+a(t,s)\frac{\partial\omega(t,s)}{\partial t}+b(t,s)\frac{\partial\omega(t,s)}{\partial s}+c(t,s)\omega(t,s)=f(t,s),\quad (3)\\&\omega(0,s)=\alpha(s),\omega(t,0)=\beta(t),\alpha(0)=\beta(0),\quad (t,s)\in[0,X]\times[0,Y],\end{aligned}$$

where a, b, c, and f are smooth functions through the domain of discussion. For the existence and uniqueness of the solution, see [15].

The Goursat PDE (3) is converted to the linear Volterra integral equation (VIE) of the second kind

$$\begin{aligned}\omega(\tau,z)=&g(\tau,z)+\int_0^\tau\kappa_1(t,z)\omega(t,z)dt+\int_0^z\kappa_2(\tau,s)\omega(\tau,s)ds\\&+\int_0^\tau\int_0^z\kappa_3(t,s)\omega(t,s)dsdt,\end{aligned}\quad (4)$$

where g,κ_1,κ_2, and κ_3 are defined in (7). Moreover, the special case of (4) for $\kappa_1=\kappa_2=0$ is considered in [19]. We develop the collocation method introduced in [3, 16, 17, 18] to solve (4). This method is based on approximating the exact solution of a given integral equation with a suitable function be-

longing to a chosen finite-dimensional space (8) such that the approximated solution satisfies the integral equation on collocation points (9). An important aspect of this method is that the number of subintervals and the degree of Taylor polynomials can be changed to get the best possible result. It is also easy to implement, and the approximate solution is based on iterative formulas without needing to solve any algebraic equations.

The paper is organized as follows: In section 2, the converting of PDE to the two-dimensional VIE. In Section 3, the approximating solution of (6) in each collocation point by a Taylor polynomial. The convergence analysis is investigated in section 4. Numerical examples are provided in section 5 to illustrate the theoretical results. Finally, a conclusion is given in section 6.

2 Converting PDE to two-dimensional VIE

In this section, we study the technique that will convert PDE (3) to an equivalent VIE (4).

Integrating both sides of (3) with respect to s yields

$$\frac{\partial\omega(t,z)}{\partial t}-\frac{\partial\omega(t,0)}{\partial t}+\int_0^z a(t,s)\frac{\partial\omega(t,s)}{\partial t}ds+\int_0^z b(t,s)\frac{\partial\omega(t,s)}{\partial s}ds + \int_0^z c(t,s)\omega(t,s)ds=\int_0^z f(t,s)ds,$$

which implies

$$\begin{aligned}\int_0^z f(t,s)ds =&\frac{\partial\omega(t,z)}{\partial t}-\frac{\partial\omega(t,0)}{\partial t}+\int_0^z a(t,s)\frac{\partial\omega(t,s)}{\partial t}ds+b(t,z)\omega(t,z)\\ &-b(t,0)\omega(t,0)-\int_0^z\frac{\partial b(t,s)}{\partial s}\omega(t,s)ds+\int_0^z c(t,s)\omega(t,s)ds.\end{aligned} \tag{5}$$

Integrating again both sides of (5) with respect to t yields

$$\begin{aligned}\int_0^\tau\int_0^z f(t,s)dsdt =&\omega(\tau,z)-\omega(0,z)-\omega(\tau,0)+\omega(0,0)\\ &+\int_0^\tau\int_0^z a(t,s)\frac{\partial\omega(t,s)}{\partial t}dsdt\\ &+\int_0^\tau b(t,z)\omega(t,z)dt-\int_0^\tau b(t,0)\omega(t,0)dt\end{aligned}$$

$$- \int_0^\tau \int_0^z \frac{\partial b(t,s)}{\partial s}\omega(t,s)dsdt + \int_0^\tau \int_0^z c(t,s)\omega(t,s)dsdt.$$

Hence,

$$\begin{aligned}\int_0^\tau \int_0^z f(t,s)dsdt =&\omega(\tau,z) - \alpha(z) - \beta(\tau) + \omega(0,0)\\ &+ \int_0^z a(\tau,s)\omega(\tau,s)ds + \int_0^\tau b(t,z)\omega(t,z)dt\\ &- \int_0^z a(0,s)\alpha(s)ds - \int_0^\tau b(t,0)\beta(t)dt\\ &- \int_0^\tau \int_0^z \frac{\partial a(t,s)}{\partial t}\omega(t,s)dsdt\\ &- \int_0^\tau \int_0^z \frac{\partial b(t,s)}{\partial s}\omega(t,s)dsdt + \int_0^\tau \int_0^z c(t,s)\omega(t,s)dsdt,\end{aligned}$$

which is equivalent to the following 2D-VIE:

$$\begin{aligned}\omega(\tau,z) =&g(\tau,z) + \int_0^\tau \kappa_1(t,z)\omega(t,z)dt + \int_0^z \kappa_2(\tau,s)\omega(\tau,s)ds\\ &+ \int_0^\tau \int_0^z \kappa_3(t,s)\omega(t,s)dsdt,\end{aligned} \tag{6}$$

where

$$\begin{aligned}\kappa_1(t,z) :&= -b(t,z),\\ \kappa_2(\tau,s) :&= -a(\tau,s),\\ \kappa_3(t,s) :&= \frac{\partial a(t,s)}{\partial t} + \frac{\partial b(t,s)}{\partial s} - c(t,s),\\ g(\tau,z) :&= \alpha(z) + \beta(\tau) - \omega(0,0) + \int_0^z a(0,s)\alpha(s)ds\\ &+ \int_0^\tau b(t,0)\beta(t)dt + \int_0^\tau \int_0^z f(t,s)dsdt.\end{aligned} \tag{7}$$

3 Taylor collocation method

In this section, we approximate solutions of 2D-VIE (6) in the space

$$S_{p-1}^{(-1)}(\Pi_{N,M}) = \{v : v_{n,m} = v|_{D_{n,m}} \in \pi_{p-1}, n = 0,1,\ldots,N-1; m = 0,1,\ldots,M-1\} \tag{8}$$

of the real bivariate polynomial spline functions of degree (at most) $p-1$ in τ and z. Its dimension is NMp^2. Here, $\Pi_N = \{\tau_i = ih, i = 0,1,\ldots,N\}$ and

$\Pi_M = \{z_j = jk, j = 0, 1, \ldots, M\}$ denote, respectively, uniform partitions of the intervals $[0, T]$ and $[0, Z]$ with the step-sizes given by $h = \frac{T}{N}$ and $k = \frac{Z}{M}$. These partitions are define a grid for D

$$\Pi_{N,M} = \Pi_N \times \Pi_M = \{(\tau_n, z_m), 0 \leq n \leq N, 0 \leq m \leq M\}. \tag{9}$$

Set the subintervals

$$\sigma_n = [\tau_n; \tau_{n+1}), n = 0, 1, \ldots, N-2; \quad \sigma_{N-1} = [\tau_{N-1}, \tau_N],$$

$$\delta_m = [z_m; z_{m+1}), m = 0, 1, \ldots, M-2; \quad \delta_{M-1} = [z_{M-1}, z_M],$$

and $D_{n,m} := \sigma_n \times \delta_m$ for all $n = 0, 1, \ldots, N-1; m = 0, 1, \ldots, M-1$.

To define the collocation solution, we use the Taylor polynomial on each rectangle $D_{n,m}; n = 0, 1, \ldots, N-1; m = 0, 1, \ldots, M-1$. Note that the solution ω of (6) is known at point $(0, 0)$: $\omega(0, 0) = g(0, 0)$.

First, we approximate ω in the rectangle $D_{0,0}$ by the polynomial

$$v_{0,0}(\tau, z) = \sum_{i+j=0}^{p-1} \frac{1}{i!j!} \frac{\partial^{i+j}\omega(0,0)}{\partial\tau^i \partial z^j} \tau^i z^j \; ; \quad (\tau, z) \in D_{0,0}, \tag{10}$$

where $\dfrac{\partial^{i+j}\omega(0,0)}{\partial\tau^i\partial z^j}$ is the exact value of $\dfrac{\partial^{i+j}\omega}{\partial\tau^i\partial z^j}$ at point $(0, 0)$.

Differentiating equation (6) j-times with respect to z and i-times with respect to τ, we obtain

$$\begin{aligned}
\frac{\partial^{i+j}\omega(\tau,z)}{\partial\tau^i\partial z^j} =& \partial_1^{(i)}\partial_2^{(j)} g(\tau, z) \\
&+ \sum_{l=0}^{j}\sum_{\eta=0}^{i-1} \binom{j}{l}\binom{i-1}{\eta} \frac{\partial^{i-1-\eta}}{\partial\tau^{i-1-\eta}} \left[\partial_2^{(j-l)}\kappa_1(\tau, z)\right] \frac{\partial^{\eta+l}\omega(\tau,z)}{\partial\tau^\eta\partial z^l} \\
&+ \sum_{l=0}^{j-1}\sum_{\eta=0}^{i} \binom{r}{l}\binom{i}{\eta} \frac{\partial^{i-\eta}}{\partial\tau^{i-\eta}} \left[\frac{\partial^{j-1-l}}{\partial z^{j-1-l}}\kappa_2(\tau, z)\right] \frac{\partial^{\eta+l}\omega(\tau,z)}{\partial\tau^\eta\partial z^l} \\
&+ \sum_{l=0}^{j-1}\sum_{\eta=0}^{i-1} \binom{j-1}{l}\binom{i-1}{\eta} \frac{\partial^{i-1-\eta}}{\partial\tau^{i-1-\eta}} \left[\frac{\partial^{j-1-l}}{\partial z^{j-1-l}}\kappa_3(\tau, z)\right] \frac{\partial^{\eta+l}\omega(\tau,z)}{\partial\tau^\eta\partial z^l}.
\end{aligned} \tag{11}$$

Second, we approximate ω in the rectangles $D_{n,0}, n = 1, \ldots, N-1$ by the polynomials

$$v_{n,0}(\tau, z) = \sum_{i+j=0}^{p-1} \frac{1}{i!j!} \frac{\partial^{i+j}\hat{v}_{n,0}(\tau_n, 0)}{\partial\tau^i\partial z^j} (\tau - \tau_n)^i z^j \; ; \quad (\tau, z) \in D_{n,0}, \tag{12}$$

where $\hat{v}_{n,0}$ is the exact solution of the integral equation:

$$\begin{aligned}\hat{v}_{n,0}(\tau,z) =& g(\tau,z)+\int_0^z \kappa_2(\tau,s)\hat{v}_{n,0}(\tau,s)ds\\ &+\sum_{\xi=0}^{n-1}\int_{\tau_\xi}^{\tau_{\xi+1}}\kappa_1(t,z)v_{\xi,0}(t,z)dt+\int_{\tau_n}^{\tau}\kappa_1(t,z)\hat{v}_{n,0}(t,z)dt\\ &+\sum_{\xi=0}^{n-1}\int_{\tau_\xi}^{\tau_{\xi+1}}\int_0^z\kappa_3(t,s)v_{\xi,0}(t,s)dsdt+\int_{\tau_n}^{\tau}\int_0^z\kappa_3(t,s)\hat{v}_{n,0}(t,s)dsdt,\end{aligned} \tag{13}$$

and $\frac{\partial^{i+j}\hat{v}_{n,0}(\tau_n,0)}{\partial\tau^i\partial z^j}$ is the exact value of $\frac{\partial^{i+j}\hat{v}_{n,0}}{\partial\tau^i\partial z^j}$ at point $(\tau_n,0)$.

Differentiating equation (13) j-times with respect to z and i-times with respect to τ, we obtain

$$\begin{aligned}\frac{\partial^{i+j}\hat{v}_{n,0}(\tau,z)}{\partial\tau^i\partial z^j} =& \partial_1^{(i)}\partial_2^{(j)}g(\tau,z)\\ &+\sum_{l=0}^{j}\sum_{\eta=0}^{i-1}\binom{j}{l}\binom{i-1}{\eta}\frac{\partial^{i-1-\eta}}{\partial\tau^{i-1-\eta}}\left[\partial_2^{(j-l)}\kappa_1(\tau,z)\right]\frac{\partial^{\eta+l}\hat{v}_{n,0}(\tau,z)}{\partial\tau^\eta\partial z^l}\\ &+\sum_{l=0}^{j-1}\sum_{\eta=0}^{i}\binom{j-1}{l}\binom{i}{\eta}\frac{\partial^{i-\eta}}{\partial\tau^{i-\eta}}\left[\frac{\partial^{j-1-l}}{\partial z^{j-1-l}}\kappa_2(\tau,z)\right]\frac{\partial^{\eta+l}\hat{v}_{n,0}(\tau,z)}{\partial\tau^\eta\partial z^l}\\ &+\sum_{l=0}^{j-1}\sum_{\eta=0}^{i-1}\binom{j-1}{l}\binom{i-1}{\eta}\frac{\partial^{i-1-\eta}}{\partial\tau^{i-1-\eta}}\left[\frac{\partial^{j-1-l}}{\partial z^{j-1-l}}\kappa_3(\tau,z)\right]\frac{\partial^{\eta+l}\hat{v}_{n,0}(\tau,z)}{\partial\tau^\eta\partial z^l}.\end{aligned} \tag{14}$$

Third, we approximate ω by $v_{n,m}$ in the rectangles $D_{n,m}, n=0,1,\ldots,N-1$ and $m=1,\ldots,M-1$ such that

$$v_{n,m}(\tau,z)=\sum_{i+j=0}^{p-1}\frac{1}{i!j!}\frac{\partial^{i+j}\hat{v}_{n,m}(\tau_n,z_m)}{\partial\tau^i\partial z^j}(\tau-\tau_n)^i(z-z_m)^j\ ;\quad (\tau,z)\in D_{n,m}, \tag{15}$$

where $\hat{v}_{n,m}$ is the exact solution of the integral equation:

$$\begin{aligned}\hat{v}_{n,m}(\tau,z) =& g(\tau,z)+\sum_{\xi=0}^{n-1}\int_{\tau_\xi}^{\tau_{\xi+1}}\kappa_1(t,z)v_{\xi,m}(t,z)dt+\int_{\tau_n}^{\tau}\kappa_1(t,z)\hat{v}_{n,m}(t,z)dt\\ &+\sum_{\rho=0}^{m-1}\int_{z_\rho}^{z_{\rho+1}}\kappa_2(\tau,s)v_{n,\rho}(\tau,s)ds+\int_{z_m}^{z}\kappa_2(\tau,s)\hat{v}_{n,m}(\tau,s)ds\\ &+\sum_{\xi=0}^{n-1}\sum_{\rho=0}^{m-1}\int_{\tau_\xi}^{\tau_{\xi+1}}\int_{z_\rho}^{z_{\rho+1}}\kappa_3(t,s)v_{\xi,\rho}(t,s)dsdt\end{aligned}$$

$$
\begin{aligned}
&+\sum_{\xi=0}^{n-1}\int_{\tau_\xi}^{\tau_{\xi+1}}\int_{z_m}^{z}\kappa_3(t,s)v_{\xi,m}(t,s)dsdt\\
&+\sum_{\rho=0}^{m-1}\int_{\tau_n}^{\tau}\int_{z_\rho}^{z_{\rho+1}}\kappa_3(t,s)v_{n,\rho}(t,s)dsdt\\
&+\int_{\tau_n}^{\tau}\int_{z_m}^{z}\kappa_3(t,s)\hat{v}_{n,m}(t,s)dsdt,
\end{aligned}
\tag{16}
$$

and $\frac{\partial^{i+j}\hat{v}_{n,0}(\tau_n,z_m)}{\partial\tau^i\partial z^j}$ is the exact value of $\frac{\partial^{i+j}\hat{v}_{n,0}}{\partial\tau^i\partial z^j}$ at point (τ_n,z_m).

Differentiating equation (16) j-times with respect to z and i-times with respect to τ, we obtain

$$
\begin{aligned}
&\frac{\partial^{i+j}\hat{v}_{n,m}(\tau,z)}{\partial\tau^i\partial z^j}\\
&=\partial_1^{(i)}\partial_2^{(j)}g(\tau,z)\\
&\quad+\sum_{l=0}^{j}\sum_{\eta=0}^{i-1}\binom{j}{l}\binom{i-1}{\eta}\frac{\partial^{i-1-\eta}}{\partial\tau^{i-1-\eta}}\left[\partial_2^{(j-l)}\kappa_1(\tau,z)\right]\frac{\partial^{\eta+l}\hat{v}_{n,m}(\tau,z)}{\partial\tau^\eta\partial z^l}\\
&\quad+\sum_{l=0}^{j-1}\sum_{\eta=0}^{i}\binom{j-1}{l}\binom{i}{\eta}\frac{\partial^{i-\eta}}{\partial\tau^{i-\eta}}\left[\frac{\partial^{j-1-l}}{\partial z^{j-1-l}}\kappa_2(\tau,z)\right]\frac{\partial^{\eta+l}\hat{v}_{n,m}(\tau,z)}{\partial\tau^\eta\partial z^l}\\
&\quad+\sum_{l=0}^{j-1}\sum_{\eta=0}^{i-1}\binom{j-1}{l}\binom{i-1}{\eta}\frac{\partial^{i-1-\eta}}{\partial\tau^{i-1-\eta}}\left[\frac{\partial^{j-1-l}}{\partial z^{j-1-l}}\kappa_3(\tau,z)\right]\frac{\partial^{\eta+l}\hat{v}_{n,m}(\tau,z)}{\partial\tau^\eta\partial z^l}.
\end{aligned}
\tag{17}
$$

4 Convergence analysis

We consider the space $L^\infty(D)$ with the norm

$$\|\varphi\|_{L^\infty(D)}=\inf\{C\in\mathbb{R}:|\varphi(\tau,z)|\le C\ \text{ for a.e. }(\tau,z)\in D\}<\infty.$$

The following lemmas will be used in proving the convergence of the presented method.

Lemma 1 (Taylor's theorem for functions of two independent variables)**.** Let f be p times continuously differentiable on $D=[a,b]\times[c,d]$ and let $(\tau_0,z_0)\in D$. Then for all $(\tau,z)\in D$, we have

$$f(\tau,z)=\sum_{i+j=0}^{p-1}\frac{1}{i!j!}\frac{\partial^{i+j}f(\tau_0,z_0)}{\partial\tau^i\partial z^j}(\tau-\tau_0)^i(z-z_0)^j + \sum_{i+j=p}\frac{1}{i!j!}\frac{\partial^{i+j}f(\tau_1,z_1)}{\partial\tau^i\partial z^j}(\tau-\tau_0)^i(z-z_0)^j,$$

where

$$\begin{cases}\tau_1=\theta\tau+(1-\theta)\tau_0\in[a,b],\\ z_1=\theta z+(1-\theta)z_0\in[c,d],\end{cases}\quad\theta\in(0,1).$$

Lemma 2 (Gronwall-type inequality [21]). Let $\omega(\tau,z)$ and $p(\tau,z)$ be nonnegative continuous functions in $\Omega=[a,b]\times[c,d]$, and let $p(\tau,z)$ be nondecreasing in each of the variables in Ω and satisfy the following inequality:

$$\omega(\tau,z)\leq p(\tau,z)+\kappa\int_a^\tau\omega(t,z)dt+\kappa\int_c^z\omega(\tau,s)ds + \kappa\int_a^\tau\int_c^z\omega(t,s)dsdt,\qquad(\tau,z)\in\Omega,$$

where κ is positive constant. Then there exists a positive constant ν such that

$$\omega(\tau,z)\leq\nu p(\tau,z).$$

Lemma 3 (Discrete Gronwall-type inequality [5]). Let $\{\kappa_j\}_{j=0}^n$ be a given nonnegative sequence and let the sequence $\{\varepsilon_n\}$ satisfy$\varepsilon_0\leq p_0$ and

$$\varepsilon_n\leq p_0+\sum_{j=0}^{n-1}\kappa_j\varepsilon_j,\quad n\geq1,$$

with $p_0\geq0$. Then

$$\varepsilon_n\leq p_0\exp\left(\sum_{j=0}^{n-1}\kappa_j\right),\quad n\geq1.$$

Lemma 4 (Discrete Gronwall-type inequality of two variables [23]). Let $\omega_{n,m}$ be a given nonnegative sequence, and let b_1,b_2,b_3 and β be independent of h and k and strictly positive. If the sequence $\omega_{n,m}$ satisfies

$$\omega_{n,m} \leq hb_1 \sum_{\xi=0}^{n-1} \omega_{\xi,m} + kb_2 \sum_{\rho=0}^{m-1} \omega_{n,\rho} + hkb_3 \sum_{\xi=0}^{n-1} \sum_{\rho=0}^{m-1} \omega_{\xi,\rho} + \beta, \tag{18}$$

for all $n = 0, 1, \ldots, N, m = 0, 1, \ldots, M$, then

$$\omega_{n,m} \leq \beta \exp(\gamma(Nh + Mk)), \tag{19}$$

where $\gamma = \frac{1}{2}\left(b_1 + b_2 + \sqrt{(b_1 + b_2)^2 + 4b_3}\right)$.

Lemma 5. Let g, κ_1,κ_2, and κ_3 be p times continuously differentiable on their respective domains. Then, there exists a positive number $\alpha(p)$ such that for all $n = 0, 1, \ldots, N-1$, $m = 0, \ldots, M-1$, and $i + j = 0, 1, \ldots, p$, we have

$$\left\| \frac{\partial^{i+j} \hat{v}_{n,m}}{\partial \tau^i \partial z^j} \right\|_{L^\infty(D_{n,m})} \leq \alpha(p),$$

where $\hat{v}_{0,0}(\tau, z) = \omega(\tau, z)$ for $(\tau, z) \in D_{0,0}$.

Proof. Let $a_{n,m}^{i+j} = \| \frac{\partial^{i+j} \hat{v}_{n,m}}{\partial \tau^i \partial z^j} \|_{L^\infty(D_{n,m})}$.

First, we have for all $i + j = 0, 1, \ldots, p$,

$$a_{0,0}^{i+j} \leq \max \left\{ \left\| \frac{\partial^{i+j} \omega}{\partial \tau^i \partial z^j} \right\|_{L^\infty(D_{0,0})}, i + j = 0, 1, \ldots, p \right\} = \alpha_1(p). \tag{20}$$

Second, for $i + j = 0$, from (13), we have for all $n = 1, \ldots, N-1$

$$\begin{aligned} a_{n,0}^{0+0} \leq & c + c_{1,1} \sum_{\xi=0}^{n-1} \int_{\tau_\xi}^{\tau_{\xi+1}} \sum_{\alpha+\beta=0}^{p-1} a_{\xi,0}^{\alpha+\beta} dt + c_{1,1} \int_{\tau_n}^{\tau} a_{n,0}^{0+0} dt + c_{2,2} \int_0^z a_{n,0}^{0+0} ds \\ & + c_{3,2} \sum_{\xi=0}^{n-1} \int_{\tau_\xi}^{\tau_{\xi+1}} \int_0^z \sum_{\alpha+\beta=0}^{p-1} a_{\xi,0}^{\alpha+\beta} ds dt + c_{3,2} \int_{\tau_n}^{\tau} \int_0^z a_{n,0}^{0+0} ds dt, \end{aligned}$$

and for $i + j = 1, \ldots, p$, from (14), we have for all $n = 1, \ldots, N-1$,

$$a_{n,0}^{i+j} \leq c + c_{1,2} \sum_{l=0}^{j} \sum_{\eta=0}^{i-1} a_{n,0}^{\eta+l} + c_{2,1} \sum_{l=0}^{j-1} \sum_{\eta=0}^{i} a_{n,0}^{\eta+l} + c_{3,3} \sum_{l=0}^{j-1} \sum_{\eta=0}^{i-1} a_{n,0}^{\eta+l},$$

where, the constants $c, c_{1,1}, c_{1,2}, c_{2,1}, c_{2,2}, c_{3,2}$, and $c_{3,3}$ are positive and independent of N and M.

Hence, for all $i + j = 0, 1, \ldots, p$,

$$\begin{aligned} a_{n,0}^{i+j} \leq & c + c_{1,1}h \sum_{\xi=0}^{n-1} \sum_{\alpha+\beta=0}^{p-1} a_{\xi,0}^{\alpha+\beta} + c_{1,2} \sum_{l=0}^{j} \sum_{\eta=0}^{i-1} a_{n,0}^{\eta+l} + c_{1,1}h \sum_{l=0}^{j} a_{n,0}^{0+l} \\ & + c_{2,1} \sum_{l=0}^{j-1} \sum_{\eta=0}^{i} a_{n,0}^{\eta+l} + c_{2,2}k \sum_{\eta=0}^{i} a_{n,0}^{\eta+0} \\ & + c_{3,2}hk \sum_{\xi=0}^{n-1} \sum_{\alpha+\beta=0}^{p-1} a_{\xi,0}^{\alpha+\beta} + c_{3,3} \sum_{l=0}^{j-1} \sum_{\eta=0}^{i-1} a_{n,0}^{\eta+l} + c_{3,2}hk a_{n,0}^{0+0}, \end{aligned}$$

which implies that

$$a_{n,0}^{i+j} \leq c + c_1 h \sum_{\xi=0}^{n-1} \sum_{\eta+l=0}^{p-1} a_{\xi,0}^{\eta+l} + c_2 \sum_{\eta+l=0}^{i+j-1} a_{n,0}^{\eta+l}, \tag{21}$$

where $c_1 = c_{1,1} + c_{3,2}k$, $c_2 = c_{1,2} + c_{1,1}h + c_{2,1} + c_{2,2}k + c_{3,3} + c_{3,2}hk$.

Now, we consider the sequence $\Gamma_n = \max\{a_{n,0}^{i+j}, i+j = 0, \ldots, p\}$, $n = 0, 1, \ldots, N-1$, from (21), we obtain

$$a_{n,0}^{i+j} \leq c + c_1 p^2 h \sum_{\xi=0}^{n-1} \Gamma_\xi + c_2 \sum_{\eta+l=0}^{i+j-1} a_{n,0}^{\eta+l}.$$

Using Lemma 3 with the following notations

$$\varepsilon_{i+j} = a_{n,0}^{i+j}, \quad p_0 = c + c_1 p^2 h \sum_{\xi=0}^{n-1} \Gamma_\xi, \quad k_{\eta+l} = c_2,$$

we obtain

$$\begin{aligned} a_{n,0}^{i+j} &\leq \left(c + c_1 p^2 h \sum_{\xi=0}^{n-1} \Gamma_\xi \right) \exp\left(\sum_{\eta+l=0}^{i+j-1} c_2 \right) \\ &\leq \underbrace{c \exp(p^2 c_2)}_{c_3} + \underbrace{c_1 p^2 \exp(p^2 c_2)}_{c_4} h \sum_{\xi=0}^{n-1} \Gamma_\xi, \end{aligned}$$

which implies that

$$\Gamma_n \leq c_3 + c_4 h \sum_{\xi=0}^{n-1} \Gamma_\xi.$$

Again, using Lemma 3, for all $n = 0, 1, \ldots, N-1$ and $i+j = 0, \ldots, p$

$$a_{n,0}^{i+j} \leq \Gamma_n \leq \underbrace{c_3 \exp(a c_4)}_{\alpha_2(p)}. \tag{22}$$

Third, for $i+j=0$, from (16), we have for all $n=0,1,\ldots,N-1$, $m=1,\ldots,M-1$,

$$\begin{aligned}a_{n,m}^{0+0} \leq & c + c_{1,1}h\sum_{\xi=0}^{n-1}\sum_{\alpha+\beta=0}^{p-1} a_{\xi,m}^{0+0} + c_{1,1}ha_{n,m}^{0+0} + c_{2,2}k\sum_{\rho=0}^{m-1}\sum_{\alpha+\beta=0}^{p-1} a_{n,\rho}^{0+0} + c_{2,2}ka_{n,m}^{0+0} \\ & + c_{3,2}hk\sum_{\xi=0}^{n-1}\sum_{\rho=0}^{m-1}\sum_{\alpha+\beta=0}^{p-1} a_{\xi,\rho}^{0+0} + c_{3,2}hk\sum_{\xi=0}^{n-1}\sum_{\alpha+\beta=0}^{p-1} a_{\xi,m}^{0+0} \\ & + c_{3,2}hk\sum_{\rho=0}^{m-1}\sum_{\alpha+\beta=0}^{p-1} a_{n,\rho}^{0+0} + c_{3,2}hka_{n,m}^{0+0},\end{aligned}$$

and for $i+j=1,\ldots,p$, we have from (17) that

$$a_{n,m}^{i+j} \leq c + c_{1,2}\sum_{l=0}^{j}\sum_{\eta=0}^{i-1} a_{n,m}^{\eta+l} + c_{2,1}\sum_{l=0}^{j-1}\sum_{\eta=0}^{i} a_{n,m}^{\eta+l} + c_{3,3}\sum_{l=0}^{j-1}\sum_{\eta=0}^{i-1} a_{n,m}^{\eta+l}.$$

Hence, for all $i+j=0,1,\ldots,p$,

$$\begin{aligned}a_{n,m}^{i+j} \leq & c + c_{1,1}h\sum_{\xi=0}^{n-1}\sum_{\eta+l=0}^{p-1} a_{\xi,m}^{\eta+l} + c_{1,2}\sum_{\eta+l=0}^{i+j-1} a_{n,m}^{\eta+l} + c_{1,1}h\sum_{\eta+l=0}^{i+j-1} a_{n,m}^{\eta+l} \\ & + c_{2,2}k\sum_{\rho=0}^{m-1}\sum_{\eta+l=0}^{p-1} a_{n,\rho}^{\eta+l} + c_{2,1}\sum_{\eta+l=0}^{i+j-1} a_{n,m}^{\eta+l} + c_{2,2}k\sum_{\eta+l=0}^{i+j-1} a_{n,m}^{\eta+l} \\ & + c_{3,2}hk\sum_{\xi=0}^{n-1}\sum_{\rho=0}^{m-1}\sum_{\eta+l=0}^{p-1} a_{\xi,\rho}^{\eta+l} + c_{3,2}hk\sum_{\xi=0}^{n-1}\sum_{\eta+l=0}^{p-1} a_{\xi,m}^{\eta+l} \\ & + c_{3,2}hk\sum_{\rho=0}^{m-1}\sum_{\eta+l=0}^{p-1} a_{n,\rho}^{\eta+l} \\ & + c_{3,3}\sum_{\eta+l=0}^{i+j-1} a_{n,m}^{\eta+l} + c_{3,2}hk\sum_{\eta+l=0}^{i+j-1} a_{n,m}^{\eta+l}. \end{aligned} \tag{23}$$

Consider, the sequence $\Gamma_{n,m} = \max\{a_{n,m}^{i+j}, i+j=0,\ldots,p\}$, $n=0,1,\ldots,N-1; m=0,1,\ldots,M-1$. Then by (23), we have

$$a_{n,m}^{i+j} \leq c + hb_{1,1}\sum_{\xi=0}^{n-1}\Gamma_{\xi,m} + kb_{1,2}\sum_{\rho=0}^{m-1}\Gamma_{n,\rho}$$

$$+ hkb_{1,3} \sum_{\xi=0}^{n-1} \sum_{\rho=0}^{m-1} \Gamma_{\xi,\rho} + b_{1,4} \sum_{\eta+l=0}^{i+j-1} a_{n,m}^{\eta+l}, \tag{24}$$

where $b_{1,1} = (c_{1,1} + c_{3,2}k)p^2$, $b_{1,2} = (c_{2,2} + c_{3,2}h)p^2$, $b_{1,3} = c_{3,2}p^2$ and $b_{1,4} = c_{1,2} + c_{1,1}h + c_{2,1} + c_{2,2}k + c_{3,3} + c_{3,2}hk$.

Using Lemma 3 with the following notations

$$\varepsilon_{i+j} = a_{n,m}^{i+j},$$
$$p_0 = c + hb_{1,1} \sum_{\xi=0}^{n-1} \Gamma_{\xi,m} + kb_{1,2} \sum_{\rho=0}^{m-1} \Gamma_{n,\rho} + hkb_{1,3} \sum_{\xi=0}^{n-1} \sum_{\rho=0}^{m-1} \Gamma_{\xi,\rho},$$
$$k_{\eta+l} = b_{1,4},$$

we obtain from (24) that

$$a_{n,m}^{i+j} \leq \underbrace{c \exp(pb_{1,4})}_{b_4} + h \underbrace{b_{1,1} \exp(pb_{1,4})}_{b_1} \sum_{\xi=0}^{n-1} \Gamma_{\xi,m} + k \underbrace{b_{1,2} \exp(pb_{1,4})}_{b_2} \sum_{\rho=0}^{m-1} \Gamma_{n,\rho}$$
$$+ hk \underbrace{b_{1,3} \exp(pb_{1,4})}_{b_3} \sum_{\xi=0}^{n-1} \sum_{\rho=0}^{m-1} \Gamma_{\xi,\rho}.$$

It follows that, for all $n = 0, 1, \ldots, N-1; m = 0, 1, \ldots, M-1$,

$$\Gamma_{n,m} \leq b_4 + hb_1 \sum_{\xi=0}^{n-1} \Gamma_{\xi,m} + kb_2 \sum_{\rho=0}^{m-1} \Gamma_{n,\rho} + hkb_3 \sum_{\xi=0}^{n-1} \sum_{\rho=0}^{m-1} \Gamma_{\xi,\rho}.$$

Using Lemma 4, we obtain, for all $n = 0, 1, \ldots, N-1, m = 0, 1, \ldots, N-1$, that

$$\Gamma_{n,m} \leq b_4 \exp(\gamma_1(Nh + Mk)) \leq \underbrace{b_4 \exp(\gamma_1(a+b))}_{\alpha_3(p)}, \tag{25}$$

where $\gamma_1 = \frac{1}{2}\left(b_1 + b_2 + \sqrt{(b_1+b_2)^2 + 4b_3}\right)$.

Hence, from (20), (22), and (25), by setting

$$\alpha(p) = \max\{\alpha_1(p), \alpha_2(p), \alpha_3(p)\},$$

the proof of Lemma 5 is completed. □

Theorem 1. Let g, κ_1, κ_2, and κ_3 be p times continuously differentiable on their respective domains. Then equations (10), (12), (15) define a unique approximation $v \in S_{p-1}^{(-1)}(\Pi_{N,M})$, and the resulting error function $e(\tau, z) = \omega(\tau, z) - v(\tau, z)$ satisfies

$$\|e\|_{L^\infty(D)} \leq C(h+k)^p,$$

where C is a finite constant independent of h and k.

Proof. Define the error $e(\tau, z)$ on $D_{n,m}$ by $e_{n,m}(\tau, z) = \omega(\tau, z) - v_{n,m}(\tau, z)$ for all $n = 0, 1, \ldots, N-1$ and $m = 0, 1, \ldots, M-1$.

The proof is split into three steps.

Claim 1. There exists a constant C_1 independent of h and k such that

$$\|e_{0,0}\|_{L^\infty(D_{0,0})} \leq C_1(h+k)^p.$$

Let $(\tau, z) \in D_{0,0}$. By using Lemma 1, we obtain from (10) that

$$|e_{0,0}(\tau, z)| \leq \sum_{i+j=p} \frac{1}{i!j!} \left\| \frac{\partial^{i+j}\omega}{\partial\tau^i \partial z^j} \right\| h^i k^j.$$

Hence, by Lemma 5, we have

$$|e_{0,0}(\tau, z)| \leq \alpha(p) \sum_{i+j=p} \frac{1}{i!j!} h^i k^j = \underbrace{\frac{\alpha(p)}{p!}}_{C_1} (h+k)^p. \tag{26}$$

Claim 2. There exists a constant C_2 independent of h and k such that

$$\|e_{n,0}\|_{L^\infty(D_{n,0})} \leq C_2(h+k)^p,$$

for all $n = 1, \ldots, N-1$. Let $(\tau, z) \in D_{n,0}$, we have from (13) that

$$\begin{aligned}\omega(\tau,z)-\hat{v}_{n,0}(\tau,z)=&\int_0^z \kappa_2(\tau,s)(\omega(\tau,s)-\hat{v}_{n,0}(\tau,s))ds\\&+\sum_{\xi=0}^{n-1}\int_{\tau_\xi}^{\tau_{\xi+1}}\kappa_1(z,t)e_{\xi,0}(t,z)dt\\&+\int_{\tau_n}^{\tau}\kappa_1(z,t)(\omega(t,z)-\hat{v}_{n,0}(t,z))dt\\&+\sum_{\xi=0}^{n-1}\int_{\tau_\xi}^{\tau_{\xi+1}}\int_0^z\kappa_3(t,s)e_{\xi,0}(t,s)dsdt\\&+\int_{\tau_n}^{\tau}\int_0^z\kappa_3(t,s)(\omega(t,s)-\hat{v}_{n,0}(t,s))dsdt.\end{aligned}$$

Hence,

$$\begin{aligned}|\omega(\tau,z)-\hat{v}_{n,0}(\tau,z)|\le&\sum_{\xi=0}^{n-1}h\kappa\|e_{\xi,0}\|_{L^\infty(D_{\xi,0})}+\sum_{\xi=0}^{n-1}hk\kappa\|e_{\xi,0}\|_{L^\infty(D_{\xi,0})}\\&+\kappa\int_{\tau_n}^{\tau}|\omega(t,z)-\hat{v}_{n,0}(t,z)|dt\\&+\kappa\int_0^z|\omega(\tau,s)-\hat{v}_{n,0}(\tau,s)|ds\\&+\kappa\int_{\tau_n}^{\tau}\int_0^z|\omega(t,s)-\hat{v}_{n,0}(t,s)|dsdt,\end{aligned}$$

where $\kappa=\max\{\|\kappa_i\|_{L^\infty(D)},i=1,2,3\}$. Then by Lemma 2, we have

$$\begin{aligned}|\omega(\tau,z)-\hat{v}_{n,0}(\tau,z)|&\le\left(\sum_{\xi=0}^{n-1}h\kappa\|e_{\xi,0}\|_{L^\infty(D_{\xi,0})}+\sum_{\xi=0}^{n-1}hk\kappa\|e_{\xi,0}\|_{L^\infty(D_{\xi,0})}\right)\nu\\&\le\sum_{\xi=0}^{n-1}h\underbrace{\kappa(1+b)\nu}_{\lambda_1}\|e_{\xi,0}\|_{L^\infty(D_{\xi,0})},\end{aligned}$$

which implies, by using Lemma 1, that

$$\begin{aligned}\|e_{n,0}\|_{L^\infty(D_{n,0})}&\le\|\omega-\hat{v}_{n,0}\|+\|\hat{v}_{n,0}-v_{n,0}\|\\&\le\sum_{\xi=0}^{n-1}h\lambda_1\|e_{\xi,0}\|_{L^\infty(D_{\xi,0})}+\sum_{i+j=p}\frac{1}{i!j!}\left\|\frac{\partial^{i+j}\hat{v}_{n,0}}{\partial\tau^i\partial z^j}\right\|h^ik^j.\end{aligned}$$

Hence, by Lemma 5, we obtain

$$\|e_{n,0}\|_{L^\infty(D_{n,0})} \leq \sum_{\xi=0}^{n-1} h\lambda_1 \|e_{\xi,0}\|_{L^\infty(D_{\xi,0})} + \frac{\alpha(p)}{p!}(h+k)^p.$$

Then, by Lemma 3, we have

$$\|e_{n,0}\|_{L^\infty(D_{n,0})} \leq \underbrace{\frac{\alpha(p)}{p!}\exp(T\lambda_1)}_{C_2}(h+k)^p.$$

Claim 3. There exists a constant C_3 independent of h and k such that

$$\|e_{n,m}\|_{L^\infty(D_{n,m})} \leq C_3(h+k)^p,$$

for all $n = 0, 1, \ldots, N-1$ and $m = 1, \ldots, M-1$. Let $(\tau, z) \in D_{n,m}$. Then from (16) we have

$$\begin{aligned}
|\omega(\tau,z) - \hat{v}_{n,m}(\tau,z)| \leq & \sum_{\xi=0}^{n-1} h\kappa\|e_{\xi,m}\| + \sum_{\rho=0}^{m-1} k\kappa\|e_{n,\rho}\| \\
& + \sum_{\xi=0}^{n-1}\sum_{\rho=0}^{m-1} hk\kappa\|e_{\xi,\rho}\| + \sum_{\xi=0}^{n-1} hk\kappa\|e_{\xi,m}\| + \sum_{\rho=0}^{m-1} hk\kappa\|e_{n,\rho}\| \\
& + \kappa\int_{\tau_n}^{\tau} |\omega(t,z) - \hat{v}_{n,m}(t,z)|dt \\
& + \kappa\int_{z_m}^{z} |\omega(\tau,s) - \hat{v}_{n,m}(\tau,s)|ds \\
& + \kappa\int_{\tau_n}^{\tau}\int_{z_m}^{z} |\omega(t,s) - \hat{v}_{n,m}(t,s)|dsdt.
\end{aligned}$$

Then by Lemma 2,

$$\begin{aligned}
|\omega(\tau,z) - \hat{v}_{n,m}(\tau,z)| \leq & \sum_{\xi=0}^{n-1} h\underbrace{\kappa(1+k)\nu}_{\lambda_2}\|e_{\xi,m}\| + \sum_{\rho=0}^{m-1} k\underbrace{\kappa(1+h)\nu}_{\lambda_3}\|e_{n,\rho}\| \\
& + \sum_{\xi=0}^{n-1}\sum_{\rho=0}^{m-1} hk\underbrace{\kappa\nu}_{\lambda_4}\|e_{\xi,\rho}\|,
\end{aligned}$$

which implies, by using Lemma 1, that

$$\begin{aligned}\|e_{n,m}\|_{L^\infty(D_{n,0})} \leq & \|\omega - \hat{v}_{n,m}\| + \|\hat{v}_{n,m} - v_{n,m}\| \\ \leq & \sum_{\xi=0}^{n-1} h\lambda_2 \|e_{\xi,m}\| + \sum_{\rho=0}^{m-1} k\lambda_3 \|e_{n,\rho}\| + \sum_{\xi=0}^{n-1}\sum_{\rho=0}^{m-1} hk\lambda_4 \|e_{\xi,\rho}\| \\ & + \sum_{i+j=p} \frac{1}{i!j!} \left\| \frac{\partial^{i+j} \hat{v}_{n,m}}{\partial \tau^i \partial z^j} \right\| h^i k^j.\end{aligned}$$

Hence, by Lemma 5, we obtain

$$\begin{aligned}\|e_{n,m}\| \leq & \sum_{\xi=0}^{n-1} h\lambda_2 \|e_{\xi,m}\| + \sum_{\rho=0}^{m-1} k\lambda_3 \|e_{n,\rho}\| + \sum_{\xi=0}^{n-1}\sum_{\rho=0}^{m-1} hk\lambda_4 \|e_{\xi,\rho}\| \\ & + \frac{\alpha(p)}{p!}(h+k)^p.\end{aligned} \tag{27}$$

Using Lemma 4, we obtain

$$\|e_{n,m}\| \leq \underbrace{\frac{\alpha(p)}{p!} \exp(\gamma_3(T+Z))}_{C_3} (h+k)^p, \tag{28}$$

where $\gamma_3 = \frac{1}{2}\left(\lambda_2 + \lambda_3 + \sqrt{(\lambda_2+\lambda_3)^2 + 4\lambda_3}\right)$.

Thus, the proof is completed by taking $C = \max\{C_1, C_2, C_3\}$. □

5 Numerical examples

The method presented in this paper is used to find numerical solutions to some illustrative examples. Our results are compared with the exact solutions by calculating the absolute error function $|e_{n,m}| = |\omega - v_{n,m}|$ for all $n = 0, 1, \ldots, N-1$ and $m = 0, 1, \ldots, M-1$, where ω and v are the exact and approximate solution, respectively. The values of errors are computed for different values of p, N, M and collected in Tables 1, 3, and 4, which are displayed in Figure 1 for Example 1. In Table 5, the presented method is compared with the numerical results obtained by using the Chelyshkov polynomials method (2D-CPs) [20] and the two-dimensional block-pulse functions method (2D-BPFs) [22]. Moreover, the exact and approximate solution over the region $([0,1] \times [0,1])$ are displayed in Figure 2 for Example 2. The re-

sults in these examples confirm the theoretical estimates and suggest that the experimental order of convergence (EOC) is p (see Table 2).

Example 1. Considering the Goursat problem, which is linear and homogeneous in hyperbolic PDE [9],

$$3\frac{\partial^2\omega(t,s)}{\partial s\partial t} = \frac{\partial\omega(t,s)}{\partial t} + \frac{\partial\omega(t,s)}{\partial s} + \omega(t,s), \quad (t,s) \in [0,1]\times[0,1],$$

with initial conditions $\omega(t,0) = e^t$ and $\omega(0,s) = e^s$. This equation is equivalent to the linear two-dimensional VIE defined as follows:

$$\begin{aligned}\omega(\tau,z) =& \frac{2}{3}(e^\tau + e^z) - \frac{1}{3} \\ &+ \frac{1}{3}\left(\int_0^\tau \omega(t,z)dt + \int_0^z \omega(\tau,s)ds + \int_0^\tau\int_0^z \omega(t,s)dsdt\right). \quad (29)\end{aligned}$$

The exact solution is $\omega(\tau,z) = e^{\tau+z}$.

The numerical results for $p = 3, 4$ and $h = k = 0.05, 0.025$ of the Taylor collocation method are presented in Table 1 and Figure 1.

Table 1: Comparison between the approximate and the exact solution for Example 1

(τ,z)	$N=M=20$,$p=3$	$N=M=20$,$p=4$	$N=M=40$, $p=3$
$(0.1,0.1)$	$1.62e-05$	$1.57e-05$	$4.05e-06$
$(0.2,0.2)$	$3.82e-05$	$3.65e-05$	$9.47e-06$
$(0.3,0.3)$	$6.67e-05$	$6.37e-05$	$1.66e-05$
$(0.4,0.4)$	$1.05e-04$	$9.89e-05$	$2.59e-05$
$(0.5,0.5)$	$1.56e-04$	$1.44e-04$	$3.80e-05$
$(0.6,0.6)$	$2.21e-04$	$2.01e-04$	$5.35e-05$
$(0.7,0.7)$	$3.04e-04$	$2.74e-04$	$7.34e-05$
$(0.8,0.8)$	$4.11e-04$	$3.66e-04$	$9.86e-05$
$(0.9,0.9)$	$5.48e-04$	$4.82e-04$	$1.30e-04$
$(1.0,1.0)$	$1.99e-03$	$6.39e-04$	$3.38e-04$
CPU time/sec	30.81	47.25	742.68

Example 2. Consider the linear nonhomogeneous Goursat problem [8]

$$\frac{\partial^2\omega(t,s)}{\partial s\partial t} = 4ts - t^2s^2 + \omega(t,s), \quad (t,s) \in [0,1]\times[0,1],$$

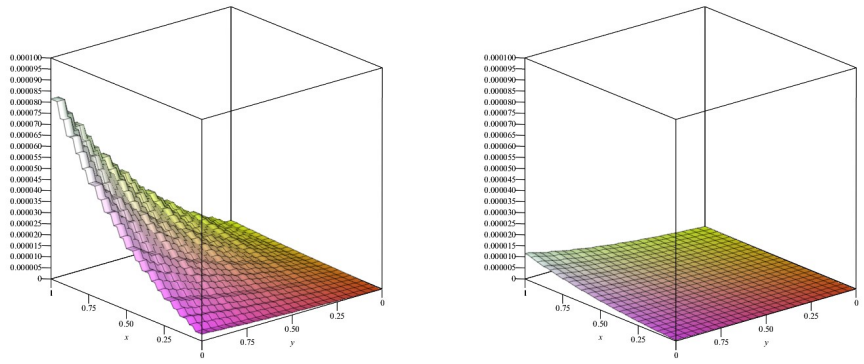


Figure 1: Plot of the absolute error function for Example 1; left: $h = k = 0.05$, right:$h = k = 0.025$.

Table 2: EOC for Example 1

(N, M)	$(2, 2)$	$(4, 4)$	$(8, 8)$	$(16, 16)$	$(32, 32)$	$(64, 64)$
$p = 2$	/	1.45	1.71	1.85	1.94	1.96
$p = 3$	/	2.40	2.69	2.84	2.92	2.96

with initial conditions $\omega(t, 0) = e^t$ and $\omega(0, s) = e^s$. This equation is equivalent to the linear two-dimensional VIE defined as follows:

$$\omega(\tau, z) = e^{\tau} + e^{z} + \tau^2 z^2 - \frac{\tau^3 z^3}{9} + \int_0^{\tau} \int_0^{z} \omega(t, s) ds dt.$$

The exact solution is $\omega(\tau, z) = \tau^2 z^2 + e^{\tau + z}$.

The numerical results for $p = 3$ and $h = k = 0.05, 0.025$ of the Taylor collocation method are presented in Table 3 and Figure 2.

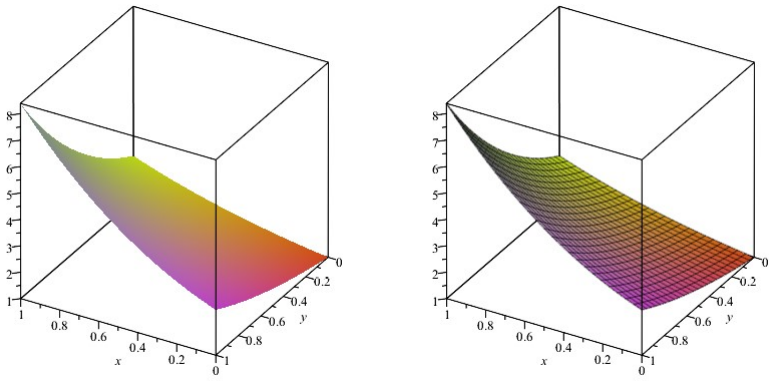
Example 3. Consider the PDE with variable coefficients

$$\begin{aligned} \frac{\partial^2 \omega(t, s)}{\partial s \partial t} =& (t + s^2) \frac{\partial \omega(t, s)}{\partial t} + (t^2 + s) \frac{\partial \omega(t, s)}{\partial s} \\ &+ ts\omega(t, s) + f(t, s), \qquad (t, s) \in [0, 1] \times [0, 1], \end{aligned}$$

with initial conditions $\omega(t, 0) = \cos(t) + e^t$ and $\omega(0, s) = \cos(s) + 1 + s^2$. This equation is equivalent to the linear two-dimensional VIE:

Table 3: Comparison between the approximate and the exact solution for Example 2

(τ, z)	$N = M = 20, p = 3$	$N = M = 40, p = 3$
$(0.1, 0.1)$	$3.64e-07$	$4.80e-08$
$(0.2, 0.2)$	$1.77e-06$	$2.29e-07$
$(0.3, 0.3)$	$4.77e-06$	$6.13e-07$
$(0.4, 0.4)$	$1.00e-05$	$1.28e-06$
$(0.5, 0.5)$	$1.83e-05$	$2.32e-06$
$(0.6, 0.6)$	$3.07e-05$	$3.93e-06$
$(0.7, 0.7)$	$4.85e-05$	$6.21e-06$
$(0.8, 0.8)$	$7.32e-05$	$9.36e-06$
$(0.9, 0.9)$	$1.06e-04$	$1.36e-05$
$(1.0, 1.0)$	$1.76e-03$	$2.28e-04$
CPU time/sec	27.71	701.59



(a) Exact solution (b) Approximate solution

Figure 2: Numerical results of Example 2.

$$\omega(\tau, z) = g(\tau, z) + \int_0^\tau (t^2 + z)\omega(t, z)dt + \int_0^z (\tau + s^2)\omega(\tau, s)ds + \int_0^\tau \int_0^z (-2 - ts)\omega(t, s)dsdt,$$

where $g(\tau, z)$ is chosen so that the exact solution is $\omega(\tau, z) = \cos(\tau + z) + e^\tau + z^2$.

The numerical results for $p = 3$ and $h = k = 0.05, 0.025$ of the Taylor collocation method are presented in Table 4.

Table 4: Comparison between the approximate and the exact solution for Example 3

(τ, z)	$N = M = 20, p = 3$	$N = M = 40, p = 3$
$(0.1, 0.1)$	$1.32e-05$	$1.66e-06$
$(0.2, 0.2)$	$2.84e-05$	$3.73e-06$
$(0.3, 0.3)$	$4.72e-05$	$6.54e-06$
$(0.4, 0.4)$	$7.13e-05$	$1.05e-05$
$(0.5, 0.5)$	$1.03e-04$	$1.62e-05$
$(0.6, 0.6)$	$1.46e-04$	$2.44e-05$
$(0.7, 0.7)$	$2.07e-04$	$3.64e-05$
$(0.8, 0.8)$	$2.98e-04$	$5.47e-05$
$(0.9, 0.9)$	$4.44e-04$	$8.47e-05$
$(1.0, 1.0)$	$1.04e-03$	$1.89e-04$
CPU time/sec	62.96	1207.40

Example 4. Consider the linear nonhomogeneous Goursat problem

$$\frac{\partial^2 u(t,s)}{\partial s \partial t} = -\frac{\partial u(t,s)}{\partial t} - \frac{\partial u(t,s)}{\partial s} - u(t,s) + f(t,s), \quad (t,s) \in [0,1] \times [0,1],$$

with initial conditions $u(t,0) = e^t$ and $u(0,s) = e^{2s}$. This equation is equivalent to the nonlinear two-dimensional VIE of the first kind [22]:

$$\frac{1}{9}(e^{\tau+z} - e^{\tau+4z} - e^{7\tau+z} + e^{7\tau+4z}) = \int_0^\tau \int_0^z 2e^{\tau+z}\omega^3(t,s)dsdt.$$

It is also equivalent to the following linear 2D-VIE of the second kind:

$$u(\tau,z) = g(\tau,z) - \int_0^\tau u(t,z)dt - \int_0^z u(\tau,s)ds - \int_0^\tau \int_0^z u(t,s)dsdt,$$

where $u = \omega^3$ and $g(\tau,z)$ is chosen such that the exact solution is $u(\tau,z) = e^{\tau+2z}$.

In Table 5, the numerical results for $p = 3$ and $N = M = 64$ of the Taylor collocation method are compared with the numerical results obtained by using 2D-CPs [20] and 2D-BPFs [22].

Table 5: Comparison of the absolute errors of Example 4

$(2^{-i}, 2^{-i})$	2D-BPFs [22]	2D-CPs [20]	Present method
$i = 1$	$1.0e-1$	$3.5e-5$	$6.1e-6$
$i = 2$	$4.6e-2$	$2.0e-6$	$2.6e-6$
$i = 3$	$2.9e-2$	$1.5e-5$	$1.3e-6$
$i = 4$	$2.3e-2$	$1.2e-5$	$7.2e-7$
$i = 5$	$2.0e-2$	$5.9e-5$	$3.7e-7$
$i = 6$	$3.1e-2$	$9.6e-5$	$1.9e-7$

6 Conclusion

In this paper, a collocation approach was presented that involves utilizing Taylor polynomials to find the solution to a two-dimensional linear VIE of the second kind, which is a conversion of a hyperbolic linear PDE Goursat problem. The method's convergence and error were investigated, and several numerical examples were provided to demonstrate its efficiency and accuracy. The results showed that the method is convergent with a high level of precision, and the numerical results match the theoretical estimates. It is recognized that this approach can be expanded and used to solve three-dimensional Goursat problems in linear hyperbolic equations of the third-order

$$\frac{\partial^3 \omega}{\partial \tau_1 \partial \tau_2 \partial \tau_3} + \sum_{i,j=1, i<j}^{3} \psi_{i,j} \frac{\partial \omega}{\partial \tau_i \partial \tau_j} + \sum_{i=1}^{3} \psi_i \frac{\partial \omega}{\partial \tau_i} + \psi \omega = F,$$

where $\psi_{i,j}, \psi_i, , i < j, i, j = 1, 2, 3, \psi$ and F are given real functions.

Acknowledgements

The authors extend their gratitude to the anonymous referees and editor for their constructive comments, which greatly enhanced the quality of this manuscript. Additionally, heartfelt appreciation is extended to Professor Azzeddine Bellour for their guidance and suggestions throughout this work.

References

[1] Asaduzzaman, M., Özger, F. and Kilicman, A. *Analytical approximate solutions to the nonlinear Fornberg –Whitham type equations via modified variational iteration method*, Partial Differ. Equ. Appl. Math. 9 (2024), 100631.

[2] Bai, D. and Zhang, L. *The quadratic B-spline finite-element method for the coupled schrödinger –Boussinesq equations*, Int. J. Comput. Math. 88 (8) (2011), 1714–1729.

[3] Bellour, A., Bousselsal, M. and Laib, H. *Numerical solution of second-order linear delay differential and integro-differential equations by using Taylor collocation method*, Int. J. Comput. Methods, 17 (09) (2020), 1950070.

[4] Boutayeb, A. and Twizell, E. *Finite-difference methods for the solution of special eighth-order boundary-value problems*, Int. J. Comput. Math. 48 (1-2) (1993), 63–75.

[5] Brunner, H. *Collocation methods for Volterra integral and related functional differential equations*, Cambridge university press, 15, 2004.

[6] Bülbül, B. and Sezer, M. *Taylor polynomial solution of hyperbolic type partial differential equations with constant coefficients*, Int. J. Comput. Math. 88 (3) (2011), 533–544.

[7] Çelik, A. and Düzgün, A. *Approximate solution of ordinary linear differential equations with analytical complex functions coefficient by means of Taylor matrix method*, Int. J. Comput. Math. 82 (6) (2005), 765–775.

[8] Datta, M., Alam, M.S., Hahiba, U., Sultana, N. and Hossain, M.B. *Exact solution of Goursat Problem with Linear and Non-linear Partial Differential Equations by Double Elzaki Decomposition Method*, Applied Mathematics 11 (1) (2021), 5–11.

[9] Day, J.T. *A Runge-Kutta method for the numerical solution of the Goursat problem in hyperbolic partial differential equations*, Comput. J. 9 (1) (1966), 81–83.

[10] Deniz, S., Ozger, F., Ozger, Z.O., Mohiuddine, S.A. and Ersoy, M.T. *Numerical solution of fractional Volterra integral equations based on rational chebyshev approximation*, Miskolc Math. Notes, 24 (3) (2023), 1287–1305.

[11] Drignei, M.C. *A numerical algorithm for a coupled hyperbolic boundary value problem*, Int. J. Comput. Methods 19 (10) (2022), 2250027.

[12] Evans, D. and Sanugi, B. *Numerical solution of the Goursat problem by a nonlinear trapezoidal formula*, Applied Mathematics Letters 1 (3) (1988), 221–223.

[13] Heydari, M.H., Hooshmandasl, M. and Ghaini, F.M. *A new approach of the Chebyshev wavelets method for partial differential equations with boundary conditions of the telegraph type*, Appl. Math. Model. 38 (5-6) (2014), 1597–1606.

[14] Ismail, M. and Alamri, S. *Highly accurate finite difference method for coupled nonlinear schrödinger equation*, Int. J. Comput. Math. 81 (3) (2004), 333–351.

[15] Jeffrey, A. and Taniuti, T. *Non-Linear Wave Propagation With Applications to Physics and Magnetohydrodynamics by A Jeffrey and T Taniuti*, Elsevier, 2000.

[16] Laib, H., Bellour, A. and Boulmerka, A. *Taylor collocation method for a system of nonlinear Volterra delay integro-differential equations with application to COVID-19 epidemic*, Int. J. Comput. Math. (2021), 1–25.

[17] Laib, H., Bellour, A. and Boulmerka, A. *Taylor collocation method for high-order neutral delay Volterra integro-differential equations*, J. Innov. Appl. Math. Comput. Sci. 2 (1) (2022), 53–77.

[18] Laib, H., Bellour, A. and Bousselsal, M. *Numerical solution of high-order linear Volterra integro-differential equations by using Taylor collocation method*, Int. J. Comput. Math. 96 (5) (2019), 1066–1085.

[19] Laib, H., Boulmerka, A., Bellour, A. and Birem, F. *Numerical solution of two-dimensional linear and nonlinear Volterra integral equations using Taylor collocation method*, J. Comput. Appl. Math. 417 (2023), 114537.

[20] Mahdy, A.M.S., Shokry, D. and Lotfy, K. *Chelyshkov polynomials strategy for solving 2-dimensional nonlinear Volterra integral equations of the first kind*, Comput. Appl. Math. 41 (2022), 257.

[21] Maleknejad, K. and Shahabi, M. *Numerical solution of two-dimensional nonlinear Volterra integral equations of the first kind using operational matrices of hybrid functions*, Int. J. Comput. Math. (2022), 1–18.

[22] Maleknejad, K., Sohrabi, S. and Baranji, B. *Application of 2d-bpfs to nonlinear integral equations*, Commun. Nonlinear Sci. Numer. Simul. 15(3) (2010), 527–535.

[23] Mckee, S., Tang, T. and Diogo, T. *An Euler-type method for two-dimensional Volterra integral equations of the first kind*, IMA J. Numer. Anal. 20 (3) (2000), 423–440.

[24] Pandey, P.K. *A finite difference method for numerical solution of Goursat problem of partial differential equation*, Open Access Libr J. 1 (2014), 1–6.

[25] Scott, E. *Determination of the Riemann function*, Amer. Math. Monthly 80 (8) (1973), 906–909.

[26] Wollkind, D.J. *Applications of linear hyperbolic partial differential equations: Predator-prey systems and gravitational instability of nebulae*, Math. Model. 7 (2-3) (1986), 413–428.

[27] Zheng, X. and Wei, Z. *Discontinuous Legendre wavelet Galerkin method for reaction –diffusion equation*, Int. J. Comput. Math. 94 (9) (2017), 1806–1832.

[28] Zheng, X., Yang, X., Su, H. and Qiu, L. *Mixed discontinuous Legendre wavelet Galerkin method for solving elliptic partial differential equations*, Int. J. Comput. Math. 88 (17) (2011), 3626–3645.

Aims and scope

Iranian Journal of Numerical Analysis and Optimization (IJNAO) is published twice a year by the Department of Applied Mathematics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad. Papers dealing with different aspects of numerical analysis and optimization, theories and their applications in engineering and industry are considered for publication.

Journal Policy

All submissions to IJNAO are first evaluated by the journal's Editor-in-Chief or one of the journal's Associate Editors for their appropriateness to the scope and objectives of IJNAO. If deemed appropriate, the paper is sent out for review using a single blind process. Manuscripts are reviewed simultaneously by reviewers who are experts in their respective fields. The first review of every manuscript is performed by at least two anonymous referees. Upon the receipt of the referee's reports, the paper is accepted, rejected, or sent back to the author(s) for revision. Revised papers are assigned to an Associate Editor who makes an evaluation of the acceptability of the revision. Based upon the Associate Editor's evaluation, the paper is accepted, rejected, or returned to the author(s) for another revision. The second revision is then evaluated by the Editor-in-Chief, possibly in consultation with the Associate Editor who handled the original paper and the first revision, for a usually final resolution.

The authors can track their submissions and the process of peer review via: http://ijnao.um.ac.ir

All manuscripts submitted to IJNAO are tracked by using "iThenticate" for possible plagiarism before acceptance.

Instruction for Authors

The Journal publishes all papers in the fields of numerical analysis and opti-

mization. Articles must be written in English.

All submitted papers will be refereed and the authors may be asked to revise their manuscripts according to the referee's reports. The Editorial Board of the Journal keeps the right to accept or reject the papers for publication.

The papers with more than one authors, should determine the corresponding author. The e-mail address of the corresponding author must appear at the end of the manuscript or as a footnote of the first page.

It is strongly recommended to set up the manuscript by Latex or Tex, using the template provided in the web site of the Journal. Manuscripts should be typed double-spaced with wide margins to provide enough room for editorial remarks.

References should be arranged in alphabetical order by the surname of the first author as examples below:

[1] Brunner, H. *A survey of recent advances in the numerical treatment of Volterra integral and integro-differential equations*, J. Comput. Appl. Math. 8 (1982), 213-229.

[2] Stoer, J. and Bulirsch, R. *Introduction to Numerical Analysis*, Springer-verlag, New York, 2002.

Iranian Journal of Numerical Analysis and Optimization

CONTENTS Vol. 14, No. 2, pp 330-637

Quantum solutions of a nonlinear Schrödinger equation 330
S. Arfaoui and A. Ben Mabrouk

Discontinuous Galerkin approach for two-parametric convection-diffusion equation with discontinuous source term 347
K. R. Ranjan and S. Gowrisankar

High order second derivative multistep collocation methods for ordinary differential equations 367
S. Fazeli

Jarratt and Jarratt-variant families of iterative schemes for scalar and system of nonlinear equations 391
O. Ogbereyivwe, E.J. Atajeromavwo and S.S. Umar

An innovative particle physics optimization algorithm for efficient test case minimization in software testing 417
U. Jaiswal and A. Prajapati

Dynamics of Cholera Pathogen Carriers and Effect of Hygiene Consciousness in Cholera Outbreaks 449
S.F. Abubakar and M.O. Ibrahim

Drug release from a two-layer stent coating considering the viscoelastic property of the arterial wall: A mathematical and numerical study 475
E. Azhdari and A. Emami

Mathematical analysis and forecasting of controlled Spatio-temporal dynamics of the EG.5 Virus 500
E.M. Moumine, O. Balatif and M. Rachik

Optimizing natural gas liquids (NGL) production process: A multi-objective approach for energy-efficient operations using genetic algorithm and artificial neural networks 522
N. Sabeghi, B. Kiani Talaei, R. Ghanbari and K. Ghorbani-Moghadam

A new bi-level data envelopment analysis model to evaluate the Human Development Index 545
E. Hajinezhad, H. Hajinezhad and M.R. Alirezaee

Modeling individual mobility's impact on COVID-19 transmission: Insights from a two-patch SEIR-V approach . . 580
M. Bouziane, M.A. Boubekeur, M.E.B. Keddar and O. Belhamiti

Goursat problem in Hyperbolic partial differential equations with variable coefficients solved by Taylor collocation method 613
F. Birem, A. Boulmerka, H. Laib and C. Hennous

web site: https://ijnao.um.ac.ir
Email: ijnao@um.ac.ir
ISSN-Print: 2423-6977
ISSN-Online: 2423-6969