ISSN: 2327-4352 Volume 5, Number 11, November 2017

Scientific Research Publishing

Journal of Applied Mathematics and Physics

ISSN: 2327-4352

www.scirp.org/journal/jamp

JOURNAL EDITORIAL BOARD

ISSN 2327-4352 (Print) ISSN 2327-4379 (Online)
http://www.scirp.org/journal/jamp

Editor-in-Chief

Editorial Board (According to Alphabet)

Scientific Research Publishing

Journal of Applied Mathematics and Physics, 2017, 5, 2093-2310
http://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Table of Contents

Volume 5 Number 11 November 2017

Journal of Applied Mathematics and Physics, 2017, 5, 2093-2310
http://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Journal of Applied Mathematics and Physics (JAMP)
Journal Information

SUBSCRIPTIONS

The *Journal of Applied Mathematics and Physics* (Online at Scientific Research Publishing, www.SciRP.org) is published monthly by Scientific Research Publishing, Inc., USA.

Subscription rates:
Print: $39 per issue.
To subscribe, please contact Journals Subscriptions Department, E-mail: sub@scirp.org

SERVICES

Advertisements
Advertisement Sales Department, E-mail: service@scirp.org

Reprints (minimum quantity 100 copies)
Reprints Co-ordinator, Scientific Research Publishing, Inc., USA.
E-mail: sub@scirp.org

COPYRIGHT

PRODUCTION INFORMATION

For manuscripts that have been accepted for publication, please contact:
E-mail: jamp@scirp.org

Scientific Research Publishing

Journal of Applied Mathematics and Physics, 2017, 5, 2093-2103
http://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

On the Crucial Role of the Variational Principle in Quantum Theories

Eliahu Comay

Charactell Ltd., Tel-Aviv, Israel
Email: elicomay@post.tau.ac.il

How to cite this paper: Comay, E. (2017) On the Crucial Role of the Variational Principle in Quantum Theories. *Journal of Applied Mathematics and Physics*, **5**, 2093-2103.
https://doi.org/10.4236/jamp.2017.511171

Received: July 11, 2017
Accepted: October 31, 2017
Published: November 3, 2017

Abstract

The paper shows that the variational principle serves as an element of the mathematical structure of a quantum theory. The experimentally confirmed properties of the corpuscular-wave duality of a quantum particle are elements of the analysis. A Lagrangian density that yields the equations of motion of a given quantum theory of a massive particle is analyzed. It is proved that if this Lagrangian density is a Lorentz scalar whose dimension is $\left[L^{-4}\right]$ then the associated action consistently defines the required phase of the quantum particle. The $\left[L^{-4}\right]$ dimension of this Lagrangian density proves that also the quantum function $\psi\left(x^{\mu}\right)$ has dimension. This result provides new criteria for the acceptability of quantum theories. An examination of the first order Dirac equation demonstrates that it satisfies the new criteria whereas the second order Klein-Gordon equation fails to do that.

Keywords

Quantum Theories, Lagrangian Density, Corpuscular-Wave Duality, Dimension of the Quantum Function, The Correspondence Principle

1. Introduction

A physical theory has two primary elements: it has a self-consistent mathematical structure and it describes adequately data which are obtained from experiments that are included in the theory's domain of validity. The present work concentrates on the mathematical structure of quantum theories of electromagnetic interactions. Like any other physical theory, it takes few experimental data as elements that the theory must satisfy. The discussion shows that these requirements lead to a quite unique mathematical structure of the theory. The results provide another example

of Wigner's well known statement about *the unreasonable effectiveness of mathematics in the natural sciences* [1].

Special relativity is a well established theory. In particular, modern accelerators produce particles whose velocity is very close to the speed of light. The design of these machines and the data which are obtained from them are consistent with the laws of special relativity. It means that accelerators provide an astronomical number of experimental tests which are consistent with special relativity. Therefore, it is assumed here that the required quantum theory must take a relativistic covariant form.

This work aims to derive the structure of a quantum theory of an *elementary massive* particle. Historically, the first purpose of quantum theories was to describe experimental data of the electron. As a matter of fact, the electron is the most well-known elementary massive particle and it provides many kinds of experimental data. Hence, the ample electronic data enable to carry out many different tests of the validity of its quantum theory. This issue is very useful because a physical theory becomes unacceptable if it is inconsistent with even one kind of well established experiment that is included within its domain of validity.

Physical principles play an important role in the search for new physical theories because they provide general requirements that should be satisfied by any new theoretical candidate. The title of this work indicates that it discusses the variational principle. The correspondence principle is also used in the following discussion and the meaning of this principle is explained before it is applied.

Units where $\hbar = c = 1$ are used. Greek indices run from 0 to 3 and Latin indices run from 1 to 3. The metric is diag. (1, −1, −1, −1). Relativistic expressions are written in the standard notation. Square brackets $[\,]$ denote the dimension of the enclosed expression. In a system of units where $\hbar = c = 1$ there is just one dimension, and the dimension of length, denoted by $[L]$, is used. In particular, energy and momentum take the dimension $\left[L^{-1}\right]$ and the electric charge is a dimensionless pure number. The value of the electron's charge is $e^2 \simeq 1/137$. The second section discusses hierarchical relations between physical theories and the significance of the correspondence principle. The role of the variational principle in the structure of quantum theories is explained in the third section. The experimental information used in the analysis is shown in the fourth section. The fifth section proves the validity of a new reason for the need of the variational principle. The sixth section describes specific results which are derived from the variational principle. The last section contains concluding remarks.

2. Hierarchical Relationships between Physical Theories

An essential feature of an acceptable physical theory is the existence of a domain of validity where the theory describes properly experimental results. For example,

it is well known that Newtonian mechanics yields good predictions in cases where the particles' velocity is much smaller than the speed of light and if quantum effects can be ignored. These restrictions define the domain of validity of Newtonian mechanics.

The founders of quantum mechanics have recognized that the classical limit of quantum mechanics should be consistent with classical physics. And indeed, a proof showing that the classical limit of quantum mechanics agrees with classical physics was published in 1927 (see the Ehrenfest theorem in [2], pp. 25-27, 136-138). This matter can be found in many textbooks. For example: "classical mechanics must therefore be a limiting case of quantum mechanics" (see [3], p. 84). A general discussion of this topic is presented in pp. 1-6 of [4].

Let A,B denote two physical theories and D_A, D_B denote their domain of validity, respectively. If $D_A \subset D_B$ then these domains of validity can be used for a definition of hierarchical relationships between A,B. It means that theory B is good in *all* cases where theory A is good, but not vice versa. In this case the rank of theory B is higher than that of theory A. For example, the rank of special relativity is higher than that of Newtonian mechanics.

Generally the hierarchical relationship between two theories is obtained in cases where D_A is relevant to a limit of a certain variable. For example, the domain of validity of Newtonian mechanics is relevant to the limit $v_i \to 0$, where v_i is the velocity of the ith particle. In this case, formulas of special relativity boil down to corresponding formulas of Newtonian mechanics. The domain of validity of Newtonian mechanics holds not only for the case where $v_i = 0$ because the continuity of expressions indicates that if $v_i \ll c$ then errors of Newtonian mechanics are smaller than measurements' errors and this theory is acceptable. The limit process used for the definition of D_A means that the correspondence between theories A,B relies on a solid mathematical basis.

The relationship $D_A \subset D_B$ means that theory B has a more profound meaning than that of theory A. However, the merits of theory A should not be underestimated because an appropriate limit of expressions of theory B must be consistent with the corresponding expressions of theory A. It means that *theory A provides theoretical constraints on the acceptability of theory B.* These constraints are useful in an examination of the acceptability of new theoretical ideas. They are related to a certain limit of a variable that defines the domains of validity of the two theories. Therefore, these constraints belong to the mathematical structure of the theories.

The hierarchical relationships between the following quantum theories are discussed in this work: non-relativistic quantum mechanics (QM), relativistic quantum mechanics (RQM) and quantum field theory (QFT). Here the hierarchical rank of RQM is higher than that of QM because QM is restricted to cases where the particles' velocity v_i (or in a quantum parlance p_i/m_i) is much smaller than the speed of light. RQM is restricted to cases where the number of particles can be regarded as a constant of the motion whereas QFT

discusses cases where additional particle-antiparticle pairs are included in the system. For example, experiments show that a non-negligible probability of the existence of quark-antiquark pairs is found in the proton (see [5], p. 282). Therefore, QFT should be used for a description of the proton structure. **Figure 1** illustrates the hierarchical relations between these theories. Here the domains of validity of the three theories are represented by the corresponding rectangles which satisfy the following relations $\mathrm{QM} \subset \mathrm{RQM} \subset \mathrm{QFT}$.

The relationships between QFT and QM is recognized in the literature. For example: "First, some good news: quantum field theory is based on the same quantum mechanics that was invented by Schroedinger, Heisenberg, Pauli, Born, and others in 1925-26, and has been used ever since in atomic, molecular, nuclear and condensed matter physics" (see [6], p. 49). In this work, these constraints are called Weinberg correspondence principle.

In the physical literature the relationship between QM and classical physics is sometimes called the Bohr correspondence principle (see [2], p. 4). The philosophical literature discusses general aspects of the correspondence between theories and this topic is called the generalized correspondence principle [7].

3. The Role of the Variational Principle in Quantum Theories

Items of the following list mention briefly examples that point out the relevance of the variational principle, its Lagrangian density $\mathcal{L}$ and the associated action S to quantum theories. These items are not new and it is shown here that they can be found in textbooks. Furthermore, the variational principle has a mathematical structure and it means that one can prove the correctness of these items.

• The variational principle is used in a demonstration of the consistence of the classical limit of quantum mechanics with classical physics (see e.g. [3], section 32; [8], pp. 19-21).

• The discussions in the previous references also show that in the classical limit, the wave function of a quantum particle takes the following form

$$\psi = A\mathrm{e}^{iS/\hbar}, \tag{1}$$

where S is the action of the given Lagrangian. In the units used herein $\hbar = 1$ and it can be removed from (1).

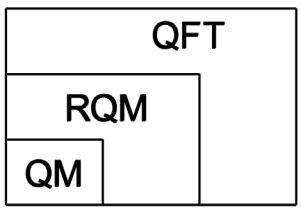


Figure 1. A chain of three rectangles that represent domains of validity of three quantum theories, respectively. The figure shows that a smaller rectangle is included in a larger rectangle (see text).

• An application of the Lagrangian that is used in the variational principle yields a definition of canonical momenta, where each of which is related to a generalized coordinates (see e.g. [9], p. 16). It can be shown that the Poisson brackets of a classical Hamiltonian and a dynamical variable correspond to an appropriate commutation relations of quantum mechanics (see e.g. [3], section 21 ; [8], pp. 26-28).

• The Noether theorem (see [10]) proves that a conservation law of a physical quantity corresponds to an appropriate invariance of the Lagrangian with respect to a certain transformation. Analogous relations are found for the Lagrangian density of QFT (see [6], pp. 306-314; [11], pp. 17-22).

• Relativistically covariant QFT equations are obtained from a Lagrangian density that is a Lorentz scalar (see [6], p. 300). This property emphasizes the significance of this kind of Lagrangian density.

• The variational principle is also used in other fields of physics. For example, textbooks prove that Newtonian mechanics can be derived from this principle (see [9, 12]).

The main objective of this work is to show that the foregoing items do not cover all aspects of the relevance of the variational principle to quantum theories. Consequences of the new applications of the variational principle are discussed.

4. Experimental Elements Used in the Analysis

The following fundamental experimental data are a combination of two features of a quantum particle. Some experiments show that it has corpuscular properties whereas other experiments show that it has wave properties (see [2], pp. 1-3; [13] p. 59). The combination of these properties is called corpuscular-wave duality. In classical physics this duality is a contradiction. Indeed, in classical physics an elementary particle is point-like (see [14], pp. 46-47) whereas a wave has a spatial distribution. Therefore, a new theory which is consistent with this duality is required.

The following discussion shows how these requirements lead to the structure of quantum theories. In particular, the primary role of the variational principle is derived.

5. Properties of Relativistic Quantum theories

The primary experimental properties of a quantum particle which are described in the previous section are used here for a construction of the main elements of a quantum theory. This task is done stepwisely.

• In order to be consistent with the pointlike property of an elementary quantum particle, the theory describes it by means of a wave function whose form is

$$\psi\left(x^{\mu}\right), \tag{2}$$

where x^{μ} denotes a *single* set of four space-time coordinates. The following

argument explains why the form of (2) describes an elementary pointlike particle.

Take for example the ground state of the positronium, which is a bound state of an electron and a positron. This object, which has a non-vanishing volume, is described by a function of the form $\psi(t, \boldsymbol{x}_1, \boldsymbol{x}_2)$, where $\boldsymbol{x}_1, \boldsymbol{x}_2$ denote the three spatial coordinates of the electron and the positron, respectively. This example shows that in order to describe a composite non-pointlike particle one needs more than four space-time degrees of freedom.

As a matter of fact, the form of (2) is used in QFT textbooks in expressions for the Lagrangian density of any elementary quantum particle [6] [11] [15]. This issue means that there is a consensus about this requirement.

• In order to be consistent with the wave properties of a quantum particle, the function $\psi(x^\mu)$ must have a phase factor. The de Broglie work has proven that the argument of the phase factor of a free quantum particle takes the form $(\boldsymbol{k}\cdot\boldsymbol{x}-\omega t)$, where

$$\boldsymbol{k}=\boldsymbol{p}/\hbar;\quad \omega=E/\hbar \tag{3}$$

and $\boldsymbol{p}, E$ denote the particle's linear momentum and energy, respectively. Historically, de Broglie published his relations (3) before they were experimentally confirmed (see [2], p. 3; [8], pp. 48-49). This is certainly an example of a successful theoretical work.

• The wave properties of the quantum function (2) show that it must satisfy a wave equation.

• A wave function of a free particle that travels in the x-direction and satisfies the de Broglie relations can be written as a linear combination of the following expressions (see [2], p. 18)

$$\cos(kx-\omega t),\ \sin(kx-\omega t),\ \exp(\pm i(kx-\omega t)) \tag{4}$$

The first and the second expressions of (4) are real functions whereas the last expression is a complex function.

The following argument proves that real functions cannot describe a massive quantum particle. Let us use the real functions of (4) and examine a massive quantum particle which is in a field free region. In a particular inertial frame this particle is at rest and its linear momentum $\boldsymbol{p}=0$. Substituting this value and (3) into (4), one finds that in this frame the general form of a real wave function of a massive quantum particle is

$$\psi(t,x)=A\sin(\omega t)+B\cos(\omega t)=C\sin(\omega t-\delta), \tag{5}$$

where A, B, C and δ are appropriate real constant numbers. It follows that for every integer N, there is an instant when $\omega t-\delta=N\pi$ and the wave function of (5) vanishes throughout the entire 3-dimensional space. The following argument proves that this function is unacceptable. In the non-relativistic QM the particle's density is $\psi^*\psi$. It means that if ψ vanished throughout the entire 3-dimensional space then the particle does not exist. The Weinberg

correspondence principle and the limiting process prove that this result also holds for higher quantum theories because these theories must use ψ as a factor for the definition of density.

For these reasons, the phase factor of the quantum function of a motionless particle must be complex and the last expression of (4) shows that this function takes the form

$$\psi = \mathrm{e}^{i\Phi}\chi(x,y,z). \tag{6}$$

The foregoing argument holds for a motionless particle. Such a particle is in a state which is the limit of states of particles that move inertially in a field-free region and their velocity tends to zero. Hence, a quantum function of particles that move inertially in a field-free region must be complex, because the limit of real functions is a real function.

Furthermore, a field-free region is the limit of regions where the intensity of the interaction tends to zero. Therefore, using a similar argument, one finds that the general state of a quantum particle is described by a complex function. This issue is also discussed in [16].

• Let us examine the power series of the phase factor of the quantum function (6)

$$\mathrm{e}^{i\Phi} = 1 + i\Phi + \cdots \tag{7}$$

Now, a very well known law of physics states that all terms of a physically acceptable expression must have the same dimension, and, in a relativistic theory, they must also satisfy covariance. Since each of the numbers $1, i$ on the right hand side of (7) is a dimensionless Lorentz scalar, one concludes that also the phase Φ must be a dimensionless Lorentz scalar. The following argument explains how these constraints are satisfied.

• Let us use the variational principle and a Lagrangian density $\mathcal{L}$ that yields the required equations of the given quantum particle. The case of a Dirac particle is used here as an illustration of this issue

$$\mathcal{L}_D = \bar{\psi}\left[\gamma^\mu i\partial_\mu - m\right]\psi - \frac{1}{16\pi}F^{\mu\nu}F_{\mu\nu} - e\bar{\psi}\gamma^\mu A_\mu\psi. \tag{8}$$

Here $\bar{\psi} \equiv \psi^\dagger\gamma^0$. The first term of (8) represents a free Dirac particle, the second term represents free electromagnetic fields and the last term represents the interaction between a Dirac charged particle and electromagnetic fields (see [11], p. 84, [15], p. 78).

The action of (8) is obtained from its 4-dimensional integral

$$S = \int \mathcal{L}_D \mathrm{d}^4x. \tag{9}$$

The action S and $\hbar$ have the same dimension (see e.g. [8], p. 20, Equation (6.1)). It follows that in the unit system used herein where $\hbar = 1$, the action is a dimensionless Lorentz scalar. Evidently, d^4x is a Lorentz scalar whose dimension is $\left[L^4\right]$ (see [14], p. 21). Hence, the action (9) proves that the Lagrangian density $\mathcal{L}$ of a quantum particle must be a Lorentz scalar whose

dimension is $\left[L^{-4}\right]$. These constraints are imposed on an acceptable theory of a quantum particle. This kind of action can be used for the particle's phase. This general result is an extension of (1) which applies to the classical limit of the wave function.

The foregoing discussion proves that the action which is obtained from the Lagrangian density of a quantum theory can be used as the required phase of a quantum particle if the Lagrangian density is a Lorentz scalar whose dimension is $\left[L^{-4}\right]$. The Weinberg correspondence principle proves that this conclusion holds for QM, RQM and QFT. Item 5 of section 3 indicates that the need for a Lorentz scalar Lagrangian density is already documented in the literature. The required $\left[L^{-4}\right]$ dimension of the Lagrangian density is the main subject of the following discussion.

6. Discussion

This section describes some results that demonstrate powerful consequences of the requirement of a Lagrangian density whose dimension is $\left[L^{-4}\right]$.

Let us take the Lagrangian density of a Dirac electron (8). Its dimension is $\left[L^{-4}\right]$ and in the unit system used herein $c=1$ and the dimension of each component of the partial derivatives ∂_μ is $\left[L^{-1}\right]$. It follows that the dimension of the product $\bar{\psi}\psi$ is $\left[L^{-3}\right]$. This is the dimension of density and indeed, the theory of a Dirac particle yields a consistent expression for a conserved 4-current

$$j^\mu = \bar{\psi}\gamma^\mu\psi; \quad j^\mu_{,\mu} = 0. \tag{10}$$

It is well known that density is the 0-component of the 4-current $j^0 = \psi^\dagger\psi$ (see [17], pp. 23, 24).

The following product of the Schroedinger wave function $\psi^*\psi$ denotes density and it is used in a consistent expression for density-current (see [8], p. 54; [13], pp. 117-120). This result shows that the Dirac equation is consistent with the Weinberg correspondence principle of section 2. In particular, a construction of a Hilbert space is a requirement that should be satisfied by a consistent quantum theory (see [6], p. 49). The consistent expression for density of a Dirac particle (10) enables a construction of a Hilbert space for a Dirac electron, where the required inner product of two functions is $\int \psi_i^\dagger \psi_j \mathrm{d}^3x$.

In the case of quantum theories, the dimension $\left[L^{-4}\right]$ of the Lagrangian density provides a simple proof of the role of the electromagnetic 4-potential in the interaction term of an electric charge. Indeed, the laws of Maxwellian electrodynamics prove that electromagnetic interaction is proportional to the charge. Therefore, in a quantum theory it should be proportional to the charge density whose dimension is $\left[L^{-3}\right]$. Hence, due to the $\left[L^{-4}\right]$ dimension of the Lagrangian density, the dimension of the electromagnetic factor of the interaction term must be $\left[L^{-1}\right]$. This is the dimension of the electromagnetic potentials. By contrast, this argument proves that the electromagnetic field tensor $F^{\mu\nu}$ is unsuitable for this purpose because it is the 4-curl of the

4-potential and its dimension is $\left[L^{-2}\right]$.

The Dirac equation is a first order differential equation. As a matter of fact, one can find in the literature quantum equations of the second order. The Klein-Gordon (KG) equation is a quite simple example of this kind of equations. The KG function is a Lorentz scalar and its Lagrangian density is (see [18], p. 198 of the English translation)

$$\mathcal{L}=\left(\phi_{,0}^{*}-ieV\phi^{*}\right)\left(\phi_{,0}+ieV\phi\right)-\sum_{k=1}^{3}\left(\phi_{,k}^{*}+ieA_{k}\phi^{*}\right)\left(\phi_{,k}-ieA_{k}\phi\right)-m^{2}\phi^{*}\phi. \quad (11)$$

The KG equation is derived from (11) (see Equation (39) therein)

$$\left(\frac{\partial}{\partial t}-ieV\right)\left(\frac{\partial}{\partial t}-ieV\right)\phi=\sum_{k=1}^{3}\left(\frac{\partial}{\partial x^{k}}+ieA_{k}\right)\left(\frac{\partial}{\partial x^{k}}+ieA_{k}\right)\phi+m^{2}\phi \quad (12)$$

and the KG density is (see Equation (42) therein)

$$\rho=i\left(\phi^{*}\phi_{,0}-\phi_{,0}^{*}\phi\right)-2eV\phi^{*}\phi. \quad (13)$$

The product of two derivatives of the KG Lagrangian density (11) proves that the dimension of the KG function ϕ is $\left[L^{-1}\right]$. Therefore, the following inconsistencies arise.

• The dimension of the product $\phi^{*}\phi$ of the KG function is $\left[L^{-2}\right]$. On the other hand it is shown earlier in this section that the dimension of the corresponding product of the Schroedinger functions is $\left[L^{-3}\right]$. Hence, the KG function is inconsistent with the Weinberg correspondence principle because the continuity of a limit process does not alter the discrete value of the wave function's dimension.

• The dimension $\left[L^{-2}\right]$ of the product $\phi^{*}\phi$ of the KG function explains why its expression for density (13) contains a derivative with respect to time. Hence, unlike the case of the non-relativistic QM, one *cannot* use density of a KG function and construct a Hilbert space in the Heisenberg picture where the quantum function is time-independent (see [3], p. 112; [11], p. 6). This is another violation of the Weinberg correspondence principle.

• The KG Lagrangian density (11) contains a second order term of the potentials. For example, a factor V^{2} of the electric potential is obtained from the first term on the right hand side of (11). By contrast, Maxwell equations are derived from a Lagrangian density that depends *linearly* on the 4-potential (see [14], pp. 78, 79). Hence, the KG Lagrangian density (11) is inconsistent with Maxwellian electrodynamics.

These problematic results support Dirac lifelong objection to the KG equation (see [19], pp. 3,4). It can be shown that analogous problems hold in other kinds of second order quantum equations (see [20], Section 4).

7. Concluding Remarks

This work examines some aspects of the relevance of the variational principle to the mathematical structure of quantum theories. The experimentally confirmed corpuscular-wave duality is the basis of the analysis. This analysis focuses on the

role of a consistent expression for the phase factor of a given quantum function. For this end, the paper examines a Lagrangian density which depends on a quantum function whose form is $\psi\left(x^{\mu}\right)$. It is proved that each term of this Lagrangian density must be a Lorentz scalar whose dimension is $\left[L^{-4}\right]$. The results show that such a Lagrangian density yields an action that can be used as the phase of a quantum particle. Moreover, it is shown that the $\left[L^{-4}\right]$ dimension of the Lagrangian density defines a specific dimension for the quantum function $\psi\left(x^{\mu}\right)$. The correspondence principle proves that the dimension of the quantum function provides a new kind of constraint that an acceptable quantum theory must satisfy. It is shown that the Dirac first order quantum theory complies with this constraint. By contrast, problems arise in the case of second order quantum theories, like that of the KG equation.

References

[1] Wigner, E. (1960) The Unreasonable Effectiveness of Mathematics in the Natural Sciences. *Communications on Pure and Applied Mathematics*, **13**, 1-14. https://doi.org/10.1002/cpa.3160130102

[2] Schiff, L.I. (1955) Quantum Mechanics. McGraw-Hill, New York.

[3] Dirac, P.A.M. (1958) The Principles of Quantum Mechanics. Oxford University Press, London.

[4] Rohrlich, F. (2007) Classical Charged Particle. World Scientific, New Jersey. https://doi.org/10.1142/6220

[5] Perkins, D.H. (1987) Introduction to High Energy Physics. Addison-Wesley, Menlo Park, CA.

[6] Weinberg, S. (1995) The Quantum Theory of Fields. Vol. I. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139644167

[7] Radder, H. (1991) Heuristics and the Generalized Correspondence Principle. *The British Journal for the Philosophy of Science*, **42**, 195-226. https://doi.org/10.1093/bjps/42.2.195

[8] Landau, L.D. and Lifshitz, E.M. (1959) Quantum Mechanics. Pergamon, London.

[9] Landau, L.D. and Lifshitz, E.M. (1960) Mechanics. Pergamon, Oxfors.

[10] Wikipedia https://en.wikipedia.org/wiki/Noether

[11] Bjorken, J.D. and Drell, S.D. (1965) Relativistic Quantum Fields. McGraw-Hill, New York.

[12] Goldstein, H., Poole, C. and Safko, J. (2014) Classical Mechanics. Addison Wesley, San Francisco.

[13] Messiah, A. (1967) Quantum Mechanics. Vol. 1, North Holland, Amsterdam.

[14] Landau, L.D. and Lifshitz, E.M. (2005) The Classical Theory of Fields. Elsevier, Amsterdam.

[15] Peskin, M.E. and Schroeder, D.V. (1995) An Introduction to Quantum Field Theory. Addison-Wesley, Reading Mass.

[16] Comay, E. (2016) Problems with Mathematically Real Quantum Wave Functions. OALib, 3, No. 8, 1. https://www.scirp.org/journal/PaperInformation.aspx?PaperID=70117

[17] Bjorken, J.D. and Drell, S.D. (1964) Relativistic Quantum Mechanics. McGraw-Hill, New York.

[18] Pauli, W. and Weisskopf, V. (1934) Uber die Quantisierung der skalaren relativistischen Wellengleichung. [The Quantization of the Scalar Relativistic Wave Equation.] *Helvetica Physica Acta*, **7**, 709-731.

[19] Dirac, P.A.M. (1978) Mathematical Foundations of Quantum Theory. In: Marlow, A.R., Ed., *Mathematical Foundations of Quantum Theory*, Academic, New York, 1-8.

[20] Comay, E. (2009) Physical Consequences of Mathematical Principles. *Progress in Physics*, **4**, 91-98. http://www.tau.ac.il/elicomay/MathPhys.pdf

Journal of Applied Mathematics and Physics, 2017, 5, 2104-2111
http://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Research on Refined Oil Distribution Plan Based on Dynamic Time Window

Qian Liu, Lianhua Wang*, Le Yu

Beijing Wuzi University, Beijing, China
Email: 1767378571@qq.com, *lianhuawang@sina.com, yulerunning@163.com

How to cite this paper: Liu, Q., Wang, L.H. and Yu, L. (2017) Research on Refined Oil Distribution Plan Based on Dynamic Time Window. *Journal of Applied Mathematics and Physics*, **5**, 2104-2111.
https://doi.org/10.4236/jamp.2017.511172

Received: September 27, 2017
Accepted: November 4, 2017
Published: November 7, 2017

Abstract

Based on the analysis of the refined oil distribution plan which includes the various vehicles models-single oil-multiple gas stations. This paper puts forward the dynamic time window of oil supplementation based on every moment, and establishes the mathematical model of the refined oil distribution plan, using C language, taking various circumstances of the model into account to find the optimal solution through several operations. This process can make the refined oil distribution plan and the distribution route is more reasonable. At the same time, the distribution cost is lowest. Through the analysis of the experimental results, the validity and algorithm of the model are proved.

Keywords

Refined Oil Distribution, Vehicle Scheduling Problem, Time Window, Optimized Model

1. Introduction

With the rapid development of Chinese economy and society, logistics is becoming an indispensable part of enterprise economic activities, and its importance is increasing day by day. The petroleum and petrochemical industry is an important guarantee for the country economic and social development, and this profession is the pillar of Chinese economy [1] [2]. After years of development, the industry has gradually matured and formed a market competition pattern dominated by Petro China, Sinopec, China National Offshore Oil Corporation. In the face of such a diverse and competitive market environment, it is more important for petroleum and petrochemical enterprises to maintain their market share and higher profit margins. In order to improve their own management

model and reduce the logistics operation cost, as the terminal part of the supply chain of the petroleum and petrochemical enterprises, refined oil distribution is a good breakthrough. By improving the logistics link, it can greatly reduce the enterprise's logistics and distribution costs, and fundamentally improve the enterprise's core-competitiveness.

In terms of its form, the issue of refined oil distribution is a problem of transport dispatching. For this problem, a large number of studies consider more about how to meet the balance of transport capacity and less consider the time constraints in the transport process. In fact, product oil has special requirements for transportation time during distribution. If the schedule is not good, it is easy to appear the phenomenon of out of stock and waiting for unloading. On the basis of previous studies, this paper combines the characteristics of oil product distribution, and puts forward the dynamic time window of oil supplementation based on every moment, and uses the C language to solve the model, so as to make the distribution plan more reasonable, which makes the distribution problem of the refined oil more in line with the actual situation.

2. Problem Description

All kinds of resources in the process of refined oil distribution are form a very complex distribution network [3] [4], including oil depots, gas stations, transport vehicles and so on. The purpose of distribution dispatching is to generate the resource scheduling scheme which has the best benefit or the lowest cost under the restrictions of transportation resources, storage capacity, transportation time and other factors. Specifically, the route of each vehicle, the delivery time and quantity of each distribution location are shown in table [5], so as to achieve the goal of minimum transportation cost. There are a stable supply and demand relationship between the oil depot and the gas station, so this paper takes a single oil depot as the object of study. The oil depot that has two types (single or double cabins) delivery vehicles delivers the same kind of oil to a series of gas stations during the business hours. In the case of ensuring every gas station is not out of stock at any moment, according to the inventory in the tank, this paper makes the delivery plan of the day, reasonably arranging the distribution vehicle, distribution route and time to make the delivery cost lowest.

3. Model Building

Before building a model, it is necessary to define some of the constants and variables necessary to build the model in advance, some of which are summarized by long-term tests and are generic. But some of the data can be adjusted and set according to the actual situation.

3.1. Model Hypothesis

1) Assuming that all types of distribution vehicles are adequate and are parked at the depot.

2) The delivery vehicles should be fully loaded (including all compartments), that is, when the vehicles leave the depot, each compartment needs to be filled with refined oil [6] [7].

3) The refined oil in the same compartment must be unloaded to the same gas station. The refined oil in different compartments of the same car can be unloaded to different gas stations.

4) Assuming that the running time from the depot to the gas station and from one gas station to another gas station is t_0 hours, and the loading time and the unloading time are ignored.

5) Assuming that every gas station unloads refined oil at the whole point of time, and the sales per hour is constant during the business hours.

6) Each gas station can use one oil truck at most to supplement refined oil at each moment.

3.2. Parameter Setting

i: the number of gas station, $i = 1, 2, \cdots, N$;

t: the business hours of gas station, $t = 0, 1, \cdots, M$;

f_i: Sales per hour of gas station i;

q_s: the compartment capacity of the vehicle which has one cabin;

q_d: a compartment capacity of the vehicle which has two cabins;

V_i: the capacity of the gas station i;

C_s: the single-trip shipping cost of the delivery which has one cabin;

C_d: the single-trip shipping cost of the delivery which has two cabins;

W_{it}: the inventory of the gas station i at t;

Q_i: safe stock of gas station i;

T_{is}: the time that gas station i begin to supplement refined oil;

T_{ie}: the time that gas station i finish supplementing refined oil;

X_{it}: binary variable. $X_{it} = 1$, if gas station i use a delivery vehicle which has one cabin to supplementing refined oil at t. $X_{it} = 0$, otherwise;

Y_{it}: binary variable. $Y_{it} = 1$, if gas station i use a cabin of delivery vehicle which has two cabins to supplementing refined oil at t. $Y_{it} = 0$, otherwise;

Z_{it}: binary variable. $Z_{it} = 1$, if gas station i use two cabins of delivery vehicle which has two cabins to supplementing refined oil at t. $Z_{it} = 0$,otherwise.

3.3. Building the Model

Based on the above problem description, model assumptions and parameter settings, the vehicle scheduling model is established:

1) Objective function

According to the above description, this paper mainly considers the transportation costs of distribution vehicles, and takes the lowest delivery cost as the optimization target:

$$\min C = C_s \times \sum_{i=1}^{N}\sum_{t=0}^{M} X_{it} + C_d \times \frac{\sum_{i=1}^{N}\sum_{t=0}^{M} Y_{it}}{2} + C_d \times \sum_{i=1}^{N}\sum_{t=0}^{M} Z_{it}$$

2) Constraints

Gas station i can be supplemented refined oil by a delivery vehicle at most at time t [8]:

$$X_{it} + Y_{it} + Z_{it} \le 1$$

The inventory of the gas station i must be less than the capacity at time t:

$$W_{it} \le V_i$$

$$W_{it} = W_{i(t-1)} + q_s \times X_{it} + q_d \times Y_{it} + 2 \times q_d \times Z_{it} - f_i$$

The paper ensures that the gas station i is not out of stock at time t:

$$W_{it} - f_i \ge 0$$

Because the running time from the depot to the gas station and from one gas station to another gas station is t_0 hours. In principle, we should ensure that each gas station is not out of stock in the next t_0 hours when the delivery task is arranged. Therefore, it is determined that the minimum inventory of refined oil at the gas station when the delivery task is arranged.

$$Q_i = t_0 \times f_i + \delta$$

In the formula, δ means the minimum inventory of refined oil when the gas station is filling up.

When the inventory of the gas station reaches Q_i, it is necessary to issue the task of oil distribution. According to the storage and the sales of the gas station at every moment, it is necessary to predict the time period that the refined oil is expected to be delivered, that is to say, to determine the dynamic time window about supplementing refined oil.

Taking the t moment as the benchmark, the start time of the oil supplement is as follows:

$$T_{is} = \frac{W_{it} - Q_i}{f_i}$$

The end time of the oil supplement is as follows:

$$T_{ie} = T_{is} + t_0$$

At the t moment, the vehicle scheduling can be divided into three cases by the oil time window [9] [10]:

a) When the T_{is} of the two gas stations are the same or when the T_{is} of the two gas stations are the different and $\Delta T_{is} \le t_0$, this paper considers unloading the refined oil to two gas station by using a vehicle with two compartments.

b) When $\Delta T_{is} > t_0$ and $V_i - W_i > 2 \times q$, this paper considers unloading the refined oil to gas station i by using a vehicle with two compartments.

c) When $\Delta T_{is} > t_0$ and $V_i - W_i < 2 \times q_d$, this paper considers unloading the refined oil to gas station i by using a vehicle with one compartments.

In the formula, ΔT_{is} the difference between the start time of any two gas stations' time window.

4. Model Solution

C language is an important programming language of computer software, which is widely used in computer software programming. Computer software programming based on the C language can greatly simplify the difficulty of programming and improve the accuracy of program operation results, and have the characteristics of quick solution speed and so on [11] [12]. Therefore, this paper uses C language to program the model.

1) Setting the opening time of the gas stating $T = 0$.

2) Calculating the inventory and time window according to the formula.

3) Judging whether to distribute according to the time window. When the start time of time window is equal to 0, it indicates that the inventory can be still sold for t_0 hours. At this time the vehicle begins to deliver.

4) Putting the constraint conditions into the loop control distribution.

5) Distribution plan:

(a) When the number of the start time of time window is equal to 0 is 1, according to the capacity and the inventory of gas stations to decide to use the vehicle with one compartment or two compartments.

(b) When the number of the start time of time window is equal to 0 is 2, the priority will be given to the vehicle with two compartments, which will be deliver refined oil to two gas stations.

(c) When the number of the start time of time window is equal to 0 is more than 2, it's going to be distributed by (b) and then by (a).

6) After refueling, let $T = T + 1$. Go to step 2).

7) Until $T = M$, end the loop.

8) Outputting the final result and getting the delivery cost.

5. Case Analysis

There are five gas stations in the area. The daily business hours are from 8:00 to 22:00, and the gas stations have the same kind of refined oil and are delivered by the same oil depot. The basic information of each gas station is shown in **Table 1**. The oil depot has two types of delivery vehicles, and each type of distribution vehicle is abundant, meanwhile compartment information and single trip freight are shown in **Table 2**. The delivery vehicles are parked at the depot and start working at 7 o'clock in the morning. The time between the oil depot and the gas stations, and the gas stations was one hour. The running time from the depot to the gas station and from one gas station to another gas station is one hour. In order to ensure the each gas station is not out of stock, it is reasonable to arrange the delivery plan and the delivery route of each delivery vehicle to make the delivery cost lowest.

Based on the model and solution process, the distribution plan of gas stations is shown in the following **Table 3**.

According to the distribution plan, we can get the distribution route of vehicles in the following **Table 4**.

Table 1. The information of gas stations and tanks unit: liter.

Name of gas station	Tank information		
	Capacity	Hourly sales	Inventory at 8 o'clock
S1	17,000	2000	8010
S2	17,000	3000	6008
S3	19,000	5000	15,015
S4	40,000	8000	16,020
S5	26,500	4000	26,009

Table 2. The information of distribution vehicle.

Types of vehicles	Capacity (liter)		Single-trip shipping costs (RMB)
The single	8000		100
The double	10,000	10,000	120

Table 3. Distribution plan unit: liter.

Delivery time	Types of vehicles	Name of gas station	Delivery quantity
9:00	The double	S2	10,000
10:00	The double	S4	10,000
10:00	The double	S3	10,000
11:00	The double	S4	10,000
11:00	The single	S1	8000
12:00	The double	S4	20,000
12:00	The double	S2	10,000
13:00	The double	S3	10,000
13:00	The double	S5	20000
14:00	The double	S3	10,000
15:00	The double	S4	10,000
15:00	The double	S1	10,000
16:00	The double	S4	10,000
16:00	The double	S2	10,000
17:00	The double	S3	10,000
17:00	The double	S4	20,000
18:00	The double	S3	10,000
19:00	The double	S5	10,000
19:00	The double	S2	10,000
20:00	The double	S4	10,000
20:00	The double	S1	10,000
21:00	The double	S3	10,000
21:00	The single	S4	8000

Table 4. Distribution route of vehicles.

Types of vehicles	The time of the vehicle leaving the oil depot	Route of vehicles
The double 1	8:00	The oil depot→S2→S4
The double 2	9:00	The oil depot→S3→S4
The single 1	10:00	The oil depot→S1
The double 3	11:00	The oil depot→S2→S3
The double 4	11:00	The oil depot→S4
The double 5	12:00	The oil depot→S5
The double 6	13:00	The oil depot→S3→S4
The double 7	14:00	The oil depot→S1→S4
The double 8	15:00	The oil depot→S2→S3
The double 9	16:00	The oil depot→S4
The double 10	17:00	The oil depot→S3→S5
The double 11	18:00	The oil depot→S2→S4
The double 12	19:00	The oil depot→S1→S3
The single 2	20:00	The oil depot→S4

According to the distribution plan, it can be concluded that the lowest delivery cost is 2 * 100 + 12 * 120 = 1640 yuan.

From the calculation result we can see that every gas station can successfully complete the supplementary task by using the proposed model and the method of solving, which is not appear the phenomenon of lacking of oil and wait for unloading. At the same time, making the whole freight as low as possible.

6. Conclusion

In this paper, based on the characteristics of refined oil distribution under the conditions of multiple vehicles, single oil and multiple gas stations, the paper establishes the distribution planning model of refined oil with the lowest freight rates as the objective function, the inventory of the gas station is less than the capacity and the gas station is not out of stock at any moment and other conditions as constraint conditions. The designs process of C language solves the model. The results prove that a feasible scheme can be obtained by using this model and algorithm. This paper only considers the factors such as vehicle transportation cost, gas tank volume, but in practical situation, the distribution problem of refined oil is also limited by the number of vehicles and the distance between the gas stations. In the future research, we can improve the model, increasing the influence factors and constraints, so as to increase its applicability.

Fund

Teaching Master of Beijing GaoChuang project Beijing (G02040011).

References

[1] Sun, T. (2012) The Research of Vehicle Scheduling Problem in Refined-Oil Secondary Logistics Distribution. Wuhan University of Technology, Wuhan.

[2] Yu, G.Y. (2016) Research on Vehicle Scheduling Problem of the Refined Oil Secondary Logistics Distribution. Chongqing Jiaotong University, Chongqing.

[3] Song, J.W. and Rong, G. (2003) Time Window Confirmation and Transportation Arrangement of Oil Distribution. *Systems Engineering-Theory & Practice*, **4**, 63-69.

[4] Liu, W.D., Li, S.J. and Liao, M.Y. (2005) Research on Scheduling and Optimization of Oil Distribution Plan. *Logistics Technology*, **9**, 22-25

[5] Yin, Q. and Dong, L.M. (2001) Discussion on the Refined Oil Distribution by Highway. *Oil Depot and Gas Station*, **10**, 1-3.

[6] Jin, L., Li, S.J. and Tang, L. (2007) Research on Refined Oil Distribution Plan Based on Heuristics Algorithm. *Logistics Technology*, **26**, 58-60.

[7] Liu, W.B., Yang, B., Yan, G. and Li, H.B. (2017) Real-Time Vehicle Scheduling Algorithm Based on Local Optimal Strategy. *Journal of Yibin University*, **9**, 1-7.

[8] Cheng, S.X. (2015) The Research of Refined Oil Secondary Distribution Vehicle Route Optimization Problem Based on C-W Saving Algorithm. Chang'an University, Xi'an.

[9] Hu, Y.Q. (2007) Operations Research Tutorial. Third Edition, Tsinghua University Press, Beijing, 50-60.

[10] Spliet, R. and Desaulniers, G. (2015) The Discrete Time Window Assignment Vehicle Routing Problem. *European Journal of Operational Research*, **244**, 379-391. https://doi.org/10.1016/j.ejor.2015.01.020

[11] Zhang, J.W. (2017) Computer Software Programming Analysis Based on C language. *Electronics World*, **5**, 69.

[12] Wei, P.P. and Cui, Z.W. (2017) C Language Programming Course Teaching Thinking. *Computer Era*, **9**, 64-66.

Scientific Research Publishing

Journal of Applied Mathematics and Physics, 2017, 5, 2112-2126
http://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

New Results of Global Asymptotical Stability for Impulsive Hopfield Neural Networks with Leakage Time-Varying Delay

Qiang Xi

School of Mathematic and Quantitative Economics, Shandong University of Finance and Economics, Ji'nan, China
Email: xiqiang_2000@163.com

How to cite this paper: Xi, Q. (2017) New Results of Global Asymptotical Stability for Impulsive Hopfield Neural Networks with Leakage Time-Varying Delay. *Journal of Applied Mathematics and Physics*, **5**, 2112-2126.
https://doi.org/10.4236/jamp.2017.511173

Received: October 9, 2017
Accepted: November 4, 2017
Published: November 7, 2017

Abstract

In this paper, Hopfield neural networks with impulse and leakage time-varying delay are considered. New sufficient conditions for global asymptotical stability of the equilibrium point are derived by using Lyapunov-Kravsovskii functional, model transformation and some analysis techniques. The criterion of stability depends on the impulse and the bounds of the leakage time-varying delay and its derivative, and is presented in terms of a linear matrix inequality (LMI).

Keywords

Global Asymptotical Stability, Hopfield Neural Networks, Leakage Time-Varying Delay, Impulse, Lyapunov-Kravsovskii Functional, Linear Matrix Inequality

1. Introduction

As we know, time delay is a common phenomenon that describes the fact that the future state of a system depends not only on the present state but also on the past state, and often encountered in many fields such as automatic control, biological chemistry, physical engineer, neural networks, and so on [1] [2] [3] [4] [5]. Moreover, the existence of time delay in a real system may lead to instability, oscillation, and bad dynamic performance [3] [4] [5]. So, it is significant and necessary to consider the delay effects on stability of dynamical systems. In Recent years, one typical class of neural networks, Hopfield neural networks (HNN) have been successfully applied to associative memory, pattern recognition, automatic control, optimization problems, etc, and HNN with various types of delay have been widely investigated by many authors, some interesting and

important results have been reported in the literature, see [6]-[16] and the references therein.

On the other hand, impulsive phenomenon exists universally in a wide variety of evolutionary processes where the state is changed abruptly at certain moments of time, involving such fields as chemical technology, population dynamics, physics and economics [17] [18] [19]. Hopfield neural networks may experience change of the state abruptly, that is, do exhibit impulsive effects. Recently, some results for the stability of HNN with impulse as well as delays are obtained via different approaches [20]-[26].

In the past several years, a special type of time delay, namely, leakage delay (or forgetting delay), is identified and investigated due to its existence in many real systems such as neural networks, population dynamics and some fuzzy systems [1] [3]. Leakage delay is a time delay that exists in the negative feedback terms of the system which are known as forgetting or leakage terms. It has been shown that such kind of time delay has a tendency to destabilize a system [27]. In [27], Gopalsamy initially investigated the dynamics of bidirectional associative memory (BAM) network model with leakage delays by using model transformation technique, Lyapunov-Kravsovskii functional and inequalities together with some properties of M-matrices. Based on this work, several papers have considered stability of some kinds of neural networks [28]-[34]. More recently, Li *et al.* [35], initially studies the impulsive effects on existence-uniqueness and stability problems of recurrent neural networks with leakage delay via some analysis techniques on impulsive functional differential equations. However, it is worth noting that in those existing results, the leakage delay considered is usually a constant. Stability research on leakage time-varying delay has been hardly considered in the literature. In [36], Li *et al.* studied the effect of leakage time-varying delay on stability of nonlinear differential systems, but ignored impulsive effect. It is interesting to consider neural networks with leakage time-varying delay as well as impulse, which describes more realistic models [37]-[40].

With the above motivation, in this paper, we consider Hopfield neural networks with leakage time-varying delay and impulse. By using Lyapunov-Kravsovskii functional, model transformation and some analysis techniques, New sufficient conditions for global asymptotical stability of the equilibrium point are derived. The criterion depends on the impulse and the bounds or length of the leakage time-varying delay and its derivative, and is given in terms of a linear matrix inequality (LMI). The developed results generalize the corresponding results in reference [36]. The work is organized as follows. In Section 2, we introduce the model, some basic notations and lemmas. In Section 3, we present the main results. Finally, the paper is concluded in Section 4.

2. Preliminaries

Notations. Let $\mathbb{R}$ denote the set of real numbers, $\mathbb{R}_+$ the set of nonnegative

real numbers, $\mathbb{Z}_+$ the set of positive integers, $\mathbb{R}^n$ the n-dimensional real space and $\mathbb{R}^{n\times m}$ $n\times m$-dimensional real space equipped with the Euclidean norm $|\cdot|$, respectively. For $\mathcal{S}=(s_{ij})\in\mathbb{R}^{n\times n}$, set $\|\mathcal{S}\|^2=\sum_{i=1}^n\sum_{j=1}^n s_{ij}^2$. $\mathcal{A}>0$ or $\mathcal{A}<0$ denotes that the matrix $\mathcal{A}$ is a symmetric and positive definite or negative definite matrix. The notation $\mathcal{A}^{\mathrm{T}}$ and $\mathcal{A}^{-}$ denote the transpose and the inverse of $\mathcal{A}$, respectively. If $\mathcal{A},\mathcal{B}$ are symmetric matrices, $\mathcal{A}>\mathcal{B}$ means that $\mathcal{A}-\mathcal{B}$ is positive definite matrix. $\lambda_{\max}(\mathcal{A})$ and $\lambda_{\min}(\mathcal{A})$ denote the maximum eigenvalue and the minimum eigenvalue of matrix $\mathcal{A}$, respectively. E denotes the identity matrix with appropriate dimensions and $\Lambda=\{1,2,\cdots,n\}$. For any $J\subseteq\mathbb{R}, S\subseteq\mathbb{R}^k\ (1\le k\le n)$, set $\mathbb{C}(J,S)=\{\phi:J\to S \text{ is continuous}\}$ and $\mathbb{PC}^1(J,S)=\{\phi:J\to S$ is continuously differentiable everywhere except at finite number of points t at which $\phi(t^+),\phi(t^-),\dot{\phi}(t^+),\dot{\phi}(t^-)$ exist and $\phi(t^+)=\phi(t)$, $\dot{\phi}(t^+)=\dot{\phi}(t)$ where $\dot{\phi}$ denotes the derivative of $\phi\}$. For any $t\in\mathbb{R}_+$, x_t is defined by $x_t=x(t+s)$, $x_{t^-}=x(t^-+s)$, $s\in[-\sigma,0]$. The notation $\star$ always denotes the symmetric block in one symmetric matrix.

Consider the following impulsive hopfield neural networks with leakage time-varying delay:

$$\begin{cases}\dot{x}(t)=-Cx(t-\sigma(t))+Af(x(t))+Bg(x(t-\tau(t)))+J, t>0, t\neq t_k,\\ \Delta x(t_k)=x(t_k)-x(t_k^-)=J_k(x(t_k^-),x_{t_k^-}), k\in\mathbb{Z}_+,\\ x(t)=\varphi(t), t\in[-\eta,0],\end{cases}\tag{1}$$

where $x(t)=(x_1(t),\cdots,x_n(t))^{\mathrm{T}}$ is the neuron state vector of the neural networks; $C=diag(c_1,\cdots,c_n)$ is a diagonal matrix with $c_i>0, i\in\Lambda$; A and B are the connection weight matrix and the delayed weight matrix, respectively; J is an external input; f and g represent the neuron activation functions. Through-out this paper, we make the following assumptions:

(H_1) $\sigma(t)$ and $\tau(t)$ denote the time-varying leakage delay and time-varying transmission delay, respectively, and satisfies $0\le\sigma(t)\le\sigma$, $0\le\tau(t)\le\tau$ and $|\dot{\sigma}(t)|\le\rho_\sigma<1$, $\dot{\tau}(t)\le\rho_\tau<1$, where $\sigma,\tau,\rho_\sigma,\rho_\tau$ are some real constants;

(H_2) $J_k(\cdot,\cdot):\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}^n, k\in\mathbb{Z}_+$, are some continuous functions;

(H_3) The impulsive times t_k satisfy $0=t_0<t_1<\cdots<t_k\to\infty$ and $\inf_{k\in\mathbb{Z}_+}\{t_k-t_{k-1}\}>0$.

(H_4) $\varphi\in\mathbb{PC}^1\doteq\mathbb{PC}^1([-\eta,0],\mathbb{R}^n)$, where $\eta\doteq\max\{\sigma,\tau\}$. For $\varphi\in\mathbb{PC}^1$, define $\|\varphi\|_\eta=\sup_{\theta\in[-\eta,0]}|\varphi(\theta)|$.

The following Lemmas will be used to derive our main results.

Lemma 2.1. [41] Given any real matrices $\Sigma_1,\Sigma_2,\Sigma_3$ of appropriate dimensions and a scalar $\epsilon>0$ such that $0<\Sigma_3=\Sigma_3^{\mathrm{T}}$. Then the following inequality holds:

$$\Sigma_1^{\mathrm{T}}\Sigma_2+\Sigma_2^{\mathrm{T}}\Sigma_1\le\epsilon\Sigma_1^{\mathrm{T}}\Sigma_3\Sigma_1+\epsilon^{-1}\Sigma_2^{\mathrm{T}}\Sigma_3^{-1}\Sigma_2.$$

Lemma 2.2. [42] Given any real matrix $M=M^{\mathrm{T}}>0$ of appropriate dimension and a vector function $\omega(\cdot):[a,b]\to\mathbb{R}^n$, such that the integrations

concerned are well defined, then

$$\left[\int_a^b \omega(s)\mathrm{d}s\right]^{\mathrm{T}} M\left[\int_a^b \omega(s)\mathrm{d}s\right] \le (b-a)\int_a^b \omega^{\mathrm{T}}(s) M\omega(s)\mathrm{d}s.$$

Lemma 2.3. [43] Let $X \in \mathbb{R}^{n\times n}$, then

$$\lambda_{\min}(X)a^{\mathrm{T}}a \le a^{\mathrm{T}}Xa \le \lambda_{\max}(X)a^{\mathrm{T}}a$$

for any $a \in \mathbb{R}^n$ if X is a symmetric matrix.

Lemma 2.4. [44] A given matrix $S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} > 0$, where $S_{11}^{\mathrm{T}} = S_{11}$, $S_{22}^{\mathrm{T}} = S_{22}$, is equivalent to any one of the following conditions:

(1) $S_{22} > 0, S_{11} - S_{12}S_{22}^{-1}S_{12}^{\mathrm{T}} > 0$;

(2) $S_{11} > 0, S_{22} - S_{12}^{\mathrm{T}}S_{11}^{-1}S_{12} > 0$.

In the following, we assume that some normal conditions, such as Lipschitz continuity of f and g, etc, are satisfied so that the equilibrium point of system (1) does exist, see [13] [21] etc, in which the existence results of equilibrium point are established by employing contraction mapping theorem, Brouwer's fixed point theorem and some functional method. Note that these results are independent of time delays, so it is easy to extend the results in the literatures to an impulsive neural network with leakage time-varying delays and other delays, we omit the details and investigate the global asymptotic stability of the equilibrium point mainly in next section. As usual, we assume that $x^* = (x_1^*, x_2^*, \cdots, x_n^*)^{\mathrm{T}}$ is an equilibrium point of system (1), *i.e.*

$$-Cx^* + Af(x^*) + Bg(x^*) + J = 0,\ J_k(x^*, x^*) = 0, k \in \mathbb{Z}_+.$$

3. Global Asymptotic Stability

In this section, we investigate the global asymptotic stability of the unique equilibrium point of system (1). For this purpose, the impulsive function J_k which is viewed as a perturbation of the equilibrium point x^* of model (1) without impulses is defined by

$$J_k\left(x(t_k^-), x_{t_k^-}\right) = -D_k\left\{x(t_k^-) - x^* - C\int_{t_k-\sigma(t_k)}^{t_k}\left(x(s)-x^*\right)\mathrm{d}s\right\}, k \in \mathbb{Z}_+,$$

where $D_k, k \in \mathbb{Z}_+$ are some $n\times n$ real symmetric matrices. It is clear that $J_k(x^*, x^*) = 0, k \in \mathbb{Z}_+$. Such a type of impulse describes the fact that the instantaneous perturbations encountered depend not only on the state of neurons at impulse times t_k but also the state of neurons in recent history, which reflects a more realistic dynamics. Similar impulsive perturbations have also been investigated by some researchers recently [22] [23] [25].

For convenience, we let $y(t) = x(t) - x^*$, then system (1) can be rewritten as

$$\begin{cases} \dot{y}(t) = -Cy(t-\sigma(t)) + A\Omega(y(t)) + B\Gamma(y(t-\tau(t))), t > 0, t \ne t_k, \\ \Delta y(t_k) = y(t_k) - y(t_k^-) = -D_k\left\{y(t_k^-) - C\int_{t_k-\sigma(t_k)}^{t_k} y(s)\mathrm{d}s\right\}, k \in \mathbb{Z}_+, \\ y(t) = \varphi(t) - x^*, t \in [-\eta, 0], \end{cases} \quad (2)$$

where

$$\Omega(y(t))=\left[\Omega_1(y_1(t)),\Omega_2(y_2(t)),\cdots,\Omega_n(y_n(t))\right]^{\mathrm{T}},$$
$$\Omega_j(y_j(t))=f_j(x_j^*+y_j(t))-f_j(x_j^*),$$
$$\Gamma(y(t-\tau(t)))=\left[\Gamma_1(y_1(t-\tau(t))),\Gamma_2(y_2(t-\tau(t))),\cdots,\Gamma_n(y_n(t-\tau(t)))\right]^{\mathrm{T}},$$
$$\Gamma_j(y_j(t-\tau(t)))=g_j(x_j^*+y_j(t-\tau(t)))-g_j(x_j^*).$$

Obviously, $y\equiv 0$ is a solution of system (2). Therefore, to consider the stability of the equilibrium point of system (1), it is equal to consider the stability of zero solution of system (2).

In this paper, we assume that there exist constants $M\geq 0, N\geq 0$ such that

(H_5) $\Omega^{\mathrm{T}}(y)\Omega(y)\leq My^{\mathrm{T}}y, \Gamma^{\mathrm{T}}(y)\Gamma(y)\leq Ny^{\mathrm{T}}y,$

which is a very important assumption for activation functions f and g. Using a model transformation, system (2) has an equivalent form as follows:

$$\begin{cases}\dfrac{\mathrm{d}}{\mathrm{d}t}\left[y(t)-C\int_{t-\sigma(t)}^{t}y(s)\mathrm{d}s\right]\\ =-Cy(t)-Cy(t-\sigma(t))\dot{\sigma}(t)+A\Omega(y(t))+B\Gamma(y(t-\tau(t))), t>0, t\neq t_k,\\ \Delta y(t_k)=y(t_k)-y(t_k^-)=-D_k\left\{y(t_k^-)-C\int_{t_k-\sigma(t_k)}^{t_k}y(s)\mathrm{d}s\right\}, k\in\mathbb{Z}_+,\\ y(t)=\varphi(t)-x^*, t\in[-\eta,0],\end{cases}\tag{3}$$

In the following, we shall establish a theorem which provides sufficient conditions for global asymptotical stability of the zero solution of system (3). It implies that, if system (1) has an equilibrium point, then it is unique and globally attractive.

Theorem 3.1. *Assume that system* (1) *has one equilibrium and that assumptions* (H_1)-(H_5) *hold. Then the equilibrium of system* (1) *is unique and is globally asymptotically stable if there exist* $n\times n$ *matrices* $P>0, Q_i>0, i=1,2,\cdots,7$ *such that the following LMI holds:*

$$\begin{bmatrix}\prod & \sqrt{\rho_\sigma}PC & PA & PB & \sigma C^{\mathrm{T}}PC & \sqrt{\rho_\sigma}\sigma C^{\mathrm{T}}PC & \sigma C^{\mathrm{T}}PA & \sigma C^{\mathrm{T}}PB\\ \star & -Q_1 & 0 & 0 & 0 & 0 & 0 & 0\\ \star & \star & -Q_2 & 0 & 0 & 0 & 0 & 0\\ \star & \star & \star & -Q_3 & 0 & 0 & 0 & 0\\ \star & \star & \star & \star & -Q_4 & 0 & 0 & 0\\ \star & \star & \star & \star & \star & -Q_5 & 0 & 0\\ \star & \star & \star & \star & \star & \star & -Q_6 & 0\\ \star & \star & \star & \star & \star & \star & \star & -Q_7\end{bmatrix}<0,\tag{4}$$

and

$$\begin{bmatrix}P & (E-D_k)^{\mathrm{T}}P\\ \star & P\end{bmatrix}>0, k\in\mathbb{Z}_+,\tag{5}$$

where

$$\prod = -2PC + \left[\lambda_{\max}(Q_2) + \lambda_{\max}(Q_6)\right]ME + Q_4 + \frac{\rho_\sigma}{1-\rho_\sigma}\left[Q_1 + Q_5\right] + \frac{1}{1-\rho_\tau}\lambda_{\max}(Q_3 + Q_7)NE.$$

Proof. Let $y(t) = y(t,0,\varphi)$ be a solution of system (2) through $(0,\varphi)$, where $\varphi \in \mathbb{C}$. Construct a Lyapunov-Krasovskii functional in the form

$$V(t,y) = V_1(t,y) + V_2(t,y) + V_3(t,y) + V_4(t,y), \tag{6}$$

where

$$V_1 = \left[y(t) - C\int_{t-\sigma(t)}^{t} y(s)\mathrm{d}s\right]^{\mathrm{T}} P\left[y(t) - C\int_{t-\sigma(t)}^{t} y(s)\mathrm{d}s\right],$$

$$V_2 = \frac{\rho_\sigma}{1-\rho_\sigma}\int_{t-\sigma(t)}^{t} y^{\mathrm{T}}(s)\left[Q_1 + Q_5\right]y(s)\mathrm{d}s,$$

$$V_3 = \frac{1}{1-\rho_\tau}\int_{t-\tau(t)}^{t} \Gamma^{\mathrm{T}}\left(y(s)\right)\left[Q_3 + Q_7\right]\Gamma\left(y(s)\right)\mathrm{d}s,$$

$$V_4 = \sigma\int_{t-\sigma}^{t}\int_{s}^{t} y^{\mathrm{T}}(u)Q_8 y(u)\mathrm{d}u\mathrm{d}s,$$

$$Q_8 = C^{\mathrm{T}}PCQ_4^{-1}C^{\mathrm{T}}PC + \rho_\sigma C^{\mathrm{T}}PCQ_5^{-1}C^{\mathrm{T}}PC + C^{\mathrm{T}}PAQ_6^{-1}A^{\mathrm{T}}PC + C^{\mathrm{T}}PBQ_7^{-1}B^{\mathrm{T}}PC.$$

Calculating the upper right derivative of $V(t,y)$ along the solution of system (2) at the continuous interval $[t_{k-1}, t_k), k \in \mathbb{Z}_+$, and considering the Lemma 2.1-2.3, it can be deduced that

$$\begin{aligned}
D^+V_1 &= 2\left[y(t) - C\int_{t-\sigma(t)}^{t} y(s)\mathrm{d}s\right]^{\mathrm{T}} P\left[-Cy(t) - Cy\left(t-\sigma(t)\right)\dot{\sigma}(t)\right.\\
&\quad \left. + A\Omega\left(y(t)\right) + B\Gamma\left(y\left(t-\tau(t)\right)\right)\right]\\
&= -2y^{\mathrm{T}}(t)PCy(t) - 2y^{\mathrm{T}}(t)PCy\left(t-\sigma(t)\right)\dot{\sigma}(t) + 2y^{\mathrm{T}}(t)PA\Omega\left(y(t)\right)\\
&\quad + 2y^{\mathrm{T}}(t)PB\Gamma\left(y\left(t-\tau(t)\right)\right) + 2\left[\int_{t-\sigma(t)}^{t} y(s)\mathrm{d}s\right]^{\mathrm{T}} C^{\mathrm{T}}PCy(t)\\
&\quad + 2\left[\int_{t-\sigma(t)}^{t} y(s)\mathrm{d}s\right]^{\mathrm{T}} C^{\mathrm{T}}PCy\left(t-\sigma(t)\right)\dot{\sigma}(t) - 2\left[\int_{t-\sigma(t)}^{t} y(s)\mathrm{d}s\right]^{\mathrm{T}} C^{\mathrm{T}}PA\Omega y(t)\\
&\quad - 2\left[\int_{t-\sigma(t)}^{t} y(s)\mathrm{d}s\right]^{\mathrm{T}} C^{\mathrm{T}}PB\Gamma\left(y\left(t-\tau(t)\right)\right)\\
&\le -2y^{\mathrm{T}}(t)PCy(t) + y^{\mathrm{T}}(t)PCQ_1^{-1}C^{\mathrm{T}}Py(t)\rho_\sigma + \rho_\sigma y^{\mathrm{T}}\left(t-\sigma(t)\right)Q_1 y\left(t-\sigma(t)\right)\\
&\quad + \Omega^{\mathrm{T}}\left(y(t)\right)Q_2\Omega\left(y(t)\right) + y^{\mathrm{T}}(t)PAQ_2^{-1}A^{\mathrm{T}}Py(t)\\
&\quad + \Gamma^{\mathrm{T}}\left(y\left(t-\tau(t)\right)\right)Q_3\Gamma\left(y\left(t-\tau(t)\right)\right) + y^{\mathrm{T}}(t)PBQ_3^{-1}B^{\mathrm{T}}Py(t)\\
&\quad + \left[\int_{t-\sigma(t)}^{t} y(s)\mathrm{d}s\right]^{\mathrm{T}} C^{\mathrm{T}}PCQ_4^{-1}C^{\mathrm{T}}PC\left[\int_{t-\sigma(t)}^{t} y(s)\mathrm{d}s\right]\\
&\quad + y^{\mathrm{T}}(t)Q_4 y(t) + \rho_\sigma\left[\int_{t-\sigma(t)}^{t} y(s)\mathrm{d}s\right]^{\mathrm{T}} C^{\mathrm{T}}PCQ_5^{-1}C^{\mathrm{T}}PC\left[\int_{t-\sigma(t)}^{t} y(s)\mathrm{d}s\right]\\
&\quad + \rho_\sigma y^{\mathrm{T}}\left(t-\sigma(t)\right)Q_5 y\left(t-\sigma(t)\right) + \left[\int_{t-\sigma(t)}^{t} y(s)\mathrm{d}s\right]^{\mathrm{T}} C^{\mathrm{T}}PAQ_6^{-1}A^{\mathrm{T}}PC\left[\int_{t-\sigma(t)}^{t} y(s)\mathrm{d}s\right]\\
&\quad + \Omega^{\mathrm{T}}\left(y(t)\right)Q_6\Omega\left(y(t)\right) + \left[\int_{t-\sigma(t)}^{t} y(s)\mathrm{d}s\right]^{T} C^{\mathrm{T}}PBQ_7^{-1}B^{\mathrm{T}}PC\left[\int_{t-\sigma(t)}^{t} y(s)\mathrm{d}s\right]\\
&\quad + \Gamma^{\mathrm{T}}\left(y\left(t-\tau(t)\right)\right)Q_7\Gamma\left(y\left(t-\tau(t)\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&\le -2y^{\mathrm{T}}(t)PCy(t)+y^{\mathrm{T}}(t)PCQ_1^{-1}C^{\mathrm{T}}Py(t)\rho_\sigma+\rho_\sigma y^{\mathrm{T}}\left(t-\sigma(t)\right)Q_1y\left(t-\sigma(t)\right)\\
&+y^{\mathrm{T}}(t)\lambda_{\max}(Q_2)MEy(t)+y^{\mathrm{T}}(t)PAQ_2^{-1}A^{\mathrm{T}}Py(t)\\
&+\Gamma^{\mathrm{T}}\left(y\left(t-\tau(t)\right)\right)Q_3\Gamma\left(y\left(t-\tau(t)\right)\right)+y^{\mathrm{T}}(t)PBQ_3^{-1}B^{\mathrm{T}}Py(t)\\
&+\left[\int_{t-\sigma(t)}^{t}y(s)\mathrm{d}s\right]^{\mathrm{T}}C^{\mathrm{T}}PCQ_4^{-1}C^{\mathrm{T}}PC\left[\int_{t-\sigma(t)}^{t}y(s)\mathrm{d}s\right]\\
&+y^{\mathrm{T}}(t)Q_4y(t)+\rho_\sigma\left[\int_{t-\sigma(t)}^{t}y(s)\mathrm{d}s\right]^{\mathrm{T}}C^{\mathrm{T}}PCQ_5^{-1}C^{\mathrm{T}}PC\left[\int_{t-\sigma(t)}^{t}y(s)\mathrm{d}s\right]\\
&+\rho_\sigma y^{\mathrm{T}}\left(t-\sigma(t)\right)Q_5y\left(t-\sigma(t)\right)+\left[\int_{t-\sigma(t)}^{t}y(s)\mathrm{d}s\right]^{\mathrm{T}}C^{\mathrm{T}}PAQ_6^{-1}A^{\mathrm{T}}PC\left[\int_{t-\sigma(t)}^{t}y(s)\mathrm{d}s\right]\\
&+y^{\mathrm{T}}(t)\lambda_{\max}(Q_6)NEy(t)+\left[\int_{t-\sigma(t)}^{t}y(s)\mathrm{d}s\right]^{\mathrm{T}}C^{\mathrm{T}}PBQ_7^{-1}B^{\mathrm{T}}PC\left[\int_{t-\sigma(t)}^{t}y(s)\mathrm{d}s\right]\\
&+\Gamma^{\mathrm{T}}\left(y\left(t-\tau(t)\right)\right)Q_7\Gamma\left(y\left(t-\tau(t)\right)\right),
\end{aligned}
\tag{7}
$$

$$
\begin{aligned}
D^+V_2&=\frac{\rho_\sigma}{1-\rho_\sigma}y^{\mathrm{T}}(t)\left[Q_1+Q_5\right]y(t)\\
&\quad-y^{\mathrm{T}}\left(t-\sigma(t)\right)\left[Q_1+Q_5\right]y\left(t-\sigma(t)\right)\frac{\rho_\sigma\left(1-\dot\sigma(t)\right)}{1-\rho_\sigma}\\
&\le\frac{\rho_\sigma}{1-\rho_\sigma}y^{\mathrm{T}}(t)\left[Q_1+Q_5\right]y(t)\\
&\quad-y^{\mathrm{T}}\left(t-\sigma(t)\right)\left[Q_1+Q_5\right]y\left(t-\sigma(t)\right)\rho_\sigma,
\end{aligned}
\tag{8}
$$

$$
\begin{aligned}
D^+V_3&=\frac{1}{1-\rho_\tau}\Gamma^{\mathrm{T}}\left(y(t)\right)\left[Q_3+Q_7\right]\Gamma\left(y(t)\right)\\
&\quad-\Gamma^{\mathrm{T}}\left(y\left(t-\tau(t)\right)\right)\left[Q_3+Q_7\right]\Gamma\left(y\left(t-\tau(t)\right)\right)\frac{1-\dot\tau(t)}{1-\rho_\tau}\\
&\le\frac{1}{1-\rho_\tau}\Gamma^{\mathrm{T}}\left(y(t)\right)\left[Q_3+Q_7\right]\Gamma\left(y(t)\right)\\
&\quad-\Gamma^{\mathrm{T}}\left(y\left(t-\tau(t)\right)\right)\left[Q_3+Q_7\right]\Gamma\left(y\left(t-\tau(t)\right)\right),
\end{aligned}
\tag{9}
$$

$$
\begin{aligned}
D^+V_4&=\sigma^2y^{\mathrm{T}}(t)Q_8y(t)-\sigma\int_{t-\sigma}^{t}y^{\mathrm{T}}(s)Q_8y(s)\mathrm{d}s\\
&\le\sigma^2y^{\mathrm{T}}(t)Q_8y(t)-\sigma(t)\int_{t-\sigma(t)}^{t}y^{\mathrm{T}}(s)Q_8y(s)\mathrm{d}s\\
&\le\sigma^2y^{\mathrm{T}}(t)Q_8y(t)-\left[\int_{t-\sigma(t)}^{t}y(s)\mathrm{d}s\right]^{\mathrm{T}}Q_8\left[\int_{t-\sigma(t)}^{t}y(s)\mathrm{d}s\right],
\end{aligned}
\tag{10}
$$

where

$$
Q_8=C^{\mathrm{T}}PCQ_4^{-1}C^{\mathrm{T}}PC+\rho_\sigma C^{\mathrm{T}}PCQ_5^{-1}C^{\mathrm{T}}PC+C^{\mathrm{T}}PAQ_6^{-1}A^{\mathrm{T}}PC+C^{\mathrm{T}}PBQ_7^{-1}B^{\mathrm{T}}PC.
$$

Combining (6)-(10), one may deduce that

$$
\begin{aligned}
D^+V&\le y^{\mathrm{T}}(t)\Big[-2PC+\rho_\sigma PCQ_1^{-1}C^{\mathrm{T}}P+\lambda_{\max}(Q_2)ME+PAQ_2^{-1}A^{\mathrm{T}}P+PBQ_3^{-1}B^{\mathrm{T}}P\\
&\quad+Q_4+\lambda_{\max}(Q_6)ME+\frac{\rho_\sigma}{1-\rho_\sigma}(Q_1+Q_5)+\frac{1}{1-\rho_\tau}\lambda_{\max}(Q_3+Q_7)NE\\
&\quad+\sigma^2C^{\mathrm{T}}PCQ_4^{-1}C^{\mathrm{T}}PC+\rho_\sigma\sigma^2C^{\mathrm{T}}PCQ_5^{-1}C^{\mathrm{T}}PC\\
&\quad+\sigma^2C^{\mathrm{T}}PAQ_6^{-1}A^{\mathrm{T}}PC+\sigma^2C^{T}PBQ_7^{-1}B^{\mathrm{T}}PC\Big]y(t)\\
&=y^{\mathrm{T}}(t)\Sigma y(t),
\end{aligned}
$$

where

$$\begin{aligned}\Sigma = &-2PC + \rho_\sigma PCQ_1^{-1}C^{\mathrm{T}}P + \lambda_{\max}(Q_2)ME + PAQ_2^{-1}A^{\mathrm{T}}P + PBQ_3^{-1}B^{\mathrm{T}}P\\ &+ Q_4 + \lambda_{\max}(Q_6)ME + \frac{\rho_\sigma}{1-\rho_\sigma}(Q_1+Q_5) + \frac{1}{1-\rho_\tau}\lambda_{\max}(Q_3+Q_7)NE\\ &+ \sigma^2 C^{\mathrm{T}}PCQ_4^{-1}C^{\mathrm{T}}PC + \rho_\sigma\sigma^2 C^{\mathrm{T}}PCQ_5^{-1}C^{\mathrm{T}}PC\\ &+ \sigma^2 C^{\mathrm{T}}PAQ_6^{-1}A^{\mathrm{T}}PC + \sigma^2 C^{\mathrm{T}}PBQ_7^{-1}B^{\mathrm{T}}PC\end{aligned}$$

By the well known Schur complements, we know that $\Sigma < 0$ if and only if the LMI (4) holds. Hence, one may derive that

$$D^+V(t,y) \le -y^{\mathrm{T}}(t)\Sigma^* y(t), t \in [t_{k-1}, t_k), k \in \mathbb{Z}_+, \tag{11}$$

where $\Sigma^* = -\Sigma > 0$.

Suppose that $t \in [t_{n-1}, t_n)$, for some $n \in \mathbb{Z}_+$. Then integrating inequality (11) at each interval $[t_{k-1}, t_k), 1 \le k \le n-1$, we derive that

$$\begin{aligned}V(t_1^-) &\le V(0) - \int_0^{t_1} y^{\mathrm{T}}(s)\Sigma^* y(s)\mathrm{d}s,\\ V(t_2^-) &\le V(t_1) - \int_{t_1}^{t_2} y^{\mathrm{T}}(s)\Sigma^* y(s)\mathrm{d}s,\\ &\vdots\\ V(t_{n-1}^-) &\le V(t_{n-2}) - \int_{t_{n-2}}^{t_{n-1}} y^{\mathrm{T}}(s)\Sigma^* y(s)\mathrm{d}s,\\ V(t) &\le V(t_{n-1}) - \int_{t_{n-1}}^{t} y^{\mathrm{T}}(s)\Sigma^* y(s)\mathrm{d}s,\end{aligned}$$

which implies that

$$V(t) \le V(0) - \int_0^t y^{\mathrm{T}}(s)\Sigma^* y(s)\mathrm{d}s + \sum_{0<t_k\le t}\left[V(t_k) - V(t_k^-)\right], t \ge 0. \tag{12}$$

In order to analyze (12), we need consider the change of V at impulse times.

Firstly, it follows from (5) that

$$\begin{aligned}\begin{bmatrix} P & (E-D_k)^{\mathrm{T}}P \\ \star & P \end{bmatrix} > 0 &\Leftrightarrow \begin{bmatrix} E & O \\ O & P^{-1} \end{bmatrix}\begin{bmatrix} P & (E-D_k)^{\mathrm{T}}P \\ \star & P \end{bmatrix}\begin{bmatrix} E & O \\ O & P^{-1} \end{bmatrix} > 0\\ &\Leftrightarrow \begin{bmatrix} P & (E-D_k)^{\mathrm{T}} \\ \star & P^{-1} \end{bmatrix} > 0\\ &\Leftrightarrow P - (E-D_k)^{\mathrm{T}}P(E-D_k) > 0,\end{aligned} \tag{13}$$

in which the last equivalent relation is obtained by Lemma 2.4.

Secondly, from system (3), it can be obtained that

$$\begin{aligned}&y(t_k) - C\int_{t_k-\sigma(t_k)}^{t_k} y(s)\mathrm{d}s\\ &= y(t_k^-) - D_k\left[y(t_k^-) - C\int_{t_k-\sigma(t_k)}^{t_k} y(s)\mathrm{d}s - C\int_{t_k-\sigma(t_k)}^{t_k} y(s)\mathrm{d}s\right]\\ &= (E-D_k)\left[y(t_k^-) - C\int_{t_k-\sigma(t_k)}^{t_k} y(s)\mathrm{d}s\right],\end{aligned}$$

which together with (13) yields

$$V_1(t_k)=\left[y(t_k)-C\int_{t_k-\sigma(t_k)}^{t_k}y(s)\mathrm{d}s\right]^T P\left[y(t_k)-C\int_{t_k-\sigma(t_k)}^{t_k}y(s)\mathrm{d}s\right]$$
$$=\left[y(t_k^-)-C\int_{t_k-\sigma(t_k)}^{t_k}y(s)\mathrm{d}s\right]^{\mathrm{T}}(E-D_k)^{\mathrm{T}}P(E-D_k)\left[y(t_k^-)-C\int_{t_k-\sigma(t_k)}^{t_k}y(s)\mathrm{d}s\right]$$
$$<\left[y(t_k^-)-C\int_{t_k-\sigma(t_k)}^{t_k}y(s)\mathrm{d}s\right]^{\mathrm{T}}P\left[y(t_k^-)-C\int_{t_k-\sigma(t_k)}^{t_k}y(s)\mathrm{d}s\right]=V_1(t_k^-).$$

Obviously, we have $V_i(t_k)\le V_i(t_k^-), i=2,3,4, k\in\mathbb{Z}_+$.

Thus, we can deduce that

$$V(t_k)\le V(t_k^-), k\in\mathbb{Z}_+.$$

Substituting the above inequality in (12) yields

$$V(t)+\int_0^t y^{\mathrm{T}}(s)\Sigma^* y(s)\mathrm{d}s\le V(0), t\ge 0. \tag{14}$$

By simple calculation, it can be deduced that

$$V(0)\le\left\{\lambda_{\max}(P)\left(1+\|C\|\sigma\right)^2+\frac{\rho_\sigma\sigma}{1-\rho_\sigma}\lambda_{\max}(Q_1+Q_5)\right.$$
$$\left.+\frac{\tau\lambda_{\max}(Q_3+Q_7)N}{1-\rho_\tau}+\sigma^3\lambda_{\max}(Q_8)\right\}\|\varphi\|_\eta^2$$
$$=\Delta\|\varphi\|_\eta^2,$$

where

$$\Delta=\lambda_{\max}(P)\left(1+\|C\|\sigma\right)^2+\frac{\rho_\sigma\sigma}{1-\rho_\sigma}\lambda_{\max}(Q_1+Q_5)$$
$$+\frac{\tau\lambda_{\max}(Q_3+Q_7)N}{1-\rho_\tau}+\sigma^3\lambda_{\max}(Q_8).$$

It follows that

$$\sqrt{\lambda_{\min}(P)}\left\|y(t)-C\int_{t-\sigma(t)}^t y(s)\mathrm{d}s\right\|\le\sqrt{V_1}\le\sqrt{V}\le\sqrt{V(0)}\le\sqrt{\Delta}\|\varphi\|_\eta,$$

which implies that

$$\|y(t)\|\le\|C\|\int_{t-\sigma(t)}^t y(s)\mathrm{d}s+\sqrt{\frac{\Delta}{\lambda_{\min}(P)}}\|\varphi\|_\eta.$$

Employing Gronwall inequality, we get

$$\|y(t)\|\le\sqrt{\frac{\Delta}{\lambda_{\min}(P)}}\|\varphi\|_\eta\,\mathrm{e}^{\sigma(t)\|C\|}$$
$$\le\sqrt{\frac{\Delta}{\lambda_{\min}(P)}}\mathrm{e}^{\sigma\|C\|}\|\varphi\|_\eta<\infty,$$

which implies that the equilibrium point of system (2) is locally stable, and uniformly bounded on $[0,\infty)$.

Thus, considering the continuity of the activation function *f* and *g*, it can be deduced from system (2) that there exists some constant $R>0$ such that $\|\dot{y}(t)\|\le R, t\in[t_{k-1},t_k), k\in\mathbb{Z}_+$, where $\dot{y}$ denotes the right-hand derivative of *y* at impulse times $t_{k-1}, k\in\mathbb{Z}_+$.

In the following, we shall prove that $\|y(t)\|\to 0$ as $t\to\infty$.

We first show that

$$\left\|y\left(t_k\right)\right\| \to 0, t_k \to \infty. \tag{15}$$

It is equivalent to prove that $\left|y_i\left(t_k\right)\right| = 0$ as $t_k \to \infty, i \in \Lambda$. Note that $\left|\dot{y}_i(t)\right| \le R, t \in \left[t_{k-1}, t_k\right), k \in \mathbb{Z}_+$, then for any $\epsilon > 0$, there exists a $\delta = \frac{\epsilon}{2R} > 0$ such that, for any $t', t'' \in \left[t_{k-1}, t_k\right), k \in \mathbb{Z}_+, \left|t' - t''\right| < \delta$ implies that

$$\left|y_i\left(t'\right) - y_i\left(t''\right)\right| \le R\left|t' - t''\right| = \frac{\varepsilon}{2}, i \in \Lambda. \tag{16}$$

By (H_3), we define $\bar{\delta} = \min\left\{\delta, \frac{1}{2}\theta\right\}$, where $\theta = \inf_{k \in \mathbb{Z}_+}\left\{t_k - t_{k-1}\right\} > 0$. From (14), it can be obtained that

$$\int_0^t \left|y_i(s)\right|^2 \mathrm{d}s \le \int_0^t y(s)^{\mathrm{T}} y(s) \mathrm{d}s \le \frac{1}{\lambda_{\min}\left(\Sigma^*\right)} \int_0^t y(s)^{\mathrm{T}} \Sigma^* y(s) \mathrm{d}s < \infty, t > 0,$$

which implies that $\int_{t_k}^{t_k+\bar{\delta}} \left|y_i(s)\right|^2 \mathrm{d}s \to 0$ as $t_k \to \infty$.

Applying Lemma 2.2, we get

$$\int_{t_k}^{t_k+\bar{\delta}} \left|y_i(s)\right| \mathrm{d}s \le \sqrt{\bar{\delta} \int_{t_k}^{t_k+\bar{\delta}} \left|y_i(s)\right|^2 \mathrm{d}s} \to 0, t_k \to \infty. \tag{17}$$

So for the above-given ϵ, there exists a $T = T(\epsilon) > 0$ such that $t_k > T$ implies that

$$\int_{t_k}^{t_k+\bar{\delta}} \left|y_i(s)\right| \mathrm{d}s < \frac{\epsilon}{2}\bar{\delta}.$$

From the continuity of $\left|y_i(t)\right|$ on $\left[t_k, t_k + \bar{\delta}\right]$, and using the integral mean value theorem, there exists some constant $\xi_k \in \left[t_k, t_k + \bar{\delta}\right]$ such that

$$\left|y_i\left(\xi_k\right)\right|\bar{\delta} = \int_{t_k}^{t_k+\bar{\delta}} \left|y_i(s)\right| \mathrm{d}s < \frac{\epsilon}{2}\bar{\delta},$$

which leads to

$$\left|y_i\left(\xi_k\right)\right| < \frac{\epsilon}{2}. \tag{18}$$

Combining (16) and (18), one may deduce that, for any $\epsilon > 0$, there exists a $T = T(\epsilon) > 0$ such that $t_k > T$ implies that

$$\left|y_i\left(t_k\right)\right| \le \left|y_i\left(t_k\right) - y_i\left(\xi_k\right)\right| + \left|y_i\left(\xi_k\right)\right| \le \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

This completes the proof of (15).

Now we are in a position to prove that $\left|y_i(t)\right| \to 0$ as $t \to \infty, i \in \Lambda$. In fact, it follows from (16) that, for any $\epsilon > 0$, there exists a $\delta = \frac{\epsilon}{2M} > 0$ such that, for any $t', t'' \in \left[t_{k-1}, t_k\right), k \in \mathbb{Z}_+, \left|t' - t''\right| < \delta$ implies that

$$\left|y_i\left(t'\right) - y_i\left(t''\right)\right| \le \frac{\epsilon}{2}, i \in \Lambda. \tag{19}$$

Since (15) holds, there exists a constant $T_1 = T_1(\epsilon) > 0$ such that

$$\left|y_i(t_k)\right| < \frac{\epsilon}{2}, t_k > T_1. \tag{20}$$

In addition, applying the same argument as in (17), we can deduce that

$$\int_t^{t+\bar{\delta}} \left|y_i(s)\right| \mathrm{d}s \to 0, t \to \infty,$$

where $\bar{\delta} = \min\left\{\delta, \frac{1}{2}\theta\right\}, \theta = \inf_{k\in\mathbb{Z}_+}\{t_k - t_{k-1}\} > 0.$

So, for the above-given ϵ, there exists a constant $T_2 = T_2(\epsilon) > 0$ such that

$$\int_{t-\bar{\delta}}^{t} \left|y_i(s)\right| \mathrm{d}s < \frac{\epsilon}{2}\bar{\delta}, t > T_2. \tag{21}$$

Set $T^* = \min\left\{t_q \middle| t_q \geq \max\{T_1, T_2\}, q \in \mathbb{Z}_+\right\}$. Now we claim that $\left|y_i(t)\right| \leq \epsilon, t > T^*$. In fact, for any $t > T^*$ and without loss of generality assume that $t \in \left[t_p, t_{p+1}\right), p \geq q$. We consider the following two cases.

Case1. $t \in \left[t_p, t_p + \bar{\delta}\right]$. In this case, it is obvious from (19) and (20) that

$$\left|y_i(t)\right| \leq \left|y_i(t) - y_i(t_p)\right| + \left|y_i(t_p)\right| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

Case2. $t \in \left[t_p + \bar{\delta}, t_{p+1}\right)$. In this case, we know that $y_i(s)$ is continuous on $\left[t-\bar{\delta}, t\right] \subseteq \left[t_p, t_{p+1}\right)$. By the integral mean value theorem, there exists at least one point $\upsilon_t \in \left[t-\bar{\delta}, t\right]$ such that

$$\int_{t-\bar{\delta}}^{t} \left|y_i(s)\right| \mathrm{d}s = \left|y_i(\upsilon_t)\right| \bar{\delta},$$

which together with (21) yields $\left|y_i(\upsilon_t)\right| < \frac{\epsilon}{2}$. Then, in view of $\upsilon_t \in \left[t-\bar{\delta}, t\right]$, we obtain

$$\left|y_i(t)\right| \leq \left|y_i(t) - y_i(\upsilon_t)\right| + \left|y_i(\upsilon_t)\right| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

So we have proved that $\left|y_i(t)\right| \leq \epsilon, t > T^*$. Therefore, the zero solution of system (2) or (3) is globally asymptotically stable, which implies that system (1) has a unique equilibrium point which is globally asymptotically stable. The proof of Theorem 3.1 is therefore complete. □

Remark 3.1. Theorem 3.1 provides some delay-dependent conditions for the global asymptotical stability of the unique equilibrium point of impulsive Hopfield neural networks with leakage time-varying delay. We would like to note that such a result has not been reported in other literatures.

In particular, when the leakage delay and transmission delay are all constants, *i.e.*, $\sigma(t) \equiv \sigma, \tau(t) \equiv \tau$, system (1) becomes

$$\begin{cases} \dot{x}(t) - Cx(t-\sigma) + Af\left(x(t)\right) + Bg\left(x(t-\tau)\right) + J, t > 0, t \neq t_k, \\ \Delta x(t_k) = x(t_k) - x(t_k^-) = J_k\left(x(t_k^-), x_{t_k^-}\right), k \in \mathbb{Z}_+, \\ x(t) = \varphi(t), t \in [-\eta, 0]. \end{cases} \tag{22}$$

For system (22), we have the following result by Theorem 3.1.

Corollary 3.1. *Assume that system* (22) *has one equilibrium and that assumptions* (H_2)-(H_5) *hold. Then the equilibrium of system* (22) *is unique and is globally asymptotically stable if there exist* $n\times n$ *matrices* $P>0, Q_i>0, i=1,2,\cdots,5$ *such that the following LMI holds:*

$$\begin{bmatrix} \prod & PA & PB & \sigma C^{\mathrm{T}}PC & \sigma C^{\mathrm{T}}PA & \sigma C^{\mathrm{T}}PB \\ \star & -Q_1 & 0 & 0 & 0 & 0 \\ \star & \star & -Q_2 & 0 & 0 & 0 \\ \star & \star & \star & -Q_3 & 0 & 0 \\ \star & \star & \star & \star & -Q_4 & 0 \\ \star & \star & \star & \star & \star & -Q_5 \end{bmatrix} < 0,$$

and

$$\begin{bmatrix} P & \left(E-D_k\right)^{\mathrm{T}}P \\ \star & P \end{bmatrix} > 0, k\in\mathbb{Z}_+,$$

where

$$\prod = -2PC + \left[\lambda_{\max}\left(Q_1\right) + \lambda_{\max}\left(Q_4\right)\right]ME + Q_3 + \lambda_{\max}\left(Q_2+Q_5\right)NE.$$

Remark 3.2. The conditions in Corollary 3.1 are independent on transmission delay and dependent only on leakage delay as $\rho_\tau = 0$ in Theorem 3.1. So, based on our results, we would like to say that the stability of system (1) is more sensitive to leakage delay, leakage time-varying delay or leakage constant delay. In other words, we should control not only the bound of leakage delay but also the bound of derivative of leakage delay, to obtain the stability of system (1), while the bound of transmission delay τ or $\tau(t)$ do not affect the stability of system in our results.

Remark 3.3. So far, there are many papers to study the dynamics of time delay systems and impulsive systems, many effective methods and results have been developed [19]-[26]. But, those results cannot be applied to systems with leakage time-varying delay and impulse which could affect the dynamics of system essentially. In this paper, we investigate the stability of impulsive Hopfield neural networks with leakage time-varying delay by model transformation technique and a certain Lyapunov-Krasovskii functional combined with LMI technique and construct a new criterion. How to improve the dynamics of systems with leakage time-varying delay and impulse may be an interesting problem and requires further research.

4. Conclusion

We have studied the global asymptotic stability of the equilibrium point of impulsive Hopfield neural networks with leakage time-varying delay. Via an appropriate Lyapunov-Krasovskii functional and model transformation technique, a new stability criterion which depends on the impulse and the bounds of leakage time-varying delay and its derivative has been presented in

terms of a linear matrix inequality. To the best of our knowledge, so far, few authors have considered the dynamics of systems with leakage time-varying delay and impulse which could affect the dynamics of neural networks essentially. How to further improve the conservation of the developed results is still a difficult problem and need consideration in the future work.

Fund

Supported by National Natural Science Foundation of China (11601269) and Project of Shan-dong Province Higher Educational Science and Technology Program (J15LI02).

References

[1] Gopalsamy, K. (1992) Stability and Oscillations in Delay Differential Equations of Population Dynamics. Kluwer, Dordrecht.

[2] Hale, J. and Verduyn Lunel, S. (1993) Introduction to Functional Differential Equations. Springer-Verlag, New York.

[3] Haykin, S. (1994) Neural Networks. Prentice-Hall, NJ.

[4] Kolmanovskii, V. and Myshkis, A. (1999) Introduction to the Theory and Applications of Functional Differential Equations. Kluwer, Dordrecht, The Netherlands.

[5] Niculescu, S. (2001) Delay Effects on Stability: A Robust Control Approach. Springer-Verlag, New York.

[6] Van den Driessche, P. and Zou, X. (1998) Global Attractivity in Delayed Hopfield Neural Network Models. *SIAM Journal on Applied Mathematics*, **58**, 1878-1890. https://doi.org/10.1137/S0036139997321219

[7] Liao, X. and Yu, J. (1998) Robust Stability for Interval Hopfield Neural Networks with Time Delay. *IEEE Transactions on Neural Networks*, **9**, 1042-1045. https://doi.org/10.1109/72.712187

[8] Liao, X. and Xiao D. (2000) Global Exponential Stability of Hopfield Neural Networks with Time Varying Delays. *Acta Electronica Sinica*, **28**, 87-90.

[9] Chen, T. (2001) Global Exponential Stability of Delayed Hopfield Neural Networks. *Neural Networks*, **14**, 977-980.

[10] Wang, L. and Xu, D. (2002) Stability Analysis of Hopfield Neural Networks with Time Delays. *Applied Mathematics and Mechanics* (*English Edition*), **23**, 65-70. https://doi.org/10.1007/BF02437731

[11] Chen, G., Pu, Z. and Zhang, J. (2005) Global Exponential Stability and Global Attractivity for Variably Delayed Hopfield Neural Network Models. *Chinese Journal of Engineering Mathematics*, **22**, 821-826.

[12] Hou, Y., Liao, T. and Yan, J. (2007) Global Asymptotic Stability for a Class of Nonlinear Neural Networks with Multiple Delays. *Nonlinear Analysis*, *Theory*, *Methods*, *Applications*, **67**, 3037-3040.

[13] Zhang, Q., Wei, X. and Xu, J. (2007) Delay-Dependent Global Stability Results for Delayed Hopfield Neural Networks. *Chaos*, *Solitons Fractals*, **34**, 662-668.

[14] Lou, X. and Cui, B. (2007) Novel Global Stability Criteria for High-Order Hopfield-Type Neural Networks with Time-Varying Delays. *Journal of Mathematical Analysis and Applications*, **330**, 144-158.

[15] Zhou, Q. and Wan, L. (2008) Exponential Stability of Stochastic Delayed Hopfield Neural Networks. *Applied Mathematics and Computation*, **206**, 818-824.

[16] Wu, H. (2009) Global Exponential Stability of Hopfield Neural Networks with Delays and Inverse Lipschitz Neuron Activations. *Nonlinear Analysis*: *Real World Applications*, **14**, 2776-2783.

[17] Bainov, D. and Simennov, P. (1989) Systems with Impulsive Effect Stability Theory and Applications. Ellis Horwood Limited, New York.

[18] Fu, X., Yan, B. and Liu, Y. (2005) Introduction of Impulsive Differential Systems. Science Press, Beijing.

[19] Liu, X. and Wang, Q. (2007) The Method of Lyapunov Functionals and Exponential Stability of Impulsive Systems with Time Delay. *Nonlinear Analysis*, *Theory*, *Methods*, *Applications*, **66**, 1465-1484.

[20] Liu, B., Liu, X. and Liao, X. (2003) Robust H-Stability of Hopfield Neural Networks with Impulsive Effects and Design of Impulsive Controllers. *Control Theory Applications*, **20**, 168-172.

[21] Yang, Z. and Xu, D. (2006) Global Exponential Stability of Hopfield Neural Networks with Variable Delays and Impulsive Effects. *Applied Mathematics and Machanics*, **27**, 1517-1522. https://doi.org/10.1007/s10483-006-1109-1

[22] Shen, J., Liu, Y. and Li, J. (2007) Asymptotic Behavior of Solutions of Nonlinear Neural Differential Equations with Impulses. *Journal of Mathematical Analysis and Applications*, **332**, 179-189.

[23] Zhou, Q. (2009) Global Exponential Stability of BAM Neural Networks with Distributed Delays and Impulses. *Nonlinear Analysis*: *Real World Applications*, **10**, 144-153.

[24] Li, X. and Chen, Z. (2009) Stability Properties for Hopfield Neural Networks with Delays and Impulsive Perturbations. *Nonlinear Analysis*, *Theory*, *Methods*, *Applications*, **10**, 3253-3265.

[25] Fu, X. and Li, X. (2009) Global Exponential Stability and Global Attractivity of Impulsive Hopfield Neural Networks with Time Delays. *Journal of Computational and Applied Mathematics*, **231**, 187-199.

[26] Chen, J., Li, X. and Wang, D. (2013) Asymptotic Stability and Exponential Stability of Impulsive Delayed Hopfield Neural Networks. *Abstract and Applied Analysis*, **2013**, Article ID: 638496.

[27] Gopalsamy, K. (2007) Leakage Delays in BAM. *Journal of Mathematical Analysis and Applications*, **325**, 1117-1132.

[28] Peng, S. (2010) Attractive Periodic Solutions of BAM Neural Networks with Continuously Distributed Delays in the Leakage Terms. *Nonlinear Analysis*: *Real World Applications*, **11**, 2141-2151.

[29] Li, X. and Cao, J. (2010) Delay-Dependent Stability of Neural Networks of Neutral Type with Time Delay in the Leakage Term. *Nonlinearity*, **23**, 1709-1726. https://doi.org/10.1088/0951-7715/23/7/010

[30] Balasubramaniam, P., Kalpana, M. and Rakkiyappan, R. (2011) State Estimation for Fuzzy Cellular Neural Networks with Time Delay in the Leakage Term, Discrete and Unbounded Distributed Delays. *Computers and Mathematics with Applications*, **62**, 3959-3972.

[31] Lakshmanan, S., Park, J., Jung, H. and Balasubramaniam, P. (2012) Design of State Estimator for Neural Networks with Leakage, Discrete and Distributed Delays. *Ap-*

plied Mathematics and Computation, **218**, 11297-11310.

[32] Li, Z. and Xu, R. (2012) Global Asymptotic Stability of Stochastic Reaction-Diffusion Neural Networks with Time Delays in the Leakage Terms. *Communications in Nonlinear Science and Numerical Simulation*, **17**, 1681-1689.

[33] Lakshmanan, S., Park, J., Lee, T., Jung, H. and Rakkiyappan, R. (2013) Stability Criteria for BAM Neural Networks with Leakage Delays and Probabilistic Time-Varying Delays. *Applied Mathematics and Computation*, **219**, 9408-9423.

[34] Li, X. and Rakkiyappan, R. (2013) Stability Results for Takagi CSugeno Fuzzy Uncertain BAM Neural Networks with Time Delays in the Leakage Term. *Neural Computing and Applications*, **22**, 203-219.

[35] Li, X., Fu, X., Balasubramaniam, P. and Rakkiyappan, R. (2010) Existence, Uniqueness and Stability Analysis of Recurrent Neural Networks with Time Delay in the Leakage Term under Impulsive Perturbations. *Nonlinear Analysis*: *Real World Applications*, **11**, 4092-4108.

[36] Li, X. and Fu, X. (2013) Effect of Leakage Time-Varying Delay on Stability of Nonlinear Differential Systems. *Journal of the Franklin Institute*, **350**, 1335-1344.

[37] Rakkiyappan, R., Chandrasekar, A. and Lakshmanan, S. (2013) Effects of Leakage Time-Varying Delays in Markovian Jump Neural Networks with Impulse Control. *Neurocomputing*, **121**, 365-378.

[38] Lu, C. and Wang, L. (2014) Robust Exponential Stability of Impulsive Stochastic Neural Networks with Leakage Time-Varying Delay. *Abstract and Applied Analysis*, **2014**, Article ID: 831027.

[39] Suresh Kumar, R., Sugumaran, G., Raja, R., Zhu, Q. and Karthik Raja, U. (2016) New Stability Criterion of Neural Networks with Leakage Delays and Impulses: A Piecewise Delay Method. *Cognitive Neurodynamics*, **10**, 85-98. https://doi.org/10.1007/s11571-015-9356-y

[40] Balasundaram, K., Raja, R., Zhu, Q., Chandrasekaran, S. and Zhou, H. (2016) New Global Asymptotic Stability of Discrete-Time Recurrent Neural Networks with Multiple Time-Varying Delays in the Leakage Term and Impulsive Effects. *Cognitive Neurodynamics*, **214**, 420-429.

[41] Li, C. and Huang, T. (2009) On the Stability of Nonlinear Systems with Leakage Delay. *Journal of the Franklin Institute*, **346**, 366-377.

[42] Gu, K. (2000) An Integral Inequality in the Stability Problem of Time-Delay Systems. *Proceedings of the* 39*th IEEE Conference on Decision and Control Sydney*, December, 2805-2810.

[43] Berman, A. and Plemmons, R.J. (1979) Nonnegative Matrics in Mathematical Sciences. Academic Press, New York.

[44] Boyd, S., Ghaoui, L. E., Feron, E. and Balakrishnan, V. (1994) Linear Matrix Inequalities in System and Control Theory. SIAM, Philadelphia.

Journal of Applied Mathematics and Physics, 2017, 5, 2127-2139
http://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Variational Formulations Yielding High-Order Finite-Element Solutions in Smooth Domains without Curved Elements

Vitoriano Ruas[1,2,3]

[1]Sorbonne Universités, UPMC Univ Paris 06 & CNRS, UMR 7190, IJRDA, Paris, France
[2]CNRS, UMR 7190, Institut Jean Le Rond d'Alembert, Paris, France
[3]CNPq Research Grant Holder, PUC-Rio, Rio de Janeiro, Brazil
Email: vitoriano.ruas@upmc.fr, ruas.vitoriano@gmail.com, vitoriano.ruas@pq.cnpq.br

How to cite this paper: Ruas, V. (2017) Variational Formulations Yielding High-Order Finite-Element Solutions in Smooth Domains without Curved Elements. *Journal of Applied Mathematics and Physics*, **5**, 2127-2139.
https://doi.org/10.4236/jamp.2017.511174

Received: September 24, 2017
Accepted: November 4, 2017
Published: November 7, 2017

Abstract

One of the reasons for the great success of the finite element method is its versatility to deal with different types of geometries. This is particularly true of problems posed in curved domains. Nevertheless it is well-known that, for standard variational formulations, the optimal approximation properties known to hold for polytopic domains are lost, if meshes consisting of ordinary elements are still used in the case of curved domains. That is why method's isoparametric version for meshes consisting of curved triangles or tetrahedra has been widely employed, especially in case Dirichlet boundary conditions are prescribed all over a curved boundary. However, besides geometric inconveniences, the isoparametric technique helplessly requires the manipulation of rational functions and the use of numerical integration. In this work we consider a simple alternative that bypasses these drawbacks, without eroding qualitative approximation properties. More specifically we work with a variational formulation leading to high order finite element methods based only on polynomial algebra, since they do not require the use of curved elements. Application of the new approach to Lagrange methods of arbitrary order illustrates its potential to take the best advantage of finite-element discretizations in the solution of wide classes of problems posed in curved domains.

Keywords

Curved Domain, Dirichlet, Finite Elements, Interpolated Boundary Condition, Polynomial Algebra

1. Introduction

This work deals with a new variational formulation designed for finite-element

solution methods of boundary value problems with Dirichlet boundary conditions, posed in a two- or three-dimensional domain having a smooth curved boundary of arbitrary shape. The principle it is based upon is related to the technique called *interpolated boundary conditions* studied in [1] for two-dimensional problems. Although the latter technique is very intuitive and has been known since the seventies (cf. [2] [3]), it has been of limited use so far. Among the reasons for this we could quote its difficult implementation, the lack of an extension to three-dimensional problems, and most of all, restrictions on the choice of boundary nodal points to reach optimal convergence rates. In contrast our method is simple to implement in both in two- and three-dimensional geometries. Moreover optimality is attained very naturally in both cases for various choices of boundary nodal points.

In order to allow an easier description of our methodology, thereby avoiding non essential technicalities, we consider as a model the Poisson equation in an N-dimensional smooth domain Ω with boundary Γ, for $N=2$ or $N=3$, with Dirichlet boundary conditions, namely,

$$\begin{cases} -\Delta u = f \text{ in } \Omega \\ u = d \text{ on } \Gamma, \end{cases} \tag{1}$$

where f and d are given functions defined in Ω and on Γ, having suitable regularity properties.

Here (1) is supposed to be solved by different N-simplex based finite element methods, incorporating degrees of freedom other than function values at the mesh vertices. For instance, if standard quadratic Lagrange finite elements are employed, it is well-known that approximations of an order not greater than 1.5 in the energy norm are generated (cf. [4]), in contrast to the second order ones that apply to the case of a polygonal or polyhedral domain, assuming that the solution is sufficiently smooth. If we are to recover the optimal second order approximation property something different has to be done. Since long the isoparametric version of the finite element method for meshes consisting of curved triangles or tetrahedra (cf. [3] [4]), has been considered as the ideal way to achieve this. It turns out that, besides a more elaborated description of the mesh, the isoparametric technique inevitably leads to the integration of rational functions to compute the system matrix, which raises the delicate question on how to choose the right numerical quadrature formula in the master element. In contrast, in the technique to be introduced in this paper exact numerical integration can always be used for this purpose, since we only have to deal with polynomial integrands. Moreover the element geometry remains the same as in the case of polygonal or polyhedral domains. It is noteworthy that both advantages are conjugated with the fact that no erosion of qualitative approximation properties results from the application of our technique, as compared to the equivalent isoparametric one.

An outline of the paper is as follows. In Section 2 we present our method to solve the model problem with Dirichlet boundary conditions in a smooth curved

two-dimensional domain with conforming Lagrange finite elements based on meshes with straight triangles, in connection with the standard Galerkin formulation. Numerical examples illustrating technique's potential are given. In Section 3 we extend the approach adopted in Section 2 to the three-dimensional case including also numerical experimentation. We conclude in Section 4 with some comments on possible extensions of the methodology studied in this work.

In the remainder of this paper we will be given partitions $\mathcal{T}_h$ of Ω into (closed) ordinary triangles or tetrahedra, according to the value of *N*, satisfying the usual compatibility conditions (see e.g. [4]). Every $\mathcal{T}_h$ is assumed to belong to a uniformly regular family of partitions. We denote by Ω_h the set $\bigcup_{T\in\mathcal{T}_h} T$ and by Γ_h the boundary of Ω_h. Letting h_T be the diameter of $T\in\mathcal{T}_h$, we set $h := \max_{T\in\mathcal{T}_h} h_T$, as usual. We also recall that if Ω is convex Ω_h is a proper subset of Ω.

2. The Two-Dimensional Case

To begin with we describe our methodology in the case where $N=2$. In order to simplify the presentation in this section we assume that $d\equiv 0$, leaving for the next one its extension to the case of an arbitrary *d*.

2.1. Method Description

Here we make the very reasonable assumption on the mesh that no element in $\mathcal{T}_h$ has more than one edge on Γ_h.

We also need some definitions regarding the skin $(\Omega\setminus\Omega_h)\cup(\Omega_h\setminus\Omega)$. First of all, in order to avoid non essential difficulties, we assume that the mesh is constructed in such a way that convex and concave portions of Γ correspond to convex and concave portions of Γ_h. This property is guaranteed if the points separating such portions of Γ are vertices of polygon Ω_h. In doing so, let $\mathcal{S}_h$ be the subset of $\mathcal{T}_h$ consisting of triangles having one edge on Γ_h. Now for every $T\in\mathcal{S}_h$ we denote by Δ_T the set delimited by Γ and the edge e_T of T whose end-points belong to Γ and set $T' := T\cup\Delta_T$ if Δ_T is not a subset of T and $T' := \overline{T\setminus\Delta_T}$ otherwise (see **Figure 1**).

Notice that if e_T lies on a convex portion of Γ_h, T is a proper subset of T', while the opposite occurs if e_T lies on a concave portion of Γ_h. With such a definition we can assert that there is a partition $\mathcal{T}_h'$ of Ω associated with $\mathcal{T}_h$ consisting of non overlapping sets T' for $T\in\mathcal{S}_h$, besides the elements in $\mathcal{T}_h\setminus\mathcal{S}_h$.

For convenience henceforth we refer to the n_k points in a triangle T which are vertices of the k^2 equal triangles T can be subdivided into, where $n_k := (k+2)(k+1)/2$ for $k>1$ as the *lagrangian nodes of order k* (cf. [4]).

Next we introduce two function spaces V_h and W_h associated with $\mathcal{T}_h$.

V_h is the standard Lagrange finite element space consisting of continuous functions v defined in Ω_h that vanish on Γ_h, whose restriction to every $T\in\mathcal{T}_h$ is a polynomial of degree less than or equal to k for $k\geq 2$. For

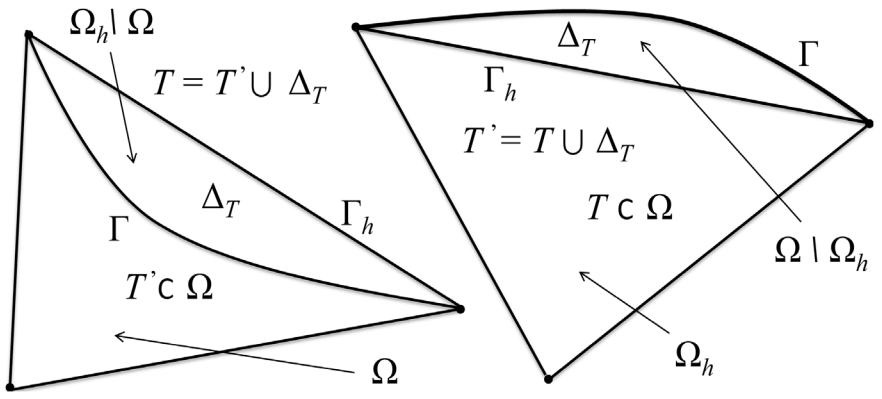


Figure 1. Skin Δ_T related to a mesh triangle T next to a convex (right) or a concave (left) portion of Γ.

convenience we extend by zero every function $v \in V_h$ to $\Omega \setminus \Omega_h$.

W_h in turn is the space of functions defined in $\Omega_h \cup \Omega$ having the properties listed below.

1) The restriction of $w \in W_h$ to every $T \in \mathcal{T}_h$ is a polynomial of degree less than or equal to k;

2) Every $w \in W_h$ is continuous in Ω_h and vanishes at the vertices of Γ_h;

3) A function $w \in W_h$ is also defined in $\Omega \setminus \Omega_h$ in such a way that its polynomial expression in $T \in \mathcal{S}_h$ also applies to points in Δ_T;

4) $\forall T \in \mathcal{S}_h$, $w(P) = 0$ for every P among the $k-1$ intersections with Γ of the line passing through the vertex O_T of T not belonging to Γ and the points M different from vertices of T subdividing the edge opposite to O_T into k segments of equal length (cf. **Figure 2**).

Remark *The construction of the nodes associated with W_h located on Γ advocated in item* 4 *is not mandatory. Notice that it differs from the intuitive construction of such nodes lying on normals to edges of Γ_h commonly used in the isoparametric technique. The main advantage of this proposal is an easy determination of boundary node coordinates by linearity, using a supposedly available analytical expression of Γ. Nonetheless the choice of boundary nodes ensuring our method's optimality is really wide, in contrast to the restrictions inherent to the interpolated boundary condition method (*cf. [1]*).*

The fact that W_h is a non empty finite-dimensional space was established in [5]. Furthermore the following result was also proved in the same reference:

Proposition 1 *(*cf. [5]*).*

Let $\mathcal{P}_k(T)$ be the space of polynomials defined in $T \in \mathcal{S}_h$ of degree less than or equal to k. Provided h is small enough $\forall T \in \mathcal{S}_h$, given a set of m_k real values $b_i, i = 1, \cdots, m_k$ with $m_k = (k+1)k/2$, there exists a unique function $w_T \in \mathcal{P}_k(T)$ that vanishes at both vertices of T located on Γ and at the $k-1$ points P of Γ defined in accordance with item 4. of the above definition of W_h, and takes value b_i respectively at the m_k lagrangian nodes of T not located on Γ_h.

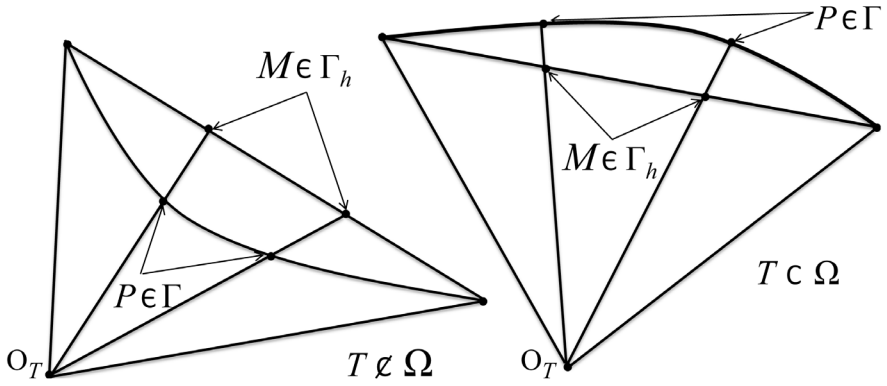


Figure 2. Construction of nodes $P \in \Gamma$ for space W_h related to lagrangian nodes $M \in \Gamma_h$ for $k = 3$.

Now let us set the problem associated with spaces V_h and W_h, whose solution is an approximation of *u*, that is, the solution of (1). Denoting by f' a sufficiently smooth extension of *f* to $\Omega_h \setminus \Omega$ in case this set is not empty, and renaming *f* by f' in Ω, we wish to solve,

$$\begin{cases} \text{Find } u_h \in W_h \text{ such that } a_h(u_h, v) = F_h(v) \ \forall v \in V_h \\ \text{where } a_h(w, v) := \int_{\Omega_h} grad\ w \cdot grad\ v \text{ and } F_h(v) := \int_{\Omega_h} f'v \end{cases} \tag{2}$$

The following result is borrowed from [5]:

Proposition 2 *Provided h is sufficiently small problem (2) has a unique solution.*

2.2. Method Assessment

In order to illustrate the accuracy and the optimal order of the method described in the previous subsection rigorously demonstrated in [5], we implemented it taking $k = 2$. Then we solved Equation (1) for several test-cases already reported in [5] and [6]. Here we present the results for the following one:

Ω is the ellipse delimited by the curve $(x/e)^2 + y^2 = 1$ with $e > 0$ for an exact solution *u* given by $u = \left(e^2 - e^2x^2 - y^2\right)\left(e^2 - x^2 - e^2y^2\right)$. Thus we take $f := -\Delta u$ and $d \equiv 0$ and owing to symmetry we consider only the quarter domain given by $x > 0$ and $y > 0$ by prescribing Neumann boundary conditions on $x = 0$ and $y = 0$. We take $e = 0.5$ and compute with quasi-uniform meshes defined by a single integer parameter *J*, constructed by the procedure described in [5]. Roughly speaking the mesh of the quarter domain is the polar coordinate counterpart of the standard uniform mesh of the unit square $(0,1)\times(0,1)$ whose edges are parallel to the coordinate axes and to the line $x = y$.

Hereunder, and in the remainder of this work we denote by $\|\mathcal{F}\|_h$ the standard mean-square norm in Ω_h of a function or a vector field $\mathcal{F}$.

In **Table 1** we display the absolute errors in the energy norm, namely $\|grad(u - u_h)\|_h$ and in the mean-square norm, that is $\|u - u_h\|_h$ for increasing

Table 1. Absolute errors in different senses for the new approach (with ordinary triangles).

J	$\rightarrow$	8	16	32	64
$\|\|grad(u-u_h)\|\|_h$	$\rightarrow$	0.143615 E−2	0.367543 E−3	0.927840 E−4	0.232998 E−4
$\|\|u-u_h\|\|_h$	$\rightarrow$	0.183918 E−4	0.230310 E−5	0.289312 E−6	0.363247 E−7
$\max\|u-u_h\|$ at the nodes	$\rightarrow$	0.172473 E−2	0.446493 E−3	0.112615 E−3	0.282163 E−4

values of J; more precisely $J=2^m$ for $m=3,4,5,6$. We also show the evolution of the maximum absolute errors at the mesh nodes.

As one infers from Table 1, the approximations obtained with our method perfectly conform to the theoretical estimate given in [5]. Indeed as J increases the errors in the energy norm decrease roughly as $(1/J)^2$, as predicted therein. The error in the mean-square norm in turn tends to decrease as $(1/J)^3$, while the maximum absolute error at the nodes seem to behave like an $O(h^2)$.

Now in order to rule out any particularity inherent to the above test-problem, we also solved it using the classical approach, that is, by replacing W_h with V_h in (2). In Table 2 we display the same kind of results as in Table 1 for this approach.

Table 2 confirms the error decrease in the energy norm like an $O(h^{1.5})$ as predicted in classical texts (cf. [4]). The behavior of the classical approach deteriorates even more, as compared to the new approach, when the errors are measured in the mean-square norm, whose order seem to decrease from three to two. The quality of the maximum nodal absolute errors in turn are not affected at all by the way boundary conditions are handled. Actually in both cases this error is roughly an $O(h^2)$, while in case Ω is a polygon it is known to be an $O(h^3)$ for sufficiently smooth solutions (see e.g. [7]).

3. The Three-Dimensional Case

In this section we consider the solution of (1) by our method in case $N=3$.

In order to avoid non essential difficulties we make the assumption that no element in $\mathcal{T}_h$ has more than one face on Γ_h, which is nothing but reasonable.

3.1. Method Description

First of all we need some definitions regarding the set $(\Omega\setminus\Omega_h)\cup(\Omega_h\setminus\Omega)$.

Let $\mathcal{S}_h$ be the subset of $\mathcal{T}_h$ consisting of tetrahedra having one face on Γ_h and $\mathcal{R}_h$ be the subset of $\mathcal{T}_h\setminus\mathcal{S}_h$ of tetrahedra having exactly one edge on Γ_h. Notice that, owing to our initial assumption, no tetrahedron in $\mathcal{T}_h\setminus[\mathcal{S}_h\cup\mathcal{R}_h]$ has a non empty intersection with Γ_h.

To every edge e of Γ_h we associate a plane skin δ_e containing e, and delimited by Γ and e itself. Except for the fact that each skin contains an edge of Γ_h, its plane can be arbitrarily chosen. In Figure 3 we illustrate one out of three such skins corresponding to the edges of a face F_T or $F_{T'}$ contained in

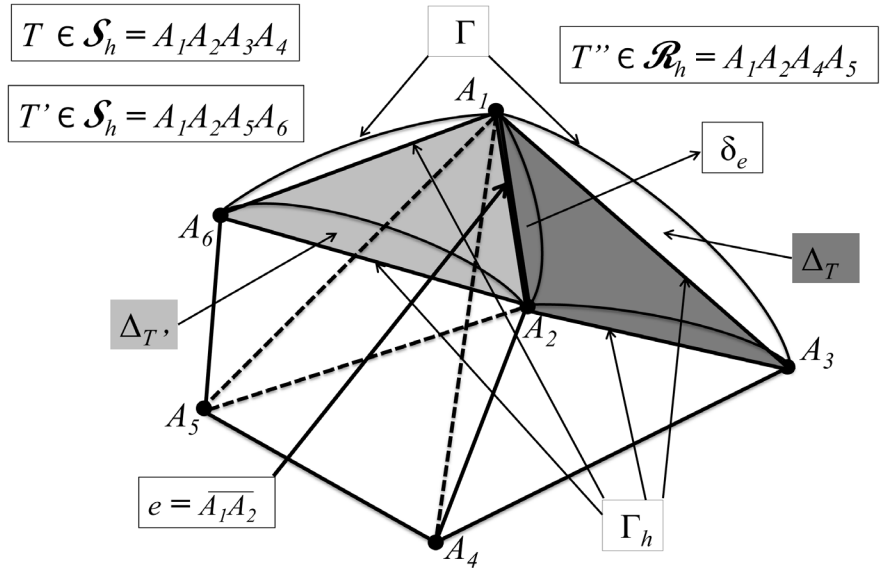


Figure 3. Sets Δ_T, $\Delta_{T'}$, δ_e for tetrahedra $T, T' \in \mathcal{S}_h$ with a common edge e and a tetrahedron $T'' \in \mathcal{R}_h$.

Table 2. Absolute errors in different senses for the classical approach with ordinary triangles.

J	$\rightarrow$	8	16	32	64
$\|grad(u-u_h)\|_h$	$\rightarrow$	0.243341 E−2	0.829697 E−3	0.285754 E−3	0.994597 E−4
$\|u-u_h\|_h$	$\rightarrow$	0.179737 E−3	0.469229 E−4	0.119436 E−4	0.300920 E−5
$\max\|u-u_h\|$ at the nodes	$\rightarrow$	0.172473 E−2	0.446493 E−3	0.112615 E−3	0.282163 E−4

Γ_h, of two tetrahedra T and T' belonging to $\mathcal{S}_h$. More precisely in **Figure 3** we show the skin δ_e, e being the edge common to F_T and $F_{T'}$. Further, for every $T \in \mathcal{S}_h$, we define a set Δ_T delimited by Γ, the face F_T and the three plane skins associated with the edges of F_T, as illustrated in **Figure 3**. In this manner we can assert that, if Ω is convex, Ω_h is a proper subset of Ω and moreover Ω is the union of the disjoint sets Ω_h and $\bigcup_{T\in\mathcal{T}_h}\Delta_T$ (cf. **Figure 3**). Otherwise $\Omega_h \setminus \Omega$ is a non empty set that equals the union of certain parts of the sets Δ_T corresponding to non convex portions of Γ.

Next we introduce two sets of functions V_h and W_h^d, both associated with $\mathcal{T}_h$.

V_h is the standard Lagrange finite element space consisting of continuous functions v defined in Ω_h that vanish on Γ_h, whose restriction to every $T \in \mathcal{T}_h$ is a polynomial of degree less than or equal to k for $k \geq 2$. For convenience we extend by zero every function $v \in V_h$ to $\Omega \setminus \Omega_h$. We recall that a function in V_h is uniquely defined by its values at the points which are vertices of the partition of each mesh tetrahedron into k^3 equal tetrahedra (see e.g. [4]). Akin to the two-dimensional case these points will be referred to as the *lagrangian nodes* of order k of the mesh.

W_h^d in turn is a linear manifold consisting of functions defined in $\Omega_h \cup \Omega$ satisfying the following conditions:

1) The restriction of $w \in W_h^d$ to every $T \in \mathcal{T}_h$ is a polynomial of degree less than or equal to k;

2) Every $w \in W_h^d$ is single-valued at all the inner lagrangian nodes of the mesh, that is all its lagrangian nodes of order k except those located on Γ_h;

3) A function $w \in W_h^d$ is also defined in $\Omega \setminus \Omega_h$ in such a way that its polynomial expression in $T \in \mathcal{S}_h$ also applies to points in Δ_T;

4) A function $w \in W_h^d$ takes the value $d(S)$ at any vertex S of Γ_h;

5) $\forall T \in \mathcal{S}_h$ and for $k > 2$, $w(P) = d(P)$ for every point P among the $(k-1)(k-2)/2$ intersections with Γ of the line passing through the vertex O_T of T not belonging to Γ and the $(k-1)(k-2)/2$ points M not belonging to any edge of F_T among the $(k+2)(k+1)/2$ points of F_T that subdivide this face (opposite to O_T) into k^2 equal triangles (see illustration in **Figure 4** for $k = 3$);

6) $\forall T \in \mathcal{S}_h \cup \mathcal{R}_h$, $w(Q) = d(Q)$ for every Q among the $k-1$ intersections with Γ of the line orthogonal to e in the skin δ_e, passing through the points $M \in e$ different from vertices of T, subdividing e into k equal segments, where e represents a generic edge of T contained in Γ_h (see illustration in **Figure 5** for $k = 3$).

Remark *The construction of the nodes associated with* W_h^d *located on* Γ *advocated in items* 5. *and* 6. *is not mandatory. Notice that it differs from the intuitive construction of such nodes lying on normals to faces of* Γ_h *commonly used in the isoparametric technique. The main advantage of this proposal is the determination by linearity of the coordinates of the boundary nodes* P *in the case of item* 5. *Nonetheless, akin to the two-dimensional case, the choice of boundary nodes ensuring our method's optimality is absolutely very wide.*

The fact that W_h^d is a non empty set is a trivial consequence of the two following results proved in [8], where $\mathcal{P}_k(T)$ represents the space of polynomials defined in $T \in \mathcal{S}_h \cup \mathcal{R}_h$ of degree not greater than k.

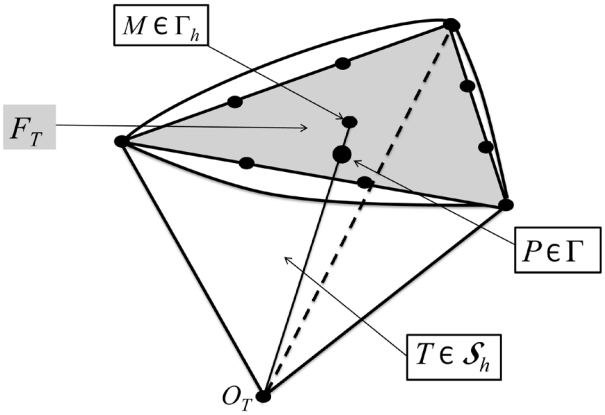


Figure 4. Construction of node $P \in \Gamma$ of W_h^d related to the Lagrange node M in the interior of $F_T \subset \Gamma_h$.

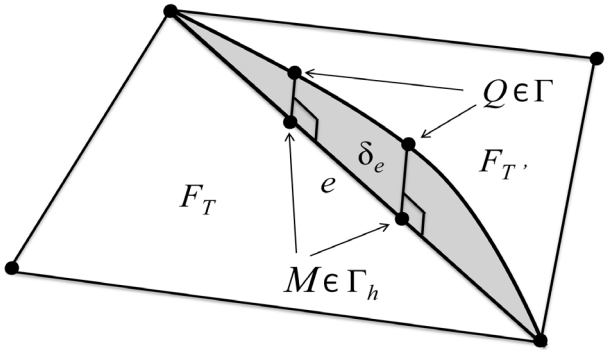


Figure 5. Construction of nodes $Q \in \Gamma \cap \overline{\delta_e}$ of W_h^d related to the Lagrange nodes $M \in e \subset \Gamma_h$.

Proposition 3 *(*cf. [8]*).*

Provided h is small enough $\forall T \in \mathcal{S}_h \bigcup \mathcal{R}_h$ given a set of m_k real values b_i, $i = 1, \cdots, m_k$ with $m_k = k(k+1)(k+2)/6$ for $T \in \mathcal{S}_h$ and $m_k = (k+1)(k+2)(k+3)/6 - (k+1)$ for $T \in \mathcal{R}_h$, there exists a unique function $w_T \in \mathcal{P}_k(T)$ that takes the value of d at the vertices S of T located on Γ, at the points P of Γ defined in accordance with item 5. for $T \in \mathcal{S}_h$ only, and at the points Q of Γ defined in accordance with item 6. of the above definition of W_h^d, and takes the value b_i respectively at the m_k lagrangian nodes of T not located on Γ_h.

A well-posedness result analogous to Proposition 2.2 holds for problem (3), according to [8], namely,

Proposition 4 *(*cf. [8]*)*

As long as h is sufficiently small problem (3) has a unique solution.

Remark *It is important to stress that, in contrast to its two-dimensional counterpart, the set W_h^d does not necessarily consist of continuous functions. This is because of the interfaces between elements in $\mathcal{R}_h$ and $\mathcal{S}_h$. Indeed a function $w \in W_h^d$ is not forcibly single-valued at all the lagrangian nodes located on one such an interface, owing to the enforcement of the boundary condition at the points $Q \in \Gamma$ instead of the corresponding lagrangian node $M \in \Gamma_h$, in accordance with item* 6. *in the definition of W_h^d. On the other hand w is necessarily continuous over all other faces common to two mesh tetrahedra.*

Next we set the problem associated with the space V_h and the manifold W_h^d, whose solution is an approximation of u, that is, the solution of (1). Extending f by a smooth f' in $\Omega_h \setminus \Omega$ if necessary, and renaming f by f' in any case, we wish to solve,

$$\begin{cases} \text{Find } u_h \in W_h^d \text{ such that } a_h(u_h, v) = F_h(v) \; \forall v \in V_h \\ \text{where } a_h(w, v) := \int_{\Omega_h} grad\, w \cdot grad\, v \text{ and } F_h(v) := \int_{\Omega_h} f' v \end{cases} \tag{3}$$

3.2. Method Assessment

In this sub-section we assess the behavior of the new method, by solving the

Poisson equation in a non convex domain. Now we consider a non polynomial exact solution. More precisely (1) is solved in the torus Ω with minor radius r_m and major radius r_M. This means that the torus' inner radius r_i equals $r_M - r_m$ and its outer radius r_e equals $r_M + r_m$. Hence Γ is given by the equation $\left(r_M - \sqrt{x^2+y^2}\right)^2 + z^2 = r_m^2$. We consider a test-problem with symmetry about the *z*-axis, and with respect to the plane $z=0$. For this reason we may work with a computational domain given by $\left\{(x,y,z)\in\Omega \middle| z\geq 0; 0\leq\theta\leq\pi/4 \text{ with } \theta = a\tan(y/x)\right\}$. A family of meshes of this domain depending on a single even integer parameter I containing $6I^3$ tetrahedra is generated by the following procedure. First we generate a partition of the cube $(0,1)\times(0,1)\times(0,1)$ into $I^3/2$ equal rectangular boxes by subdividing the edges parallel to the *x*-axis, the *y*-axis and the *z*-axis into 2*I*, *I*/2 and *I*/2 equal segments, respectively. Then each box is subdivided into six tetrahedra having an edge parallel to the line $4x = y = z$. This mesh with $3I^3$ tetrahedra is transformed into the mesh of the quarter cylinder $\left\{(x,y,z)\middle| 0\leq x\leq 1,\ y\geq 0,\ z\geq 0,\ y^2+z^2\leq 1\right\}$, following the transformation of the mesh consisting of $I^2/2$ equal right triangles formed by the faces of the mesh elements contained in the unit cube's section given by $x = j/(2I)$, for $j = 0,1,\cdots,2I$. The latter transformation is based on the mapping of the cartesian coordinates (y,z) into the polar coordinates (r,φ) with $r=\sqrt{y^2+z^2}$, using a procedure described in [8] (cf. **Figure 6**). Then the resulting mesh of the quarter cylinder is transformed into the mesh with $6I^3$ the trahedra of the half cylinder $\left\{(x,y,z)\middle| 0\leq x\leq 1,\ -1\leq y\leq 1,\ z\geq 0,\ y^2+z^2\leq 1\right\}$ by symmetry with respect to the plane $y=0$. Finally this mesh is transformed into the computational mesh (of an eighth of half-torus) by first mapping the cartesian coordinates (x,y) into polar coordinates (ρ,θ), with $\rho = r_M + yr_m$

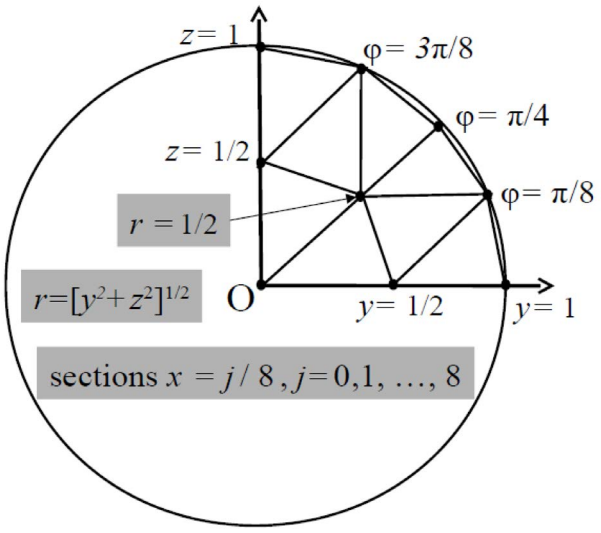


Figure 6. Trace of the intermediate mesh of 1/4 cylinder on sections $x = j/(2I)$, $0\leq j\leq 2I$, for $I=4$.

and $\theta = x\pi/4$, and then the latter coordinates into new cartesian coordinates (x, y) using the relations $x = \rho\cos\theta$ and $y = \rho\sin\theta$. Notice that the faces of the final tetrahedral mesh on the sections of the torus given by $\theta = j\pi/(8I)$, for $j = 0,1,\cdots,2I$, form a triangular mesh of a disk with radius equal to r_m, having the pattern illustrated in Figure 6 for a quarter disk, taking $I = 4$, $\theta = 0$ and $r_m = 1$.

Recalling that here $\rho = \sqrt{x^2 + y^2}$, we take $r_M = 5/6$, $r_m = 1/6$ and $f' = 6 - 5/(3\rho)$. For $d \equiv 0$ the exact solution is given by $u = 1/36 - z^2 - (5/6 - \rho)^2$. In order to enable the calculation of mean-square norms of the error in Ω, we extend u to u' in a neighborhood of Γ lying outside Ω, taking the same expression as above. In Table 3 we display the absolute errors in the energy norm, that is $\|grad(u' - u_h)\|_h$ and in the mean-square norm $\|u' - u_h\|_h$, for increasing values of I, namely $I = 2^m$ for $m = 1,2,3,4$. Now we take as a reference $h = \pi/(8I)$.

As one can observe from Table 3, here again the quality of the approximations obtained with the new method is in very good agreement with the theoretical result in [8], for as I increases the errors in the energy norm decrease roughly as $1/I^2$, as predicted. On the other hand here again the mean-square norm of the error function $u' - u_h$ tends to decrease as $1/I^3$. Likewise the two-dimensional case, Table 4 in turn shows a qualitative erosion of the solution errors if W_h^d is replaced by V_h in (3).

4. Final Comments

1) The method addressed in this work to solve the Poisson equation with Dirichlet boundary conditions in curved domains with classical Lagrange finite elements provides a simple and reliable manner to overcome technical difficulties brought about by more complicated problems and interpolations. For example, Hermite finite element methods to solve fourth order problems in curved domains with normal derivative degrees of freedom can also be dealt with very easily by means of our new method. The author intends to show this in

Table 3. Absolute errors for the new approach (with ordinary tetrahedra) in two different norms.

h	$\rightarrow$	$\pi/32$	$\pi/64$	$\pi/128$	$\pi/256$
$\|grad(u' - u_h)\|_h$	$\rightarrow$	0.786085 E−3	0.205622 E−3	0.522963 E−4	0.131844 E−4
$\|u' - u_h\|_h$	$\rightarrow$	0.133794 E−4	0.171222 E−5	0.214555 E−6	0.269187 E−7

Table 4. Absolute errors for the classical approach with ordinary tetrahedra in two different norms.

h	$\rightarrow$	$\pi/32$	$\pi/64$	$\pi/128$	$\pi/256$
$\|grad(u' - u_h)\|_h$	$\rightarrow$	0.829181 E−2	0.327176 E−2	0.119077 E−2	0.425739 E−3
$\|u' - u_h\|_h$	$\rightarrow$	0.579150 E−3	0.143425 E−3	0.343823 E−4	0.834136 E−5

a paper to appear shortly.

2) The technique studied in this paper is also particularly handy, to treat problems posed in curved domains in terms of vector fields, such as the linear elasticity system and Maxwell's equations of electromagnetism. The same remark applies to multi-field systems such as the Navier-Stokes equations, and more generally mixed formulations of several types with curved boundaries, to be approximated by the finite element method.

3) As for the Poisson equation with homogeneous Neumann boundary conditions $\partial u/\partial n = 0$ on Γ (provided f satisfies the underlying scalar condition) our method coincides with the standard Lagrange finite element method. Notice that if inhomogeneous Neumann boundary conditions are prescribed, optimality can only be recovered if the linear form F_h is modified, in such a way that boundary integrals for elements $T \in \mathcal{S}_h$ are shifted to the curved boundary portion sufficiently close to Γ of an extension or reduction of T. But this is an issue that has nothing to do with our method, which is basically aimed at resolving those related to the prescription of degrees of freedom in the case of Dirichlet boundary conditions.

4) Finally we note that our method leads to linear systems of equations with a non symmetric matrix, even when the original problem is symmetric. Moreover in order to compute the element matrix and right side vector for an element T in $\mathcal{S}_h$ or in $\mathcal{R}_h$, the inverse of an $n_k \times n_k$ matrix has to be computed, where n_k is the dimension of $P_k(T)$. However this extra effort is not really a problem nowadays, in view of the significant progress already accomplished in Computational Linear Algebra.

Acknowledgements

The author is thankful for the support provided by CNPq-Brazil, through grant 307996/2008-5, to carry out this work, which was first presented in August 2017 at both the MFET Conference in Bad Honnef, Germany, and the French Congress of Mechanics in Lille.

Sincere thanks are due to the managing editor *Hellen XU* for her precious assistance.

References

[1] Brenner, S.C. and Scott, L.R. (2008) The Mathematical Theory of Finite Element Methods. Texts in Applied Mathematics, Vol. 15, Springer, Berlin. https://doi.org/10.1007/978-0-387-75934-0

[2] Nitsche, J.A. (1972) On Dirichlet Problems Using Subspaces with Nearly Zero Boundary Conditions. In: Aziz, A.K., Ed., *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, Academic Press, Cambridge, Massachusetts.

[3] Scott, L.R. (1973) Finite Element Techniques for Curved Boundaries. PhD Thesis, MIT, Cambridge, Massachusetts,.

[4] Ciarlet, P.G. (1978) The Finite Element Method for Elliptic Problems. North Hol-

land, Amsterdam.

[5] Ruas, V. (2017) Optimal Simplex Finite-Element Approximations of Arbitrary Order in Curved Domains Circumventing the Isoparametric Technique. arXiv Numerical Analysis. https://arxiv.org/abs/1701.00663

[6] Ruas, V. (2017) A Simple Alternative for Accurate Finite-Element Modeling in Curved Domains. Congrès Français de Mécanique, Lille, France. https://cfm2017.sciencesconf.org/133073/document

[7] Nitsche, J.A. (1979) L^{∞}-Convergence of Finite Element Galerkin Approximations for Parabolic Problems. *ESAIM*: *Mathematical Modelling and Numerical Analysis*, **13**, 31-54.

[8] Ruas, V. (2017) Methods of Arbitrary Optimal Order with Tetrahedral Finite-Element Meshes Forming Polyhedral Approximations of Curved Domains. arXiv Numerical Analysis. https://arxiv.org/abs/1706.08004

Journal of Applied Mathematics and Physics, 2017, 5, 2140-2161
http://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Air Pollutant Emissions in the Fukui-Ishibashi and Nagel-Schreckenberg Traffic Cellular Automata

Alejandro Salcido, Susana Carreón-Sierra

Programa de Sustentabilidad Ambiental, Instituto Nacional de Electricidad y Energías Limpias, Cuernavaca, Mexico
Email: salcido@ineel.mx, susana.carreon@iie.org.mx

How to cite this paper: Salcido, A. and Carreón-Sierra, S. (2017) Air Pollutant Emissions in the Fukui-Ishibashi and Nagel-Schreckenberg Traffic Cellular Automata. *Journal of Applied Mathematics and Physics*, **5**, 2140-2161.
https://doi.org/10.4236/jamp.2017.511175

Received: October 9, 2017
Accepted: November 7, 2017
Published: November 10, 2017

Abstract

Vehicular traffic is a hard problem in big cities. Internal combustion vehicles are the main fossil fuel consumers and frame the main source of urban air pollutants, such as particulate matter, nitrogen oxides, and volatile organic compounds. Vehicular traffic is also a promoter of climate change due to its greenhouse gas emissions, such as CO and CO_2. Awareness of the spatiotemporal distribution of urban traffic, including the velocity distribution, allows knowing the spatiotemporal distribution of the air pollutant vehicular emissions required to understand urban air pollution. Although no well-established traffic theory exists, some models and approaches, like cellular automata, have been proposed to study the main aspects of this phenomenon. In this paper, a simple approach for estimating the space-time distribution of the air pollutant emission rates in traffic cellular automata is proposed. It is discussed with the Fukui-Ishibashi (FI) and Nagel-Schreckenberg (NS) models for traffic flow of identical vehicles in a single lane. We obtained the steady-state emission rates of the FI and NS models, being larger those produced by the first one, with relative differences of up to 45% in hydrocarbons, 56% in carbon monoxide, and 77% in nitrogen oxides.

Keywords

Cellular Automata, Mobile Source Emissions, Traffic Emission Rates, Traffic Models, Fukui-Ishibashi, Nagel-Schreckenberg

1. Introduction

Big cities are suffering severe problems because of the growing number of vehicles moving over their streets. In Mexico City (CDMX), for example, the reg-

istered vehicular fleet was estimated close to 5 million in 2015. Figure 1 describes information published by the Mexico's National Institute of Statistics and Geography (Instituto Nacional de Estadística y Geografía, INEGI) in relation to the increasing number of registered vehicles between 1980 and 2014 [1].

The vehicular fleet of CDMX is composed, in a great majority, by internal combustion vehicles that consume fossil fuels (gasoline, diesel, and gas); therefore, vehicular flow through the city streets is one of the main responsible for urban air pollution. In fact, the 2014 emissions inventory of CDMX [2] reported that the contributions of the mobile sources to the air pollutant emissions in the city were as described in Table 1.

In Table 1, we observe that the vehicular traffic contributed with the 44% of the CDMX air pollution, in average. These emissions, of course, are not uniformly distributed in the region because there is no a uniform distribution of traffic in the city and winds either are not uniform either perennial. This means that the most polluted areas of a city are not necessarily those ones where more pollutants are released to the atmosphere.

Addressing the urban air pollution problems depends on the knowledge of the distribution modes of the urban vehicles in space, time, and over the possible speeds because these modes determine how the emissions of gases and particulate matter by the vehicles will result in a spatiotemporal distribution of emission rates in the city.

Once in the atmosphere, the air pollutants will be transported by the wind and dispersed by the atmospheric turbulence. Figure 2 shows the basic scheme of coupling of the models required for simulating the impacts of traffic emissions on air quality. It comprises a traffic flow model, which estimates the spatiotemporal distribution of the vehicles and their velocity distribution; an emission model, which allows determining the spatiotemporal distribution of the emission rates of the pollutants produced by the traffic flow; an atmospheric

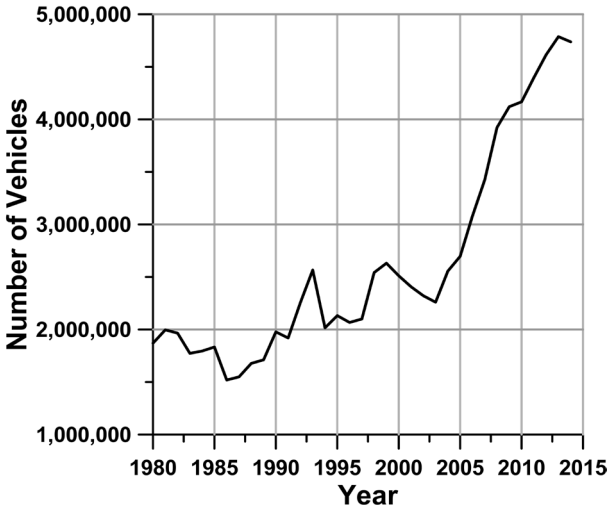


Figure 1. Growing of the MCMA's vehicular fleet from 1980 to 2014.

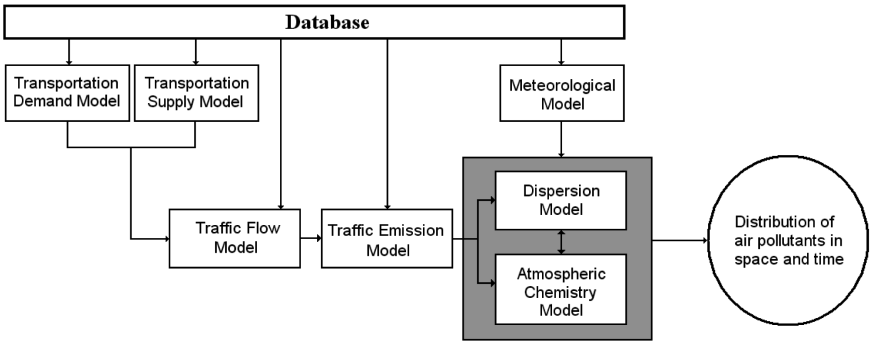


Figure 2. Basic scheme of coupling of the models required for simulating the impacts of the emissions of mobile sources on air quality.

Table 1. Mobile source air pollutant emissions [%] according to the 2014 emissions inventory of CDMX.

PM_{10}	$PM_{2.5}$	SO_2	CO	NO_X	TOC	VOC	CO_2	N_2O	HFC	$CO_{2\text{-eq}}$	Black Carbon
20.7	28.7	16.5	96.1	78.5	11.3	20.0	61.5	50.0	98.0	49.0	83.7

transport and dispersion modelling system of the emissions; and an atmospheric chemistry model, which addresses the possible transformations of the pollutants in the atmosphere.

There is no a complete theory for traffic flow phenomena. However, several models and approaches for analyzing traffic phenomena, such as traffic jamming and some other common modes of traffic flow, have been developed from the macroscopic, mesoscopic and microscopic standpoints.

The scientific treatment of the traffic flow phenomena began with Robert Herman in 1956, and some years later, in the early 1960s, Herman and Prigogine started to study vehicular traffic as a collective flow phenomenon, developing a kinetic theory for multi-lane traffic flow using a Boltzmann like model for the vehicle interactions [3].

In the second half of the 1980s, an alternative line of research emerged for traffic flow simulation based on cellular automata [4], but its proper development started in the early 1990s with the models proposed by K. Nagel and M. Schreckenberg [5] and by M. Fukui and Y. Ishibashi [6], hereafter referred as NS and FI models, respectively. They defined cellular automata models for the microscopic simulation of vehicular traffic. The initial NS and FI models were formulated for identical vehicles moving on a single lane highway. In these models, each vehicle can be at rest or be hopping from site to site in a 1D lattice with a positive integer speed which does not exceed a given maximum. The dynamic rules of these cellular automata are probabilistic and control the propagation, acceleration, and braking of the model vehicles, although conserving its number and preventing them from collisions and overtaking. Several variants and extensions of the NS and FI basic models have been developed for simulating traffic flow in double-sense and multi-lane highways [7] [8] [9], and also for

2D complex traffic networks similar to that of a city [10] [11] [12] [13] [14].

In this paper, we propose a simple approach for estimating the spatiotemporal distribution of the emission rates for the traffic flow phenomena described by the NS and FI models. This approach assumes it is possible to know the velocity distribution of the system, *i.e.* how many vehicles are moving with each one of the possible velocities (from zero to a maximum speed). The velocity distribution of the traffic cellular automata can be obtained always by computer simulations, but also theoretically, at least for a class of models. In this paper, for obtaining the velocity distributions, we used computer simulations and the statistical mechanics approach proposed in [15] [16] [17] for the maximum entropy states of 1D traffic cellular automata. The results of this work show that due to the transition rules of the FI model, which favor the highest speeds, the steady state emission rates of this model are higher than those ones of the NS model, with relative differences as large as 45% for hydrocarbons, 56% for carbon monoxide, and 77% for nitrogen oxides.

The rest of the paper is organized as follows. In Section 2, we described our methodological approach. First, we presented the basic NS and FI traffic cellular automata, discussing, in particular, how the velocity distributions of these models can be obtained from computer simulations and from the theoretical approach proposed by Salcido and collaborators [15] [16] [17]. We described also the estimation model for the pollutants emission rates from mobile sources [18], and the extension we proposed for traffic cellular automata. In Section 3, we presented and discussed the results of the application of our methodological approach to the problem of estimating the distribution of the pollutant emissions from the NS and FI traffic cellular automata.

2. Methodology

In this section, we provide first a brief introduction to cellular automata; then we describe and discuss the NS and FI traffic cellular automata models and the approaches to frame their velocity distributions, and, finally, we present the approach to estimate the model cars emissions.

2.1. Cellular Automata

Cellular automata (henceforth: CA) are a class of spatially and temporally discrete, complex dynamical systems characterized by local interaction and an inherently parallel form of evolution. Following a suggestion of Stanislaw Ulam, cellular automata were first introduced by John von Neumann in the early 1950s to act as simple models of biological self-reproduction [19]. Cellular automata can be considered as prototypical models for complex systems and processes consisting of a large number of identical, simple, locally interacting components [20]. The study of CA has generated great interest over the years because of their ability to generate a rich spectrum of very complex patterns of behavior out of sets of relatively simple underlying rules [20] [21] [22] [23]. Moreover, CA ap-

pear to capture many essential features of complex self-organizing cooperative behavior observed in real systems.

There exists a wide variety of particular CA models; however, most of them usually possess the following common generic characteristics. The system substrate consists of a one-, two- or three-dimensional lattice of cells; all cells are equivalent; each cell takes on one of a finite number of possible discrete states; each cell interacts only with cells that are in its local neighborhood; and at each discrete time step, each cell updates its current state according to a transition rule taking into account the states of cells in its neighborhood.

If $\psi(x,t)$ denotes the state at cell x at time t, $V(x)$ is the neighborhood of this cell (in a well-defined sense of proximity), and $\{\psi(\tilde{x},t)|\tilde{x}\in V(x)\}$ is the set of the states of the cells in the neighborhood, then the state at cell x at time $t+1$ will be given by

$$\psi(x,t+1)=F\left(\{\psi(\tilde{x},t)|\tilde{x}\in V(x)\}\right) \tag{1}$$

Here F represents the transition rules of the system dynamics. Note that both the neighborhood and the transition rule have the same definitions for all the lattice cells. Usually, neighborhoods contain the first nearest neighbors (von Neumann), or the first and second nearest neighbors (Moore). Some widely known cellular automata are the Wolfram's 1D elementary cellular automata [24] and the Conway's Game of Life [25].

2.2. One-Dimensional Traffic Cellular Automata

The basic one-dimensional traffic cellular automata (B1DTCA) are concerned with the traffic flow of identical vehicles (cars) on a single lane highway with no anticipation. This class of CA models shares the following properties:

- The system can be considered as a lattice gas of N indistinguishable unit mass particles, which evolves in a 1D lattice with L cells (or sites).
- The particles of the system obey an exclusion principle, which establishes that no more than one particle can be in one lattice cell.
- Each particle can be at rest or be moving with a positive integer velocity that cannot exceed a given maximum $v_{\max}>0$: $v_k=k$ with $k=0,1,2,3,\cdots,v_{\max}$. This means that the particles move always in the same direction (say, from left to right), and never can go in the reverse direction. The velocity $v_{\max}$ is interpreted as a speed limit that drivers have to respect inexcusably.
- The dynamics of the system is defined by a set of local transition rules. The same rules are applied simultaneously to all the lattice cells. These rules allow no particle collisions neither overtaking. Traffic accidents never occur and each car follows always same another car.
- The local transition rules preserve the number of particles, but not necessarily momentum neither the energy.
- The system evolution occurs in discrete time steps. Time increases in one unit only once all the cells of the system have been updated according to the

transition dynamical rules.

In Figure 3, we illustrate a possible spatial distribution of the system cars in the lattice. Here, the different car velocities are evidenced with different background colors. It must be noted that the no anticipation condition implies that each car with velocity v occupies $v+1$ lattice sites.

The distance among adjacent cells is usually defined as the unit, but for the purpose of real traffic simulations, it is assumed to be the average front-bumper-to-front-bumper distance of adjacent vehicles under conditions of strongly jammed traffic and set equal to 7.5 m. In this case, the time step is set equal to one second, and the velocity increases in steps of 27 km/h.

We can describe the state of the system indicating the number of lattice cells (L), the total number of particles (N), and the numbers of particles N_k which move with velocity $v_k = k$ ($k = 0,1,2,\cdots,v_{\max}$). In general, however, we will use the intensive properties (densities) defined as

$$n = \frac{N}{L}, \quad n_k = \frac{N_k}{L} \tag{2}$$

The density of particles (*i.e.* the number of particles per cell) is equal to the sum of the partial densities

$$n = \sum_k n_k \tag{3}$$

and the densities of momentum (traffic flow) and kinetic energy are given by

$$q = nv = \sum_k v_k n_k, \quad \varepsilon = \sum_k \varepsilon_k n_k \tag{4}$$

respectively, where v is the average speed of the traffic flow and $\varepsilon_k = v_k^2/2$ is the kinetc energy of a particle with speed v_k.

The traffic models developed by Nagel and Schreckenberg [5] and by Fukui and Ishibashi [6] [26] belong to the class of B1DTCA.

2.2.1. The Nagel and Schreckenberg Model

The dynamics of the Nagel-Schreckenberg model [5] is defined by the following set of local transition rules. If one vehicle is located at the cell c ($c = 1,2,3,\cdots,L$) at time t, and it is moving with velocity $v(c,t)$, then

- Rule 1. Acceleration: $v(c,t)$ is replaced by $u(c,t) = \min\{v(c,t)+1, v_{\max}\}$.
- Rule 2. Braking: $u(c,t)$ is replaced by $w(c,t) = \min\{d(c,t), u(c,t)\}$, where $d(c,t)$ is the number of empty cells ahead the cell c, at time t.
- Rule 3. Randomization: the velocity of the vehicle located at cell c is updated to $v(c,t+1) = \max\{w(c,t)-1, 0\}$ with probability p, or to $v(c,t+1) = w(c,t)$ with probability $1-p$.
- Rule 4. Flow: the vehicle jumps from cell c to cell $c + v(c,t+1)$.

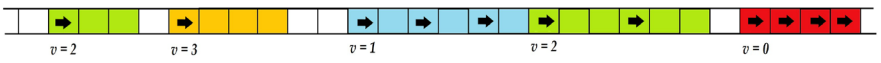


Figure 3. One instantaneous spatial distribution of the system cars. Colors make evident the possible different car velocities.

These rules are applied simultaneously to all the non-empty lattice cells; time increases by one only when all the lattice cells have been updated.

These rules have widely accepted simple interpretations. Rule 1 mimics the fact that drivers like to go as fast as allowed. Rule 2 takes into account that one driver has to reduce its car's velocity to avoid the collision against the vehicle ahead. Rule 3 aims to take into account some effects which produce velocity fluctuations, even in the free flow case; for example, the road conditions (slopes, potholes, and speed humps, among others), the impact of climatic conditions on traffic flow, and psychological effects. Consequently, this rule can produce braking overreaction, which may give rise to spontaneous jamming [27]. Finally, Rule 4 displaces the vehicles in the lattice. It is worthy of comment that Brilon and Wu [28] have questioned Rule 3; they argue that it has no theoretical basis. Nevertheless, Rule 3 is essential in simulating realistic traffic flow since otherwise the model dynamics would be completely deterministic [5].

Collectively, these four rules enable the NS model to reproduce the basic phenomena of real traffic, such as the occurrence of the phantom traffic jams. These rules define a minimal model in the sense that any further simplification of them no longer produces nontrivial and realistic behavior. For proper modelling of the fine structure of traffic, however, it is necessary the introduction of additional rules and/or the modification of the transition rules above-presented.

2.2.2. The Fukui and Ishibashi Model

In the Fukui-Ishibashi model [6] [26] [29] the cars can move by at most $v_{\max}$ lattice sites in one time step if vehicles in front do not block them. Specifically, if at time t the number of empty sites h in front of a car is larger than $v_{\max}$ then, in the next time-step, it can move forward $v_{\max}$ sites with probability $1-p$, or $v_{\max}-1$ sites with probability p. Here, the randomization probability p represents the degree of stochastic delay. Within the framework of this model, drivers do not like to use brakes if they are far away from the vehicle ahead. For a large density of cars, the stochastic delay in the FI model represents the assurance of the avoidance of crashes. When the stochastic delay is null ($p=0$), this cellular automaton is referred to as the deterministic FI model with the maximum velocity $v_{\max}$. On the other hand, the case $p=1$ defines the deterministic FI model with the maximum velocity $v_{\max}-1$. If $h<v_{\max}$ at time t, then the car can only move by h sites in the next time-step. Important differences of the FI model with respect the NS model are that the acceleration of cars may occur abruptly and that stochastic delay only affects the high-speed cars.

2.2.3. The Maximum Entropy States

Many of the cellular automata models proposed for traffic flow are based on the NS and FI models that we described in the previous sections. These models, in general, have been developed as computational systems for simulating traffic phenomena, and there are no analytical theoretical formulations to describe them. In fact, up today, very few efforts have been made to establish a unified

theoretical formalism for the traffic cellular automata. In this section, we provide a brief description of a statistical mechanics' analysis carried out by Salcido *et al.* [15] [16] [17] for obtaining the equilibrium states of the B1DTCA (such as the NS and FI models) from a maximum entropy principle.

It is important to stress that the dynamical rules of the models like NS and FI are not microscopically reversible (they do not satisfy the principle of detailed balance [30] [31]), and, consequently, the system is always far from equilibrium. In fact, the Nagel-Schreckenberg and Fukui-Ishibashi models have been considered as variants of the well-known asymmetric exclusion process (ASEP), a paradigm of non-equilibrium systems [32]. In spite of this fact, an entropy function can be defined for the class of B1DTCA, and the velocity distribution that corresponds to the maximum-entropy states may be determined [15] [16] [17].

Such as detailed in [17], we assume that our system belongs to the class of B1DTCA (defined in Section 2.2). In addition, we assume that it has periodic boundary conditions, so that when one particle leaves the lattice by one end, it appears immediately in the other end. Moreover, it is observed that each particle of the system, which is moving with the velocity v_i can be considered as a block that occupy v_i+1 cells in the 1D lattice. This observation allows showing that the entropy per cell of the system of blocks is

$$s=(\lambda+n)\ln(\lambda+n)-\lambda\ln(\lambda)-\sum_i n_i\ln(n_i) \tag{5}$$

where λ is the vacancy (the number of cells per cell that remain empty after accommodating all the blocks of the system in the lattice) and n_i is the partial density of the particles with velocity v_i (the number of blocks each one occupying v_i+1 cells, per cell of the system). We observe that

$$\lambda=1-\sum_i (v_i+1)n_i\geq 0 \tag{6}$$

Equations (3) and (4) give the densities of particles, momentum, and kinetic energy of the system.

Under this context, the maximum entropy states of the system are given by

$$n_i=\lambda\left(\frac{\lambda}{\lambda+n}\right)^{v_i}\mathrm{e}^{-\alpha-\beta\varepsilon_i}=n_0\left(\frac{\lambda}{\lambda+n}\right)^{v_i}\mathrm{e}^{-\beta\varepsilon_i} \tag{8}$$

Here α and β are Lagrange multipliers, and it has been defined $n_0\equiv\lambda\mathrm{e}^{-\alpha}$ [17]. Equation (8), for each velocity $v_i=v_1,v_2,\cdots,v_{\max}$, gives the number of particles (cars) per cell which are moving with that velocity; *i.e.*, this equation gives the velocity distribution of the system. This maximum entropy approach describes, as a particular case, the low-density behavior of the FI model with a very good agreement [17], and also reproduce, approximately, at least, the steady states of the NS model [15] [16] [17].

2.3. Pollutant Emission Rates of Traffic Cellular Automata

Let us assume that $e(\alpha,v_i)$ is the emission rate of the pollutant α of one particle (a model car of a traffic cellular automaton) which is moving with veloc-

ity v_i. Let us assume also that at the cell x, the average number of particles per cell which are moving with the velocity v_i at time t, is $n_i(x,t)$. Then, the partial emission rate of the pollutant α due to the vehicles with velocity v_i is given by

$$\mu(\alpha, v_i, x, t) = e(\alpha, v_i) n_i(x,t) \tag{9}$$

And the total emission rate of the pollutant α at time t, due to all the particles of the system is

$$Q(\alpha,t) = \sum_x \sum_i \mu(\alpha, v_i, x, t) = \sum_x \sum_i e(\alpha, v_i) n_i(x,t) \tag{10}$$

where the sums extend over all the lattice cells and over all the possible velocities.

For traffic cellular automata, the velocity distributions that we need to estimate their pollutant emissions can be obtained in general from computer simulations, but also from a theoretical standpoint such as the maximum entropy approach that we described in the previous section.

The emission rate $e(\alpha, v_i)$, on the other hand, must be determined experimentally, using emission factors, or with a proper emission model. This function represents a subset of a mobile source emission inventory disaggregated by pollutant, type of vehicle, and speed of movement of the vehicle. This emission rate, of course, will depend also on the characteristics and conditions of the vehicle, on driving habits, and on the weather conditions. In general, the reference data for estimating the emissions of road vehicles is obtained by measuring the emissions of a representative vehicle in a controlled ambient and simulating specific driving condition. The results of the observations are usually aggregated either by estimating a functional relationship (e.g., the German recommendations for economic assessment of road infrastructure investments (EWS) [33]) or by clustering the data into typical driving situations (e.g., the Workbook on Emission Factors for Germany and Switzerland [18]).

The EWS has the advantage that the full functional relationship on the vehicle's velocity v is given for a specific pollutant α and vehicle type [34]:

$$e_f(\alpha, v) = \begin{cases} c_0 + c_1 v^2 + \dfrac{c_2}{v} & \text{for } v > 20 \text{ km/h} \\ \min\left\{ c_{SG}, \left(c_0 + c_1 v^2 + \dfrac{c_2}{v} \right) \right\} & \text{for } v \le 20 \text{ km/h} \end{cases} \tag{11}$$

with parameters c_0, c_1 and c_2 for free flow, and parameter c_{SG} for stop-and-go traffic conditions. These parameters are differentiated by vehicle type and pollutant. A reduction factor is applied for each pollutant in order to take account of advanced pollution reduction technologies. From the emission factor, $e_f(\alpha, v)$, the emission rate $e(\alpha, v)$ is calculated as follows:

$$e(\alpha, v) = \frac{e_f(\alpha, v) v}{3600} \tag{12}$$

Here, the emission rate is expressed in [g/s] if the velocity and the emission

factor are expressed in [km/h] and [g/km], respectively.

Extending EWS [33] [34], we assume that the emission factor and the emission rate of the pollutant α for a particle with velocity v_i in traffic cellular automata, can be estimated as

$$e_f(\alpha, v_i) = A_0(\alpha) + A_1(\alpha) v_i^2 + \frac{A_2(\alpha)}{v_i} + \frac{A_3(\alpha)}{\sqrt[5]{v_i}} \tag{13}$$

$$e(\alpha, v_i) = B_0(\alpha) + B_1(\alpha) v_i + B_2(\alpha) v_i^3 + B_3(\alpha) \sqrt[5]{v_i^4} \tag{14}$$

for $v_i = v_0, v_1, \cdots, v_{\max}$. The parameters A_r and B_r depend on the pollutant α and on the characteristics of the vehicle. The parameter B_0 represents the emission rate of one vehicle at rest (stopped, but with its motor running). Note the additional term in Equation (13) in comparison with Equation (11). In Section 2.5, we will see that this term allows a very good fitting to the available data reported in [34].

3. Results and Discussion

The main goal of this work is to estimate and compare the emissions rates of the Nagel-Schreckenberg and Fukui-Ishibashi traffic cellular automata. For simplicity, we considered only simulations of the steady states for models with $v_{\max} = 5$ and randomization probability $p = 0.25$. They were carried out with an 800-cells lattice with periodic boundary conditions. Particle densities from 0 to 1 in steps of 0.01 were considered. In each simulation, the system was allowed to evolve during 600 time steps, starting from an initially random spatial distribution of the particles. The simulation was repeated 1000 times for each particle density value. In this case, the ensemble average of the local velocity distribution at each lattice cell is the same as the ensemble average of the global one.

3.1. The Velocity Distributions

In **Figure 4** and **Figure 5**, we present some results of the space-time evolution of the NS and FI models, which we obtained from computer simulations. In **Figure 4**, it is shown the evolution of the NS model for particle densities $n = 0.10, 0.12, 0.17$ and 0.30. Each row (horizontal line) contains an instantaneous spatial distribution of the particles. Time increases vertically from top to bottom. It is observed that $n = 0.12$ defines a transition between two different flow regimes: from free to congested flow.

Figure 5 shows the spatiotemporal evolution of the FI model for densities $n = 0.18, 0.20, 0.22$ and 0.30. In this case, $n = 1/5$ defines a transition between the free flow and congested flow regimes.

The graphs of **Figure 6** show, for the NS and FI models ($v_{\max} = 5$, $p = 0.25$), the steady state partial densities $n_0, n_1, \cdots, n_5$ of the particles with velocities $v_0, v_1, \cdots, v_5$, respectively; and the densities of kinetic energy, ε, and momentum, q, and the velocity v, of the traffic flow, expressed as functions of the vehicular density n. The partial densities $n_i(n)$ were obtained as ensemble averages

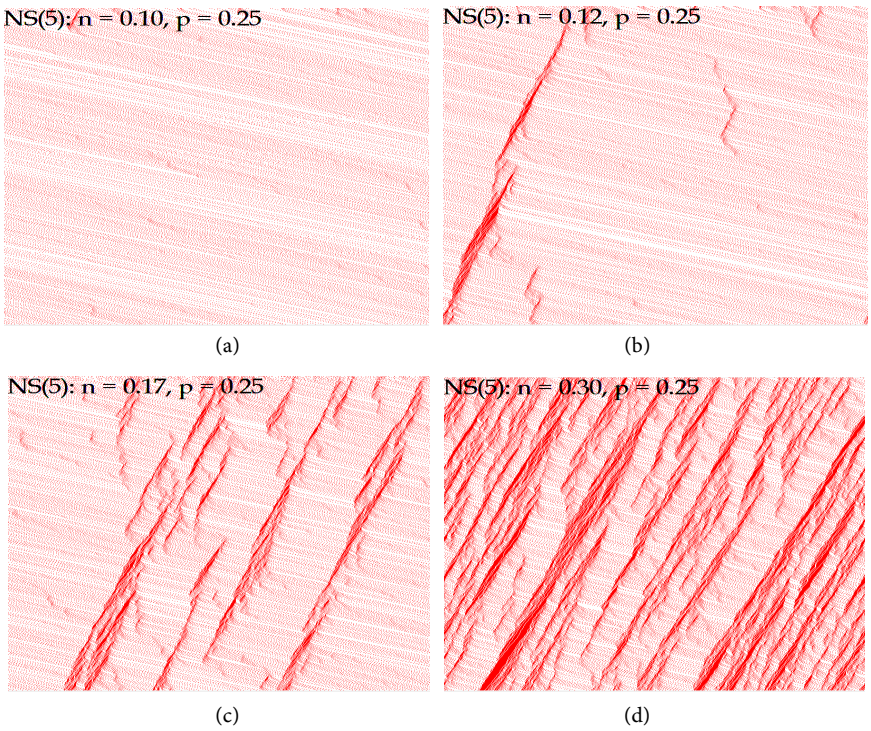


Figure 4. Computer simulations with the NS model with $v_{\max} = 5$ and $p = 0.25$, for several values of the particle density. In particular, for $n = 0.12$, a transition between two different flow regimes is observed. (a) $n = 0.10$; (b) $n = 0.12$; (c) $n = 0.17$; (d) $n = 0.30$.

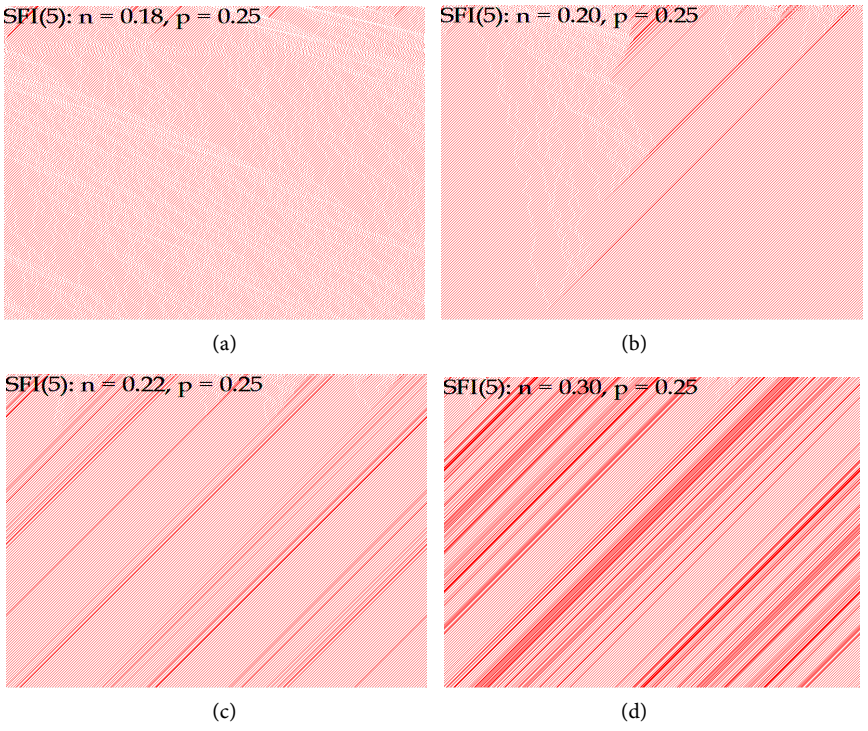


Figure 5. Computer simulations of the FI model with $v_{\max} = 5$ and $p = 0.25$, for several values of the particle density. A transition from the free to the congested flow is observed around $n = 1/5$. (a) $n = 0.18$; (b) $n = 0.20$; (c) $n = 0.22$; (d) $n = 0.30$.

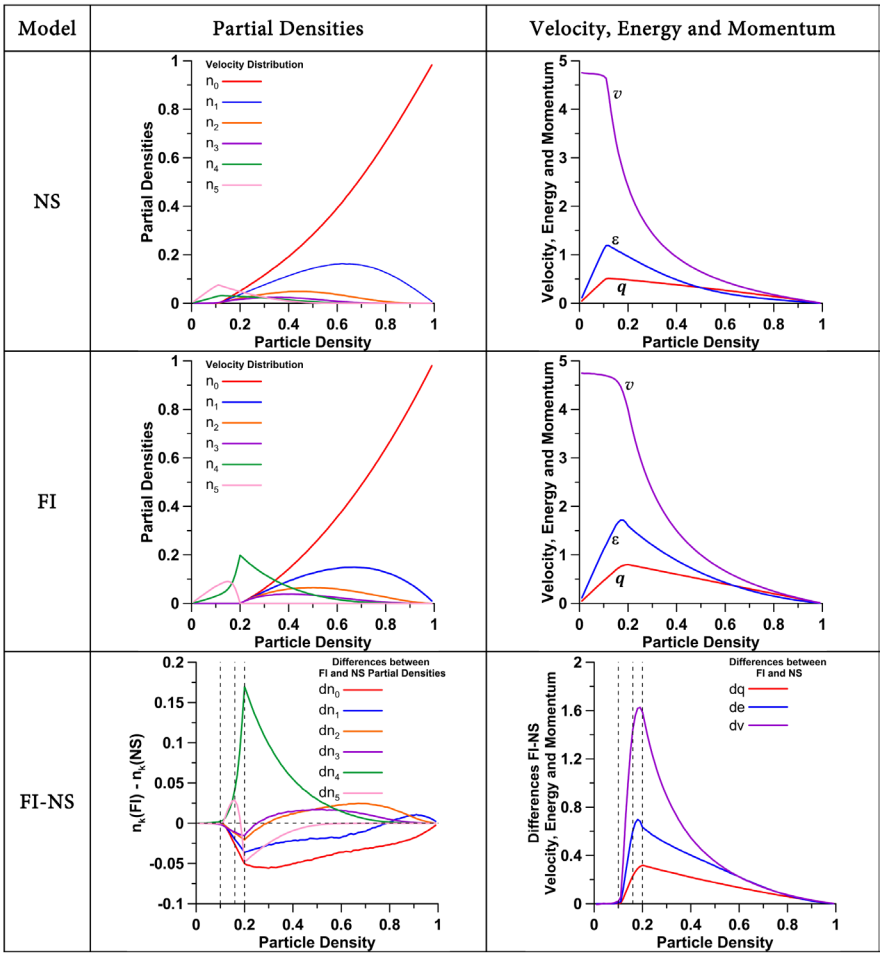


Figure 6. Steady state properties of the NS and FI models with $v_{\max}=5$ and $p=0.25$. The partial densities $n_0, n_1, \cdots, n_5$ and the densities of kinetic energy ε, and momentum q, and the velocity v of the traffic flow, are shown in the first and second rows. The last row shows the differences of these properties between the FI and NS models.

over the 1000 repetitions of the simulations. The properties ε, q, and v were calculated from Equations (4). In the traffic science jargon, the plot of q is known as the fundamental diagram. In the bottom row, we presented graphs which show the differences of these properties between the FI and NS models.

Here, it is observed that all the partial densities, $n_0, n_1, \cdots, n_5$ of the NS model are different from zero in the interval $0<n<1$, although only the partial densities n_4 and n_5 have non-negligible values in the interval $0<n<0.12$. For the FI model, otherwise, only the partial densities n_4 and n_5 are greater than zero in the low-density regime $0<n<1/5$, and for $1/5<n<1$, all partial densities, except n_5, are different from zero. Then, for $v_{\max}=5$ and $p=0.25$, the free flow regime in the FI model extends up to densities close to $n=1/5$, while in the NS model this regime extends only up to $n=0.12$. This is clear in the plots of the average velocity of the traffic flow and in the densities of momentum and kinetic energy, which are shown in the right column of **Figure 6**.

The left column of the last row of Figure 6 shows the plots of the differences between the partial densities of the FI and NS models:

$$dn_k(n) = n_k(NI, n) - n_k(NS, n), \quad k = 0,1,\cdots,5 \tag{15}$$

These plots show that these differences are negligible in the interval $0 < n < 0.12$, that the partial densities n_4 and n_5 in the FI model are larger than in the NS model in the density interval $0.12 < n < 0.18$, but in the same interval the partial densities of the smaller velocities are larger in the NS model than in the FI model. For particle densities $0.18 < n < 0.26$, only the partial density n_4 of the FI model is larger than in the NS model, and for $0.3 < n < 1$, the numbers of particles with velocities v_2, v_3, and v_4 are larger in the FI model. In the high density region $0.79 < n < 1$, also the number of particles with velocity v_1 is larger in the FI model than in the NS model. These observations underline that, in general, the average velocity of traffic flow is larger in the FI model, such as it is shown in the plots we presented in the right column of the last row of Figure 6. These results are consequences of the dynamical rules of the FI traffic cellular automaton, where the particles can increase their velocities faster than in the NS model, and where the stochastic delay only applies to the high-speed cars. As we will show in Section 3.2, this behavior has an important consequence in relation with the air pollutant emissions of the traffic flows described by these models.

For concluding this section, in the graphs of the Figure 7 we showed the maximum entropy states of the NS and FI models for the same set of couples of particle density and kinetic energy per cell, (n, ε), of the simulations we described previously for these traffic models. Again, the partial densities $n_0, n_1, \cdots, n_5$, and the densities of kinetic energy ε and momentum q, and the velocity v, of the traffic flow, are shown. The partial densities presented in this figure (left column) were obtained by numerical solution of Equation (8) using the points (n, ε) of the curves $\varepsilon(n)$ presented in the first and second rows of the right column of Figure 6, as input data. An exception was the case of the low density behavior of the FI model, $0 < n < 1/5$, because the following exact analytical solution of Equation (8) exists for this case [17]:

$$n_5 = \frac{1}{2}\left\{1 - 4n - \sqrt{(1-4n)^2 - 4n(1-5n)(1-p)}\right\} \tag{16a}$$

$$n_4 = n - n_5 \tag{16b}$$

$$v = 4 + \frac{1}{2n}\left\{1 - 4n - \sqrt{(1-4n)^2 - 4n(1-5n)(1-p)}\right\} \tag{16c}$$

For each model, important differences can be observed between the plots of the partial densities obtained from computer simulations (first and second rows of Figure 6) and from the maximum entropy approach (Figure 7), mainly for the high-density regimes. The main reason for these behavior differences is due to the dynamical transition rules of the NS and FI traffic cellular automata, which do not satisfy the principle of detailed balance [30] [31], and, therefore, both systems are always driven out of equilibrium.

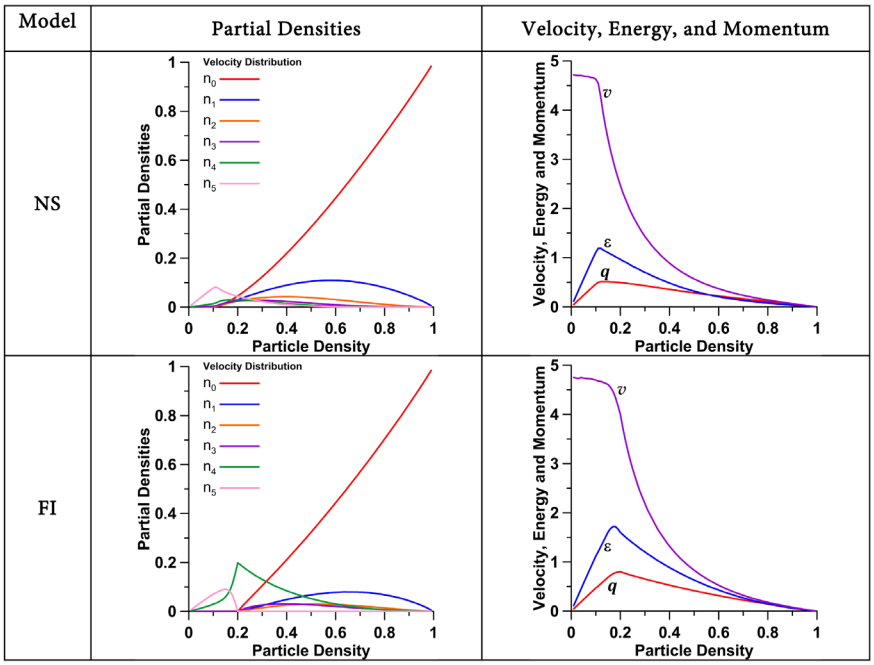


Figure 7. Maximum entropy states of the NS and FI models with $v_{\max} = 5$ and $p = 0.25$. The partial densities $n_0, n_1, \cdots, n_5$, the velocity v, and the densities of kinetic energy ε, and momentum q, of the traffic flow, are shown.

3.2. The Emission Rates

With reference to Section 2.2, we underline that the distance among adjacent cells is usually assumed as the average front-bumper-to-front-bumper distance of adjacent vehicles under conditions of strongly jammed traffic, and it is set equal to 7.5 m. Then, if the time step is set equal to one second, the velocity of a vehicle will change in steps of 27 kph. Therefore, when comparing with real traffic data, the interpretations of the model velocities will be as follows,

$$v_0 = 0, v_1 = 27, v_2 = 54, v_3 = 81, v_4 = 108 \text{ and } v_5 = 135 \text{ kph} \qquad (17)$$

The emission factors we used in this work are based on [18]. In the data base, the emission factors are given for traffic situations which are characterized by a mean speed (beside other dependencies). In order to obtain an effortless mapping between the velocity and the amount of emission, the different traffic situations were aggregated into bins of size 10 km/h [34] [35].

Figure 8 shows the emission factors and the emission rates for three different pollutants: carbon monoxide (CO), hydrocarbons (HC), and nitrogen oxides (NO_x).

In this figure, we observed that the emission behavior is rather different for the distinct pollutants, and that their amount strongly depends on velocity. For estimating the pollutant emissions in the NS and FI traffic cellular automata, we used the emission rates of CO, HC, and NO_x shown in **Figure 8**. The associated best fitting functions and their determination coefficients (R^2) are presented in **Table 2**.

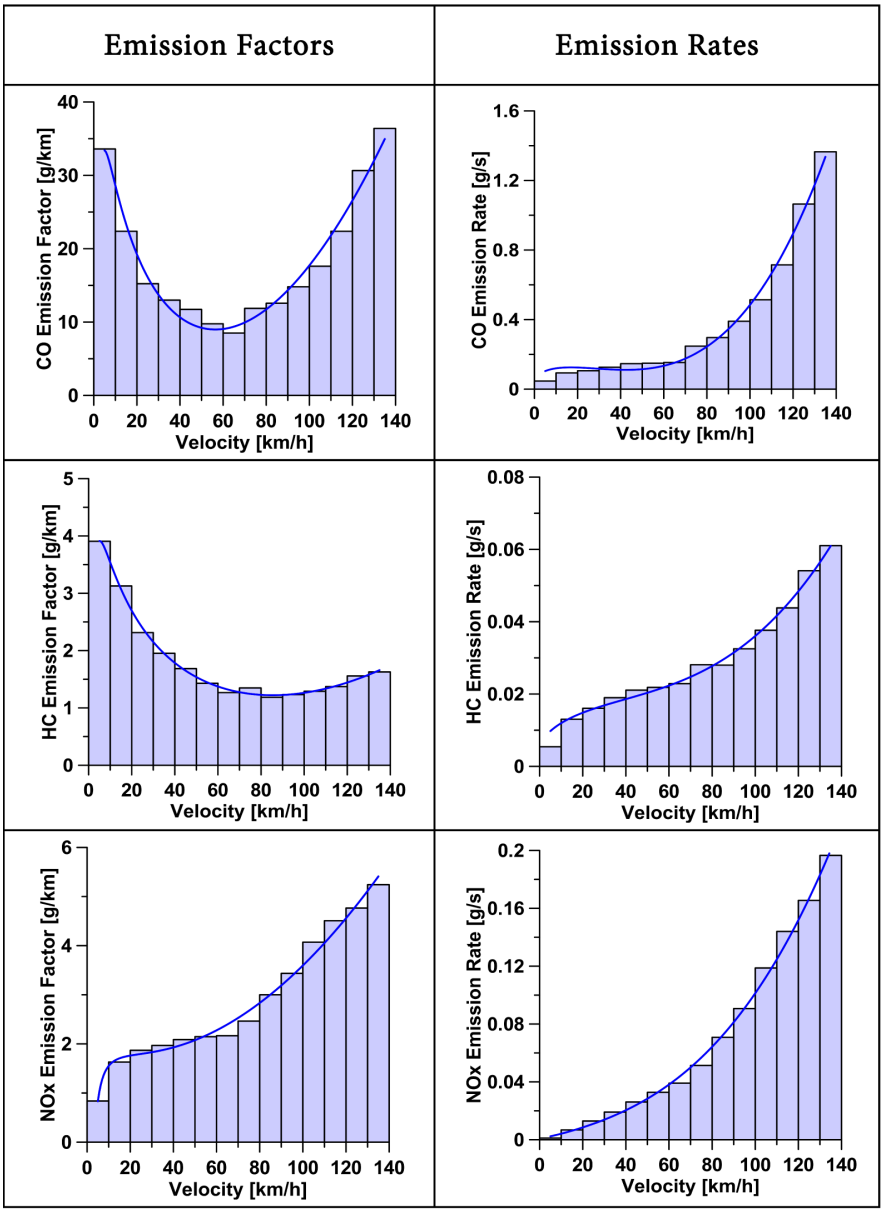


Figure 8. Vehicular emission factors and emission rates for the pollutants CO, HC, and NO_x as functions of velocity. We observe that the amount of emission is strongly dependent on the vehicle's velocity and on the kind of pollutant. The best fitting curves (blue solid lines) of the model functions (13) and (14) to the data available in [34] [35] are also shown.

Figure 9 shows the partial emission rates produced by the simulations with the NS and FI models for vehicle emissions of carbon monoxide (CO), hydrocarbons (HC), and nitrogen oxides (NO_x). The differences between these models are also shown. These results were obtained with the Equation (9), using the NS and FI velocity distributions shown in Figure 6, and the CO, HC and NO_x emission rates given in Table 2 with the allowed particle velocities given by the Equation (17).

The differences between the emission rates produced by the FI and NS models

(Figure 9) are shown in Figure 10. Here we note that the FI traffic model produces the larger emission rates, particularly those associated with the velocity $v_{\max}-1$. However, when the system is in the high-density regime, the NS model produces an emission rate larger than the FI model does associated with the largest velocity $v_{\max}$.

In Figure 11, we show the partial emission rates obtained with the maximum entropy velocity distributions (see Figure 7). In Figure 12, the differences between the FI and NS emission rates are shown.

The graphs of Figure 9 and Figure 11 show important qualitative similarities between the partial emission rates estimated with the velocity distributions obtained by computer simulation and by means of the maximum entropy approach. However, there exist non-negligible numerical differences which are reflected also in the total emission rates, particularly for the CO and HC pollutants, as it is shown in the Figure 13.

In Figure 13, we present the total emission rates for the models NS and FI. It includes the results obtained with computer simulations and with the maximum

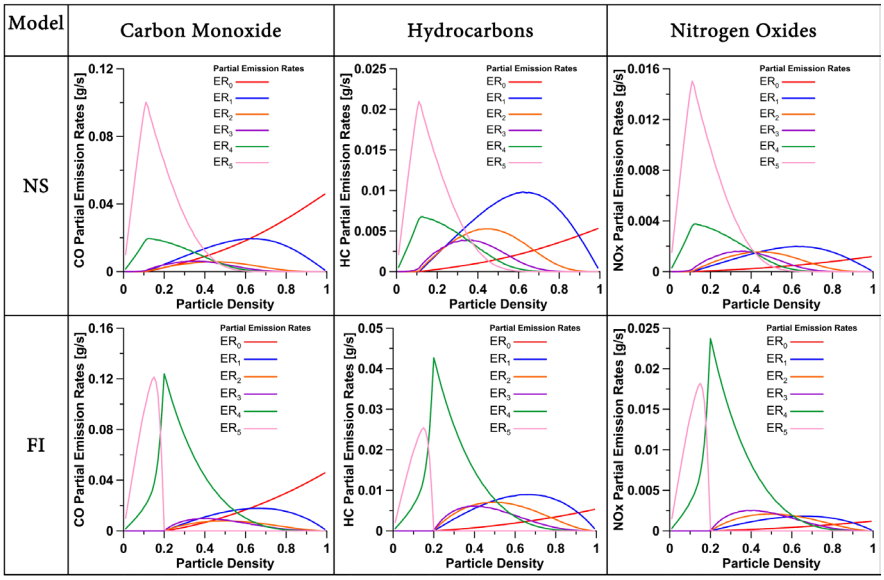


Figure 9. Partial emission rates of the NS and FI traffic models for the pollutants CO, HC and NO_x.

Table 2. Best fitting functions for the traffic emission rates of CO, HC, and NO_x as dependent on the vehicle velocity. Estimated from data available in [34] [35]. Determination coefficients (R^2) are also shown.

Pollutant	Emission Rate [g/s]	R^2
CO	$e_{CO}(v_i)=0.0467-0.020966v_i+7.551701\times10^{-7}v_i^3+0.044694\sqrt[5]{v_i^4}$	0.991621
HC	$e_{HC}(v_i)=0.0054-0.000810v_i+1.931618\times10^{-8}v_i^3+0.002321\sqrt[5]{v_i^4}$	0.988483
NO_x	$e_{NO_x}(v_i)=0.0012+0.000703v_i+5.577680\times10^{-8}v_i^3-0.000653\sqrt[5]{v_i^4}$	0.996947

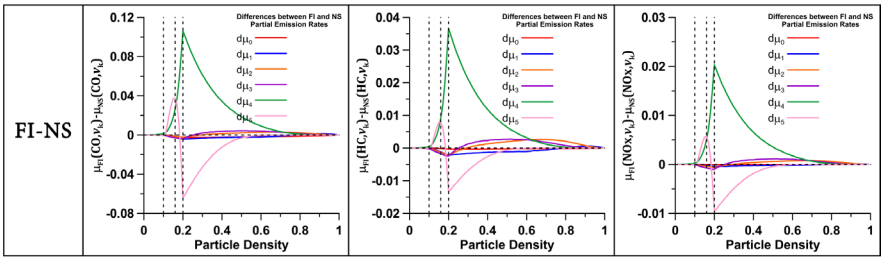


Figure 10. Differences between the FI and NS partial emission rates of **Figure 9**.

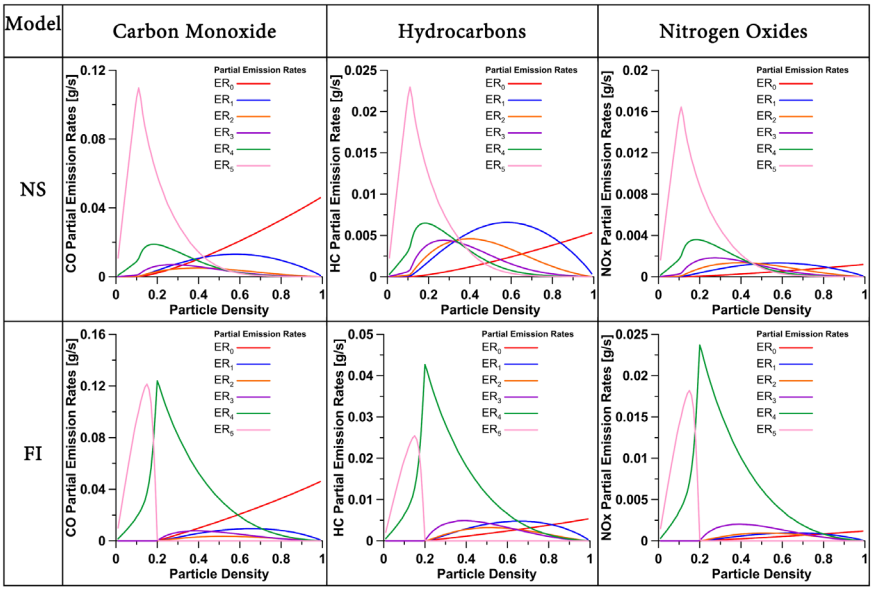


Figure 11. Partial emission rates of the NS and FI traffic cellular automata estimated with the maximum entropy velocity distribution.

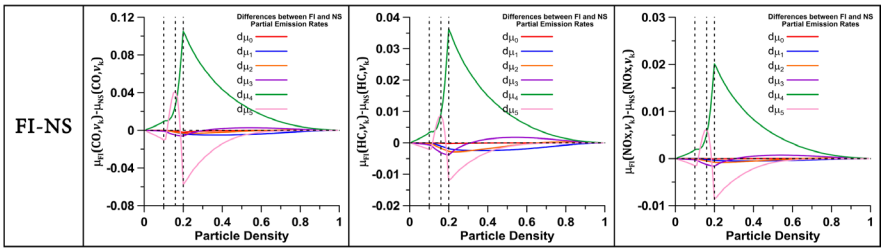


Figure 12. Differences between the FI and NS emission rates of Figure 11.

entropy approach. The plots presented in Figure 13 were obtained by summing, respectively, the partial emission rates of Figure 9 and Figure 11, such as it is indicated by the Equation (10).

In **Figure 13**, we observe: for densities $0 < n < 0.11$, both traffic cellular automata produced the same total emission rates for each pollutant; for densities $n > 0.11$, the FI traffic model produced total emission rates of CO, HC and NO_x larger than the NS model did, respectively. In the limit $n \to 1$, the emission rates of both models become the same because all the particles become at rest, remaining only the emissions in the idle conditions.

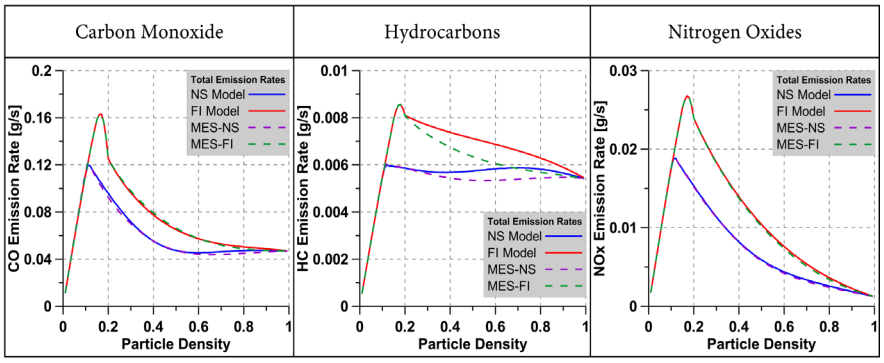


Figure 13. Total Emission rates of the NS and FI traffic cellular automata ($v_{\max}=5$ and $p=0.25$) as functions of the particle density. The solid line curves correspond to the computer simulations of steady state conditions. The dotted line curves correspond to the maximum entropy states.

Figure 13 shows also the plots of the emission rates estimated with the maximum entropy state velocity distributions. We observe that the deviations with respect the estimations with velocity distributions obtained from computer simulations only result important in the case of the hydrocarbons, because in this case the differences and the emission rates themselves are of the same order. The larger emission rates of the FI model are a consequence of its dynamic rules because the stochastic delay is applied only to the highest speed vehicles, extending its free flow regime up to particle densities higher than in the NS model.

In Figure 14, the relative differences

$$\delta(\alpha,n)\equiv 100\left(\frac{Q_{FI}(\alpha,n)-Q_{NS}(\alpha,n)}{Q_{NS}(\alpha,n)}\right), \tag{18}$$

between the total emission rates of the FI and NS models for steady state conditions and the selected pollutants (Figure 13), are shown. Here we can underline three interesting density intervals:

1) $0<n<0.11$: This is the interval of the low density behavior of the FI and NS traffic models with $v_{\max}=5$ and $p=0.25$. Here, almost all the particles are moving with one of the two highest velocities, $v=5$ (*i.e.* $v_{\max}$) or $v=4$ (*i.e.* $v_{\max}-1$). It is a free flow regime. In this density region, the relative differences between the estimations of the emission rates of the NS and FI models are negligible for all the pollutants we considered:

$$\delta(\mathrm{CO},n)\cong\delta(\mathrm{HC},n)\cong\delta(\mathrm{NO_x},n)\cong 0.$$

2) $0.11<v<1/5$: In this interval, while the numbers of particles at rest and with the lower velocities in the system start to be non-negligible in the NS model, all the particles persist in the free flow regime, with the highest velocities in the FI model; however, the number of particles con velocity $v=5$ decreases to zero at $n=1/5$. Because of the velocity distribution (Figure 6, left column) and of the dependence of emission rate on velocity (Figure 8, right column), the emission rates, in this interval, also reach their highest values for each pollutant

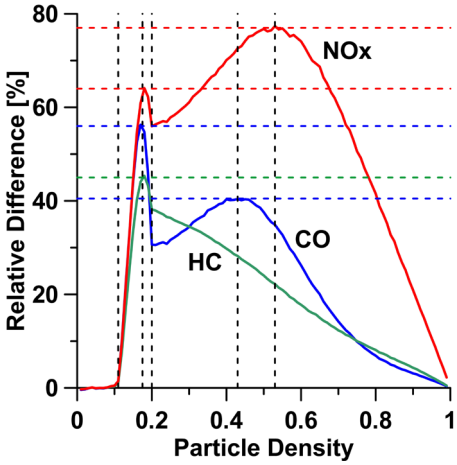


Figure 14. Relative difference (%) between the FI and NS estimations of the emission rates.

(**Figure 11**), with larger values for the FI model than for the NS model. Here, all the relative differences $\delta(\mathrm{HC},n)$, $\delta(\mathrm{CO},n)$ and $\delta(\mathrm{NO_x},n)$ show sharp peaks (45.36%, 56.27% and 64.10%, respectively) around $n=0.175$, just where the highest values of the FI emission rates occur.

3) $1/5<n<1$: In this interval, both models exhibit a congested flow regime. As the particle density increases, the numbers n_4 and n_5 of the particles that move with the highest velocities, decrease monotonically; the number of particles at rest (n_0) increases monotonically; and the numbers of particles with velocities v_1, v_2 and v_3 grow up to a maximum and then drop to zero. Because of this, the total emission rates diminish monotonically up to their idle condition values, when all the particles become at rest. On the other hand, the relative difference between the emission rates of the FI and NS models (**Figure 14**) seems to decrease monotonically for hydrocarbons, but for carbon monoxide and nitrogen oxides grows up and then drops to zero, reaching their maximum values, 40.41% and 76.87%, at $n=0.43$ and $n=0.55$, respectively.

4. Concluding Remarks

There exists a growing interest in using cellular automata to model traffic flow phenomena from a microscopic standpoint. The possibility of using these models to simulate traffic in the cities brings out the attention to the problem of assessing the contributions of this phenomenon to the urban air pollution. To do it, the velocity distribution of the traffic network has to be known, spatially and temporally disaggregated. It is also required the engine's emission factors or emission rates as functions of the vehicle velocity. In this work, we used computer simulations and a maximum entropy approach for obtaining the velocity distributions of the traffic cellular automata of Nagel-Schreckenberg and Fukui-Ishibashi under steady state conditions. The engine emissions were obtained from data available in [18] [33] [34] [35], which allowed us to estimate and compare the emission rates of CO, HC, and NO_x produced by the NS and FI

traffic models.

Although the dynamical rules of the NS and FI models are not microscopically reversible and, therefore, these systems are always far from equilibrium, our estimations of the total traffic emission rates with the maximum entropy velocity distributions resulted very similar to those we obtained using the velocity distributions from computer simulations with these traffic cellular automata.

In general, the emission rates in the FI traffic flow resulted larger than in the NS model. The relative differences $\delta(\alpha, n)$ reached values of up to 45% in HC, 56% in CO, and 77% in NO_x. These results are consequences of the differences between the FI and NS dynamic rules: In the NS model, the acceleration of the particles is gradual, while in the FI model, a particle can accelerate from rest up to the maximum velocity in a single time step. Moreover, the stochastic delay is applied only to the particles with the highest velocities in the FI model.

The ideas of this study can be extended easily to other 1D or 2D traffic cellular automata for estimating the traffic flow contributions to air pollution.

Acknowledgements

We thank Ana Teresa Celada Murillo (Instituto Nacional de Electricidad y Energías Limpias) for beneficial comments and for her help to make the paper more comprehensible.

Fund

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing Interest Statement

The authors declare no conflict of interest.

References

[1] INEGI (2015) Instituto Nacional de Estadística y Geografía (INEGI) [Motor Vehicles Registered in Circulation. Administrative Records of the National Institute of Statistics and Geography.] Data available at the website: http://www.inegi.org.mx/est/contenidos/proyectos/registros/economicas/vehiculos (in Spanish). Last Access: November 3, 2017.

[2] IE-CDMX (2016) Inventario de Emisiones de la CDMX 2014. Contaminantes Criterio, Tóxicos y de Efecto Invernadero. Secretaría del Medio Ambiente del Gobierno de la Ciudad de México. [Emission Inventory of Mexico City 2014. Criteria, Toxic, and Greenhouse Effect Pollutants.] (in Spanish)

[3] Prigogine, I. and Herman, R.C. (1971) Kinetic Theory of Vehicular Traffic. 1st Edition (July 23, 1971), Elsevier, New York, London, Amsterdam.

[4] Cremer, M. and Ludwig, J. (1986) A Fast Simulation Model for Traffic Flow on the Basis of Boolean Operations. *Mathematics and Computers in Simulation*, **28**, 297-303. https://doi.org/10.1016/0378-4754(86)90051-0

[5] Nagel, K. and Schreckenberg, M. (1992) A Cellular Automaton Model for Freeway

Traffic. *Journal de Physique Archives*, **2**, 2221-2229. https://doi.org/10.1051/jp1:1992277

[6] Fukui, M. and Ishibashi, Y. (1996) Traffic Flow in 1D Cellular Automaton Model Including Cars Moving with High Speed. *Journal of the Physical Society of Japan*, **65**, 1868-1870. https://doi.org/10.1143/JPSJ.65.1868

[7] Wagner, P., Nagel, K. and Wolf, D.E. (1997) Realistic Multi-Lane Traffic Rules for Cellular Automata. *Physica A*, **234**, 687-698.

[8] Nagel, K., Wolf, D.E., Wagner, P. and Simon, P. (1998) Two-Lane Traffic Rules for Cellular Automata: A Systematic Approach. *Physical Review E*, **58**, 1425-1437. https://doi.org/10.1103/PhysRevE.58.1425

[9] Chowdhury, D., Santen, L. and Schadschneider, A. (2000) Statistical Physics of Vehicular Traffic and Some Related Systems. *Physics Reports*, **329**, 199-329.

[10] Rickert, M. and Nagel, K. (1997) Experiences with a Simplified Microsimulation for the Dallas/Fort Worth Area. *International Journal of Modern Physics C*, **8**, 483-503.

[11] Nagel, K. and Barrett, C.L. (1997) Using Microsimulation Feedback for Trip Adaptation for Realistic Traffic in Dallas. *International Journal of Modern Physics C*, **8**, 505-525.

[12] Tonguz, O.K. and Viriyasitavat, W. (2009) Modeling Urban Traffic: A Cellular Automata Approach. IEEE Communications Magazine, May 2009, 142-150. https://doi.org/10.1109/MCOM.2009.4939290

[13] Zhou, T. and Lijuan, P. (2013) Cellular Automata Simulation of Urban Traffic Flow Considering Bus Lane. *International Journal of Online Engineering*, **9**, 65-70. https://doi.org/10.3991/ijoe.v9iS7.3193

[14] Kurnaz, İ. (2016) Urban Traffic Modeling with Microscopic Approach using Cellular Automata. *Technical Gazette*, **23**, 1565-1570.

[15] Salcido, A. (2007) The Maximum Entropy States of 1D Cellular Automata Traffic Models. *Proceedings of the* 18*th IASTED International Conference on Modelling and Simulation*, Montreal, 30 May-1 June 2007, 160-165.

[16] Salcido, A. (2011) Equilibrium Properties of the Cellular Automata Models for Traffic Flow in a Single Lane. In: Salcido, A., Ed., *Cellular Automata*, *Simplicity behind Complexity*, InTech, 159-192. https://doi.org/10.5772/15371

[17] Salcido, A., Hernández-Zapata, E. and Carreón-Sierra, S. (2017) Exact Results of 1D Traffic Cellular Automata: The Low-Density Behavior of the Fukui-Ishibashi Model. (under Review in Physica A)

[18] HBEFA (1999) Handbuch Emissionsfaktoren des Straßenverkehrs. [Handbook of Emission Factors of Road Traffic.] Version 1.2, on Behalf of the Swiss Ministry of Environment, Forestry and Agriculture, Bern and the German Environmental Agency Berlin. (in German)

[19] Von Neumann, J. (1966) The Theory of Self-Reproducing Automata. University of Illinois Press, Urbana.

[20] Ilachinski, A. (2001) Cellular Automata. A Discrete Universe. World Scientific, Singapore.

[21] Salcido, A. (2011) Cellular Automata. Simplicity behind Complexity. InTech.

[22] Salcido, A. (2011) Cellular Automata. Innovative Modelling for Science and Engineering. InTech.

[23] Salcido, A. (2013) Emerging Applications of Cellular Automata. InTech.

[24] Wolfram, S. (1984) Computation Theory of Cellular Automata. *Communications in*

Mathematical Physics, **96**, 15-57.

[25] Gardner, M. (1970) The Fantastic Combinations of John Conway's New Solitaire Game "Life". *Scientific American*, **223**, 120.

[26] Fukui, M. and Ishibashi, Y. (1993) Evolution of Traffic Jam in Traffic Flow Model. *Journal of the Physical Society of Japan*, **62**, 3841-3844. https://doi.org/10.1143/JPSJ.62.3841

[27] Schadschneider, A. (1999) The Nagel-Schreckenberg Model Revisited. *The European Physical Journal B*, **10**, 573-582. https://doi.org/10.1007/s100510050888

[28] Brilon, W. and Wu, N. (1999) Evaluation of Cellular Automata for Traffic Flow Simulation on Freeway and Urban Streets. In: Brilon, W., Huber, F., Schreckenberg, M. and Wallentowitz, H., Eds., *Traffic and Mobility*, Springer, Berlin, Heidelberg.

[29] Wagner, P., Nagel, K. and Wolf, D.E. (1997) Realistic Multi-Lane Traffic Rules for Cellular Automata. *Physica A*, **234**, 687-698.

[30] Klein, M.J. (1955) Principle of Detailed Balance. *Physical Review*, **97**, 1446-1447. https://doi.org/10.1103/PhysRev.97.1446

[31] Gorban, A.N. (2014) Detailed Balance in Micro- and Macrokinetics and Micro-Distinguish Ability of Macro-Processes. *Results in Physics*, **4**, 142-147.

[32] Schütz, G.M. (2001) Exactly Solvable Models for Many-Body Systems Far from Equilibrium. In: Domb, C. and Lebowitz, J.L., Eds., *Phase Transitions and Critical Phenomena*, Vol. 19, Academic Press, San Diego, 1-251.

[33] Walther, C. (2000) Geschwindigkeiten und Schadstoffemissionen des motorisierten Straßenverkehrs in innerörtlichen Netzen auf der Grundlage der EWS-Berechnungsvorschriften. [Speeds and Pollutant Emissions of Motorized Road Transport in Intra-Urban Networks Based On EWS (Environmental Protection) Calculation Rules.] *Straßenverkehrstechnik*, **1**, 19-25. (in German)

[34] Eissfeldt, N.G. (2004) Vehicle-Based Modelling of Traffic. Theory and Application to Environmental Impact Modelling. Inaugural-Dissertation zur Erlangung des Doktorgrades der Mathematisch-Naturwissenschaftlichen Fakultät der Universität zu Köln. [Vehicle-Based Modelling of Traffic. Theory and Application to Environmental Impact Modelling. Inaugural-Dissertation to Obtain the Doctor Degree of the Faculty of Mathematics and Natural Sciences of the University of Cologne.]

[35] Eissfeldt, N. and Schrader, R. (2002) Calculation of Street Traffic Emissions with a Queuing Model. *Journal of Computational Technologies*, **7**, 5-15. http://e-archive.informatik.uni-koeln.de/id/eprint/412

Scientific Research Publishing

Journal of Applied Mathematics and Physics, 2017, 5, 2162-2171
http://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Simplest Method for Calculating the Lowest Achievable Uncertainty of Model at Measurements of Fundamental Physical Constants

Boris Menin

Refrigeration Consultancy Ltd., Beer-Sheba, Israel
Email: meninbm@gmail.com

How to cite this paper: Menin, B. (2017) Simplest Method for Calculating the Lowest Achievable Uncertainty of Model at Measurements of Fundamental Physical Constants. *Journal of Applied Mathematics and Physics*, **5**, 2162-2171.
https://doi.org/10.4236/jamp.2017.511176

Received: October 16, 2017
Accepted: November 7, 2017
Published: November 10, 2017

Abstract

The CODATA procedure for calculating the recommended relative uncertainty of the measured fundamental physical constants is complex and is based on the use of powerful computers and modern mathematical statistical methods. In addition, the expert's opinion caused by accumulated knowledge, life experience and intuition of researchers is applied at each stage of the calculations. In this article, the author continues to advocate a theoretically grounded information method as the most effective tool for testing and achieving the minimum possible relative uncertainty for any measurements of experimental physics and engineering. The introduced fundamental limit characterizing discrepancy between a model and the observed object cannot be overcome by any improvement of instruments, methods of measurement and the model's computerization. Examples are given.

Keywords

Fundamental Physical Constants, Information Theory, Mathematical Modeling, Similarity Theory, Uncertainty

1. Introduction

In the past century, an increased attention to the accuracy of fundamental physical constants has led to the development of new physical theories and many technological improvements [1]. We are, however, possibly gradually reaching the limits of accuracy, in spite of the fact of usage of powerful computers and unique newest mathematical methods. Is there a reasonable limit to our exact

understanding surrounding the world? This limit is unquestionably manifested in the restructuring of the International System of Units (SI) and in the identification of the difficulties that now appear in electrical measurements [2].

Why is it important to know the limits of our ability to measure fundamental physical constants? First, a precise definition of the fundamental constants allows us to verify the consistency and correctness of the basic physical theories. Second, the quantitative predictions of the basic physical theories depend on the numerical values of the constants involved in these theories: each new sign can lead to the discovery of a previously unknown inconsistency or, conversely, can eliminate the existing inconsistency in our description of the physical world. However, by formulating the model of the experiment, scientists, on the one hand, somehow break off the connections with possible, potentially influencing variables, which are hidden from our eyes at the moment. On the other hand, it is very difficult to measure all the variables that are taken into account in the measurement model with high accuracy and not 'miss' the significant effect. In addition to these problems, it often turns out that a new series of measurements using more advanced methods shows that the former value was erroneous and that its new value differs from the old 'conventional' by an amount many times greater than the uncertainty attributed to the previous value.

These 'flowers' do not end the field of activity of scientists to clarify the true meaning of fundamental physical constants. A difficult task in negotiating the values of constants is the estimation of uncertainties. In most experiments, physicists try to collect as much data as possible in order to reduce the random measurement uncertainty to a negligible value. In this case, the final uncertainty attributed to the result of the measurement is determined only based on an assessment of systematic uncertainties. These uncertainties are associated with effects, for which little is known. Therefore, the corresponding estimates are somewhat subjective and are usually obtained essentially intuitively [1]. The problem is aggravated by the fact that different experimenters approach the evaluation of systematic uncertainties from completely different positions. Some of them cautiously attribute their data to an overestimated uncertainty in the hope that subsequent measurements will not reject their results as incorrect. Others, on the contrary, underestimate the sources of systematic uncertainties in their experiments, apparently proceeding from an unconscious (and perhaps even intentional) desire to conduct 'the best experiment'. Such factors, so far from scientific objectivity, are inevitable, since in the end, people who have different life experiences and are endowed with greater or lesser abilities make science. That is why the measurement uncertainties presented by different researchers are not at all easy to compare with each other.

Undoubtedly, the coordinated set of physical constants of 2015 [3] is closer to the truth than the previous ones. However, being realists, we cannot reject the possibility that a new, theoretically grounded approach to the calculation of the recommended value of the relative uncertainty is required or can exist, by reali-

zation of which measuring of the true-target values of fundamental physical constants is possible.

That is why, from the point of view of the author, the recently developed information approach may to some extent facilitate the process of calculating the amount of relative uncertainty to be measured by physicists. It is based on the proposition that a multi-physics model contains a certain quantity of information about the object under study, depending on the quantitative and qualitative set of physical variables to be taken into account. By this way the optimal number of selected parameters can be calculated. It allows reaching the lowest discrepancy between a model and the observed object. This approach defines a limit of accuracy that cannot be overcome by any improvement of instruments, methods of measurement and the model's computerization. It has physical meaning and its value is much higher than the Heisenberg uncertainty relation provides. It is important to mention that applying information theory allows giving a theoretical explanation and grounding of experimental results, which determine precision of fundamental constants.

The information approach plays a decisive role in a new view of the achievement of the least uncertainty in the model of a physical phenomenon. Unfortunately, it is still little known to most scientists and engineers. In this paper, we continue to apply it for measurements of fundamental physical constants.

2. Suggested Applied Tools

In [4], it was shown that a certain uncertainty exists before starting an experiment or computer simulations. It is caused due only to the known number of recorded variables. The value of this uncertainty can be calculated by the following formula

$$\Delta_{\text{pmm}} = S \cdot \left[\left(z' - \beta' \right) / \mu_{\text{SI}} - \left(z'' - \beta'' \right) / \left(z' - \beta' \right) \right], \tag{1}$$

where u is the dimensionless researched variable; Δu is the dimensionless uncertainty of the physical-mathematical model describing the experiment with the apriority chosen number of variables; S is the predetermined dimensionless interval of u variations; z'' is the given number of selected physical dimensional variables; β'' is the number of primary physical variables recorded in a model; $\varepsilon = \Delta_{\text{pmm}}/S$ is the comparative uncertainty. Equation (1), surprisingly, is very simple. Absolute and relative uncertainties are familiar to physicists. As for the comparative uncertainty, it is rarely mentioned. Nevertheless, the importance of comparative uncertainty is of great importance for the application of information theory in physics and engineering; z' is the number of physical dimensional variables in the selected class of phenomena (*CoP*-see below), β' is the number of primary physical dimensional variables in the selected *CoP*; μ_{SI} is the total number of possible dimensionless criteria with $\xi = 7$ main dimensional variables for SI-see below.

The relation (1), which follows from the general provisions of information theory, is accurate and does not depend on the conditions of experience, the

concrete implementation of the test stand, the expert opinion of scientists and the selected statistical mathematical methods.

An overall uncertainty of the model including inaccurate input data, physical assumptions, the approximate solution of the integral-differential equations, etc., will be larger than Δ_{pmm}. Thus, Δ_{pmm} is the ***first-born*** and ***least component*** of a possible mismatch of a real object and its modeling results.

Equation (1) has physical meaning. It testifies that in nature there is a fundamental limit to the accuracy of measuring any process, which cannot be surpassed by any improvement of instruments, methods of measurement and the model's computerization. The value of this limit is much higher and stronger than the Heisenberg uncertainty relation provides. In addition, this fundamental limit places severe restrictions on the micro-physics.

SPV and SI are a fantasy generated by collective imagination. However, without SPV, the simulation of the phenomenon is impossible. You can interpret SPV as the basis of all the available knowledge that people have about the surrounding nature at the moment.

SI includes the primary and secondary variables used for descriptions of different classes of phenomena (*CoP*). For example, in mechanics of SI there is used a basis {*L*-length, *M*-mass, *T*-time}, *i.e.* $CoP_{\text{SI}} \equiv LMT$.

It is known [5] [6] that the dimension of any secondary variable can be expressed as a unique function of the product of primary variables *L*, *M*, *T*, *I*, Θ, *J*, and *F* with certain exponents *l*, *m*, *t*, *i*, *θ*, *j*, *f*, which can take only integer values and vary in specific ranges

$$q \ni L^l \cdot M^m \cdot T^t \cdot I^i \cdot \Theta^\theta \cdot J^j \cdot F^f, \tag{2}$$

$$\begin{aligned} &-3 \le l \le +3,\ \ -1 \le m \le +1,\ \ -4 \le t \le +4,\ \ -2 \le i \le +2, \\ &-4 \le \theta \le +4,\ \ -1 \le j \le +1,\ \ -1 \le f \le +1, \end{aligned} \tag{3}$$

$$e_l = 7;\ e_m = 3;\ e_t = 9;\ e_i = 5;\ e_\theta = 9;\ e_j = 3;\ e_f = 3 \tag{4}$$

where $e_l, \cdots, e_f$ are the numbers of variants of the dimension for each variable. For example, I^{-3} is used in the density formula; θ^4 is used in the Stefan-Boltzmann law.

We can calculate the total number of possible dimensionless criteria μ_{SI} with $\xi = 7$ main dimensional variables for SI [4]

$$\mu_{\text{SI}} = \Psi - \xi = 38272 - 7 = 38265 \tag{5}$$

where Ψ = 38,272 is the total number of dimensional options of physical variables in SI; μ_{SI} corresponds to the maximum amount of information contained in the SPV; each variable allows the researcher to obtain a certain amount of information about the studied object; the main definitions and estimates of the amount of information used in the experiment were clearly formulated by L. Brillouin [7] and generalized by M. Burgin [8].

Equating the derivative of Δ_{pmm}/S from Equation (1) with respect to $z' - \beta'$ to zero, we obtain the condition for achieving the minimum comparative uncertainty for a particular COP:

$$\left(z'-\beta'\right)^2 \Big/ \left(\Psi-\xi\right)=\left(z''-\beta''\right). \tag{6}$$

For the analysis of experimental data, we need to know the recommended number of selectable variables, with which we can achieve a minimum comparative uncertainty for a specific *CoP*:

1) For $CoP_{\mathrm{SI}} \equiv LMTF$, taking into account the aforementioned explanations and (6), the lowest comparative uncertainty ε_{LMTF} can be reached at the following conditions:

$$\left(z'-\beta'\right)=\left(e_l\times e_m\times e_t\times e_f-1\right)\Big/2-4=\left(7\times 3\times 9\times 3-1\right)\Big/2-4=279 \tag{7}$$

$$\left(z''-\beta''\right)=\left(z'-\beta'\right)^2\Big/\mu_{\mathrm{SI}}=279^2/38265\approx 2 \tag{8}$$

where '−1' corresponds to the case when all the primary variable exponents are zero in Formula (2); dividing by 2 indicates that there are direct and inverse variables, e.g., L^1 is length, L^{-1} is run length. Because the object can be judged knowing only one of its symmetrical parts, while others structurally duplicating this part may be regarded as information empty [4]. Therefore, the number of options of dimensions may be reduced by $\omega = 2$ times; 4 corresponds to the four primary variables *L*, *M*, *T*, *F*.

$$\begin{aligned}\left(\varepsilon_{\min}\right)_{LMTF}&=\Delta_{\mathrm{pmm}}/S=\left[\left(z'-\beta'\right)/\mu_{\mathrm{SI}}-\left(z''-\beta''\right)/\left(z'-\beta'\right)\right]\\&=279/38265+2/279=0.0073+0.0073=0.0146\end{aligned} \tag{9}$$

2) For $CoP_{\mathrm{SI}} \equiv LMT\theta I$, taking into account (6), the lowest comparative uncertainty ε_{LMTFI} can be reached at the following conditions:

$$\left(z'-\beta'\right)_{LMT\theta I}=\left(e_l\times e_m\times e_t\times e_\theta\times e_i-1\right)\Big/2-5=\left(7\times 3\times 9\times 9\times 5-1\right)\Big/2-5=4247 \tag{10}$$

$$\left(z''-\beta''\right)_{LMT\theta I}=\left(z'-\beta'\right)^2\Big/\mu_{\mathrm{SI}}=4247^2/38265\approx 471 \tag{11}$$

where '−1' corresponds to the case when all the primary variable exponents are zero in Formula (2); dividing by 2 indicates that there are direct and inverse variables, e.g., L^1-length, L^{-1}-run length, and 5 corresponds to the five primary variables *L*, *M*, *T*, Θ, *I*.

Then, one can calculate the minimum achievable comparative uncertainty $\varepsilon_{LMT\Theta I}$

$$\varepsilon_{LMT\Theta I}=\left(\Delta u/S\right)_{LMT\Theta I}=4247/38265+471/4247=0.222 \tag{12}$$

Let's speculate about applying the information approach for the measurement of several fundamental physical constants.

3. Applications

3.1. Proton Mass m_p

We analyzed several research publications and CODATA (Committee on Data for Science and Technology) recommendations over the past 19 years (**Table 1**, [9]-[15]) from the position of the reached relative uncertainty values. All studies belong to the $CoP_{\mathrm{SI}} \equiv LMT\theta I$. In none of the current experiments of the calculation of the m_p value has the prospective interval been declared, in which its true

Table 1. Proton mass and achieved relative uncertainty.

No	Year	Proton mass × 10^{27}, kg	Achieved relative uncertainty	References
1	1999	1.672621**7161**	4.6×10^{-7}	[9]
2	2005	1.672621**7129**	1.7×10^{-7}	[10]
3	2008	1.672621**6378**	5×10^{-8}	[11]
4	2008	1.672621**7162**	1.8×10^{-10}	[12]
5	2012	1.672621**7777**	4.4×10^{-8}	[13]
6	2014	1.672621**8982**	1.2×10^{-8}	[14]
7	2017	1.672621**7154**	3.2×10^{-11}	[15]

value can be placed. In other words, the exact trace of the placement of m_p is lost somewhere. Therefore, in order to apply our stated approach, as a possible measurement interval of m_p, we choose the difference of its value reached by the experimental results of two projects: $m_{p\min}=1.67262163783\times10^{-27}$ kg [11] and $m_{p\max}=1.67262189821\times10^{-27}$ kg [14]. Then, the possible observed range S_p of m_p variations equals

$$\begin{aligned} S_p &= m_{p\max}-m_{p\min}=1.67262189821\times10^{-27} \\ &\quad -1.67262163783\times10^{-27}=2.6\times10^{-34}\,(\mathrm{kg}) \end{aligned} \tag{13}$$

It is seen from the data given in **Table 1** that there was not dramatic improvement of the measurement accuracy of m_p during the last 18 years by view of the relative uncertainty, except [15]. It differs sharply from other calculated values of relative uncertainty. The question of reliability is key, since the refinement of the values of fundamental constants by innovative methods is extremely vulnerable [2]. Although specialists are highly qualified and use the latest technologies, the lack of accumulated experience in pioneering research affects and we need to wait new experiments.

It is obvious that the spread in the magnitude of the measured m_p is significant. In addition, the truthful and precise value of m_p is not known at the moment. Therefore, scientists of CODATA calculate and declare each 2 years the recommended value of the relative uncertainty, by which, in the future, it will be possible to achieve the true-target value of m_p.

We can argue about the order of the desired value of the relative uncertainty of $CoP_{\mathrm{SI}}\equiv LMT\Theta I$ that is usually used for measurements of the proton mass. For this purpose, we take into account $(\varepsilon_{\min})_{LMT\theta I}=0.222$, $S_p=2.6\times10^{-34}$ kg. Then, the lowest possible absolute uncertainty for $CoP_{\mathrm{SI}}\equiv LMT\Theta I$ equals

$$\left(\Delta_{\min}\right)_{LMT\theta I}=\left(\varepsilon_{\min}\right)_{LMT\theta I}\times S_p=0.222\times2.6\times10^{-34}=0.5772\times10^{-34}\,(\mathrm{kg}) \tag{14}$$

In this case, the lowest possible relative uncertainty $(r_{\min})_{LMT\theta I}$ for $CoP_{\mathrm{SI}}\equiv LMT\Theta I$ is as follows:

$$\begin{aligned} \left(r_{\min}\right)_{LMT\theta I} &= \left(\Delta_{\min}\right)_{LMT\theta I}\Big/\left(\left(m_{p\max}+m_{p\min}\right)/2\right) \\ &= 0.5772\times10^{-34}\Big/\left(1.672621768\times10^{-27}\right)=3.4\times10^{-8} \end{aligned} \tag{15}$$

This value is in excellent agreement with the recommendations mentioned in [13] (4.4×10^{-8}), and can be used for the new definition of the Kelvin and a significant revision of the International System of Units.

3.2. Avogadro Number Na

We performed an analogous procedure for analyzing the results of measurements of the Avogadro number over the past 15 years ([11] [13] [14] [16]-[20]). The data are summarized in **Table 2**.

All studies belong to the $CoP_{SI} \equiv LMTF$. In order to verify the desired value of the relative uncertainty $(r_{\min})_{LMTF}$ of $CoP_{SI} \equiv LMTF$ and taking into account (1) (4) (5) (7) (8), we get the following:

$$\left(z'-\beta'\right)_{LMTF}=\left(e_l\times e_m\times e_t\times e_f-1\right)/2-4=\left(7\times3\times9\times3-1\right)/2-4=279 \tag{16}$$

$$\left(z''-\beta''\right)_{LMTF}=\left(z'-\beta'\right)^2/\mu_{SI}=279^2/38265\approx2 \tag{17}$$

$$\begin{aligned}\left(\varepsilon_{\min}\right)_{LMTF}&=\Delta_{pmm}/S=\left[\left(z'-\beta'\right)/\mu_{SI}-\left(z''-\beta''\right)/\left(z'-\beta'\right)\right]\\&=279/38265+2/279=0.0073+0.0073=0.0146\end{aligned} \tag{18}$$

$$N_{a\max}=6.022141793\times10^{23}\left(\mathrm{mol}^{-1}\right)\ [14],$$

$$N_{a\min}=6.022133900\times10^{23}\left(\mathrm{mol}^{-1}\right)\ [11],$$

$$S_{Na}=\left(N_{a\max}-N_{a\min}\right)=7.9\times10^{17}\left(\mathrm{mol}^{-1}\right) \tag{19}$$

$$\begin{aligned}\left(\Delta_{\min}\right)_{LMTF}&=\left(\varepsilon_{\min}\right)_{LMTF}\times\left(N_{a\max}-N_{a\min}\right)\\&=0.0146\times7.9\times10^{17}=0.1153\times10^{17}\left(\mathrm{mol}^{-1}\right)\end{aligned} \tag{20}$$

$$\begin{aligned}\left(r_{\min}\right)_{LMTF}&=\left(\Delta_{\min}\right)_{LMTF}/\left(\left(N_{a\max}+N_{a\min}\right)/2\right)\\&=0.1153\times10^{17}/\left(6.0221378465\times10^{23}\right)=1.9\times10^{-8}\end{aligned} \tag{21}$$

The value 1.9×10^{-8} is in excellent agreement with the recommendations mentioned in [20] (2×10^{-8}), and can be used for a significant revision of the International System of Units.

It is necessary to note the fundamental difference between the described method and the CODATA method for determining the recommended value of the

Table 2. Avogadro number and achieved relative uncertainty.

N	Year	Value of $N_a \times 10^{-23}$, mol^{-1}	Achieved relative uncertainty $\times 10^8$	References
1	2001	6.022133900	46	[16]
2	2003	6.022135300	34	[17]
3	2008	6.0221417930	5	[11]
4	2011	6.02214082(18)	3	[18]
5	2011	6.02214078(18)	3	[19]
6	2012	6.02214129(27)	4.4	[13]
7	2014	6.022140857(74)	1.2	[14]
8	2015	6.02214076(12)	2	[20]

relative uncertainty of a fundamental physical constant. Within the framework of the CODATA concept, a detailed discussion of the input data and the justification and construction of tables of values sufficient for the direct use of relative uncertainty are conducted using modern advanced statistical methods and powerful computers. This, in turn, allows you to check the self-consistency of input data and output sets of values. However, at each stage of data processing, an expert conclusion based on intuition, accumulated knowledge and the life experience of scientists is also used. Within the framework of the presented approach, a theoretical and informational justification is carried out to calculate the relative uncertainty. A detailed description of the data and the processing procedures used do not require considerable time. This is a reason for the wide implementation of the μ_{SI}-hypothesis, the concept of a system of primary variables for analyzing existing experimental data on the measurement of fundamental physical constants.

4. Conclusions

The information approach for calculating the uncertainty of the model of a physical phenomenon or technological process is very promising in those areas of science and technology where it is required to predict the result of an experiment or to calculate a given basic parameter with very high accuracy, for example, reliability of an atomic power station, seismic stability of buildings, strength of a submarine's hull, thermal resistance of spacecraft's casing, measurement of fundamental physical constants and so on.

The 'new angle' is to apply an information approach to the problems associated with the origin of the choice of the most applicable value of the recommended relative uncertainty. We think that this is perhaps the only tool that does this strictly theoretically.

This is in a sense a little bit more basic as Heisenberg tenet declares. The fundamental limit places severe restrictions on the micro-physics.

We hope that the implementation of the information approach will be recreated in real experiments, possibly using appropriate test benches and various groups of variables. But even if finely tuned fixed points can be provided in the laboratory, we view the presented method as 'necessary but not sufficient', because, at this moment, it cannot accurately indicate the true set of specific variables to achieve a minimum comparative uncertainty.

In our view, the ability of our approach to respond to information quantity embedded in a model is key that can get up and walk away from the subjective environment consisting from accumulated knowledge, life experience and intuition of scientists.

Even if the information approach is on the right track about experimental physics and technology, we want and will develop more detailed information, such as a theory about which primitive 'proto-bricks' are at the core of the system of primary variables, and how this system can be extended and modified, close to the surrounding natural perfection or chaos.

References

[1] Taylor, B.N., Langenberg, D.N. and Parker, W.H. (1970) The Fundamental Physical Constants. *Scientific American*, **223**, 62-78. https://goo.gl/jZ55TD

[2] Karshenboim, S.G. (2013) Progress in the Accuracy of the Fundamental Physical Constants: 2010 CODATA Recommended Values. *UFN*, **183**, 935-962. http://www.mathnet.ru/links/fb8357c7049e641ede305b7d35c7578f/ufn4438.pdf

[3] Mohr, P.J., Taylor, B.N. and Newell D.B. (2016) CODATA Recommended Values of the Fundamental Physical Constants: 2014. *Reviews of Modern Physics*, **88**, 1-73. http://ws680.nist.gov/publication/get_pdf.cfm?pub_id=920687

[4] Menin, B. (2017) Universal Metric for the Assessing the Magnitude of the Uncertainty in the Measurement of Fundamental Physical Constants. *Journal of Applied Mathematics and Physics*, **5**, 365-385. http://file.scirp.org/pdf/JAMP_2017021711410991.pdf

[5] NIST Special Publication 330 (SP330), the International System of Units (SI), 2008, 1-97. http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication330e2008.pdf

[6] Sonin, A.A. (2001) The Physical Basis of Dimensional Analysis. 2nd Edition, Department of Mechanical Engineering, MIT, Cambridge, 1-57. http://web.mit.edu/2.25/www/pdf/DA_unified.pdf

[7] Brillouin, L. (1964) Scientific Uncertainty and Information. Academic Press, New York, 1-155. https://goo.gl/JbPFvF

[8] Burgin, M. (2009) Theory of Information—Fundamentality, Diversity and Unification. World Scientific Publishing Co., USA 2009, ch. 3, 255-300.

[9] Van Dyck, R.S., Farnham, Jr. D.L., Zafonte, S.L. and Schwinberg, P.B. (1999) High precision Penning Trap Mass Spectroscopy and a New Measurement of the Proton's "Atomic Mass". In: Dubin D.E. and Schneider, D., Eds., *Trapped Charged Particles and Fundamental Physics*, *AIP Conference Proceedings* 457 *AIP*, Woodbury, NY, **101**, 2-11. https://goo.gl/VFxkVu

[10] CODATA Recommended Values of the Fundamental Physical Constants: 2002, 1-107. https://goo.gl/ZJsZQL

[11] Mohr, P.J., Taylor, B.N. and Newell, D.B. (2008) CODATA Recommended Values of the Fundamental Physical Constants: 2006. *Reviews of Modern Physics*, **80**, 1-98. https://goo.gl/h5Sc1P

[12] Solders, A., Bergström, I., Nagy, Sz., Suhonen, M. and Schuch, R. (2008) Determination of the Proton Mass from a Measurement of the Cyclotron Frequencies of D^+ and H_2^+ in a Penning Trap. *Physical Review Letters A*, **78**, Article ID: 012514. https://goo.gl/5CXswV

[13] Mohr, P.J., Taylor, B.N. and Newell, D.B. (2012) CODATA Recommended Values of the Fundamental Physical Constants: 2010. *Journal of Physical and Chemical Reference Data*, **41**, Article ID: 043109. https://goo.gl/tqwqoy

[14] CODATA Recommended Values of the Fundamental Physical Constants: 2014. https://codata.org/blog/?p=451

[15] Heiße, F., Köhler-Langes, F., Rau, S., Hou, J., Junck, S., Kracke, A., Mooser, A., Quint, W., Ulmer, S., Werth, G., Blaum, K. and Sturm, S. (2017) High-Precision Measurement of the Proton's Atomic Mass. *Physical Review Letters*, **119**, Article ID: 033001. https://goo.gl/iRozRW

[16] De Bievre, P., Valkiers, S., Kessel, R., Taylor, P.D.P., Becker, P., Bettin, H., Peuto, A.,

Pettorruso, S., Fujii, K., Waseda, A., Tanaka, M., Deslattes, R.D., Peiser, H.S. and Kenny, M.J. (2001) A Reassessment of the Molar Volume of Silicon and of the Avogadro Constant. *IEEE Transactions on Instrumentation and Measurement*, **50**, 593-597. https://goo.gl/hCJkce

[17] Becker, P., Bettin, H., Danzebrink, H.-U., Gläser, M., Kuetgens, U., Nicolaus, A., Schiel, D., De Bièvre, P., Valkiers, S. and Taylor, P. (2003) Determination of the Avogadro Constant via the Silicon Route. *Metrologia*, **40**, 271-287. https://goo.gl/vp88c5

[18] (2011) International Avogadro Project. *Physical Review Letters*, **106**, Article ID: 030801. https://goo.gl/33E6Z3

[19] Andreas, B., Azuma, Y., Bartl, G., Becker, P., Bettin, H., Borys, M., Busch, I., Gray, M., Fuchs, P., Fujii, K., Fujimoto, H., Kessler, E., Krumrey, M., Kuetgens, U., Kuramoto, N., Mana, G., Manson, P., Massa, E. and Mizushima, S. (2011) Determination of the Avogadro Constant by Counting the Atoms in a ^{28}Si Crystal. *Physical Review Letters*, **106**, Article ID: 030801. https://goo.gl/4g7nkr

[20] Azuma, Y., *et al*. (2015) Improved Measurement Results for the Avogadro Constant using a ^{28}Si-Enriched Crystal. *Metrologia*, **52**, 360-375. https://goo.gl/PURKaG

Scientific Research Publishing

Journal of Applied Mathematics and Physics, 2017, 5, 2172-2182
http://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

A New Job Shop Heuristic Algorithm for Machine Scheduling Problems

Maryam Ehsaei[1], Duc T. Nguyen[2]

[1]Civil and Environmental Engineering Department, Old Dominion University, Norfolk, VA, USA
[2]Civil and Environmental Engineering (CEE) and Modeling, Simulation and Visualization Engineering (MSVE) Departments, Old Dominion University, Norfolk, VA, USA
Email: mehsa001@odu.edu, dnguyen@odu.edu

How to cite this paper: Ehsaei, M. and Nguyen, D.T. (2017) A New Job Shop Heuristic Algorithm for Machine Scheduling Problems. *Journal of Applied Mathematics and Physics*, **5**, 2172-2182.
https://doi.org/10.4236/jamp.2017.511177

Received: October 17, 2017
Accepted: November 10, 2017
Published: November 13, 2017

Abstract

The purpose of this research is to present a straightforward and relatively efficient method for solving scheduling problems. A new heuristic algorithm, with the objective of minimizing the makespan, is developed and presented in this paper for job shop scheduling problems (JSP). This method determines jobs' orders for each machine. The assessment is based on the combination of dispatching rules e.g. the "Shortest Processing Time" of each operation, the "Earliest Due Date" of each job, the "Least Tardiness" of the operations in each sequence and the "First come First Serve" idea. Also, unlike most of the heuristic algorithms, due date for each job, prescribed by the user, is considered in finding the optimum schedule. A multitude of JSP problems with different features are scheduled based on this proposed algorithm. The models are also solved with Shifting Bottleneck algorithm, known as one of the most common and reliable heuristic methods. The result of comparison between the outcomes shows that when the number of jobs are less than or equal to the number of machines, the proposed algorithm concludes smaller, and better, makespan in a significantly lower computational time, which shows the superiority of the suggested algorithm. In addition, for a category when the number of jobs are greater than the number of machines, the suggested algorithm generates more efficient results when the ratio of the number of jobs to the number of machines is less than 2.1. However, in this category for the mentioned ratio to be higher than 2.1, the smaller makespan could be generated by either of the methods, and the results do not follow any particular trend, hence, no general conclusions can be made for this case.

Keywords

Heuristic Algorithm, Job Shop Scheduling, Shifting Bottleneck, Makespan

1. Introduction

Job shop scheduling (JSP) has been one of the most critical subjects in optimization and applied mathematics in the past few decades. Its vast applicability in the industry and all economic domains, and on the other hand its complexity, specifically for large-scale problems, make this topic very critical [1] [2]. JSP is an NP-hard problem due to its computational complicacy [3] [4]. Based on the scheduling literature, a relatively small problem consists of 10 jobs and 10 machines, proposed by Muth and Thompson [5], remained unsolved for more than a quarter of a century. Also, the fact that a scheduling problem included 15 jobs and 15 machines is considered unsolvable with the exact method nowadays, clearly shows the sophistication of this kind of problems [6] [7].

In job shop scheduling problems "n" jobs are needed to be processed in "m" machines. Each job includes some operations, each of which are required to be done by a particular machine. Each machine only can process one job at a time and cannot be interrupted [8]. The order of jobs in each machine is calculated by minimizing a specific character. In this paper, the completion time of all jobs, which is called makespan, is the objective [9].

Lots of algorithms and procedures have been proposed for efficiently scheduling JSP, which are divided into three major groups: 1) the exact algorithms; such as the one proposed by Giffler and Thompson (1960), and branch and bound by Lageweg, Lenstra and Rinnooy Kan (1977) [7]. This group results are surely optimum, but they are too time-consuming to reach, 2) heuristic procedures; such as Palmer, Johnson and Shifting Bottleneck algorithms. They are not always derived optimum answer; some gives answers close to optimum. Comparing to two other groups, the computational time of algorithms in this category is relatively small, 3) meta-heuristic Algorithms; such as genetic algorithm, and SA and TS algorithms by Fattahi, Mehrabad and Jolai [10]. This group does not guaranty the optimal answer, but present better results comparing to the second group.

A new heuristic algorithm is presented in this paper for optimally scheduling JSP. This method is the result of combining different dispatching rules, so, its implementation is justly straightforward. It is also worth to mention that tardiness is the difference between the completion time of each job (or operation) and the job's related due date (or relative due date); in other words, the time that a job takes to be completed after its due date arrives is called tardiness.

Due to the complexity of developing a reliable and efficient algorithm, some heuristic algorithms just consider operations' processing times, ignoring the jobs' due dates and others are designed based on the due dates and ignored the processing time. However, the new algorithm presented in this paper takes both factors into account.

Furthermore, the new method can schedule both job shop and flow shop problems. In flow shop problems, all jobs should be operated in all machines in the same order. However, job shop scheduling is more general, and the sequence

of machines in each job may not follow a specific order [11]. So, unlike algorithms like Johnson and NEH, which are only applicable and useful in flow shop problems, the proposed algorithm is capable of handling the job shop schedule as well. The new algorithm detail is described in the following section.

Among all existing heuristic algorithms, Shifting Bottleneck algorithm is one of the most well-known and reliable ones. Therefore, to evaluate the reliability of the new algorithm, its results have been compared to the Shifting Bottleneck outcomes. Scheduling models for comparison of two algorithms are JSP problems. The comparison of outcomes is reported in the Results section. Finally, based on the observed results, conclusions will be made.

2. Materials and Methods

2.1. Proposed Algorithm Description

The proposed algorithm has been developed based on some primary dispatching rules including "earliest due date" of each job, "shortest processing time" of each operation, "least tardiness" of operations in each sequence and "first come first serve" idea. The theory behind using these simple rules is to create a heuristic method with a straightforward procedure to apply to JSP problems, concluding to acceptable results. The efficiency of the results comparing to other heuristic algorithms, is also contemplated. Besides, the proposed method is designed in such a way that the due dates' values required to be specified by the user. This advantage can equip users to affect their tendency in using specific due dates in the scheduling problems.

Moreover, the sequence of operations and their related processing times for each job are needed to be imported. Besides, the user is required to clarify the machine used for each operation. The procedures' details of the new algorithm are being described on a small example to explain the steps thoroughly.

Table 1 shows a JSP example, consisting of 3 jobs and 3 machines. The processing times and due dates are also specified. Each job is defined in each row of the table. Each job consists of some operations, required to be done by a particular machine, which each of the operations has a deterministic processing time specified in the table. For instance, Job 1 has three operations; first operation processing time is 7 minutes (instead of minutes any unit of time can be used), and it needs to be done by machine 1 (M1).

1) **Step 1**: The minimum value of all the provided due dates is selected and subtracted from the rest of the due dates; the result values are called relative due

Table 1. Job shop scheduling example.

	1st Operation	2nd Operation	3rd Operation	Due Date
Job 1	7 (M1)	8 (M3)	10 (M2)	26
Job 2	6 (M3)	4 (M1)	12 (M2)	26
Job 3	8 (M2)	8 (M1)	7 (M3)	27

dates. There are two reasons for doing so: 1st reason is that this way it is not necessary to deal with values of due dates, which might be significantly large. And the second reason is that urgency of the jobs can be compared more clearly. The results of the subtraction for the mentioned example in Table 1, is elaborated in Table 2.

2) **Step 2**: The new algorithm is task wise, which means the priority is the first to be operated task in each job, which is derived from the primary rule "first come first serve". In the mentioned example the priority is the column by the title of "1st Operation" (second column of the table). These are tasks, which are needed to be completed in their related jobs, for jobs to be able to go to the next stage. Based on the explained idea, the first tasks are considered first. If there is no ready time for an operation (best situation), its completion time will be equal to their related processing time. Therefore, for each operation, the tardiness will be equivalent to the operation's relative due date, subtracted by the related completion time or the operation processing time. The procedure is presented in Table 3. The order of implementation will be based on the least tardiness such that, the operation with the lowest tardiness takes place first, and the operation with the most massive tardiness will take place at last. If there is a tie-breaker, the algorithm chooses the operation based on the job orders; for example in the mentioned case, between the 1st operation of job 1 and 1st operation of job 3, the priority for the algorithm is job 1.

For clarification of the proposed method, execution of each step is demonstrated in a diagram, similar to Gantt chart. Gantt chart is a bar chart used to demonstrate a project schedule, which in job shop scheduling problems it usually illustrates the order of jobs in each machine. However, in the diagrams used in this paper, called modified Gantt chart, the charts show the sequence of machines in each job (with consideration of their order and waiting time). Figure 1

Table 2. New algorithm execution: step 1 (subtraction the value of the minimum due date from other ones).

	Job 1	Job 2	Job 3
Related Due Dates	26	26	27
Minimum Due Date Value	26	26	26
Result of Subtraction (Relative Due Dates)	0	0	1

Table 3. New algorithm execution: step 2 (subtraction of the relative due dates from the completion time of each job till the end of the 1st Operations).

	(Job 1, M1)	(Job 2, M3)	(Job 3, M2)
The Least Completion Time of Each Job till the End of the 1st Operation	7	6	8
Related Relative Due Dates	0	0	1
Result (Tardiness)	7	6	7
Order of Implementation	2	1	3

Figure 1. Implementation of the 1st operations on modified Gantt chart.

shows the implementation of step 2.

3) **Step 3:** The procedure in this step is the same as step two, just the considered completion times are different. They are equal to the processing time of each job from the start of the job until the end of the current operation. For instance, the completion time of job 1 after the end of task 2 is equal to the summation of 7 and 8 (7 + 8 = 15). **Table 4** displays the detail of this step on the mentioned example. Also, the implementation of this step on modified Gantt chart is shown in **Figure 2**.

4) **Step 4:** This step is also similar to the two previous ones, by considering the completion time of each job up to the end of the current operation. The procedure detail and implementation on modified Gantt chart for this stage are described in **Table 5** and **Figure 3** consequently. Based on the modified Gantt chart it is clear that the makespan is 33.

It is also worth to mention that the completion time of all jobs (makespan) is equal to the completion time of all machines. As it is mentioned earlier the implementation of the operations should be in a way to avoid confliction between the machines; in other words, the ready time for each operation and the processing time of the machine before starting the current operation are needed to be considered. So, for the mentioned instance, the sequence of jobs in each machine is derived as follow:

a) M1: Job 1 – Job 2 – Job 3

b) M2: Job 3 – Job 2 – Job 1

c) M3: Job 2 – Job 1 – Job 3

With the same procedure implemented to the mentioned example, any number of jobs and machines can be scheduled by the proposed algorithm. The algorithm is also presented in the form of the flowchart in **Figure 4**.

For a new algorithm to be evaluated, it is necessary to be compared with a well-known and reliable existing algorithm in various models. One of the most popular and acceptable heuristic algorithms, which is known to be superior among heuristic algorithms for JSP, is Shifting Bottleneck algorithm proposed by Adams, Egon and Zawack [8].

2.2. Shifting Bottleneck Algorithm

This algorithm does not guaranty the optimum answer, but its results are so

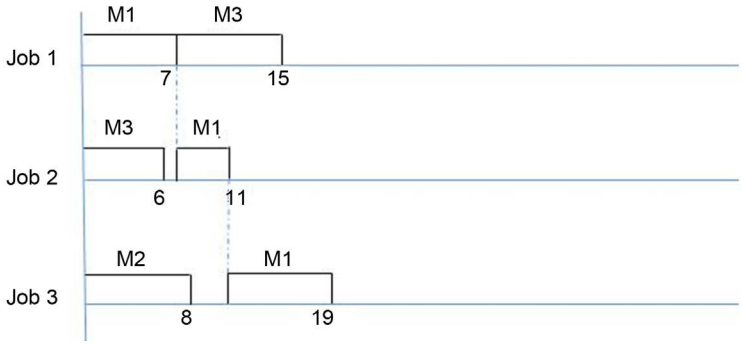


Figure 2. Implementation of the 2nd operations on modified Gantt chart.

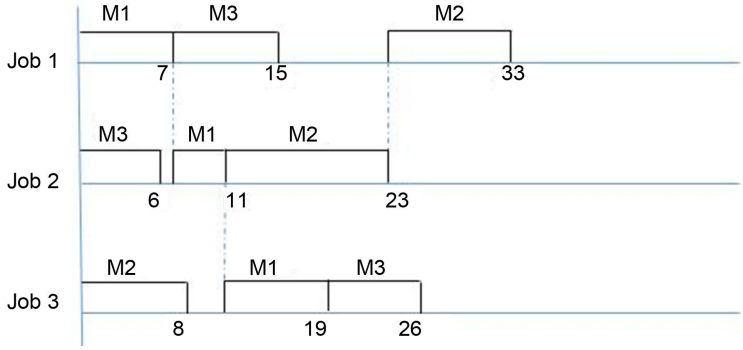


Figure 3. Implementation of the 3rd operations on modified Gantt chart.

Table 4. New algorithm execution: step 3 (subtraction of the relative due dates from the completion time of each job till the end of the 2nd Operations).

	(Job 1, M3)	(Job 2, M1)	(Job 3, M1)
The Least Completion Time of Each Job till the End of the 2nd Operations	7 + 8 = 15	6 + 4 = 10	8 + 8 = 16
Related Relative Due Dates	0	0	1
Result (Tardiness)	15	10	15
Order of Implementation	[2]	[1]	[3]

promising, especially on benchmark problem sets from the literature, such that lots of researchers like, Dauzere-Peres and Lasserre (1993) and Schutten (1995), consider it as a fundamental algorithm for their work. However, the efficiency of this algorithm may be reduced by increament of the ratio of number of machines per number of jobs [12].

Shifting bottleneck algorithm approach is machine wise. It is solved a one-machine scheduling problem at a time for all not sequenced machines. Then based on the rank of scheduled machines, it sets the job sequence for the highest rank machine and reorders the job sequence for others. This method is chosen for comparison with the new proposed algorithm in this paper.

3. Results

The proposed and Shifting Bottleneck algorithms have been coded in MATLAB

Table 5. New algorithm execution: step 4 (subtraction of the relative due dates from the completion time of each job till the end of the 3rd operations).

	(Job 1, M3)	(Job 2, M1)	(Job 3, M1)
The Least Completion Time of Each Job till the End of the 3rd Operations	7 + 8 + 10 = 25	6 + 4 +12 = 22	8 + 8 + 7 = 23
Related Relative Due Dates	0	0	1
Result (Tardiness)	25	22	22
Order of Implementation	3	1	2

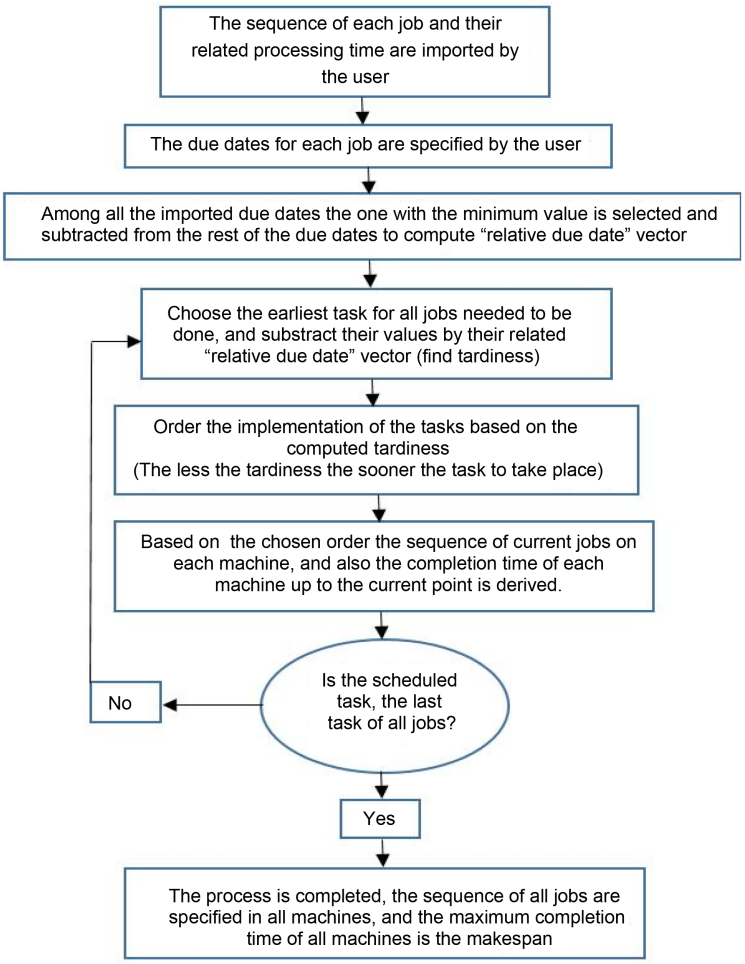


Figure 4. Flow chart of the proposed algorithm.

software and their results for different problems have been compared. The models considered for comparison could be divided to three major groups; those with equal number of jobs and machines, the ones with greater number of jobs than machines and visa versa. For each category, problems with different sizes, small, medium and large, have been examined. Each problem is solved 27 times with 27 different sets of randomly generated due dates. The results for all the 27 sets, derived by the proposed method and Shifting Bottleneck algorithm, are compared. The comparison results are presented in three different tables. Table

6 is allocated to the models with the same number of jobs and machines. Table 7 is associated with the models with the higher number of machines. Finally, the models with the higher number of jobs are presented in Table 8.

Since the proposed algorithm gets advantage of the assigned due dates in the calculation of the makespan, changing them may result in varying the outcomes consequently. However, the Shifting Bottleneck algorithm does not consider the due dates provided by users, so its results will remain unchanged. In each model 27 different cases, with a set of randomly generated due dates, have been scrutinized. For each model, the results derived by the new algorithm in all the cases have been compared to the ones derived from the Shifting Bottleneck algorithm,

Table 6. Comparison of the models with equal number of jobs and machines.

# Jobs* #Machines (Models)	Percentage of the Times that the New Algorithm Gives a Lower Makespan in 27 Iterations (%)	Makespan derived by Shifting Bottleneck Algorithm	Makespan derived by the New Algorithm (Average of 27 Iteration Results)	Average Computational Time for Shifting Bottleneck Algorithm (Second)	Average Computational Time for New Algorithm (Second)
3 * 3	88%	37	36.37037	0.061752	0.034239
10 * 10	100%	182	123.4615	0.246965	0.040610
18 * 18	100%	1489	1134.778	1.401769	0.728196
26 * 26	100%	2659	1837.185	2.556270	0.762216
35 * 35	100%	1753	1194.111	5.444633	0.842263
60 * 60	100%	3465	2605.37	37.439974	0.926187
73 * 73	100%	4143	3254.704	79.510854	1.146496
80 * 80	100%	8982	7040.077	100.277707	1.230148
100 * 100	100%	4143	3378.296	282.844018	1.662978
140 * 140	100%	6063	4822	1129.604031	4.566296

Table 7. Comparison of the models with higher number of machines.

# Jobs *#Machines (Models)	*Percentage of the Times that the New Algorithm Gives a Lower Makespan in 27 Iterations (%)*	*Makespan derived by Shifting Bottleneck Algorithm*	*Makespan derived by the New Algorithm (Average of 27 Iteration Results)*	*Average Computational Time for Shifting Bottleneck Algorithm (Second)*	*Average Computational Time for New Algorithm (Second)*
3 * 5	100%	383	231	0.754168	0.700745
4 * 10	100%	806	426.5926	0.984952	0.722524
12 * 17	100%	1423	896.5926	1.362245	0.723708
15 * 35	100%	4522	2318.704	2.888122	0.994421
12 * 60	100%	4492	2260.741	8.422953	0.706465
40 * 100	100%	12,593	7095.148	95.288633	0.994421
50 * 70	100%	11,040	7723.593	42.723094	0.922766
60 * 73	100%	7036	5024.593	66.844959	1.028324
55 * 110	100%	8741	5331.481	207.232346	1.237143
200 * 222	100%	17,837	13994.15	6768.341210	22.180659

Table 8. Comparison of the models with higher number of jobs.

#Jobs* #Machines (Models)	#Jobs/# Machines	Percentage of the Times that the New Algorithm Gives a Lower Makespan in 27 Iterations (%)	Makespan derived by Shifting Bottleneck Algorithm	Makespan derived by the New Algorithm (Average of 27 Iteration Results)	Average Computational Time for Shifting Bottleneck Algorithm (Second)	Average Computational Time for New Algorithm (Second)
7 * 5	1.4	100%	196	176.3333	0.790561	0.721444
40 * 30	1.33	100%	4153	3461.074	5.721949	0.734871
200 * 100	2.0	93%	28710	28070.37	540.440321	3.694662
20 * 10	2.0	77%	274	267.7778	0.441940	0.040610
300 * 142	2.11	81%	42,551	41999.89	2450.178286	17.044220
11 * 5	2.2	0%	1104	1276.111	0.813793	0.717175
40 * 17	2.35	0%	4405	4785.222	1.725297	0.775433
50 * 20	2.5	3%	2965	3101.222	2.459510	0.711970
30 * 10	3.0	0%	1975	2316.667	0.441940	0.040610
35 * 7	5	0%	1152	1344.481	0.954360	0.717637
13 * 2	6.5	100%	211	211	0.765863	0.712924
18 * 3	6	100%	378	378	0.837103	0.770999
26 * 4	6.5	63%	388	392.2963	0.901294	0.743331

but to save the space, for each model, just the average of the results of 27 states is presented here. The percentage of the number of the times that the new algorithm produces lower makespan is also reported here. Therefore, for each group of problems two algorithms have been compared 270 times or more. The observation from the compared models is noted in this section. It is worth to mention that all the considered processing times are chosen randomly.

As it is shown in Table 6, the new algorithm produces lower makespan in almost all iterations. Moreover, it is observed that by increasing the size of the problem in this category (number of jobs and machines) the difference, between the computational time generated by the two algorithms becomes significant. Growing the problem size is also concluded to the variation of makespans, produced by two methods for an identical problem, to be increased considerably.

When the number of machines is higher than the number of jobs in all cases the derived makesspan by the proposed method is lower than the identical ones derived from the Shifting Bottleneck algorithm. The difference between the computational time is also increased by the increment of the size of the problem. The details are presented in Table 7.

Eventually, for the category, in which the number of jobs is higher than the number of machines, there is no consistency observed in the results. Extensive testing problems have been scheduled in this case, which for the sake of saving the space only, 350 selected models have been presented in this paper. It is noted that when the ratio of the number of jobs to the number of machines, is less than or equal to 2.1, in almost all circumstances, the proposed algorithm produces

lower makespan, in a smaller computational time. However, when the mentioned ratio is higher than 2.2, the observed results do not have solidarity, which means in some situations the proposed algorithm, and in some other cases Shifting Bottleneck algorithm, generates lower makespan.

4. Conclusions

A new algorithm for scheduling job shop problems has been proposed in this article. This algorithm is based on the combination of some primary dispatching rules like the "Shortest Processing Time" of each operation, the "Earliest Due Date" of each job, the "Least Tardiness" of the operations in each sequence and the "First come First Serve" idea. Straightforward procedures and ease of implementation are two of the most significant advantages of the proposed method. The flowchart and the execution steps have been described in previous sections in detail.

For numerical evaluation and verification of the suggested method, its produced results have been compared to the outcomes derived by the Shifting Bottleneck algorithm for enormous problems. Results comparison is presented in this paper for more than 30 models with almost 900 different iterations (using random due dates).

Based on the compared models, it is observed that when the number of jobs is less than or equal to the number of machines, the proposed algorithm produces lower makespan in a significantly smaller computational time, which shows the superiority of the proposed method. Also, the larger the size of the problem, the more the difference between the identical makespans generated by two methods. Besides, in the mentioned categories, in the models with a larger size for an identical problem, the computational time by the new method is remarkably less than the computational time by the Shifting Bottleneck algorithm.

It is also observed that when the ratio of the number of jobs to the number of machines, is less than 2.1, the proposed algorithm produces lower makespan in a smaller computational time. But, when the mentioned ratio becomes greater than 2.1, the smaller makespan could be generated by either of the methods, and the results do not follow any particular trend, hence, no general conclusions can be made for this case.

It is also perceived that for all the tested cases, the computational period of the proposed method is lower than the computational time of the Shifting Bottleneck algorithm.

Acknowledgements

Authors would like to thank Old Dominion University for providing resources to facilitate this project.

References

[1] Bagheri, A. and Zandieh, M. (2003) An Artificial Immune Algorithm for the Flexible Job-Shop Scheduling Problem. *Future Generation Computer Systems*, **26**, 13-20.

[2] Nguyen, S., Zhang, M., Johnston, M. and Tan, K.C. (2012) Evolving Reusable Operation-Based Due-Date Assignment Models for Job Shop Scheduling with Genetic Programming. *European Conference on Genetic Programming*, Malaga, 11-13 April 2012, 121-133. https://doi.org/10.1007/978-3-642-29139-5_11

[3] Zhang, R. and Wu, C. (2011) An Artificial Bee Colony Algorithm for the Job Shop Scheduling Problem with Random Processing Times. *Entropy*, **13**, 1708-1729. https://doi.org/10.3390/e13091708

[4] Brucker, P., Jurisch, B. and Sievers, B. (1994) A Branch and Bound Algorithm for the Job-Shop Scheduling Problem. *Discrete Applied Mathematics*, **49**, 107-127. https://doi.org/10.1016/0166-218X(94)90204-6

[5] Muth, J.F. and Thompson, G.L. (1963) Industrial Scheduling. Prentice-Hall, Englewood Cliffs, N.J.

[6] Zhang, C.Y., Li, P., Rao, Y. and Guan, Z. (2008) A Very Fast TS/SA Algorithm for the Job Shop Scheduling Problem. *Computers & Operations Research*, **35**, 282-294. https://doi.org/10.1016/j.cor.2006.02.024

[7] Lageweg, B.J., Lenstra, J.K. and Rinnooy Kan, A.H.G. (1978) A General Bounding Scheme for the Permutation Flow-Shop Problem. *OPNS RES*, **26**, 53-67. https://doi.org/10.1287/opre.26.1.53

[8] Adams, J., Egon, B. and Zawack, D. (1988) The Shifting Bottleneck Procedure for Job Shop Scheduling. *Management Science*, **34**, 391-401. https://doi.org/10.1287/mnsc.34.3.391

[9] Bhosale, P.P. and Kalshetty, Y.R. (2016) Genetic Algorithm for Job Shop Scheduling. *International Journal of Innovations in Engineering and Technology* (*IJIET*), **7**, 357-361.

[10] Fattahi, P., Mehrabad, M.S. and Jolai, F. (2007) Mathematical Modeling and Heuristic Approaches to Flexible Job Shop Scheduling Problems. *Journal of intelligent manufacturing*, **18**, 331-342. https://doi.org/10.1007/s10845-007-0026-8

[11] Abbas, M., Abbas, A. and Khan, W.A. (2016) Scheduling Job Shop—A Case Study. *IOP Conference Series: Material Science and Engineering*, **146**, 021052. https://doi.org/10.1088/1757-899X/146/1/012052

[12] Demirkol, E., Mehta, S. and Uzsoy, R. (1997) A Computational Study of Shifting Bottleneck Procedures for Shop Scheduling Problems. *Journal of Heuristic*, **3**, 111-137. https://doi.org/10.1023/A:1009627429878

Journal of Applied Mathematics and Physics, 2017, 5, 2183-2195
http://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

A Growth Framework Using the Constant Elasticity of Substitution Model

Prabir Bhattacharya

Department of Computer Science, Morgan State University, Baltimore, MD, USA
Email: prabir_bhattacharya@yahoo.com

How to cite this paper: Bhattacharya, P. (2017) A Growth Framework Using the Constant Elasticity of Substitution Model. *Journal of Applied Mathematics and Physics*, **5**, 2183-2195.
https://doi.org/10.4236/jamp.2017.511178

Received: October 10, 2017
Accepted: November 11, 2017
Published: November 14, 2017

Abstract

Some results in growth theory based on the Cobb-Douglas production function model are generalized when the production function is chosen to be the Constant Elasticity of Substitution (CES) function. Such a generalization is of considerable interest because it is known that the Cobb-Douglas function cannot be used as a suitable model for some production technologies (like the US economy and climate changes). It is shown that in the steady state the growth rate of the output is equal to the Solow residual and that the capital deepening term becomes zero. The CES function is a homogeneous function of degree two and a result is obtained on the wage of a worker using the Euler's theorem.

Keywords

CES Function, Cobb-Douglas Function, Growth Equation, Solow Residual, Factor of Productivity, Capital Deepening Term

1. Introduction

In growth theory, a "production function" is taken to be a mathematical expression that is used to model a production technology with distinct inputs and outputs. Several types of production functions have been proposed in the past based on either empirical or theoretical considerations (for general surveys, see e.g., [1] [2]). For example, the Cobb-Douglas production function ([3]) is widely used as a simple model to study economic growth in spite of some of its limitations (that we shall discuss later).

The CES ("Constant Elasticity of Substitution") function was introduced by Solow ([4]), and later expounded by Arrow *et al.* ([5]) to synthesize several types of production functions. The CES function has been applied extensively to study

economic growth (e.g., [6]-[17]). It has also applications in high energy physics [18]. Many generalizations of the format of the CES function (including the multi-input case) have been proposed, for details see e.g. [6] [7]. The shape of the frontier of the CES function is also of significant interest in economic analysis and its connection with the differential geometry of hyperspaces has been studied. (e.g., [19] [20] [21]).

The *elasticity of substitution* between any two input variables in a production function measures how easily one variable can be substituted for the other variable and it measures the curvature of the isoquant (the concept was first introduced by Hicks [22]). So, an elasticity of substitution equal to 0 indicates no substitution between the input variables can be possible and an elasticity of substitution equal to infinity indicates the perfect substitution. More formally, the elasticity of substitution between two factors of production is an index that measures the percentage of response of the relative marginal products of the two factors to a percentage of change in the ratio of the two quantities. In order to make the paper self-contained, we shall briefly review in Appendix I the formal definition of the elasticity of substitution for the case of a production function with n input variables.

For the Cobb-Douglas function the elasticity of substitution between the input variables is always equal to 1 (for a proof see, e.g., [7]) and this fact restricts its use as a suitable production model in several applications, as claimed by many authors. For example, Antrás [23] has shown that the US economy is not amenable to the elasticity of substitution being taken as 1. Also Werf [24] has shown that it is not suitable to take the Cobb-Douglas function as a production function for modeling climate change policies. Furthermore, Young [25] has shown that the elasticity of substitution for U.S. aggregate and of most industries cannot be equal to 1 and it is estimated to be less than 0.620; thus it follows that the Cobb-Douglas production model (whose elasticity of substitution is fixed to be 1) is not suitable for such applications.

The CES function has a constant elasticity of substitution (as the name suggests) and it can have any pre-determined value as its elasticity of sub-stitution (as we shall show later). Thus, it offers a wider flexibility than the Cobb-Douglas function and is still computationally tractable, as remarked in ([6], p. 54). These reasons have partially motivated us to extend some results of the neoclassical growth theory based on the Cobb-Douglas function by using the more general setting of the CES production function.

We now briefly review some definitions and results.

Definition 1 ([5]) The *Constant Elasticity of Substitution* (CES) production function for the three factors-capital K, labor L and the total factor of productivity F, is given by

$$Y = F\left(\alpha K^{\gamma} + \beta L^{\gamma}\right)^{\frac{1}{\gamma}} \quad \text{with } \alpha + \beta = 1 \tag{1}$$

where Y is the output and K, L, F are smooth functions of time t; α is a

certain constant, called the *share parameter* between the capital and labor; and γ is another constant, called the *substitution parameter*.

The following result indicates that the CES production function is a generalization of the Cobb-Douglas function:

Proposition 1 ([5]) *When* $\gamma \to 0$, *the CES production function* (1) *approaches the Cobb-Douglas production function*

$$Y = FK^{\alpha}L^{\beta} \tag{2}$$

where α, β are constants such that $\alpha + \beta = 1$.

It is known that the elasticity of substitution of the CES production function as defined by (1), is equal to $1/(1-\gamma)$ ([5], p. 230) when $\gamma \neq 1$, and using this result we can easily construct (as indicated in Example 1) an infinite family of CES production functions each of whose members has the same elasticity of substitution equal to any given nonzero number.

Example 1 Suppose we want to construct a CES production function whose elasticity of substitution, ϵ, is equal to, say, 2. Solving the equation

$$1/(1-\gamma) = 2 \tag{3}$$

gives $\gamma = 0.5$. Substituting $\gamma = 0.5$ into (1) and taking F to be any smooth function of t, and by varying α, we get an infinite family of CES production functions given by

$$Y = F\left\{\alpha\sqrt{K} + (1-\alpha)\sqrt{L}\right\}^2 \tag{4}$$

where $0 < \alpha < 1$ and each member of the family has the same elasticity of substitution equal to 2. Similarly, if $\epsilon = 1$, then by solving the equation $1/(1-\gamma) = 1$ we get $\gamma = 0$ and this corresponds to the Cobb-Douglas production function (compare with Proposition 1).

In defining the CES production function, in the form given by (1), many authors take F to be a parameter (e.g., [5], p. 230; [7], p. 397; [26], p. 397). However, we shall consider here a more general model where F is assumed to be a function of time. Such a model would be able to handle some situations that cannot be accommodated by a CES function where F is a parameter. For example, the output of a factory may increase at a time when the production manager is replaced by a more efficient one (*i.e.* when F increases) even when there are no increases in investments in capital and labor. We note that ([6], p. 54) takes F to be a function of time, like us. Furthermore we shall exclude the cases when the output is either identically zero, or a negative number as these cases are not of interest. Thus we make the following assumption:

Assumption 1 *We assume in* (1) *that F is a function of time and that* $F > 0$.

As remarked in ([27], p. 107), the wages and salaries in USA and many other countries form about 70 percent of the national income. Consequently, the value of $\alpha = 0.3$ has been used in ([27], p. 107-109) as the share of the capital for the Cobb-Douglas production function model (2) to estimate the growth rate for a number of countries. However, such an estimate for the growth rate of countries

based solely on a Cobb-Douglas production function model may not be realistic because the Cobb-Douglas production function (2) has the unitary elasticity of substitution and as [23] has shown, it is not suitable to model the US economy with the elasticity of substitution equal to 1 (also, it is not known whether we can realistically assume that the elasticity of substitution is 1 for all the other countries involved in that study). So, it would be of significant interest to estimate the economic growth for various countries using the same data but for a more general setting involving the CES production function model for a range of values of the parameter γ in (1) with $\alpha = 0.3$, and then to estimate an optimal value of γ to fit the data set.

The structure of the rest of the paper is as follows. In Section 2 we obtain a growth equation for the CES production function and define the (generalized) Solow residual and the corresponding capital deepening term. In Section 3 we obtain some bounds for the (generalized) Solow residual and the capital deepening term. In Section 4 we investigate the growth rates corresponding to the CES production function. In Section 5 we investigate the homogeneous property of the CES production function. Section 6 gives our conclusions. Appendix I reviews the definition of the elasticity of substitution, The proofs of all the results are given in the Appendix II.

2. Growth Equation for CES Production Model

In this section, we generalize some results obtained earlier in the setting of the Cobb-Douglas production model (e.g., as in [27], Chap. 5) to the case of the CES production function. First, we shall derive a growth equation corresponding to the CES production function. Continuing with the notation introduced in Definition 1, we now define the variables y and k given by

$$y = Y/L,\ k = K/L \tag{5}$$

where we assume that L is nonzero (for, if $L = 0$, the CES production function takes the simple form $Y = FK$ and that is a special case of the Cobb-Douglas function with $\alpha = 1$ and $\beta = 0$; and so we omit this case). Thus, y and k are well-defined and y represents the output per worker (*i.e.*, per capita output) and k is the capital stock per worker. Also, L cannot be negative because we have not given any interpretation to negative labor. So, we shall make the following assumption:

Assumption 2

We assume that $L > 0$.

We note that ([6], p. 36) also makes a similar assumption.

From (1), dividing both sides of the equation by L, and using (5), we get

$$y = F\left(\alpha k^{\gamma} + \beta\right)^{\frac{1}{\gamma}} \tag{6}$$

Log differentiating both sides of (6) with respect to t and denoting $\dot{y}/y$ by G, we get the following *growth equation* when the CES production function is taken as the production model:

$$G = \frac{\dot{F}}{F} + \frac{\alpha k^{\gamma-1}\dot{k}}{\alpha k^{\gamma} + \beta} \tag{7}$$

where $\cdot$ denotes differentiation with respect t. Both the terms on the right hand side of (7) are well-defined because their denominators are nonzero (F is nonzero by the Assumption 1, and $\left(\alpha k^{\gamma} + \beta\right)$ is nonzero since otherwise it will follow from (6) that $y = 0$ and we shall exclude this trivial case when the output is identically equal to zero); also G is well-defined (since otherwise from the expression $G = \dot{y}/y$, it will follow that $y = 0$).

Taking the limit as $\gamma \to 0$ in (7) and using Proposition 1, we can easily obtain the growth equation corresponding to the Cobb-Douglas production function (2) and the resulting equation matches with the corresponding Equation (5.3) derived in ([27], p. 106).

Definition 2 For the CES production function, the expression

$$G = \dot{y}/y \tag{8}$$

is called the *growth rate of the output per worker*, and the growth rate of the total factor of productivity, $\dot{F}/F$, will be called the *Solow residual* corresponding to the CES production function.

Using (7), the Solow residual can be expressed as:

$$\frac{\dot{F}}{F} = G - D \tag{9}$$

where

$$D = \frac{\alpha k^{\gamma-1}\dot{k}}{\alpha k^{\gamma} + \beta} \tag{10}$$

is called the *capital deepening term* corresponding to the CES production function. From (9), we observe that G, the growth rate of the output per worker, is the sum of two components: (i) the Solow residual, and (ii) the capital deepening component D (we note that a similar observation is made for the Cobb-Douglas production function (2), in ([27], p. 106)).

Next, we consider the form of the growth equation (7) in the steady state.

Proposition 2 (i) *When a steady state of production is reached, the growth rate of the output per worker is equal to the Solow residual, and the capital deepening term is zero. As a partial converse, if the production is not entirely labor intensive, a steady state of production is reached when the growth rate of the output per worker is equal to the Solow residual (or equivalently when the capital deepening term is zero).*

(ii) The total factor of productivity is in a steady state if and only if the growth rate of the output per worker is equal to the capital deepening term.

3. Estimates for Solow Residual

We now obtain some bounds for the capital deepening term D and the Solow residual.

Proposition 3 *For the CES production function given by* (6)*, if the ratio of the share parameters is less than* k^{γ}*, i.e. if*

$$k > (\beta/\alpha)^{1/\gamma} \tag{11}$$

then (i) the capital deepening term is less than the growth rate of capital per worker, and

(ii) the Solow residual is greater than the difference between the growth rates per worker, of the output and the capital.

Corollary 1 (i) *When* $\gamma \to \infty$*, the estimates given in Proposition 3 hold for any* $k > 1$.

(ii) When $\gamma \to 0$, (*i.e.*, when the production function is approaching the Cobb-Douglas function (2), see Proposition 1), the estimates given in Proposition 3 hold for any $\alpha > 1/2$.

We now give some examples to illustrate what happens to the inequality (11) as we progressively increase the value of γ.

Example 2 (i) Suppose, as an illustration, we choose $\alpha = 0.3$ and so $\beta = 0.7$. In **Figure 1**, we plot the values of $u := (\beta/\alpha)^{1/\gamma}$ where the horizontal axis corresponds to γ and the vertical axis corresponds to *u*. We know from Corollary 1 (i) that as γ becomes larger and larger, the value of *u* would tend to 1 and this property is being exhibited in Figure 1. We mention that similar illustrations can also be given by taking other values of α.

(ii) If $\alpha = 0.5$ (that is, when the capital and labor are shared equally in the CES production model), the condition (11) reduces to $k > 1$ and so the estimates described in Proposition 3 hold for any $k > 1$ and for *any* value of γ. In other words, when the distributions of resources between the capital and labor in the production are equal and the capital stock per worker is greater than one, the results of Proposition 3 hold for *any* value of γ.

Now we obtain further interpretations of the estimates that were given in Proposition 3.

Proposition 4 *For any set of values of* α, β, γ *and k satisfying the condition*

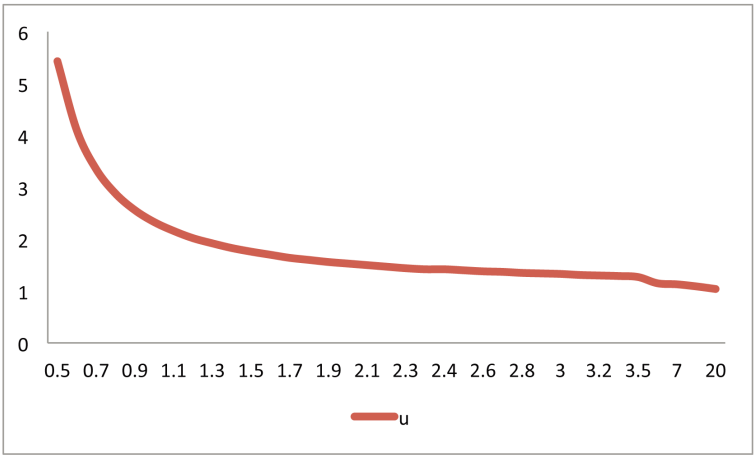


Figure 1. Plot of $u = (\beta/\alpha)^{1/\gamma}$ against γ.

(11), *the estimates given in Proposition 3 would continue to hold when we keep increasing the value of either,* α *or, k.*

4. Marginal Rate of Substitution

For any arbitrary production function with output Y and the factors of production, K, L and F, the marginal products with respect to the capital, labor and the factor of productively are defined as $\partial Y/\partial K$, $\partial Y/\partial L$ and $\partial Y/\partial F$ respectively. Also, the marginal rate of substitution (MRS) is defined as the ratio of the marginal product with respect to the capital by the marginal product with respect to labor, *i.e.*,

$$\text{MRS} = \left(\partial Y/\partial K\right)/\left(\partial Y/\partial L\right) \tag{12}$$

We now consider the growth of the marginal products with respect to the three factors appearing in the CES production function.

Proposition 5

For the CES production function, the marginal products with respect to all the three factors of production are increasing functions of time. Further, the marginal rate of substitution (MRS) is given by

$$\text{MRS} = \left(\alpha/\beta\right)k^{\gamma-1} \tag{13}$$

We note that (13) is well defined since by Assumption 2 we have $L \neq 0$ and so $\beta \neq 0$. It follows from (13) that the MRS is independent of both the total factor of productivity and the output per worker. Also, for $\gamma > 1$, (13) implies that $\text{MRS} \to +\infty$ as $k \to +\infty$. For $\gamma < 1$, (13) implies that $\text{MRS} \to 0$ as $k \to +\infty$.

For the Cobb-Douglas function (2), the marginal rate of substitution can be easily obtained by taking the limit as $\gamma \to 0$ in (13) and using Proposition 1; and the result thus obtained matches with the corresponding expression in ([27], p. 107).

5. Homogeneity of CES Function and Wages

We recall the standard definition that a function $Y = f\left(x_1, x_2, \cdots, x_n\right)$ of n independent variables $x_1, \cdots, x_n$ is a homogenous function of degree k if

$$f\left(tx_1, tx_2, \cdots, tx_n\right) = t^k f\left(x_1, x_2, \cdots, x_n\right)$$

for any positive scalar t. A classical theorem due to L. Euler (1703-1783) on homogeneous functions (see, e.g., [28] for a proof) states that if $Y = f\left(x_1, x_2, \cdots, x_n\right)$ is a homogeneous function of degree k with continuous partial derivatives then

$$\sum_{i=1}^{n} x_i \frac{\partial f}{\partial x_i} = kf\left(x_1, x_2, \cdots, x_n\right) \tag{14}$$

Now, it is easy to see that the CES production function, given by (1), is a homogeneous function of degree 2 in the variables K, L and F. So, from (14) with $k = 2, n = 3$ and $x_1 = K, x_2 = L, x_3 = F$ and writing Y for f, we have

$$\frac{\partial Y}{\partial K}K+\frac{\partial Y}{\partial L}L+\frac{\partial Y}{\partial F}F=2Y \tag{15}$$

Using (15) we now obtain a result for the wage of a worker in the context of the CES production function model (the result is closely along the lines of ([29], p. 7), Equation (15)):

Proposition 6 *Assume that in the short run the relative price of the factors adjust so that capital and labor are fully employed. Then, for the CES production function, the wage of a worker is equal to the balance remaining from the output per worker when we spend the rental price of capital times the capital per worker (assuming that there is no wage differentiation, i.e. all the workers receive the same wages).*

We remark that if F is a constant (in temporary contravention of Assumption 1), then the degree of homogeneity of the CES function is unity, and the Euler's theorem on homogeneous function (14) now gives (compare with (15))

$$\frac{\partial Y}{\partial K}K+\frac{\partial Y}{\partial L}L=Y \tag{16}$$

and it is easy to verify that the statement of Proposition 6 still holds for this case by rearranging slightly the proof of Proposition 6.

6. Conclusion

We have extended some results of the neoclassical growth theory when the production function is taken to be the CES function instead of the Cobb-Douglas function. Such generalizations are of considerable interest because the Cobb-Douglas function is not suitable for some application areas because its elasticity of substation has always the fixed value 1 whereas a CES function can be designed to have any pre-determined value as its elasticity of substitution. We assume that the total factor of productivity is a variable instead of being a parameter and under this assumption the CES production function becomes a homogenous function of degree two, and so it gives increasing returns to scale. When the total factor of productivity is steady, we show that the growth rate is equal to the Solow residual. We obtain some estimates of the Solow residual and the capital deepening term. We show that when the production is not entirely capital-intensive, an increase in capital implies an increase in the ratio of the production rate. We have considered here a CES production function where the input variables are the capital, labor and the total factor of productivity; and it would be of interest to extend our results to the more general cases of several input variables and also several output variables.

References

[1] Shepherd, R.W. (2015) Theory of Cost and Production Functions. Princeton University Press, Princeton. https://doi.org/10.1515/9781400871087

[2] Mishra, S.K. (2010) A Brief History of Production Functions. *IUP Journal of Managerial Economics*, **8**, 6-34.

[3] Cobb, C.W. and Douglas, P.H. (1928) A Theory of Production. *American Economics Review*, **18**, 139-165.

[4] Solow, R.M. (1956) A Contribution to the Theory of Economic Growth. *Quarterly Journal of Economics*, **70**, 65-94. https://doi.org/10.2307/1884513

[5] Arrow, K.J., Chenery, H.B., Minhas, B.S. and Solow, R.M. (1961) Capital-Labor Substitution and Economic Efficiency. *Review of Economics and Statistics*, **43**, 225-250. https://doi.org/10.2307/1927286

[6] Acemoglu, D. (2009) Introduction to Modern Economic Growth. Princeton University Press, Princeton.

[7] Barro, R.J. and Sala-i-Martin, X. (2004) Economic Growth. Second Edition, MIT Press, Cambridge.

[8] Daniels, G.E. and Kakar, V. (2017) Economic Growth and CES Production Function with Human Capital. *Economics Bulletin*, **37**, 930-951. https://doi.org/10.2139/ssrn.2878578

[9] Fragiadakis, K., Peoussos, L., Kouvaritakis N. and Capros P. (2013) A Multi-Country Econometric Estimation of the Constant Elasticity of Substitution. National Technical University of Athens, Institute of Communication and Computer Systems, Technical Report.

[10] Grandville, O. (2016) Economic Growth: A Unified Approach (Chapter 4: The CES Production Function as a General Mean). Second Edition, Cambridge University Press, New York, 90-113.

[11] Henningsen, A. and Henningsen, G. (2012) On the Estimation of the CES Production Function-Revisited. *Economics Letters*, **115**, 67-69.

[12] Klump, R. and Preissler, H. (2000) CES Production Functions and Economic Growth. *Scandinavian Journal of Economics*, **102**, 41-54. https://doi.org/10.1111/1467-9442.00183

[13] Klump, R. and Saam, M. (2008) Calibration of Normalized CES Production Functions in Dynamic Models. *Economics Letters*, **99**, 256-259.

[14] Rao, T.V.S.R. (2011) Contemporary Relevance and Ongoing Controversies Related to CES Production Function. *Journal of Quantitative Economics*, **9**, 36-57.

[15] Raval, D. (2010) Beyond Cobb-Douglas: Estimation of a CES Production Function with Factor Augmenting Technology. PhD Dissertation, University of Chicago, Chicago.

[16] Raval, D. (2015) The Micro Elasticity of Substitution with Non-Neutral Technology. Federal Trade Commission. White Paper, Washington DC.

[17] Uzawa, H. (1962) Production Functions with Constant Elasticities of Substitution. *Review of Economics Studies*, **9**, 291-299. https://doi.org/10.2307/2296305

[18] Zha, D. and Degun, Z. (2014) The Elasticity of Substitution and the Way of Nesting CES Production Function with Emphasis on Energy Input. *Applied Energy*, **130**, 793-798.

[19] Vîlci, A.D. and Vîlci, G.E. (2011) On some Geometric Properties of the Generalized CES Production Functions. *Applied Mathematics and Computation*, **218**, 124-129.

[20] Vîlci, A.D. and Vîlci, G.E. (2017) A Survey of the Geometry of the Production Models in Economics. *Arab Journal of Mathematical Sciences*, **23**, 18-31.

[21] Wang, X. and Fu, Y. (2013) Some Characterizations of the Cobb-Douglas and CES Production Functions in Microeconomics. *Abstract and Applied Analysis*, **2013**, Article ID: 761832.

[22] Hicks, J.R. (1932) The Theory of Wages. Macmillan, London.

[23] Antrás, P. (2004) Is the US Aggregate Production Function Cobb-Douglas? New Estimates of the Elasticity of Substitution. *Contributions to Macroeconomics*, **4**, 1-34. https://doi.org/10.2202/1534-6005.1161

[24] Werf, E. (2007) Production Functions for Climate Policy Modeling: An Empirical Analysis. *Energy Economics*, **30**, 2964-2979.

[25] Young, A.T. (2013) US Elasticities of Substitution and Factor Augmentation at the Industry Level. *Macroeconomic Dynamics*, **17**, 861-897. https://doi.org/10.1017/S1365100511000733

[26] Chiang, A.C. and Wainwright, K. (2005) Fundamental Methods of Mathematical Economics. 4th Edition, McGraw-Hill, New York.

[27] Aghion, P. and Howitt, P. (2009) The Economics of Growth. MIT Press, Cambridge.

[28] Widder, D.W. (1987) Advanced Calculus. 2nd Edition, Dover Publications, New York.

[29] Bénassy, J.-P. (2011) Macroeconomic Theory. Oxford University Press., New York. https://doi.org/10.1093/acprof:osobl/9780195387711.001.0001

APPENDIX I: Elasticity of Substitution-Review

We briefly review here the definition of the elasticity of substitution (for further details see, e.g., [6] [7]). Consider a general production function with output Y given by

$$Y = f\left(x_1, x_2, \cdots, x_n\right) \tag{17}$$

where x_i $(1 \le i \le n)$ are some independent variables and f is an arbitrary function that is differentiable partially with respect to each of the variables x_i. The *elasticity of substitution* σ_{ij} between any two distinct variables x_i and x_j measures the percentage of response of the relative marginal products of the two factors to a percentage of change in the ratio of the two quantities. It is defined as (e.g., [29], p. 509):

$$\sigma_{ij} = \frac{\partial \log_e\left(x_i/x_j\right)}{\partial \log_e\left(\left(\partial f/\partial x_i\right)/\left(\partial f/\partial x_j\right)\right)} \tag{18}$$

along the curve $f\left(x_1, x_2, \cdots, x_n\right) = \lambda$ where λ is a constant; the logarithm being taken to the base e (*i.e.*, the natural logarithm). For the CES production function, (17) takes the form given by (1) with $n = 3$ and x_1, x_2, x_3 to be the variables K, L, F respectively. For the CES function, it can be shown ([5]) that

$$\sigma_{ij} = 1/\left(1-\gamma\right) \tag{19}$$

when we take i, j to be any two of the variables K, L, F.

APPENDIX II: Proofs

1) Proof of Proposition 2:

(i) In the steady state, $\dot{k} = 0$, and we get from (7) that $G = \dot{F}/F$. Also, when $\dot{k} = 0$, we get from (10) that $D = 0$.

For the partial converse, we note that if the growth rate is equal to the rate of the total factor of productivity, *i.e.* if $G = \dot{F}/F$, then we get from (7) and (10) that

$$D = \frac{\alpha k^{\gamma-1}\dot{k}}{\alpha k^{\gamma} + \beta} = 0 \tag{20}$$

and (20) implies that $\dot{k} = 0$ because both α and k are nonzero by our assumptions; also $\left(\alpha k^{\gamma} + \beta\right)$ is nonzero since otherwise from (6) we would get $y = 0$ and this would imply from (5) that $Y = 0$, a trivial case.

(ii) It follows from (9) that the total factor of productivity is in a steady state (*i.e.*, $\dot{F} = 0$) if and only if $G = D$.

2) Proof of Proposition 3:

From (10) on expanding by Taylor's theorem and stopping after one term, we have as a linear approximation

$$D = \frac{\dot{k}}{k}\left(1 + \frac{\beta}{\alpha k^{\gamma}}\right)^{-1} \approx \frac{\dot{k}}{k}\left(1 - \frac{\beta}{\alpha k^{\gamma}}\right) \tag{21}$$

provided $\beta/\left(\alpha k^{\gamma}\right) < 1$; or, equivalently, provided (11) holds (here $\approx$ denotes

approximately). Now, the condition $\beta/(\alpha k^{\gamma}) < 1$ can be expressed as:

$$0 < \left(1 - \frac{\beta}{\alpha k^{\gamma}}\right) < 1 \tag{22}$$

Thus, from (21) and (22) we get

$$D \le \dot{k}/k \tag{23}$$

provided (11) holds. This proves (i).

(We remark that if instead of a linear approximation, we had taken a second degree approximation from (10), *i.e.* had stopped after the second term in the Taylor's expansion (21), then it is easy to verify that we would obtain the same result (23) provided (11) holds; we omit the details.)

Using (9) and (23) we get

$$\frac{\dot{F}}{F} \ge \left(G - \frac{\dot{k}}{k}\right) \tag{24}$$

provided (11) holds. This proves (ii).

3) Proof of Corollary 1:

(i) When $\gamma \to \infty$, we have $(\beta/\alpha)^{1/\gamma} \to 1$ for a fixed value of α, and the inequality (11) reduces to $k > 1$. (ii) When $\gamma \to 0$, we have that $k^{\gamma} \to 1$ and (11) gives $\alpha > \beta$, *i.e.*, $\alpha > 1/2$ since $\alpha + \beta = 1$.

4) Proof of Proposition 4:

If α_1, α_2 are any two non-zero values, then it is easy to see that for a given non-zero value of γ, we have $\alpha_2 > \alpha_1$ implies that

$$\left(\frac{1-\alpha_1}{\alpha_1}\right)^{1/\gamma} > \left(\frac{1-\alpha_2}{\alpha_2}\right)^{1/\gamma} \tag{25}$$

It follows from (25) that for any $\alpha_2 > \alpha_1$,

$$k > \left(\frac{1-\alpha_1}{\alpha_1}\right)^{1/\gamma} \Rightarrow k > \left(\frac{1-\alpha_2}{\alpha_2}\right)^{1/\gamma} \tag{26}$$

So, once the inequality (11) holds for a certain set of values of α, k, and γ, the estimates (23)-(24) in the statement of Proposition 3 would continue to hold if we keep on increasing the value of α while keeping γ fixed. Further, if we increase the value of k, the estimates (23)-(24) would still hold.

5) Proof of Proposition 5:

From (1) we get after some simplifications

$$R_K := \partial Y/\partial K = \alpha F^{\gamma} (K/Y)^{(\gamma-1)} = \alpha F^{\gamma} (k/y)^{\gamma-1} \tag{27}$$

If the production is not entirely labor intensive, then α and K are not identically zero (also $F > 0$ by the Assumption 1); so it follows from (27) that $R_K > 0$; thus the production rate with respect to capital is increasing. Similarly,

$$R_L := \partial Y/\partial L = \beta F^{\gamma} (L/Y)^{(\gamma-1)} = \beta F^{\gamma} y^{1-\gamma} \tag{28}$$

By the Assumption 2, the production is not entirely capital-intensive, and so

$\beta \neq 0$ and $L \neq 0$. We thus have from (28) that $R_L := \partial Y / \partial L > 0$; thus the production rate with respect to labor is increasing (note that by the Assumption 2, $L \neq 0$).

Again, we get that

$$R_F := \partial Y / \partial F = Y / F > 0 \tag{29}$$

since otherwise $Y = 0$, a trivial case that we exclude (recall that $F > 0$ by the Assumption 1). Thus the production rate with respect to the total factor of productivity is also increasing. From (27) and (28) we now obtain

$$R_K / R_L = (\alpha / \beta) k^{\gamma - 1} \tag{30}$$

and this proves (13).

6) Proof of Proposition 6:

Assuming that the capital and labor are fully employed, in the short run the wage is given by $R_L = \partial Y / \partial L$. Also, we have

$$R_K = \partial Y / \partial K \text{ and } \partial Y / \partial F = Y / F \tag{31}$$

So, (15) can be written as

$$LR_L + KR_K = Y \tag{32}$$

and rewriting (32) by using (5), we get

$$\text{wage} = (Y - KR_K) / L = (y - kR_K) \tag{33}$$

and this implies that in the short run, the wage of a worker is equal to the balance remaining from the output per worker when we spend the rental price of capital times the capital per worker since y is the output per worker and R_K is the rental price and k is the capital stock per worker. This proves the result.

Scientific Research Publishing

Journal of Applied Mathematics and Physics, 2017, 5, 2196-2204
http://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Stability Analysis of a Numerical Integrator for Solving First Order Ordinary Differential Equation

Samuel Olukayode Ayinde[1*], Adesoji Abraham Obayomi[1], Funmilayo Sarah Adebayo[2]

[1]Department of Mathematics, Faculty of Science, Ekiti State University, Ado Ekiti, Nigeria
[2]Federal Polytechnic, Ado Ekiti, Nigeria
Email: *biskay2003@yahoo.com

How to cite this paper: Ayinde, S.O., Obayomi, A.A. and Adebayo, F.S. (2017) Stability Analysis of a Numerical Integrator for Solving First Order Ordinary Differential Equation. *Journal of Applied Mathematics and Physics*, **5**, 2196-2204.
https://doi.org/10.4236/jamp.2017.511179

Received: October 6, 2017
Accepted: November 13, 2017
Published: November 16, 2017

Abstract

In this paper, we used an interpolation function to derive a Numerical Integrator that can be used for solving first order Initial Value Problems in Ordinary Differential Equation. The numerical quality of the Integrator has been analyzed to authenticate the reliability of the new method. The numerical test showed that the finite difference methods developed possess the same monotonic properties with the analytic solution of the sampled Initial Value Problems.

Keywords

Numerical Integrator, Autonomous and Non-Autonomous, Ordinary Differential Equation, Initial Value Problems, Stability Analysis

1. Introduction

Many Scholars have derived various Numerical Integrators using various techniques including interpolating functions that include the work of [1] [2] [3] [4] among others. All these authors have employed some analytically continuous functions to create numerically stable Integrators that can be used for ordinary differential equations. In this work we use an analytically differentiable interpolating function to create a one-step Finite Difference scheme for solving Initial Value Problems of first order Ordinary Differential Equations, and we are considering the concept of Nature, of Solutions of first order Ordinary Differential Equations to assume a theoretical solution and use that assumption to derive a discrete model that can be applied to some Ordinary differential equations.

Definition 1 [5]

Consider the *n*th-order ordinary differential equation

$$F\left(x, y, y^1, \cdots, y^n\right) = 0 \tag{1}$$

where F is a real function of its $(n+2)$ arguments $x, y, y^1, \cdots, y^n$.

1) Let f be a real function defined for all x in a real interval I and having an *n*th derivative (and hence also all lower ordered derivatives) for all $x \in \mathrm{I}$. The function f is called an explicit solution of the differential Equation (1) on interval I if it fulfills the following two requirements.

$$F\left(x, f(x), f^1(x), \cdots, f^n(x)\right) \tag{2}$$

is defined for all $x \in \mathrm{I}$, and

$$F\left(x, f(x), f^1(x), \cdots, f^n(x)\right) = 0 \tag{3}$$

for all $x \in \mathrm{I}$.

That is, the substitution of $f(x)$ and its various derivatives for y and its corresponding derivatives.

2) A relation $g(x, y) = 0$ is called an implicit solution of (1) if this relation defines at least one real function f of the variable x on an internal I such that this function is an explicit solution of (1) on this interval.

3) Both explicit solutions and implicit solutions will usually be called simply Solutions.

We now consider the geometric significance of differential equations and their solutions. We first recall that a real function $F(x)$ may be represented geometrically by a Curve $y = F(x)$ in the xy plane and that the value of the derivative of F at x, $F'(x)$, may be interpreted as the slope of the curve $y = F(x)$ at x.

2. Formulation of the Interpolating Function

Consider the initial value problem of the IVP

$$y'(x) = f(x, y), \quad y(x_0) = \eta, \tag{4}$$

where η is a discrete variables in the interval $[x_n, x_{n+1}]$. In this we consider the method based on local representation of the theoretical solution $y(x)$.

Let us assume that the theoretical solution $y(x)$ to the initial value problem 4) can be locally represented in the interval $[x_n, x_{n+1}], n \geq 1$ by the non-polynomial interpolating function given by:

$$F(x) = \alpha_1 \mathrm{e}^{-2x} + \alpha_2 x^2 + \alpha_3 x + \alpha_4 \tag{5}$$

where α_1, α_2 *and* α_3 *are real undetermined coefficients, and* α_4 *is a constant.*

3. Derivation of the Integrator

We assumed that the theoretical solution $y(x)$ to the initial value problem (5) can be locally represented in the interval $[x_n, x_{n+1}], n \geq 0$ by the non-polynomial interpolating function;

$$F(x) = \alpha_1 e^{-2x} + \alpha_2 x^2 + \alpha_3 x + \alpha_4 \tag{6}$$

where α_1, α_2 *and* α_3 are real undetermined coefficients, and α_4 is a constant.

We shall assume y_n is a numerical estimate to the theoretical solution $y(x)$ and $f_n = f(x_n, y_n)$.

We define mesh points as follows:

$$x_n = a + nh,\ n = 0,1,2,\cdots \tag{7}$$

We impose the following constraints on the interpolating function (6) in order to get the undetermined coefficients:

1) The interpolating function must coincide with the theoretical solution at $x = x_n$ and $x = x_{n+1}$. Hence we required that

$$F(x_n) = \alpha_1 e^{-2x_n} + \alpha_2 x_n^2 + \alpha_3 x_n + \alpha_4 \tag{8}$$

$$F(x_{n+1}) = \alpha_1 e^{-2x_{n+1}} + \alpha_2 x_{n+1}^2 + \alpha_3 x_{n+1} + \alpha_4 \tag{9}$$

2) The derivatives of the interpolating function are required to coincide with the differential equation as well as its first, second, and third derivatives with respect to x at $x = x_n$.

We denote the i-th derivatives of $f(x, y)$ with respect to x with $f^{(i)}$ such that

$$F^1(x_n) = f_n,\ F^2(x_n) = f_n^1,\ F^3(x_n) = f_n^2, \tag{10}$$

This implies that,

$$f_n = -2\alpha_1 e^{-2x_n} + 2\alpha_2 x_n + \alpha_3 \tag{11}$$

$$f_n^1 = 4\alpha_1 e^{-2x_n} + 2\alpha_2 \tag{12}$$

$$f_n^2 = -8\alpha_1 e^{-2x_n} \tag{13}$$

Solving for α_1, α_2 *and* α_3 from Equations (11) (12) and (13), we have

$$\alpha_1 = -\frac{1}{8} f_n^2 e^{2x_n} \tag{14}$$

$$\alpha_2 = \frac{1}{2}\left(f_n^1 + \frac{1}{2} f_n^2 \right) \tag{15}$$

and

$$\alpha_3 = \left(f_n - \frac{1}{4} f_n^2 \right) - \left(f_n^1 + \frac{1}{2} f_n^2 \right) x_n \tag{16}$$

Since $F(x_{n+1}) = y(x_{n+1})$ and $F(x_n) = y(x_n)$

Implies that $y(x_{n+1}) = y_{n+1}$ and $y(x_n) = y_n$

$$F(x_{n+1}) - F(x_n) = y_{n+1} - y_n \tag{17}$$

Then we shall have from (8) and (9) into (17)

$$y_{n+1} - y_n = \alpha_1 \left[e^{-2x_{n+1}} - e^{-2x_n} \right] + \alpha_2 \left[x_{n+1}^2 - x_n^2 \right] + \alpha_3 \left[x_{n+1} - x_n \right] \tag{18}$$

Recall that $x_n = a + nh$, $x_{n+1} = a + (n+1)h$ with $n = 0,1,2,\cdots$

Substitute (14) (15) (16), into (18), and simplify we have the integrator

$$y_{n+1} = y_n - \frac{1}{8} f_n^2 \left(e^{-2h} - 1\right) + \frac{1}{2}\left(f_n^1 + \frac{1}{2} f_n^2 \right) h^2 + \left(f_n - \frac{1}{4} f_n^2 \right) h \tag{19}$$

for solution of the first order differential equation.

4. Properties of the Integration Method

4.1. Qualitative Properties of the Scheme

4.1.1. Definition 2 [6]

Define any algorithm for solving differential equations in which the approximation y_{n+1} to the solution at the point x_{n+1} can be calculated if only x_n, y_n and h are known as one-step method. It is a common practice to write the functional dependence, y_{n+1}, on the quantities x_n, y_n and h in the form:

$$y_{n+1} = y_n + h\varnothing\left(x_n, y_n; h\right) \tag{20}$$

where $\varnothing\left(x_n, y_n; h\right)$ is the increment function.

The numerical integrator can be expressed as a one-step method in the form (20) above thus:

From (19) *i.e.* $y_{n+1} = y_n - \frac{1}{8} f_n^2 \left(e^{-2h} - 1\right) + \frac{1}{2}\left(f_n^1 + \frac{1}{2} f_n^2 \right) h^2 + \left(f_n - \frac{1}{4} f_n^2 \right) h$

Expanding e^{-2h} into the fourth term, we have

$$e^{-2h} = \sum_{r=0}^{\infty} \frac{(-2h)^r}{r!} = 1 - 2h + \frac{(2h)^2}{2!} - \frac{(2h)^3}{3!} + \cdots \tag{21}$$

Put (21) into (19), then expand

$$y_{n+1} = y_n + f_n h + f_n^1 x_n h + \frac{1}{2} f_n^1 h^2 + \frac{1}{2} f_n^2 x_n h + \frac{1}{6} f_n^2 h^3 \tag{22}$$

$$= y_n + h\left\{ f_n + f_n^1\left(x_n + \frac{1}{2} h \right) + f_n^2\left(x_n + \frac{1}{6} h^2 \right) \right\} \tag{23}$$

$$\text{Let} \quad A = x_n + \frac{1}{2} h \quad \text{and} \quad B = x_n + \frac{1}{6} h^2 \tag{24}$$

Thus our integrator (19) can be written compactly as

$$y_{n+1} = y_n + h\left\{ f_n + A f_n^1 + B f_n^2 \right\} \tag{25}$$

Which is in the form

$$y_{n+1} = y_n + h\varnothing\left(x_n, y_n; h\right) \tag{26}$$

$$\text{where} \quad \varnothing\left(x_n, y_n; h\right) = \left\{ f_n + A f_n^1 + B f_n^2 \right\} \tag{27}$$

4.1.2. Theorem 1. [7]

Let the increment function of the method defined by (25) be continuous as a function of its arguments in the region defined by

$$x \in [a, b],\ y \in (-\infty, \infty);\ 0 \le h \le h_o,$$

where $h_o > 0$, and let there exists a constant L such that

$$\left|\varnothing\left(x_n, y_n^*; h\right) - \varnothing\left(x_n, y_n; h\right)\right| \le L\left|y_n^* - y_n\right| \tag{28}$$

for all $\left(x_n, y_n; h\right)$ and $\left(x_n, y_n^*; h\right)$ in the region just defined. Then the relation (28) is the Lipscitz condition and it is the necessary and sufficient condition for the convergence of our method (19).

We shall proof that (19) satisfies (28) in line with the established Fatunla's theorem.

4.1.3. Proof of Convegence of the Integrator

The increment function $\varnothing\left(x_n, y_n; h\right)$ can be written in the form

$$\varnothing\left(x_n, y_n; h\right) = \left\{ f\left(x_n, y_n\right) + Af^{(1)}\left(x_n, y_n\right) + Bf^{(2)}\left(x_n, y_n\right)\right\} \tag{29}$$

where A and B are constants defined below.

$$A = x_n + \frac{1}{2}h$$

and

$$B = x_n + \frac{1}{6}h^2$$

Consider Equation (29), we can also write

$$\begin{aligned}
&\varnothing\left(x_n, y_n^*; h\right) = \left\{ f\left(x_n, y_n^*\right) + Af^{(1)}\left(x_n, y_n^*\right) + Bf^{(2)}\left(x_n, y_n^*\right)\right\} \\
&\varnothing\left(x_n, y_n^*; h\right) - \varnothing\left(x_n, y_n; h\right) \\
&= f\left(x_n, y_n^*\right) - f\left(x_n, y_n\right) + A\left[f^{(1)}\left(x_n, y_n^*\right) - f^{(1)}\left(x_n, y_n\right)\right] \\
&\quad + B\left[f^{(2)}\left(x_n, y_n^*\right) - f^{(2)}\left(x_n, y_n^*\right)\right]
\end{aligned} \tag{30}$$

Let $\bar{y}$ be defined as a point in the interior of the interval whose points are y and y^*, applying mean value theorem, we have

$$\left.\begin{aligned}
&f\left(x_n, y_n^*\right) - f\left(x_n, y_n\right) = \frac{\partial f\left(x_n, \bar{y}\right)}{\partial y_n}\left(y_n^* - y_n\right) \\
&f^{(1)}\left(x_n, y_n^*\right) - f^{(1)}\left(x_n, y_n\right) = \frac{\partial f^{(1)}\left(x_n, \bar{y}\right)}{\partial y_n}\left(y_n^* - y_n\right) \\
&\text{and}\quad f^{(2)}\left(x_n, y_n^*\right) - f^{(2)}\left(x_n, y_n\right) = \frac{\partial f^{(2)}\left(x_n, \bar{y}\right)}{\partial y_n}\left(y_n^* - y_n\right)
\end{aligned}\right\} \tag{31}$$

We define

$$\left.\begin{aligned}
&L = \sup_{(x_n, y_n)\in D} \frac{\partial f\left(x_n, y_n\right)}{\partial y_n} \\
&L_1 = \sup_{(x_n, y_n)\in D} \frac{\partial f^{(1)}\left(x_n, y_n\right)}{\partial y_n} \\
&\text{and}\quad L_2 = \sup_{(x_n, y_n)\in D} \frac{\partial f^{(2)}\left(x_n, y_n\right)}{\partial y_n}
\end{aligned}\right\}$$

Therefore

$$\begin{aligned}
&\varnothing\left(x_n, y_n^*; h\right) - \varnothing\left(x_n, y_n; h\right) \\
&= \frac{\partial f\left(x_n, \bar{y}\right)}{\partial y_n}\left(y_n^*, y_n\right) + A\left\{\frac{\partial f^{(1)}\left(x_n, \bar{y}\right)}{\partial y_n}\left(y_n^*, y_n\right)\right\} \\
&\quad + B\left\{\frac{\partial f^{2}\left(x_n, \bar{y}\right)}{\partial y_n}\left(y_n^*, y_n\right)\right\} \\
&= L\left(y_n^* - y_n\right) + AL_1\left(y_n^* - y_n\right) + BL_2\left(y_n^* - y_n\right)
\end{aligned} \tag{32}$$

Taking the absolute value of both sides

$$\begin{aligned}
&\left|\varnothing\left(x_n, y_n^*; h\right) - \varnothing\left(x_n, y_n; h\right)\right| \\
&\le \left|L\left(y_n^* - y_n\right) + AL_1\left(y_n^* - y_n\right) + BL_2\left(y_n^* - y_n\right)\right| \\
&\le \left|L + AL_1 + BL_2\right|\left|y^* - y\right|
\end{aligned} \tag{33}$$

If we let $M = \left|L + AL_1 + BL_2\right|$

then our Equation (33) turns to

$$\left|\varnothing\left(x_n, y_n^*; h\right) - \varnothing\left(x_n, y_n; h\right)\right| \le M\left|y^* - y\right| \tag{34}$$

which is the condition for convergence.

4.2. Consistence of the Integrator

Definition 3 [8]

The integration scheme: $y_{n+1} = y_n + h\left(x_n, y_n; h\right)$ is said to be consistent with the initial-value problem $y'(x) = f\left(x, y(x)\right)$, $y(a) = y_o$, $x \in [a, b]$, $y \in R$ provided the increment function $\varnothing\left(x, y; h\right)$ satisfies the following relationship

$$\varnothing\left(x, y; h\right) = f\left(x, y\right) \tag{35}$$

The significance of the consistency of a formula is that it ensures that the method approximates the ordinary differential equation in its place.

Therefore from

$$y_{n+1} = y_n + h\left\{f_n + f_n^1\left(x_n + \frac{1}{2}h\right) + f_n^2\left(x_n + \frac{1}{6}h^2\right)\right\} \tag{36}$$

where $y_{n+1} = y_n + \theta\left(x_n, y_n; h\right)$ then

$$\theta\left(x_n, y_n; h\right) = h\left\{f\left(x_n, y_n\right) + Af^{(1)}\left(x_n, y_n\right) + Bf^{(2)}\left(x_n, y_n\right)\right\}$$

and

$$A = x_n + \frac{1}{2}h,\ B = x_n + \frac{1}{6}h^2$$

If $h = 0$, then (36) reduced to $y_{n+1} = y_n$

$$\Rightarrow \theta\left(x_n, y_n; 0\right) = f\left(x, y\right) \tag{37}$$

It is a known fact that a consistent method has order of at least one [9]. Therefore, the new numerical integrator is consistent since Equation (36) can be reduced to (37) when $h = 0$.

4.3. Stability Analysis of the Integration Method

We shall establish the stability analysis of the integrator by considering the theorem established by Lambert 1972.

Let $y_n = y(x_n)$ and $P_n = P(x_n)$ denote two different numerical solutions of initial value problem of ordinary differential Equation (35) with the initial conditions specified as $y(x_o) = \eta$ and $p(x_o) = \eta^*$ respectively, such that $|\eta - \eta^*| < \varepsilon$, $\varepsilon > 0$. If the two numerical estimates are generated by the integrator (19). From the increment function (26), we have

$$y_{n+1} = y_n + h\varnothing(x_n, y_n; h) \tag{38}$$

$$P_{n+1} = P_n + h\varnothing(x_n, p_n; h) \tag{39}$$

The condition that

$$|y_{n+1} - P_{n+1}| \le K|\eta - \eta^*| \tag{40}$$

is the necessary and sufficient condition that our new method (19) be stable and convergent.

Proof

From (27) we have

$$y_{n+1} = y_n + h\{f_n + Af_n^1 + Bf_n^2\} \tag{41}$$

Then let

$$y_{n+1} = y_n + h\{f(x_n, y_n) + Af^1(x_n, y_n) + Bf^2(x_n, y_n)\} \tag{42}$$

and

$$p_{n+1} = p_n + h\{f(x_n, p_n) + Af^1(x_n, p_n) + Bf^2(x_n, p_n)\} \tag{43}$$

Therefore,

$$\begin{aligned} y_{n+1} - p_{n+1} = y_n - p_n + h\{&f(x_n, y_n) - f(x_n, p_n) + A[f^1(x_n, y_n) - f^1(x_n, p_n)] \\ &+ B[f^2(x_n, y_n) - f^2(x_n, p_n)]\} \end{aligned} \tag{44}$$

Applying the mean value theorem as before, we have

$$\begin{aligned} y_{n+1} - p_{n+1} = y_n - p_n + h\Bigg\{&\frac{\delta f(x_n, p_n)}{\delta p_n}(x_n - p_n) + A\left[\frac{\delta f^1(x_n, p_n)}{\delta P_n}(x_n - p_n)\right] \\ &+ B\left[\frac{\delta f^2(x_n, y_n)}{\delta p_n}(x_n - p_n)\right]\Bigg\} \end{aligned} \tag{45}$$

$$\begin{aligned} y_{n+1} - p_{n+1} = y_n - p_n + h\Bigg\{&\sup_{(x_n, p_n)\in D}\frac{\delta f(x_n, p_n)}{\delta p_n}(x_n - p_n) \\ &+ A\sup_{(x_n, p_n)\in D}\frac{\delta f^1(x_n, p_n)}{\delta p_n}(x_n - p_n) \\ &+ B\sup_{(x_n, p_n)\in D}\frac{\delta f^2(x_n, p_n)}{\delta p_n}(x_n - p_n)\Bigg\} \end{aligned} \tag{46}$$

$$y_{n+1} - p_{n+1} = y_n - p_n + h\left\{L\left(x_n, p_n\right) + AL_1\left(x_n, p_n\right) + BL_2\left(x_n, p_n\right)\right\} \quad (47)$$

Taking absolute value of both sides of (47) gives

$$\left|y_{n+1} - p_{n+1}\right| \le \left|y_n - p_n\right| + h\left|L + AL_1 + BL_2\right|\left|x_n - p_n\right| \quad (48)$$

Let $N = h\left|L + AL_1 + BL_2\right|$ and $y\left(x_o\right) = \eta$, $P\left(x_o\right) = \eta^*$, given $\varepsilon > 0$, then

$$\left|y_{n+1} - p_{n+1}\right| \le N\left|y_n - p_n\right| \quad (49)$$

and

$$\left|y_{n+1} - p_{n+1}\right| \le N\left|\eta - \eta^*\right| < \varepsilon, \text{ for every } \varepsilon > 0 \quad (50)$$

Then we conclude that our method (19) is stable and hence convergent.

5. The Implementation of the Integrator

Example 1

Using the Integrator (19) to solve the initial value problem

$$y' = 2x^2 - y, \; y(0) = -1, \text{ in the interval } 0 \le x \le 1$$

The analytical solution $y(x) = -5e^{-x} + 2x^2 - 4x + 4, h = 0.1$

Xn	Numerical	Analytical	Error
Solution	Solution		
[0.00]	[−1.000000000000000]	[−1.000000000000000]	[0.00000000000000]
[0.10]	[−0.904206720673739]	[−0.904187090179798]	[1.963049394060334e−005]
[0.20]	[−0.813671527795362]	[−0.813653765389909]	[1.776240545248164e−005]
[0.30]	[−0.724107175497677]	[−0.724091103408589]	[1.607208908771529e−005]
[0.40]	[−0.631614772805788]	[−0.631600230178197]	[1.454262759126301e−005]
[0.50]	[−0.532666457276770]	[−0.532653298563167]	[1.315871360274556e−005]
[0.60]	[−0.424070086966572]	[−0.424058180470132]	[1.190649644056130e−005]
[0.70]	[−0.302937292400544]	[−0.302926518957047]	[1.077344349675879e−005]
[0.80]	[−0.166654568800905]	[−0.166644820586107]	[9.748214797572485e−006]
[0.90]	[−0.012857119252502]	[−0.012848298702996]	[8.820549506724507e−006]
[1.00]	[0.1605948129795460]	[0.160602794142788]	[7.981163242049005e−006]

Example 2

Consider the initial value problem

$$y' = 2x - y, \; y(0) = 1, \text{ in the interval } 0 \le x \le 1$$

The analytical solution $y(x) = 3e^{-x} - 2(x+1), h = 0.1$

Xn	Numerical	Analytical	Error
Solution		Solution	
[0.00]	[1.000000000000000]	[1.000000000000000]	[0.00000000000000]
[0.10]	[1.115475967595757]	[1.115512754226943]	[3.678663118633629e−005]
[0.20]	[1.264167618965549]	[1.264208274480510]	[4.065551496124087e−005]
[0.30]	[1.449531491435216]	[1.449576422728009]	[4.493129279348196e−005]
[0.40]	[1.675424436165703]	[1.675474092923811]	[4.965675810808534e−005]
[0.50]	[1.946108932895438]	[1.946163812100385]	[5.487920494617882e−005]
[0.60]	[2.266295750270214]	[2.266356401171527]	[6.065090131368578e−005]
[0.70]	[2.641191092799143]	[2.641258122411429]	[6.702961228644000e−005]
[0.80]	[3.076548706299253]	[3.076622785477404]	[7.407917815127618e−005]
[0.90]	[3.578727463317524]	[3.578809333470850]	[8.187015332561387e−005]
[1.00]	[4.154755004864622]	[4.154845485377138]	[9.048051251525635e−005]

6. Summary and Conclusion

In this paper, we have proposed a new integration for the solution of standard initial value problem of first order ordinary differential equations. The new method was found to be convergence, consistence, and stable.

References

[1] Ayinde, S.O., *et al.* (2015) A New Numerical Method for Solving First Order Differential Equations. *American Journal of Applied Mathematics and Statistics*, *USA*, **3**, 156-160.

[2] Fatunla, S.O. (1988) Numerical Methods for Initial Value Problems in Ordinary Differential. Academic Press, San Diego, USA.

[3] Ibijola, E.A., *et al.* (2010) On a New Numerical Scheme for the Solution of Initial Value Problems in ODE. *Australian Journal of Basic and Applied Sciences*, **4**, 5277-5282.

[4] Ogunrinde, R.B. (2010) A New Numerical Scheme for the Solution of Initial Value Problem (IVP) in Ordinary Differential Equations. Ph.D. Thesis, Ekiti State University, Ado Ekiti.

[5] Zwillinger, D. (1997) Handbook of Differential Equations. 3rd Ed., Academic Press, Boston, MA.

[6] Stoer, J. and Bulirsh, R. (1966) Numerical Treatment of ODEs by Extrapolation Methods. *Numerische Mathematical*, **8**, 93-104.

[7] Henrici, P. (1962) Discrete Variable Methods in ODEs. New York, John Wiley and Sons., U.S.A.

[8] Lambert, J.D. (1972) Introductory Mathematics for Scientists and Engineers. John Wiley & Sons, New York.

[9] Shepley, L.R. (1984) Differential Equations. Third Edition, John Wiley & Sons Inc., Canada.

Journal of Applied Mathematics and Physics, 2017, 5, 2205-2217
http://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Positive Radial Solutions for a Class of Semilinear Elliptic Problems Involving Critical Hardy-Sobolev Exponent and Hardy Terms

Yong-Yi Lan

School of Sciences, Jimei University, Xiamen, China
Email: 200661000135@jmu.edu.cn

How to cite this paper: Lan, Y.-Y. (2017) Positive Radial Solutions for a Class of Semilinear Elliptic Problems Involving Critical Hardy-Sobolev Exponent and Hardy Terms. *Journal of Applied Mathematics and Physics*, **5**, 2205-2217.
https://doi.org/10.4236/jamp.2017.511180

Received: September 6, 2017
Accepted: November 14, 2017
Published: November 17, 2017

Abstract

In this paper, we investigate the solvability of a class of semilinear elliptic equations which are perturbation of the problems involving critical Hardy-Sobolev exponent and Hardy singular terms. The existence of at least a positive radial solution is established for a class of semilinear elliptic problems involving critical Hardy-Sobolev exponent and Hardy terms. The main tools are variational method, critical point theory and some analysis techniques.

Keywords

Hardy Singular Terms, Hardy-Sobolev Exponent, Positive Radial Solution, Perturbation Method, Variational Approach

1. Introduction and Main Results

In this paper, we are concerned with the existence of positive radial solutions for the following semilinear elliptic problem with Hardy-Sobolev exponent and Hardy singular terms:

$$\begin{cases} -\Delta u - \mu\dfrac{u}{|x|^2} = \left[1+\delta h\left(|x|\right)\right]\dfrac{|u|^{2^*(s)-2}}{|x|^s}u, & x \in \mathbb{R}^N \\ u > 0, & x \in \mathbb{R}^N \\ u \in D_r^{1,2}\left(\mathbb{R}^N\right) = \left\{u \in D^{1,2}\left(\mathbb{R}^N\right) : u \text{ is radial}\right\}, \end{cases} \tag{1.1}$$

where $0 < s < 2, 2^*(s) = \dfrac{2(N-s)}{N-2}$ is the Hardy-Sobolev critical exponent and $2^* = 2^*(0) = \dfrac{2N}{N-2}$ is the Sobolev critical exponent, $\mu < \overline{\mu} \triangleq \dfrac{(N-2)^2}{4}$.

$D^{1,2}\left(\mathbb{R}^N\right)\left(N\geq 3\right)$ denotes the space of the functions $u\in L^{2^*}\left(\mathbb{R}^N\right)$ such that $\nabla u\in L^2\left(\mathbb{R}^N\right)$, endowed with scalar product and norm, respectively, given by

$$\left\langle u,v\right\rangle=\int_{\mathbb{R}^N}\left(\nabla u\cdot\nabla v-\mu\frac{uv}{|x|^2}\right)\mathrm{d}x,$$

$$\|u\|^2=\int_{\mathbb{R}^N}\left(|\nabla u|^2-\mu\frac{u^2}{|x|^2}\right)\mathrm{d}x,$$

that coincides with the completion of $C_0^\infty\left(\mathbb{R}^N\right)$ with respect to the L^2-norm of the gradient. By Hardy inequality [1], we easily derive that the norm is equivalent to the usual norm:

$$\|u\|_0^2=\int_{\mathbb{R}^N}|\nabla u|^2\,\mathrm{d}x$$

in $D^{1,2}\left(\mathbb{R}^N\right)$.

Clearly, $D_r^{1,2}\left(\mathbb{R}^N\right)$ is a closed subset of $D^{1,2}\left(\mathbb{R}^N\right)$, so $D_r^{1,2}\left(\mathbb{R}^N\right)$ is a Hilbert space. By the symmetric criticality principle, in view of [2], we know that the positive radial solutions of problem (1.1) correspond to the nonzero critical points of the functional $I_\delta: D_r^{1,2}\left(\mathbb{R}^N\right)\to\mathbb{R}$ defined by

$$I_\delta(u)=\frac{1}{2}\int_{\mathbb{R}^N}\left(|\nabla u|^2-\mu\frac{u^2}{|x|^2}\right)\mathrm{d}x-\frac{1}{2^*(s)}\int_{\mathbb{R}^N}\frac{\left|u^+\right|^{2^*(s)}}{|x|^s}\mathrm{d}x$$

$$-\frac{\delta}{2^*(s)}\int_{\mathbb{R}^N}h(|x|)\frac{\left|u^+\right|^{2^*(s)}}{|x|^s}\mathrm{d}x,$$

where $u^+=\max\{u,0\}$.

The reason why we investigate (1.1) is the presence of the Hardy-Sobolev exponent, the unbounded domain $\mathbb{R}^N$ and the so-called inverse square potential in the linear part, which cause the loss of compactness of embedding $D^{1,2}\left(\mathbb{R}^N\right)\to L^{2^*}\left(\mathbb{R}^N\right)$, $H^1\left(\mathbb{R}^N\right)\to L^p\left(\mathbb{R}^N\right)$ and $D^{1,2}\left(\mathbb{R}^N\right)\to L^2\left(|x|^{-2}\,\mathrm{d}x\right)$. Hence, we face a type of triple loss of compactness whose interacting with each other will result in some new difficulties. In last two decades, loss of compactness leads to many interesting existence and nonexistence phenomena for elliptic equations. There are abundant results about this class of problems. For example, by using the concentration compactness principle, the strong maximum principle and the Mountain Pass lemma, Li *et al.* [3] had obtained the existence of positive solutions for singular elliptic equations with mixed Dirichlet-Neumann boundary conditions involving Sobolev-Hardy critical exponents and Hardy terms. Bouchekif and Messirdi [4] obtained the existence of positive solution to the elliptic problem involving two different critical Hardy-Sobolev exponents at the same pole by variational methods and concentration compactness principle. Lan and Tang [5] have obtained some existence results of (1.1) with $\mu=0$ via an abstract perturbation method in critical point theory. There are some other sufficient conditions, we refer the

interested readers to ([6]-[18]) and the references therein.

In the present paper, we investigate the existence of positive radial solutions of problem (1.1). There are several difficulties in facing this problem by means of variational methods. In addition to the lack of compactness, there are more intrinsic obstructions involving the nature of its critical points. Based on a suitable use of an abstract perturbation method in critical point theory discussed in [5] [13] [14], we show that the semilinear elliptic problem with Hardy-Sobolev exponent and Hardy singular terms has at least a positive radial solution.

In this paper, we assume that h satisfies one of the following conditions:

(H) $h \in L^{\infty}\left(\mathbb{R}^N\right) \cap C^1\left(\mathbb{R}^N\right), h(x) = h(|x|) = h(r), r = |x|$, and

$$\int_1^{\infty} r^{-\alpha+N-s-1} h(r) \mathrm{d}r < \infty$$

for some $\alpha < N - s$.

(H') $h \in C^2\left(\mathbb{R}^N\right), h(x) = h(|x|) = h(r), r = |x|$, $h(r)$ is T-periodic and

$$\int_0^T h(r) \mathrm{d}r = 0.$$

The main results read as follows.

Theorem 1 Let (H) hold, and assume that $h(0) = 0$ and $h \not\equiv 0$. Then for $|\delta|$ small, problem (1.1) has a positive radial solution u_δ.

Remark 1 It is easy to check that the following function $h(r)$ satisfies the conditions of **Theorem 1**,

$$h(r) = \frac{2r}{\mathrm{e}^r}.$$

Theorem 2 If assumption (H) holds, and suppose that $h \in C^2\left(\mathbb{R}^N\right)$ and $h(0)h''(0) > 0$. Then for $|\delta|$ small, problem (1.1) has a positive radial solution u_δ.

Remark 2 It is easy to check that the following function $h(r)$ satisfies the conditions of **Theorem 2**,

$$h(r) = \frac{1-2r}{\mathrm{e}^r}.$$

Theorem 3 Assume that (H) holds, and suppose

$$\int_0^{\infty} h(r)\left(1+r^{2-s}\right)^{-\frac{2(N-s)}{2-s}} r^{N-s-1} \mathrm{d}r \neq 0$$

and $\int_0^{\infty} h(0)h(r)\left(1+r^{2-s}\right)^{-\frac{2(N-s)}{2-s}} r^{N-s-1} \mathrm{d}r \leq 0.$

Then for $|\delta|$ small, problem (1.1) has a positive radial solution u_δ.

Remark 3 It is easy to check that the following function $h(r)$ satisfies the conditions of **Theorem 3** for all $N \geq 3$ and $0 < s < 2$,

$$h(r) = \frac{r}{\mathrm{e}^r},$$

in fact,

$$\int_0^{\infty} h(r)\left(1+r^{2-s}\right)^{-\frac{2(N-s)}{2-s}} r^{N-s-1}\mathrm{d}r \neq 0$$

and $\int_0^{\infty} h(0)h(r)\left(1+r^{2-s}\right)^{-\frac{2(N-s)}{2-s}} r^{N-s-1}\mathrm{d}r = 0;$

We can also give the following example for $N=3$ and $s=1$,

$$h(r)=\frac{1-100r}{\mathrm{e}^r},$$

in fact, with the help of computers, we can get

$$\int_0^{\infty} \frac{1-100r}{\mathrm{e}^r}\left(1+r^{2-1}\right)^{-\frac{2(3-1)}{2-1}} r^{3-1-1}\mathrm{d}r \approx -4.06 \neq 0$$

and $\int_0^{\infty} h(0)\frac{1-100r}{\mathrm{e}^r}\left(1+r^{2-1}\right)^{-\frac{2(3-1)}{2-1}} r^{3-1-1}\mathrm{d}r \approx -4.06 < 0.$

Theorem 4 Suppose that assumption (H') holds, and satisfies the condition that $h(0)h''(0)>0$. Then problem (1.1) has a positive radial solution u_δ, provided $|\delta| \ll 1$.

Remark 4 It is easy to check that the following function $h(r)$ satisfies the conditions of **Theorem 4**,

$$h(r)=\mathrm{e}^{\sin\left(\frac{7\pi}{4}+r\right)}\cos\left(\frac{7\pi}{4}+r\right),$$

in fact,

$$h(0)=\mathrm{e}^{\sin\left(\frac{7\pi}{4}+0\right)}\cos\left(\frac{7\pi}{4}+0\right)=\frac{\sqrt{2}}{2}\mathrm{e}^{-\frac{\sqrt{2}}{2}}>0,$$

and by a direct computation, we have

$$h''(0)=\sqrt{2}\mathrm{e}^{-\frac{\sqrt{2}}{2}}>0.$$

Theorem 5 Let h satisfy (H'), and suppose that $h(0)=0$ and $h\not\equiv 0$. Then problem (1.1) has a positive radial solution u_δ, provided $|\delta| \ll 1$.

Remark 5 It is easy to check that the following function $h(r)$ satisfies the conditions of **Theorem 5**,

$$h(r)=\sin 2r.$$

This paper is organized as follows. After a first section we devoted to studying the unperturbed problem $-\Delta u-\mu\frac{u}{|x|^2}=\frac{|u|^{2^*(s)-2}}{|x|^s}u$. The main results are proved in Section 3. In the following discussion, we denote various positive constants as C or $C_i\,(i=0,1,2,3,\cdots)$ for convenience. $o(t)$ denote $\frac{o(t)}{t}\to 0$ as $t\to 0^+$. This idea is essentially introduced in [5] [13].

2. The Case $\delta = 0$

In this section, we will study the unperturbed problem

$$\begin{cases} -\Delta u - \mu \dfrac{u}{|x|^2} = \dfrac{|u|^{2^*(s)-2}}{|x|^s} u, & x \in \mathbb{R}^N; \\ u \in D_r^{1,2}\left(\mathbb{R}^N\right), \ u > 0, & x \in \mathbb{R}^N. \end{cases} \tag{2.1}$$

It is well-known that the nontrivial solutions of problem (2.1) are equivalent to the nonzero critical points of the energy functional

$$I_0(u) = \frac{1}{2}\int_{\mathbb{R}^N}\left(|\nabla u|^2 - \mu\frac{u^2}{|x|^2}\right)\mathrm{d}x - \frac{1}{2^*(s)}\int_{\mathbb{R}^N}\frac{|u^+|^{2^*(s)}}{|x|^s}\mathrm{d}x, \quad u \in D_r^{1,2}\left(\mathbb{R}^N\right).$$

Obviously, the energy functional $I_0(u)$ is well-defined and is of C^2 with derivatives given by

$$\left\langle I_0'(u), v\right\rangle = \int_{\mathbb{R}^N}\left(\nabla u\cdot\nabla v - \mu\frac{uv}{|x|^2}\right)\mathrm{d}x - \int_{\mathbb{R}^N}\frac{|u^+|^{2^*(s)-1}}{|x|^s}v\mathrm{d}x, \ \ u, v \in D_r^{1,2}\left(\mathbb{R}^N\right);$$

$$\left\langle I_0''(u)v, w\right\rangle = \int_{\mathbb{R}^N}\left(\nabla v\cdot\nabla w - \mu\frac{vw}{|x|^2}\right)\mathrm{d}x - \int_{\mathbb{R}^N}\frac{\left(2^*(s)-1\right)|u^+|^{2^*(s)-2}}{|x|^s}vw\mathrm{d}x$$

$$u, v, w \in D_r^{1,2}\left(\mathbb{R}^N\right).$$

For all $\varepsilon > 0$, it is well known that the function

$$z_\varepsilon(r) = \left(\frac{2\varepsilon^{\frac{(2-s)\sqrt{\bar{\mu}-\mu}}{\sqrt{\bar{\mu}}}}(N-s)\left(\bar{\mu}-\mu\right)}{\sqrt{\bar{\mu}}}\right)^{\frac{\sqrt{\bar{\mu}}}{2-s}} \Bigg/ \left(r^{\sqrt{\bar{\mu}}-\sqrt{\bar{\mu}-\mu}}\left(\varepsilon^{\frac{(2-s)\sqrt{\bar{\mu}-\mu}}{\sqrt{\bar{\mu}}}} + r^{\frac{(2-s)\sqrt{\bar{\mu}-\mu}}{\sqrt{\bar{\mu}}}}\right)^{\frac{N-2}{2-s}}\right)$$

solves the equation (2.1) and satisfies

$$\int_{\mathbb{R}^N}\left(|\nabla z_\varepsilon|^2 - \mu\frac{z_\varepsilon^2}{|x|^2}\right)\mathrm{d}x = \int_{\mathbb{R}^N}\frac{|z_\varepsilon|^{2^*(s)}}{|x|^s}\mathrm{d}x.$$

Let

$$U(r) = \left(\frac{2(N-s)\left(\bar{\mu}-\mu\right)}{\sqrt{\bar{\mu}}}\right)^{\frac{\sqrt{\bar{\mu}}}{2-s}} \Bigg/ \left(r^{\sqrt{\bar{\mu}}-\sqrt{\bar{\mu}-\mu}}\left(1 + r^{\frac{(2-s)\sqrt{\bar{\mu}-\mu}}{\sqrt{\bar{\mu}}}}\right)^{\frac{N-2}{2-s}}\right),$$

then

$$z_\varepsilon(r) = \varepsilon^{-\frac{N-2}{2}}U\left(\frac{r}{\varepsilon}\right).$$

I_0 has a (non-compact) 1-dimensional critical manifold given by

$$Z = \left\{z = z_\varepsilon(r) : \varepsilon > 0\right\}.$$

The unperturbed problem is invariant under the transformation that transforms the function $u(r)$ in the function $\varepsilon^{-\frac{N-2}{2}}u\left(\frac{r}{\varepsilon}\right)$. The purpose of this

section is to show the following lemmas.

Lemma 2.1. For all $\varepsilon > 0$, $T_{z_\varepsilon} Z = \mathrm{Ker}\left[I_0''(z_\varepsilon)\right]$.

Proof. We will prove the lemma by taking $\varepsilon = 1$, hence $z_\varepsilon = U$. The case of a general $\varepsilon > 0$ will follow immediately. It is always true that $T_U Z \subseteq \mathrm{Ker}\left[I_0''(U)\right]$. We will show the converse, *i.e.*, that if $v \in \mathrm{Ker}\left[I_0''(U)\right]$, namely v is a solution of

$$\begin{cases} -\Delta u - \mu \dfrac{u}{|x|^2} = \left(2^*(s)-1\right)\dfrac{U^{2^*(s)-2}(x)}{|x|^s}u, & x \in \mathbb{R}^N \\ u \in D_r^{1,2}\left(\mathbb{R}^N\right), \ u > 0, & x \in \mathbb{R}^N \end{cases} \tag{2.2}$$

then $v \in T_U Z$, namely $\exists a \in \mathbb{R}$ such that $v = aD_\varepsilon z_\varepsilon\big|_{\varepsilon=1}$, where D_ε denotes the derivatives with respect to the parameter ε. We look for solutions $u \in D_r^{1,2}\left(\mathbb{R}^N\right)$ of problem (2.2). One has

$$-\Psi'' - \frac{n-1}{r}\Psi' - \mu\frac{\Psi}{r^2} = \left(2^*(s)-1\right)\frac{U^{2^*(s)-2}}{|r|^s}\Psi,$$

and then a first solution is given by

$$w = D_\varepsilon z_\varepsilon\big|_{\varepsilon=1} = C\left(r^{\frac{(2-s)\sqrt{\bar{\mu}-\mu}}{\sqrt{\bar{\mu}}}} - 1\right) \Bigg/ \left(r^{\sqrt{\bar{\mu}}-\sqrt{\bar{\mu}-\mu}}\left(r^{\frac{(2-s)\sqrt{\bar{\mu}-\mu}}{\sqrt{\bar{\mu}}}} + 1\right)^{\frac{N-s}{2-s}}\right)$$

which belongs to $D_r^{1,2}\left(\mathbb{R}^N\right)$, where $C = \sqrt{\bar{\mu}-\mu}\left(\dfrac{4(N-s)(\bar{\mu}-\mu)}{N-2}\right)^{\frac{N-2}{2(2-s)}}$. If we look for a second independent solution of the form $u(r) = c(r)w(r)$, we will check that u is not in $D_r^{1,2}\left(\mathbb{R}^N\right)$. A direct computation gives

$$-\left(c''w + 2c'w' + cw''\right) - \frac{N-1}{r}\left(c'w + cw'\right) - \mu\frac{cw}{r^2} = \left(2^*(s)-1\right)\frac{U^{2^*(s)-2}(r)}{|r|^s}cw,$$

and because w is a solution, we have

$$-c''w - c'\left(2w' + \frac{N-1}{r}w\right) = 0.$$

Setting $v = c'$, we obtain

$$-\frac{v'}{v} = 2\frac{w'}{w} + \frac{N-1}{r},$$

namely

$$v(r) = \frac{1}{r^{N-1}w^2(r)} \approx Cr^{1-N+2\sqrt{\bar{\mu}}-2\sqrt{\bar{\mu}-\mu}} \quad \left(r \to 0^+\right),$$

where C is a constant. This implies $c(r) \approx Cr^{2-N+2\sqrt{\bar{\mu}}-2\sqrt{\bar{\mu}-\mu}}$ as well as

$$u(r) \approx Cr^{\frac{2-N}{2}-\sqrt{\bar{\mu}-\mu}},$$

as $r \to 0^+$. Since $\dfrac{2-N}{2} - \sqrt{\bar{\mu}-\mu} < 0$, we have $u \notin D_r^{1,2}\left(\mathbb{R}^N\right)$. This implies a

contradiction to assumption which had been made. So $T_U Z = \mathrm{Ker}\left[I_0''(U)\right]$. This completes the proof of Lemma. □

Lemma 2.2. For all $\varepsilon > 0$, $I_0''(z_\varepsilon)$ is a Fredholm operator with index zero.

Proof. Indeed, $D_r^{1,2}(\mathbb{R}^N)$ is a Hilbert space, this implies

$D_r^{1,2}(\mathbb{R}^N) \cong \left[D_r^{1,2}(\mathbb{R}^N)\right]^*$ and $T_U Z = \mathrm{Ker}\left[I_0''(U)\right]$, we have

$$I_0''(U): D_r^{1,2}(\mathbb{R}^N) \to \left[D_r^{1,2}(\mathbb{R}^N)\right]^* = D_r^{1,2}(\mathbb{R}^N);$$

$$I_0''(U)(v+w) = I_0''(U)(w), \quad \text{where } v \in T_U Z, w \in \left[T_U Z\right]^\perp;$$

$$I_0''(U)(w) = -\Delta w - \mu\frac{w}{|x|^2} - \left(2^*(s)-1\right)\frac{U^{2^*(s)-2}(x)}{|x|^s}w.$$

It is obviously that $I_0''(U)$ is a self-adjoint operator on $D_r^{1,2}(\mathbb{R}^N)$, we have $\left(\mathrm{Im}\left(I_0''(U)\right)\right)^\perp = T_U Z$, hence

$$\mathrm{codim}\left(\mathrm{Im}\left(I_0''(U)\right)\right) = \dim\left(D_r^{1,2}(R^N)\Big/\left[T_U Z\right]^\perp\right) = \dim T_U Z = 1.$$

Moreover, fox fixed $u \in D_r^{1,2}(\mathbb{R}^N)$, the map

$$v \mapsto \int_{\mathbb{R}^N} a(x)uv\mathrm{d}x$$

is a bounded linear functional in $D_r^{1,2}(\mathbb{R}^N)$, where

$a(x) = \left(2^*(s)-1\right)\frac{U^{2^*(s)-2}(x)}{|x|^s}$. So by the Riesz representation theorem, there is an element in $D_r^{1,2}(\mathbb{R}^N)$, denote it by Tu, such that

$$\langle Tu, v\rangle = \int_{\mathbb{R}^N} a(x)uv\mathrm{d}x. \tag{2.3}$$

Clearly $T: D_r^{1,2}(\mathbb{R}^N) \to D_r^{1,2}(\mathbb{R}^N)$ is linear symmetric and bounded. Moreover T is compact; indeed, let $\{u_n\}$ be a bounded sequence in $D_r^{1,2}(\mathbb{R}^N)$. Passing to a subsequence we may assume that $u_n \rightharpoonup u$ in $D_r^{1,2}(\mathbb{R}^N)$, $u_n \to u$ in $L^{2^*(s)}(\mathbb{R}^N)$. Use u replaced by $u_n - u$ and v by $Tu_n - Tu$ in (2.3), and apply Hölder's inequality with $\frac{1}{2^*(s)} + \frac{1}{2^*(s)} + \frac{1}{p} = 1\left(p = \frac{N-s}{2-s}\right)$ to get

$$\|Tu_n - Tu\|^2 \le \|a\|_{L^p}\|u_n - u\|_{L^{2^*(s)}}\|Tu_n - Tu\|_{L^{2^*(s)}}$$

$$\Rightarrow \|Tu_n - Tu\| \le c\|u_n - u\|_{L^{2^*(s)}},$$

which implies that $Tu_n \to Tu$ in $D_r^{1,2}(\mathbb{R}^N)$. This shows that T is compact. We have

$$\langle I_0''(U)u, v\rangle = \langle u, v\rangle - \langle Tu, v\rangle = \langle u - Tu, v\rangle = \langle (I-T)u, v\rangle.$$

So $I_0''(U) = I - T$, where I is an identical operator. By the fact that $\lambda I - T$ is a Fredholm operator with index zero, where $\lambda \neq 0$ and T is compact, we obtain that $I_0''(U) = I - T$ is a Fredholm operator with index zero. This completes the

proof of Lemma. □

Now, we give the abstract perturbation method, which is crucial in our proof of the main results of this paper.

Lemma 2.3. [13] **(Abstract Perturbation Method)** Let E be a Hilbert space and let $f_0, G \in C^2(E,\mathbb{R})$ be given. Consider the perturbed functional $f_\varepsilon(u) = f_0(u) - \varepsilon G(u)$.

Suppose that f_0 satisfies:

1) f_0 has a finite dimensional manifold of critical points Z, let $b = f_0(z)$, for all $z \in Z$;

2) for all $z \in Z$, $D^2 f_0(z)$ is a Fredholm operator with index zero;

3) for all $z \in Z$, $T_z Z = \mathrm{Ker} D^2 f_0(z)$.

Hereafter we denote by Γ the functional $G|_Z$.

Let f_0 satisfy (1)-(3) above and suppose that there exists a critical point $\overline{z} \in Z$ of Γ such that one of the following conditions hold:

1) $\overline{z}$ is nondegenerated;

2) $\overline{z}$ is a proper local minimum or maximum;

3) $\overline{z}$ is isolated and the local topological degree of Γ' at $\overline{z}$, $\deg_{loc}(\Gamma', 0)$ is different from zero. Then for $|\varepsilon|$ small enough, the functional f_ε has a critical point u_ε such that $u_\varepsilon \to \overline{z}$, as $\varepsilon \to 0$.

Remark 2.4. [13] If $Z_0 := \{z \in Z : \Gamma(z) = \min_Z \Gamma\}$ is compact, then one can still prove that f_ε has a critical point near Z_0. The set Z_0 could also consist of local minima and the same for maxima.

3. Proof of the Theorems

We will now solve the bifurcation equation. In order to do this, let us define the reduced functional, see [14],

$$\Phi_\delta : Z \to \mathbb{R}$$

$$\Phi_\delta(z_\varepsilon) = I_\delta(z_\varepsilon + \omega_\delta(z_\varepsilon))$$

$$= c_0 - \frac{\delta}{2^*(s)} \int_{\mathbb{R}^N} h(x) \frac{z_\varepsilon^{2^*(s)}(x)}{|x|^s} \mathrm{d}x + o(\delta),\ c_0 = I_0(U),$$

where $\omega_\delta(z_\varepsilon) \perp T_{z_\varepsilon} Z$ and verifies $\|\omega_\delta(z_\varepsilon)\delta^{-1}\| \le C$ as $\delta \to 0$. Hence we are led to study the finite-dimensional functional

$$\Gamma(\varepsilon) := \int_{\mathbb{R}^N} h(x) \frac{z_\varepsilon^{2^*(s)}(x)}{|x|^s} \mathrm{d}x = \int_{\mathbb{R}^N} h(\varepsilon x) \frac{U^{2^*(s)}(x)}{|x|^s} \mathrm{d}x,\ (\varepsilon > 0).$$

The functional $\Gamma(\varepsilon)$ can be extended by continuity to $\varepsilon = 0$ by setting

$$\Gamma(0) = h(0) \int_{\mathbb{R}^N} \frac{U^{2^*(s)}(x)}{|x|^s} \mathrm{d}x.$$

Here we will prove the existence result by showing that problem (1.1) has a positive radial solution provided that h satisfies some integrability conditions. Before giving the proof of the main results, we need the following lemma.

Lemma 3.1. If (H) holds, then $\Gamma_r(\varepsilon)\to 0$ as $\varepsilon\to+\infty$.

Proof. From the definition of $\Gamma(\varepsilon)$ and U, we have

$$\begin{aligned}\Gamma_r(\varepsilon)&=\int_0^{+\infty}h(\varepsilon r)U^{2^*(s)}(r)r^{N-1-s}\mathrm{d}r\\&=\int_0^{+\infty}h(\varepsilon r)\frac{\big((N-s)(N-2)\big)^{\frac{N-s}{2-s}}}{\left(1+|r|^{2-s}\right)^{\frac{2(N-s)}{2-s}}}r^{N-1-s}\mathrm{d}r\\&=\int_0^{+\infty}h(r)\frac{\big((N-s)(N-2)\big)^{\frac{N-s}{2-s}}}{\left(1+\left|\frac{r}{\varepsilon}\right|^{2-s}\right)^{\frac{2(N-s)}{2-s}}}\frac{r^{N-1-s}}{\varepsilon^{N-s}}\mathrm{d}r\\&=\big((N-s)(N-2)\big)^{\frac{N-s}{2-s}}\int_0^{+\infty}h(r)\frac{\varepsilon^{N-s}\cdot r^{N-1-s}}{\left(\varepsilon^{2-s}+r^{2-s}\right)^{\frac{2(N-s)}{2-s}}}\mathrm{d}r\\&\le C\varepsilon^{-(N-s)}\int_0^1 h(r)\cdot r^{N-1-s}\mathrm{d}r+C_1\varepsilon^{\alpha-(N-s)}\int_1^{+\infty}\frac{h(r)}{r^{\alpha}}\cdot r^{N-1-s}\mathrm{d}r,\end{aligned}$$

where $\alpha<N-s$. It is easy to get the first integral in the right hand side; next we show the second integral: In fact,

$$\left(1+\left(\frac{r}{\varepsilon}\right)^{2-s}\right)^{\frac{2(N-s)}{2-s}}\cdot\frac{\varepsilon^{\alpha}}{r^{\alpha}}\ge 1\quad(\alpha<N-s),$$

so we have

$$\int_1^{+\infty}h(r)\frac{\varepsilon^{N-s}\cdot r^{N-1-s}}{\left(\varepsilon^{2-s}+r^{2-s}\right)^{\frac{2(N-s)}{2-s}}}\mathrm{d}r\le\varepsilon^{\alpha-(N-s)}\int_1^{+\infty}\frac{h(r)}{r^{\alpha}}\cdot r^{N-1-s}\mathrm{d}r\ (\alpha<N-s).$$

we deduce that $\Gamma_r(\varepsilon)\to 0$ as $\varepsilon\to+\infty$.

Proof of Theorem 1. Firstly, we claim that $\Gamma_r(\varepsilon)$ is not identically equal to 0. To prove this claim we will use Fourier analysis. We introduce some notation that will be used in the following discussion. If $g\in L^1\left([0,\infty),\frac{\mathrm{d}r}{r}\right)$, we define

$$M[g](s)=\int_0^{\infty}r^{-is}g(r)\frac{\mathrm{d}r}{r},$$

M is nothing but the Mellin transform. The associated convolution is defined by

$$(g\times k)(s)=\int_0^{\infty}g(r)k\left(\frac{s}{r}\right)\frac{\mathrm{d}r}{r}.$$

From the definition, it follows that $M[g\times k]=M[g]\cdot M[k]$. Indeed,

$$\begin{aligned}&M\left[g(x)\times k(x)\right](s)=F\left[g\left(e^x\right)\times k\left(e^x\right)\right](s)\\&=\int_{-\infty}^{+\infty}\left[g\left(e^x\right)\times k\left(e^x\right)\right]e^{-ixs}\mathrm{d}x=\int_{-\infty}^{+\infty}\left[\int_0^{+\infty}g(z)k\left(\frac{e^x}{z}\right)\frac{\mathrm{d}z}{z}\right]e^{-ixs}\mathrm{d}x\ \left(z=e^t\right)\\&=\int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty}g\left(e^t\right)k\left(e^{x-t}\right)\mathrm{d}t\right]e^{-ixs}\mathrm{d}x\end{aligned}$$

$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{\infty} g\left(e^{t}\right)e^{-its}k\left(e^{x-t}\right)e^{-i(x-t)s}\mathrm{d}t\mathrm{d}x$$
$$= \int_{-\infty}^{+\infty} g\left(e^{t}\right)e^{-its}\mathrm{d}t\int_{-\infty}^{\infty} k\left(e^{x-t}\right)e^{-i(x-t)s}\mathrm{d}x$$
$$= M\left[g(x)\right](s)\cdot M\left[k(x)\right](s).$$

With this notation we can write our Γ_r in the form

$$\Gamma_r(\varepsilon) = \int_0^{\infty} h(r)U^{2^*(s)}\left(\frac{r}{\varepsilon}\right)\left(\frac{r}{\varepsilon}\right)^{N-s}\frac{\mathrm{d}r}{r}.$$

We set $m = N - s - \alpha$ and

$$g(r) = h(r)r^{m},\ k(r) = U^{2^*(s)}\left(\frac{1}{r}\right)\left(\frac{1}{r}\right)^{N-s-m}.$$

Note that $g, k \in L^1\left([0,\infty), \frac{\mathrm{d}r}{r}\right)$. We have $\Gamma_r(\varepsilon) = \varepsilon^{-m}(g\times k)(\varepsilon)$ and hence if, by contradiction, $\Gamma \equiv 0$ then $g\times k \equiv 0$ and one has

$$M[g\times k] = M[g]\cdot M[k] \equiv 0.$$

On the other hand, $M[k]$ is real analytic and so has a discrete number of zeros. By continuity it follows that $M[g] \equiv 0$. Then g and hence h are identically equal to 0. We arrive at a contradiction. This proves the claim. Since $\Gamma_r(0) = 0$, $\Gamma_r(\varepsilon) \to 0$ as $\varepsilon \to +\infty$, and $\Gamma_r \not\equiv 0$, it follows that Γ_r has a maximum or a minimum at some $\bar{\varepsilon} > 0$. By a straight application of **Lemma 2.3** jointly with **Remark 2.4**, the result follows. □

Proof of Theorem 2. Using **Lemma 3.1**, we have $\Gamma_r(\varepsilon) \to 0$ as $\varepsilon \to +\infty$. and Γ_r can be extended to $\varepsilon = 0$ by continuity setting $\Gamma_r(0) = a_0 h(0)$, where $a_0 = \int_0^{+\infty} U^{2^*(s)}(r)r^{N-1-s}\mathrm{d}r > 0$. From the assumption, we have

$$\Gamma_r'(0) = 0,\ \Gamma_r''(0) = a_1 h''(0),\ a_1 = \int_0^{+\infty} U^{2^*(s)}(r)r^{N+1-s}\mathrm{d}r > 0$$

and the condition $h(0)h''(0) > 0$ implies that Γ_r has a (global) maximum (if $h(0) > 0$) or a (global) minimum (if $h(0) < 0$), at some $\bar{\varepsilon} > 0$. This allows us to use the abstract results, yielding a radial solution of problem (1.1), for $|\delta|$ small. □

Proof of Theorem 3. It suffices to remark that

$$\Gamma_r(1) = \left[(N-s)(N-2)\right]^{\frac{N-s}{2-s}}\int_0^{\infty} h(r)\left(1+r^{2-s}\right)^{-\frac{2(N-s)}{2-s}} r^{N-s-1}\mathrm{d}r \neq 0.$$

If

$$\int_0^{\infty} h(r)\left(1+r^{2-s}\right)^{-\frac{2(N-s)}{2-s}} r^{N-s-1}\mathrm{d}r > 0$$

$$\left(\text{resp. } \int_0^{\infty} h(r)\left(1+r^{2-s}\right)^{-\frac{2(N-s)}{2-s}} r^{N-s-1}\mathrm{d}r < 0\right)$$

then $h(0) \le 0$ (resp. $h(0) \ge 0$) and, once more Γ_r has a (global) maximum (resp. a (global) minimum) at some $\bar{\varepsilon} > 0$. □

In the rest of the section we will give the proof of **Theorem 4** and **Theorem 5**.

First we give the following **Lemma**. Hypothesis (H') allows us to use the following Riemann-Lebesgue convergence result.

Lemma 3.2 [13] Let $Q=[0,T]^N$ be a cube in $\mathbb{R}^N$, and $f\in L^q(Q)$ be a T-periodic function. Consider $f_\varepsilon(x)=f(\varepsilon x)$, then

$$f_\varepsilon \rightharpoonup \overline{f}=\frac{1}{|Q|}\int_Q f\,\mathrm{d}x, \text{ weakly in } L^q_{loc}(\mathbb{R}^N), \text{ as } \varepsilon\to\infty.$$

Lemma 3.3 If (H') holds, then

$$\Gamma_r(\varepsilon)\to 0,\ \varepsilon\to+\infty.$$

Proof. Given $\varepsilon>0$, there exists $R>0$ large enough such that

$$\left|\int_R^\infty h(r)z_\varepsilon^{2^*(s)}(r)r^{N-s-1}\mathrm{d}r\right|$$
$$\le \|h(r)\|_\infty \int_R^\infty z_\varepsilon^{2^*(s)}(r)r^{N-s-1}\mathrm{d}r<\varepsilon.$$

On the other hand, the remainder integral over the interval $0\le r<R$ tends to 0 as $\varepsilon\to\infty$ because of hypothesis (H') and the Riemann-Lebesgue lemma. □

Proof of Theorem 4. Using **Lemma 3.3**, we have $\Gamma_r(\varepsilon)\to 0$ as $\varepsilon\to+\infty$. and Γ_r can be extended to $\varepsilon=0$ by continuity setting $\Gamma_r(0)=a_0h(0)$, where $a_0=\int_0^{+\infty}U^{2^*(s)}(r)r^{N-1-s}\mathrm{d}r>0$. From the assumption, we have

$$\Gamma_r'(0)=0,\ \Gamma_r''(0)=a_1h''(0),\ a_1=\int_0^{+\infty}U^{2^*(s)}(r)r^{N+1-s}\mathrm{d}r>0.$$

and the condition $h(0)h''(0)>0$ implies that Γ_r has a (global) maximum (if $h(0)>0$) or a (global) minimum (if $h(0)<0$), at some $\overline{\varepsilon}>0$. This allows us to use the abstract results, yielding a radial solution of problem (1.1), for $|\delta|$ small. □

Proof of Theorem 5. It suffices to repeat the arguments used to prove **Theorem 1** using **Lemma 3.1** instead of **Lemma 3.3**.

4. Conclusion

We study a class of semilinear elliptic problems involving critical Hardy-Sobolev exponent and Hardy terms, and obtain positive radial solutions for these problems via an abstract perturbation method in critical point theory. Extensions of nonradial solutions for these problems are being investigated by the author. Results will be submitted for publication in the near future.

Acknowledgements

We would like to thank the editor and the referee for their valuable comments which have led to an improvement of the presentation of this paper.

Fund

This work is supported by Natural Science Foundation of China (No. 11671331); Natural Science Foundation of Fujian Province (No. 2015J01585) and Scientific Research Foundation of Jimei University.

References

[1] Garcia Azorero, J.P. and Peral Alonso, I. (1998) Hardy Inequalities and Some Critical Elliptic and Parabolic Problems. *Journal of Differential Equations*, **144**, 441-476. https://doi.org/10.1006/jdeq.1997.3375

[2] Palais, R.S. (1979) The Principle of Symmetric Criticality. *Communications in Mathematical Physics*, **69**, 19-30. https://doi.org/10.1007/BF01941322

[3] Li, Y.Y., Ruf, B., Guo, Q.Q. and Niu, P.C. (2014) Positive Solutions for Singular Elliptic Equations with Mixed Dirichlet-Neumann Boundary Conditions. *Mathematische Nachrichten*, **287**, 374-397. https://doi.org/10.1002/mana.201100351

[4] Bouchekif, M. and Messirdi, S. (2015) On Elliptic Problems with Two Critical Hardy-Sobolev Exponents at the Same Pole. *Applied Mathematics Letters*, **42**, 9-14. https://doi.org/10.1016/j.aml.2014.10.012

[5] Lan, Y.Y. and Tang, C.L. (2014) Perturbation Methods in Semilinear Elliptic Problems Involving Critical Hardy-Sobolev Exponent. *Acta Mathematica Scientia*, **34B**, 703-712. https://doi.org/10.1016/S0252-9602(14)60041-2

[6] Yan, S.S. and Yang, J.F. (2013) Infinitely Many Solutions for an Elliptic Problem Involving Critical Sobolev and Hardy-Sobolev Exponents. *Calculus of Variations and Partial Differential Equations*, **48**, 587-610. https://doi.org/10.1007/s00526-012-0563-7

[7] Ding, L. and Tang, C.L. (2008) Existence and Multiplicity of Positive Solutions for a Class of Semilinear Elliptic Equations Involving Hardy Term and Hardy-Sobolev Critical Exponents. *Journal of Mathematical Analysis and Applications*, **339**, 1073-1083. https://doi.org/10.1016/j.jmaa.2007.07.066

[8] Shang, Y.Y. and Tang, C.L. (2009) Positive Solutions for Neumann Elliptic Problems Involving Critica Hardy-Sobolev Exponent with Boundary Singularities. *Nonlinear Analysis: Theory, Methods & Applications*, **70**, 1302-1320. https://doi.org/10.1016/j.na.2008.02.013

[9] Cao, D.M., He, X.M. and Peng, S.J. (2005) Positive Solutions for Some Singular Critical Growth Nonlinear Elliptic Equations. *Nonlinear Analysis: Theory, Methods & Applications*, **60**, 589-609. https://doi.org/10.1016/j.na.2004.08.042

[10] Cao, D.M. and Peng, S.J. (2003) A Note on the Sign-Changing Solutions to Elliptic Problems with Critical Sobolev and Hardy Terms. *Journal of Differential Equations*, **193**, 424-434. https://doi.org/10.1016/S0022-0396(03)00118-9

[11] Ghoussoub, N. and Kang, X.S. (2004) Hardy-Sobolev Critical Elliptic Equations with Boundary Singularities. *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, **21**, 767-793.

[12] Kang, D.S. and Peng, S.J. (2005) Solutions for Semi-Linear Elliptic Problems with Critical Hardy-Sobolev Exponents and Hardy Potential. *Applied Mathematics Letters*, **18**, 1094-1100.

[13] Wang, C. and Shang, Y.Y. (2017) Existence and Multiplicity of Positive Solutions for Elliptic Equation with Critical Weighted Hardy-Sobolev Exponents and Boundary Singularities. *Computers Mathematics with Applications*, **74**, 701-713.

[14] Wang, C. and Shang, Y.Y. (2017) Existence and Multiplicity of Positive Solutions for a Perturbed Semilinear Elliptic Equation with Two Hardy-Sobolev Critical Exponents. *Journal of Mathematical Analysis and Applications*, **451**, 1198-1215.

[15] Bhakta, M. (2017) Infinitely Many Sign-Changing Solutions of an Elliptic Problem Involving Critical Sobolev and Hardy-Sobolev Exponent. *Proceedings of the Indian Academy of Sciences-Mathematical Sciences*, **127**, 337-347.

https://doi.org/10.1007/s12044-016-0304-5

[16] Deng, Z.Y. and Huang, Y.S. (2017) Positive Symmetric Results for a Weighted Quasilinear Elliptic System with Multiple Critical Exponents in R^N, Boundary Value Problems. https://doi.org/10.1186/s13661-017-0758-0

[17] Ambrosetti, A., Garcia Azorero, J. and Peral, I. (1999) Perturbation of $\Delta u + u^{\frac{N+2}{N-2}} = 0$, the Scalar Curvature Problem in R^N, and Related Topics. *Journal of Functional Analysis*, **165**, 117-149. https://doi.org/10.1006/jfan.1999.3390

[18] Ambrosetti, A. and Malchiodi, A. (2006) Perturbation Methods and Semilinear Elliptic Problems on R^n. Birkhäuser Verlag.

Journal of Applied Mathematics and Physics, 2017, 5, 2218-2227
http://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

A Quadratic Programming with Triangular Fuzzy Numbers

Seyedeh Maedeh Mirmohseni[1], Seyed Hadi Nasseri[2*]

[1]School of Mathematics and Information Science, Key Laboratory of Mathematics and Interdisciplinary Sciences of Guangdong Higher Education Institutes, Guangzhou University, Guangzhou, China
[2]Department of Mathematics, University of Mazandaran, Babolsar, Iran
Email: *nasseri@umz.ac.ir

How to cite this paper: Mirmohseni, S.M. and Nasseri, S.H. (2017) A Quadratic Programming with Triangular Fuzzy Numbers. *Journal of Applied Mathematics and Physics*, **5**, 2218-2227.
https://doi.org/10.4236/jamp.2017.511181

Received: June 4, 2017
Accepted: November 20, 2017
Published: November 23, 2017

Abstract

Quadratic Programming (QP) is a mathematical modeling technique designed to optimize the usage of limited resources and has been widely applied to solve real world problems. In conventional quadratic programming model the parameters are known constants. However in many practical situations, it is not reasonable to require that the constraints or the objective function in quadratic programming problems be specified in precise, crisp terms. In such situations, it is desirable to use some type of Fuzzy Quadratic Programming (FQP) problem. In this paper a new approach is proposed to derive the fuzzy objective value of fuzzy quadratic programming problem, where the constraints coefficients and the right-hand sides are all triangular fuzzy numbers. The proposed method is solved using MATLAB™ toolbox and the numerical results are presented.

Keywords

Fuzzy Numbers, Quadratic Programming, Membership Function

1. Introduction

Quadratic programming is a particular kind of nonlinear programming. There are several classes of problems that are naturally expressed as quadratic problems. Examples of such problems can be found in game theory, engineering modeling, design and control, problems involving economies of scale, facility allocation and location problems, etc. Several applications and test problems for quadratic programming can be found in [1] [2] [3] [4] [5]. Some traditional methods are available in the literature [6] [7] for solving such problems. Among the several applications, the portfolio selection problem is an important research

field in modern finance. This problem was first introduced by Markowitz [8] [9], and provided a risk investment analysis. Some works about portfolio selection problem by using fuzzy approaches can be found in [10] [11] [12] [13] [14].

The classical quadratic programming problem is to find the minimum or maximum values of a quadratic function under constraints represented by linear inequality or equations. The most typical quadratic programming problem is:

$$\begin{aligned} &\text{maximaize}\left(\text{or minimize}\right)\sum_{j=1}^{n} c_j x_j + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij} x_i x_j \\ &\text{subject to } \begin{cases} \sum_{j=1}^{n} a_{ij} x_j \le b_i, \ i \in N_m \\ x_j \ge 0, \ i \in N_m \end{cases} \end{aligned} \tag{1.1}$$

The function to be minimized (or maximized) is called an objective function; let us denote it by z. The numbers $c_j, j \in N_n$ are called cost coefficients and the vector $c = (c_1, c_2, \cdots, c_n)$ is called a cost vector. The vector $b^{\mathrm{T}} = (b_1, b_2, \cdots, b_m)$ is called a right-hand side vector and the matrix $A = \left[a_{ij}\right]_{m\times n}$, where $i \in N_m$ and $j \in N_n$, is called a constraint matrix. The matrix $Q = \left[q_{ij}\right]_{m\times n}$, is called the matrix of quadratic form where $i \in N_n$ and $j \in N_n$. Using this notation, the formulation of the problem can be simplified as:

$$\begin{aligned} &\max z = cx + \frac{1}{2}x^{\mathrm{T}}Qx \\ &\text{s.t.} \begin{cases} Ax \le b \\ x \ge 0 \end{cases} \end{aligned} \tag{1.2}$$

where $x = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$ is a vector of variables and s.t. stands for "subject to". In the following example, a quadratic programming problem is addressed:

Example 1.1: Consider a problem which is formulated as follows:

$$\begin{aligned} &\max z = 2x_1 + x_2 + 2x_1^2 + x_1x_2 + 2x_2^2 \\ &\text{s.t.} \begin{cases} x_1 + x_2 \le 2 \\ 2x_1 + 3x_2 \le 1 \\ x_1, x_2 \ge 0 \end{cases} \end{aligned} \tag{1.3}$$

in which $Q = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}$, $c = (2,1)$, $A = \begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix}$ and $b^{\mathrm{T}} = (2,1)$. Hence it can be rewritten as follows:

$$\begin{aligned} &\max z = (2,1)\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &\text{s.t.} \begin{cases} \begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \le \begin{bmatrix} 2 \\ 1 \end{bmatrix} \\ x_1, x_2 \ge 0 \end{cases} \end{aligned} \tag{1.4}$$

The paper is organized in 5 sections. In the next section some necessary notations and definitions of fuzzy set theory and fuzzy arithmetic are given. In Sec-

tion 3, we first define a quadratic programming problem and then give a new approach to solve these problems. In Section 4, a numerical example is presented to illustrate how to apply the concept of this paper for solving such quadratic programming problems. Finally we conclude in Section 5.

2. Arithmetic on Fuzzy Numbers

Here, we first give some necessary definitions of fuzzy set theory which is taken from [15] [16] [17].

Definition 2.1: Let $\boldsymbol{R}$ be the real line, then a fuzzy set A in $\boldsymbol{R}$ is defined to be a set of ordered pairs $A=\left\{\left(x,\mu_A(x)\right)\middle| x\in\boldsymbol{R}\right\}$, where $\mu_A(x)$ is called the membership function for the fuzzy set. The membership function maps each element of $\boldsymbol{R}$ to a membership value between 0 and 1.

Definition 2.2: The support of a fuzzy set A is defined as fallow:

$$supp(A)=\left\{x\in\boldsymbol{R}\middle|\mu_A(x)>0\right\}$$

Definition 2.3: The core of a fuzzy set is the set of all points x in $\boldsymbol{R}$ with $\mu_A(x)=1$.

Definition 2.4: A fuzzy set A is called normal if its core is nonempty. In other words, there is at least one point $x\in\boldsymbol{R}$ with $\mu_A(x)=1$.

Definition 2.5: The α–cut or α–level set of a fuzzy set is a crisp set defined by

$$A_\alpha=\left\{x\in\boldsymbol{R}\middle|\mu_A(x)>\alpha\right\}.$$

Definition 2.6: A fuzzy set A on $\boldsymbol{R}$ is convex, if for any $x,y\in\boldsymbol{R}$ and $\lambda\in[0,1]$, we have

$$\mu_A\left(\lambda x+(1-\lambda)y\right)\geq\min\left\{\mu_A(x),\mu_A(y)\right\}$$

Definition 2.7: A fuzzy number $\bar{a}$ is a fuzzy set on the real line that satisfies the condition of normality and convexity.

Definition 2.8: A fuzzy number $\tilde{a}$ on $\boldsymbol{R}$ is said to be triangular fuzzy number, if there exist real numbers and $l,r\geq 0$ such that

$$\tilde{a}(x)=\begin{cases}\dfrac{x}{l}+\dfrac{l-s}{l}, & x\in[s-l,s]\\ \dfrac{-x}{r}+\dfrac{s+r}{r}, & x\in[s,s+r]\\ 0, & \text{o.w.}\end{cases}$$

We denote a triangular fuzzy number $\tilde{a}$ by three real numbers s,l and r as $\tilde{a}=\langle s,l,r\rangle$, whose meaning are defined in **Figure 1**. We also denote the set of all triangular fuzzy numbers with F(R).

Definition 2.9: Let $\tilde{a}=\langle s_a,l_a,r_a\rangle$ and $\tilde{b}=\langle s_b,l_b,r_b\rangle$ be two triangular numbers and $x\in\boldsymbol{R}$. Summation and multiplication of fuzzy numbers defined as [18]:

$$x\tilde{a}=\begin{cases}\langle xs_a,xl_a,xr_a\rangle, & x\geq 0\\ \langle xs_a,-xr_a,-xl_a\rangle, & x<0\end{cases}$$

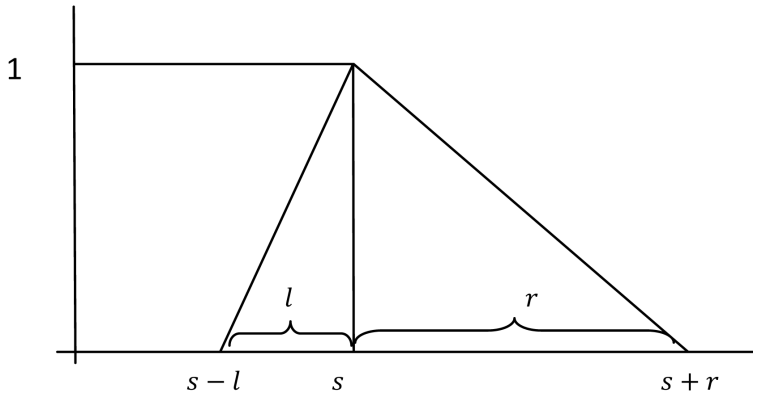


Figure 1. Fuzzy Triangular Number.

$$\tilde{a}+\tilde{b}=\langle s_a+s_b, l_a+l_b, r_a+r_b\rangle$$

$$\tilde{a}-\tilde{b}=\langle s_a-s_b, l_a-r_b, r_a-l_b\rangle$$

$\tilde{a}\le\tilde{b}$ if and only if $s_a\le s_b, s_a-l_a\le s_b-l_b, s_a+r_a\le s_b+r_b$

Definition 2.10: We let $\tilde{0}=(0,0,0)$ as a zero triangular fuzzy number.

Remark 2.1: $\tilde{a}\ge\tilde{0}$ if and only if $s_a\ge 0, s_a-l_a\ge 0, s_a+r_a\ge 0$.

Remark 2.2: $\tilde{a}\le\tilde{b}$ if and only if $-\tilde{a}\ge-\tilde{b}$.

3. Fuzzy Numbers Quadratic Programming

Here we first define the model and then propose a novel method for solving the mentioned problem.

3.1. Definition of Model

Several studies have developed efficient and effective algorithms for solving quadratic programming when the value assigned to each parameter is a known constant. However, quadratic programming models usually are formulated to find some future course of action so the parameter values used would be based on a prediction of future conditions which inevitably involves some degree of uncertainty [19] [20]. The most general type of this programming is formulated as follows:

$$\max z=\sum_{j=1}^{n}\tilde{C}_j\tilde{X}_j+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{Q}_{ij}\tilde{X}_i\tilde{X}_j$$

$$\text{s.t.}\begin{cases}\sum_{j=1}^{n}\tilde{A}_{ij}\tilde{X}_j\le\tilde{B}_i, & i\in N_m,\\ \tilde{X}\ge 0\end{cases}\tag{3.1}$$

where $\tilde{A}_{ij}$, $\tilde{B}_i$, $\tilde{C}_j$ and $\tilde{Q}_{ij}$ are fuzzy numbers, and $\tilde{X}_j$ are variables whose states are fuzzy numbers $(i\in N_m, j\in N_n)$; the operations of addition and multiplication are operations of fuzzy arithmetic, and denotes the ordering of fuzzy numbers. Here instead of discussing this general type, we consider a special case of fuzzy quadratic programming problem in which only the right-hand side parameters and constraint coefficient are triangular fuzzy numbers.

$$\begin{aligned} &\max z = \sum_{j=1}^{n} c_j x_j + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij} x_i x_j \\ &\text{s.t.} \begin{cases} \sum_{j=1}^{n} \tilde{A}_{ij} x_j \le \tilde{B}_i, \ i \in N_m, \\ x \ge 0 \end{cases} \end{aligned} \tag{3.2}$$

3.2. The New Approach

In this case we assume that all fuzzy numbers are triangular. According to Definition 2.8, the fuzzy quadratic programming (3.2) is rewritten as follows:

$$\begin{aligned} &\max z = \sum_{j=1}^{n} c_j x_j + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij} x_i x_j \\ &\text{s.t.} \begin{cases} \sum_{j=1}^{n} \langle a_{ij}, l_{ij}, r_{ij} \rangle x_j \le \langle b_i, u_i, v_i \rangle, \ i \in N_m \\ x_j \ge 0, \ j \in N_n \end{cases} \end{aligned} \tag{3.3}$$

where $\tilde{A}_{ij} = \langle a_{ij}, l_{ij}, r_{ij} \rangle$ and $\tilde{B}_i = \langle b_i, u_i, v_i \rangle$ are fuzzy numbers. According to Definition 2.9, the constraint $\sum_{j=1}^{n} \langle a_{ij}, l_{ij}, r_{ij} \rangle x_j \le \langle b_i, u_i, v_i \rangle, i \in N_m$ yields that:

$$\sum_{j=1}^{n} a_{ij} x_j \le b_i, \ \sum_{j=1}^{n} (a_{ij} - l_{ij}) x_j \le b_i - u_i \ \text{ and } \ \sum_{j=1}^{n} (a_{ij} + r_{ij}) x_j \le b_i + v_i, i \in N_m$$

Substituting these relations in to (3.3) the conventional quadratic program is derived as follows:

$$\begin{aligned} &\max z = \sum_{j=1}^{n} c_j x_j + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij} x_i x_j \\ &\sum_{j=1}^{n} a_{ij} x_j \le b_i, \ i \in N_m \\ &\sum_{j=1}^{n} (a_{ij} - l_{ij}) x_j \le b_i - u_i, \ i \in N_m \\ &\sum_{j=1}^{n} (a_{ij} + r_{ij}) x_j \le b_i + v_i, \ i \in N_m \\ &x_j \ge 0, \ j \in N_n \end{aligned} \tag{3.4}$$

However, since all numbers involved are real numbers, this is classical quadratic programming problem which can be solved using MATLAB™ toolbox. The SQP algorithm is used as an optimization method to minimize the nonlinear constrained optimization problem. This method is described as follow:

3.3. The SQP Algorithm

SQP is an iterative analytical nonlinear programming method. This technique begins from an initial point to find a solution using the gradient based information. This optimization method is found [21] faster than other population based search algorithms. Although the SQP method is highly dependent on the initial estimate of the solution [22] [23], this has successfully been applied in some [24] [25] optimal control problems. The SQP method is based on an iterative formulation together with the solution of some other quadratic programming sub-problems. An

optimization problem in the SQP method is considered as follows:

$$\begin{aligned} &\text{minimize } J(x) \\ &\text{subjected to}: \psi_i(x) \le 0,\ i = 1, \cdots, m \end{aligned} \tag{3.5}$$

where $J(x)$ is the cost function and $\psi_i(x)$ stands for the constraint. In this regard a Lagrangian function $L(x, \lambda)$ is constructed in terms of the Lagrangian multiplier λ_i. The cost function together with the above constraint is defined as follows:

$$L(x, \lambda) = J(x) + \sum_{i=1}^{m} \lambda_i \psi_i(x) \tag{3.6}$$

In fact the SQP consists of three main parts:

1- Update the Hessian of the Lagrangian function according to:

$$\begin{aligned} &H_{k+1} = H_k + \frac{q_k q_k^{\mathrm{T}}}{q_k^{\mathrm{T}} S_k} - \frac{H_k^{\mathrm{T}} H_k}{S_k^{\mathrm{T}} H_k S_k} \\ &H_0 = I \\ &S_k = X_{k+1} - X_k \\ &q_k = \nabla f(X_{k+1}) + \sum_{i=1}^{n} \lambda_i \nabla g_i(X_{k+1}) - \left(\nabla f(X_k) + \sum_{i=1}^{n} \lambda_i \nabla g_i(X_k) \right) \end{aligned} \tag{3.7}$$

2- Solve the quadratic programming sub-problem:

$$\begin{aligned} &\min \frac{1}{2} q_k^{\mathrm{T}} H_k d_k + \nabla f(x_k)^{\mathrm{T}} d_k \\ &\nabla \psi_i(x_k)^{\mathrm{T}} d_k + \psi_i(x_k) = 0,\ \ i = 1, \cdots, m \\ &\nabla \psi_i(x_k)^{\mathrm{T}} d_k + \psi_i(x_k) \ge 0,\ \ i = 1, \cdots, m \end{aligned} \tag{3.8}$$

3- A linear search to find a solution for the next iteration:

$$X_{k+1} = X_k + \alpha d_k \tag{3.9}$$

The algorithm is repeated until a stopping criterion (either maximum iteration or convergence criterion) is met. It must be mentioned that the SQP algorithm is a gradient based algorithm. Generally gradient based methods have the possibility of getting trapped at local optimum depending on the initial guess of the solution. In order to achieve a good final result, these methods require very good initial guesses of the solution. Since the matrix Q is supposed symmetric $(q_{ij} = q_{ji})$ and semidefinite, the objective function is convex and thus the SQP algorithm yields the global optimum solution. The corresponding theorem is presented as follows:

Theorem 3.1: In mathematical terminology, $f(x_1, x_2, \cdots, x_n)$ is convex if and only if its $n \times n$ Hessian matrix is positive semi definite for all possible values of $(x_1, x_2, \cdots, x_n)$. That is for any $x \ge 0$ the following relation is satisfied:

$$(x_1, x_2, \cdots, x_n) \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_n x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \ge 0,\ \forall (x_1, x_2, \cdots, x_n) \in \mathbb{R}^n$$

in which $H(x)=\begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}=\left[\frac{\partial^2 f}{\partial x_j \partial x_i}\right]_{n\times n}$ is the Hessian matrix of function f. Hence the above relation can be rewritten as follow:

$$x^{\mathrm{T}} H(x) x \geq 0, \forall x \in \mathbb{R}^n$$

Proof: See in [26].

4. An Example

In this section, we utilize an example to illustrate the solution method proposed in this paper. Consider a river from which diversions are made to three water-consuming firms that belong to the same corporation, as illustrated in **Figure 1** Each firm makes a product, and is the critical resource. Water is needed in the process of making that product, and it is critical resource. The three firms can be denoted by the index $j=1,2,3$ and their water allocations by x_j. Assume the problem is to determine the allocations x_j of water to each of three firms $(j=1,2,3)$ that maximize the total net benefits, $\sum_j NB_j(x_j)$, obtained from all three firms. The total amount of water available is constrained or limited to a quantity of *Q*. Assume the net benefits $NB_j(x_j)$, derived from water x_j allocated to each firm j, are defined by:

$$NB_1(x_1)=x_1+0.5x_1^2 \tag{4.1}$$

$$NB_2(x_2)=2x_2+x_2^2 \tag{4.2}$$

$$NB_3(x_3)=x_3+1.5x_3^2 \tag{4.3}$$

The problem is to find the allocations of water to each firm that maximize the total benefits $TB(X)$:

$$TB(X)=NB_1(x_1)+NB_2(x_2)+NB_3(x_3) \tag{4.4}$$

These allocations cannot exceed the amount of water available, *Q*, less any that must remain in the river, *R*. Assuming the available flow for allocations, $Q-R$, is 4. The crisp optimization problem is to maximize Equation (4.4) subject to the resource constraint:

$$x_1+x_2+x_3 \leq 4 \tag{4.5}$$

Thus the problem is:

$$\begin{aligned} &\max TB(X)=\left(x_1+0.5x_1^2\right)+\left(2x_2+x_2^2\right)+\left(x_3+1.5x_3^2\right) \\ &\text{s.t.}\ \begin{cases} x_1+x_2+x_3 \leq 4 \\ x_1,x_2,x_3 \geq 0 \end{cases} \end{aligned} \tag{4.6}$$

The precise quantification of many system performance criteria and parameter and decision variables is not always possible, nor is it always necessary. When the values of variable cannot be precisely specified, they are said to be uncertain

or fuzzy. If the values are uncertain, probability distributions may be used to quantify them. Alternatively, if they are best described by qualitative adjectives, such as dry or wet, hot or cold, clean or dirty, and high or low, fuzzy membership function can be used to quantify them. Both probability distribution and fuzzy membership functions of these uncertain or qualitative variables can be included in quantitative optimization models. Now we illustrate how fuzzy descriptors can be incorporated into optimization models of water resources systems. Assuming the available flow for allocations, $Q-R$, is not certainly known and is represented by an interval $\langle 4,2,1.5\rangle$. Thus the problem turns to the fuzzy quadratic programming problem as follows:

$$\begin{aligned} &\max TB(X)=\left(3x_1+x_1^2\right)+\left(2x_2+x_2^2\right)+\left(x_3+1.5x_3^2\right)\\ &\text{s.t.}\begin{cases} x_1+x_2+x_3\le\langle 4,2,1.5\rangle\\ x_1,x_2,x_3\ge 0\end{cases}\end{aligned} \tag{4.7}$$

This problem is in the form of model (3.2). Hence it can be solved using the proposed method. Since the parameter b_1, is triangular fuzzy number, thus the objective value of the problem should be fuzzy number as well. According to the proposed method the fuzzy solution obtained by solving the following program:

$$\begin{aligned} &\max TB(X)=\left(3x_1+x_1^2\right)+\left(2x_2+x_2^2\right)+\left(x_3+1.5x_3^2\right)\\ &\text{s.t.}\begin{cases} x_1+x_2+x_3\le 4\\ x_1+x_2+x_3\le 2\\ x_1+x_2+x_3\le 5.5\\ x_1,x_2,x_3\ge 0\end{cases}\end{aligned}$$

where parameter values are all known constant. Thus this model is conventional quadratic programming problems. By solving this problem using SQP algorithm the global optimum solution is obtained as:

$$X=(2,0,0)$$

The value of objective function is also achieved $z^*=10$.

5. Conclusion

This paper generalized the conventional quadratic programming of constant parameters to fuzzy parameters and the range of optimal objective values produced from the fuzzy parameters, including constraint coefficient and right-hand sides. The idea is to transform the defined fuzzy quadratic programming problem in to a conventional quadratic problem. Then the classical programming problem is solved using SQP algorithm and as a result, the optimal solution and the optimal value of the objective function are obtained. We emphasize that this work can be extended to the other practical situations as well as water resource management and etc. Also, as we mentioned in our study, in the mentioned model we assumed the type of fuzzy number is triangular while in the many real situations it is a limitation for our study and hence we suggest the interested readers focus on the other kinds of fuzzy numbers and then they may achieve some new results.

Our other suggestion is establishing a new rule for fuzzy ordering.

Acknowledgements

The authors would like to thank the anonymous referees for their valuable comments that help us to improve the earlier version of this work.

References

[1] Floudas C.A., Pardalos P., Adjiman C., Esposito W.R., Gumus Z.H., Harding S.T., Klepeis J.L., Meyer C.A. and Sshweiger, C.A. (1999) Handbook of Test Problems in Local and Global Optimization. Non convex Optimization and its Applications, Vol. 33, Kluwer Academic Publishers, Dordrecht.

[2] Gould, N.I.M. and Toint, P.L. (2000) A Quadratic Programming Bibliography. RAL Numerical Analysis Group Internal Report 2000-1, Department of Mathematics, University of Namur, Bruxelles, Belgium.

[3] Hock, W. and Schittkowski, K. (1981) Test Examples for Nonlinear Programming Codes. Lecture Notes in Economics and Mathematical Systems, Vol. 187, Spring-Verlag, Berlin.

[4] Maros, I. and Meszaros, C. (1997) A Repository of Convex Quadratic Problems. Department Technical Report DOC 97/6, Department of Computing, Imperial College, London, UK.

[5] Schittkowski, K. (1987) More Test Examples for Nonlinear Programming Codes. Lecture Notes in Economics and Mathematical Systems, Vol. 282, Spring-Verlag, Berlin.

[6] Beale, E. (1959) On Quadratic Programming. *Naval Research Logistics Quarterly*, **6**, 227-244. https://doi.org/10.1002/nav.3800060305

[7] Bellman, R.E. and Zadeh, L.A. (1970) Decision-Making in a Fuzzy Environment. *Management Science*, **17**, 164. https://doi.org/10.1287/mnsc.17.4.B141

[8] Markowitz, H. (1952) Portfolio Selection. *The Journal of Finance*, **7**, 77-91.

[9] Markowitz, H.M. (1991) Portfolio Selection: Efficient Diversification of Investments. 2nd Edition, Blackwell Publisher.

[10] Inuiguchi, M. and Ramik, J. (2000) Possibilistic Linear Programming: A Brief Review of Fuzzy Mathematical Programming and a Comparison with Stochastic Programming in Portfolio Selection Problem. *Fuzzy Sets and Systems*, **111**, 3-28.

[11] Leon, T. and Liercher, E. (2000) Viability of Infeasible Portfolio Selection Problems: A Fuzzy Approach. *European Journal of Operational Research*, **139**, 178-189.

[12] Tanaka, H., Guo, P. and Turksen, B. (2000) Portfolio Selection Based on Fuzzy Probabilities and Possibility Distributions. *Fuzzy Sets and Systems*, **111**, 387-397.

[13] Verchur, E., Bermudez, J.D. and Segura, J.V. (2007) Fuzzy Portfolio Optimization under Downside Risk Measures. *Fuzzy Sets and Systems*, **158**, 769-782.

[14] Watada, J. (1997) Fuzzy Portfolio Selection and Its Applications to Decision Making. *Tatra Mountains Mathematics Publications*, **13**, 219-248.

[15] Mahdavi-Amiri, N., Nasseri, S.H. and Yazdani Cherati, A. (2009) Fuzzy Primal Simplex Algorithm for Solving Fuzzy Linear Programming Problems. *Iranian Journal of Operational Research*, **2**, 68-84.

[16] Nasseri, S.H. and Ebrahimnejad, A. (2010) A Fuzzy Primal Simplex Algorithm and Its Application for Solving Flexible Linear Programming Problems. *European Jour-*

nal of Industrial Engineering, **4**, 327-389. https://doi.org/10.1504/EJIE.2010.033336

[17] Nasseri, S.H. and Mahdavi-Amiri, N. (2009) Some Duality Results on Linear Programming Problems with Symmetric Fuzzy Numbers. *Fuzzy Information and Engineering*, **1**, 59-66. https://doi.org/10.1007/s12543-009-0004-2

[18] Nasseri, S.H., Attari, H. and Ebrahimnejad, A. (2012) Revised Simplex Method and Its Application for Solving Fuzzy Linear Programming Problems. *European Journal of Industrial Engineering*, **6**, 259-280. https://doi.org/10.1504/EJIE.2012.046670

[19] Lio, S.T. (2009) Quadratic Programming with Fuzzy Parameters. *Chaos*, *Solitons and Fractals*, **40**, 237-245.

[20] Lio, S.T. (2009) A Revisit to Quadratic Programming with Fuzzy Parameters. *Chaos*, *Solitons and Fractals*, **41**, 1401-1407.

[21] Modares, H. and NaghibiSistani, M.B. (2011) Solving Nonlinear Optimal Control Problems using a Hybrid IPSO-SQP Algorithm. *Journal of Engineering Applications of Artificial Intelligence*, **24**, 476-484.

[22] Bayon, L., Grau, J.M., Ruiz, P.M. and Suarez, M.M. (2010) Initial Guess of the Solution of Dynamic Optimization of Chemical Processes. *Journal of Mathematical Chemistry*, **48**, 27-38. https://doi.org/10.1007/s10910-009-9614-5

[23] Costa, C.B.B., Maciel, A.C. and Filho, R. (2005) Mathematical Modeling and Optimal Control Strategy Development for an Acidic Crystallization Process. *Chemical Engineering and Processing*, **44**, 737-753.

[24] Hu, G. and Xu, S. (2009) Optimization Design of Microchannel Heat Sink Based on SQP Method and Numerical Simulation. *Applied Conference on Superconductivity and Electromagnetic Devises*, ASEMD, 89-92.

[25] Wang, R., Wang, P. and Zhu, Y. (2011) Study on Optimization of Isolated Heat Pipe Heat Exchange System Based on SQP. *International Conference on Electric Information and Control Engineering*.

[26] Luenberger, G.D. (2008) Linear and Nonlinear Programming. 3rd Edition, Springer, New York.

Journal of Applied Mathematics and Physics, 2017, 5, 2228-2243
http://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Sign-Changing Solutions for Discrete Dirichlet Boundary Value Problem

Yuhua Long, Baoling Zeng

School of Mathematics and Information Science, Guangzhou University, Guangzhou, China
Email: longyuhua214@163.com

How to cite this paper: Long, Y.H. and Zeng, B.L. (2017) Sign-Changing Solutions for Discrete Dirichlet Boundary Value Problem. *Journal of Applied Mathematics and Physics*, **5**, 2228-2243.
https://doi.org/10.4236/jamp.2017.511182

Received: September 7, 2017
Accepted: November 20, 2017
Published: November 23, 2017

Abstract

Using invariant sets of descending flow and variational methods, we establish some sufficient conditions on the existence of sign-changing solutions, positive solutions and negative solutions for second-order nonlinear difference equations with Dirichlet boundary value problem. Some results in the literature are improved.

Keywords

Sign-Changing Solution, Difference Equation, Dirichlet Boundary Value Problem, Invariant Sets of Descending Flow

1. Introduction

Let N, Z and R denote sets of all natural numbers, integers and real numbers, respectively. We consider the existence of sign-changing solutions, positive solutions and negative solutions for the following second-order nonlinear difference equation with Dirichlet boundary value problem (BVP for short)

$$\begin{cases} -\Delta^2 x(k-1) = f(k, x(k)), k \in [1,T] \\ x(0) = x(T+1) = 0 \end{cases} \tag{1.1}$$

where $T \geq 2$ is a given integer and $[1,T] := \{1,2,\cdots,T\}$, $f : [1,T] \times R \to R$ is continuous in the second variable, Δ denotes the forward difference operator defined by $\Delta x(k) = x(k+1) - x(k)$, $\Delta^2 x(k) = \Delta(\Delta x(k))$.

In recent years, many authors devoted to the study of (1.1) by employing various methods and obtained some interesting results. Here we mention a few. Employing critical point theory, Agarwal [1] established the existence results of multiple positive solutions. While the nonlinearity is discontinuous, Zhang [2] gained another new multiple solutions. Zhang and Sun [3] obtained two exis-

tence results of multiple solutions. By aid of algebra and Krasnoselskii fixed point theorem, Luo [4] investigated the existence of positive solutions.

Study on the sign-changing solutions is a very important research field both in differential equations and difference equations. As to the sign-changing solutions for differential equations, many scholars achieved excellent results [5]-[14] by making using of a variety of methods and techniques, such as Leray-Schauder degree theory, fixed point index theory, topological degree theory, invariant sets of descending flow, critical point theory and etc.. Among them, invariant sets of descending flow play an important role, which was first used by Sun [10]. However, to the authors' knowledge, there are few literatures that considered sign-changing solutions for difference equations. Making use of invariant sets of descending flow, [15] studied periodic boundary value problem

$$\begin{cases} -\Delta\left[p(k-1)\Delta x(k-1)\right]+q(k)x(k)=f\left(k,x(k)\right), & k\in[1,T] \\ x(0)=x(T), \quad \Delta x(0)=\Delta x(T). \end{cases}$$

In this paper, our purpose is to establish some sufficient conditions for the existence of solutions for (1.1). First, we will construct a functional I such that solutions of (1.1) correspond to critical points of I. Then, by using invariant sets of descending flow and Mountain pass lemma, we obtain sign-changing solutions, negative solutions and positive solutions for (1.1).

2. Preliminaries and Main Results

Given $m\geq 0$, let $G=\left\{x:[0,T+1]\to R \,\middle|\, x(0)=x(T+1)=0\right\}$ be a T-dimensional Hilbert space which is equipped with the inner product

$$\langle x,y\rangle_m=\sum_{k=1}^{T+1}\Delta x(k-1)\Delta y(k-1)+\sum_{k=1}^{T}mx(k)y(k),$$

then the norm $\|\cdot\|_m$ can be induced by

$$\|x\|_m=\left(\sum_{k=1}^{T+1}\left|\Delta x(k-1)\right|^2+\sum_{k=1}^{T}m\left|x(k)\right|^2\right)^{\frac{1}{2}}.$$

Let H be the T-dimensional Hilbert space equipped with the usual inner product $(\cdot,\cdot)$ and the usual norm $\|\cdot\|$. It is not difficult to see that G is isomorphic to H, $\|\cdot\|_m$ and $\|\cdot\|$ are equivalent. Denote $x^+=\max\{x,0\}$, $x^-=\min\{x,0\}$. Then for any $x\in H$, we find $\langle\cdot,\cdot\rangle_m\geq 0$.
Define functional $I:H\to R$ as

$$I(x)=\frac{1}{2}\sum_{k=1}^{T+1}\left|\Delta x(k-1)\right|^2-\sum_{k=1}^{T}F\left(k,x(k)\right). \tag{2.1}$$

For any $x=\left(x(1),x(2),\cdots,x(T)\right)^\tau\in H$, $I(x)$ can be rewritten as

$$I(x)=\frac{1}{2}(Ax,x)-\sum_{k=1}^{T}F\left(k,x(k)\right), \tag{2.2}$$

Here α^τ is the transpose of the vector α on H and

$$A=\begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{pmatrix}_{T\times T}.$$

In the following, we first consider the linear eigenvalue problem corresponding to (1.1)

$$\begin{cases} -\Delta^2 x(k-1)=\lambda x(k), \ k\in[1,T] \\ x(0)=x(T+1)=0 \end{cases} \tag{2.3}$$

By direct computation, we get eigenvalues of (2.3) as

$$\lambda_k=4\sin^2\frac{k\pi}{2(T+1)}, \ k=1,2,\cdots,T. \tag{2.4}$$

Denote $\{z_k\}_{k=1}^T$ be the corresponding eigenvectors of $\{\lambda_k\}_{k=1}^T$, where

$$z_k(j)=\sqrt{\frac{2}{T+1}}\sin\frac{kj\pi}{T+1}, \ k,j=1,2,\cdots,T.$$

It is obvious that $0<\lambda_k<4$, $\lambda_1=4\sin^2\frac{\pi}{2(T+1)}>0$ and $z_1(j)=\sqrt{\frac{2}{T+1}}\sin\frac{\pi j}{T+1}>0$ for all $k,j\in[1,T]$. Note that $\lambda_1,\lambda_2,\cdots,\lambda_T$ are also eigenvalues of matrix A.

Next, for $m\geq 0$, we consider BVP

$$\begin{cases} -\Delta^2 x(k-1)+mx(k)=h(k), \quad k\in[1,T] \\ x(0)=x(T+1)=0 \end{cases} \tag{2.5}$$

where $h:[1,T]\to R$. It is not hard to know that (2.5) and the system of linear algebra equations $(A+mI)x=h$ are equivalent, then the unique solution of (2.5) can be expressed by

$$x=(A+mI)^{-1}h \tag{2.6}$$

On the other side, we have

Lemma 2.1 The unique solution of (2.5) is

$$x(k)=\sum_{s=1}^{T}G_m(k,s)h(s), \ k\in[0,T+1],$$

here $G_m(k,s)$ can be written as

$$G_m(k,s)=\begin{cases} \dfrac{\left(P^{s-T-1}-P^{T+1-s}\right)\left(P^k-P^{-k}\right)}{W}, 0\leq k\leq s\leq T+1, \\ \dfrac{\left(P^{k-T-1}-P^{T+1-k}\right)\left(P^s-P^{-s}\right)}{W}, 0\leq s\leq k\leq T+1, \end{cases}$$

$$W=\left(P^{T+1}-P^{-T-1}\right)\left(P^{-1}-P\right).$$

Proof. First consider the homogeneous equation of (2.5)

$$\begin{cases} -\Delta^2 x(k-1) + mx(k) = 0, & k \in [1,T] \\ x(0) = x(T+1) = 0 \end{cases} \tag{2.7}$$

then the corresponding characteristic equation of (2.7) is

$$p^2 - (2+m)p + 1 = 0$$

Since $(2+m)^2 - 4 > 0$, which means we have

$$p_1 = \frac{2+m+\sqrt{4m+m^2}}{2}, \quad p_2 = \frac{2+m-\sqrt{4m+m^2}}{2}.$$

Two independent solutions of (2.7) can be expressed by $x_1(k) = p_1^k$ and $x_2(k) = p_2^k$. Therefore, the general solution of (2.5) is $x(k) = a_1(k)p_1^k + a_2(k)p_2^k$.

The next step is to determine coefficients $a_1(k), a_2(k)$. Now using the method of variation of constant, it follows

$$\begin{cases} \Delta a_1(k-1)p_1^k + \Delta a_2(k-1)p_2^k = 0, \\ \Delta a_1(k-1)p_1^{k+1} + \Delta a_2(k-1)p_2^{k+1} = -h(k). \end{cases}$$

Then

$$\Delta a_1(k-1) = \frac{p_2^k h(k)}{p_2 - p_1},$$

$$\Delta a_2(k-1) = -\frac{p_1^k h(k)}{p_2 - p_1}.$$

Moreover

$$a_1(k) = a_1(0) + \sum_{s=1}^{k} \frac{p_2^s h(s)}{p_2 - p_1},$$

$$a_2(k) = a_2(0) - \sum_{s=1}^{k} \frac{p_1^s h(s)}{p_2 - p_1}.$$

Thus, the general solution of (2.5) is

$$x(k) = \left[a_1(0) + \sum_{s=1}^{k} \frac{p_2^s h(s)}{p_2 - p_1} \right] p_1^k + \left[a_2(0) - \sum_{s=1}^{k} \frac{p_1^s h(s)}{p_2 - p_1} \right] p_2^k.$$

Using initial conditions, we find $a_1(0) = -a_2(0)$ and

$$a_1(0) = \sum_{s=1}^{T} \frac{\left(p_1^s p_2^{T+1} - p_2^s p_1^{T+1}\right)}{\left(p_1^{T+1} - p_2^{T+1}\right)(p_2 - p_1)} h(s)$$

Write $W = \left(p_1^{T+1} - P_2^{T+1}\right)(p_2 - p_1)$, $p_1 = p_2^{-1} = P$, then

$$x(k) = \frac{1}{W}\left[\sum_{s=1}^{T} \left(P^{s-T-1} - P^{T+1-s}\right)\left(P^k - P^{-k}\right) + \sum_{s=1}^{k} \left(P^{k-s} - P^{s-k}\right)\left(P^{T+1} - P^{-T-1}\right) \right] h(s).$$

Hence, we achieve the unique solution of (2.5)

$$x(k) = \sum_{s=1}^{T} G_m(k,s)h(s), \quad k \in [0,T+1],$$

here $G_m(k,s)$ can be written as

$$G_m(k,s)=\begin{cases}\dfrac{\left(P^{s-T-1}-P^{T+1-s}\right)\left(P^{k}-P^{-k}\right)}{W}, 0\le k\le s\le T+1,\\ \dfrac{\left(P^{k-T-1}-P^{T+1-k}\right)\left(P^{s}-P^{-s}\right)}{W}, 0\le s\le k\le T+1.\end{cases}$$

Remark 2.1 From Lemma 2.1, we have

$$G_m(k,s)=G_m(s,k)>0,\quad k,s\in[1,T].$$

Define $K_m, f_m, A_m: H\to H$ as follows

$$\begin{aligned}&(K_m x)(k)=\sum_{s=1}^{T}G_m(k,s)x(s),\ x\in H, k\in[1,T];\\ &(f_m x)(k)=f\left(k,x(k)\right)+mx(k),\ x\in H, k\in[1,T];\\ &A_m=K_m f_m,\end{aligned}$$

where $A_m: H\to H$ is a completely continuous operator. Combining (2.6) with Lemma 2.1, we achieve that $K_m=(A+mI)^{-1}$.

Remark 2.2 According to Lemma 2.1, it is not difficult to know that $\{x(k)\}_{k=0}^{T+1}$ is a solution of (1.1) if and only if $\{x(k)\}_{k=1}^{T}$ is a fixed point of A_m.

Lemma 2.2 The functional I defined by (2.1) is *Frechet* differentiable on H and $I'(x)$ has the expression $I'(x)=x-K_m f_m x$ for $x\in H$.

Proof. For any $x, y\in H$, using the mean value theorem, it follows

$$\begin{aligned}&I(x+y)-I(x)\\ &=\frac{1}{2}\sum_{k=1}^{T}\left|\Delta y(k-1)\right|^2+\sum_{k=1}^{T+1}\Delta x(k-1)\Delta y(k-1)-\sum_{k=1}^{T}f\left(k,x(k)+\theta(k)y(k)\right)y(k),\end{aligned}$$

Here $\theta(k)\in(0,1), k\in[1,T]$. As f is continuous in x, we find

$$\begin{aligned}&I(x+y)-I(x)-\langle x,y\rangle_m+\sum_{k=1}^{T}\left(f\left(k,x(k)\right)+mx(k)\right)y(k)\\ &=\sum_{k=1}^{T}\left[f\left(k,x(k)\right)-f\left(k,x(k)+\theta(k)y(k)\right)\right]y(k)+\frac{1}{2}\|y\|_m^2-\frac{1}{2}m\|y\|^2\\ &=\|y\|_m\,o(1)\end{aligned}$$

which leads to

$$\lim_{\|y\|_m\to 0}\frac{1}{\|y\|_m}\left(\sum_{k=1}^{T}\left[f\left(k,x(k)\right)-f\left(k,x(k)+\theta(k)y(k)\right)\right]y(k)+\frac{1}{2}\|y\|_m^2-\frac{1}{2}m\|y\|^2\right)=0$$

thus we can immediately conclude that I is *Frechet* differentiable on H and

$$\langle I'(x),y\rangle_m=\langle x,y\rangle_m-\sum_{k=1}^{T}\left(f\left(k,x(k)\right)+mx(k)\right)y(k)\tag{2.8}$$

On the other side, for all $x=\{x(k)\}\in H$ and $z=\{z(k)\}\in H$, there holds

$$\sum_{k=1}^{T}\Delta^2 x(k-1)z(k)=-\sum_{k=1}^{T+1}\Delta x(k-1)\Delta z(k-1).$$

Making use of the definition of inner product and Lemma 2.1, it follows

$$\begin{aligned}
&\langle x - K_m f_m x, y\rangle_m \\
&= \langle x, y\rangle_m - \sum_{k=1}^{T+1} \Delta\left(K_m f_m x\right)(k-1)\Delta y(k-1) - \sum_{k=1}^{T} m\left(K_m f_m x\right)(k) y(k) \\
&= \langle x, y\rangle_m - \sum_{k=1}^{T} \left\{-\Delta^2\left(K_m f_m x\right)(k-1) + m\left(K_m f_m x\right)(k)\right\} y(k) \\
&= \langle x, y\rangle_m - \sum_{k=1}^{T} \left(f\left(k, x(k)\right) + mx(k)\right) y(k)
\end{aligned}$$

Then $\langle I'(x), y\rangle_m = \langle x - K_m f_m x, y\rangle_m$ for all $x, y \in H$, that is to say, $I'(x) = x - K_m f_m x$.

Remark 2.3 According to Lemma 2.2 and Remark 2.2, we find that critical points of I defined on H are precisely solutions of (1.1).

Now, we give some necessary lemmas and definitions.

Definition 2.1 ([16]) Let $I \in C^1(H, R)$, I is said to be satisfied Palais-Smale condition((PS)condition for short) if every sequence $\{x_n\} \subset H$ such that $I(x_n)$ is bounded and $I'(x_n) \to 0 (n \to \infty)$ has a convergent subsequence in H.

Definition 2.2 ([17]) Assume $I \in C^1(H, R)$. If any sequence $\{x_n\}$ for which $I(x_n)$ is bounded and $\left(1 + \|x_n\|_m\right)\|I'(x_n)\|_m \to 0 (n \to \infty)$ possesses a convergent subsequence in H, then we say that I satisfies the Cerami condition ((C) condition for short).

Lemma 2.3 (Mountain pass lemma [16]) Let H be a real Hilbert space, assume that $I \in C^1(H, R)$ satisfies the (PS) condition and the following conditions:

(H_1) There exist constants $\rho > 0$ and $\alpha > 0$ such that $I(x) \geq a$ for all $x \in \partial B_\rho$.

(H_2) There exists $x_0 \notin B_\rho$ such that $I(x_0) \leq 0$.

Then I has a critical value $c \geq a$, moreover, c can be characterized as

$$c = \inf_{h \in \Gamma} \max_{s \in [0,1]} I\left(h(s)\right),$$

here

$$\Gamma = \left\{h \in \left(C[0,1], H\right) \middle| h(0) = 0, h(1) = x_0\right\},$$

B_ρ be the open ball in H with radius ρ and centered at 0, ∂B_ρ denote boundary of B_ρ.

Lemma 2.4 ([11]) Let H be a Hilbert space, there are two open convex subsets B_1 and B_2 on H with $A_m(\partial B_1) \subset B_1$, $A_m(\partial B_2) \subset B_2$ and $B_1 \cap B_2 \neq \phi$. If $I \in C^1(H, R)$ satisfies the (PS) condition and $I'(x) = x - A_m x$ for all $x \in H$. Assume there is a path $g : [1, T] \to H$ such that

$g(0) \in B_1 \setminus B_2$, $g(1) \in B_2 \setminus B_1$

and

$$\inf_{x \in \overline{B_1} \cap \overline{B_2}} I(x) > \sup_{\tau \in [0,1]} I\left(g(\tau)\right)$$

then I has at least four critical points, one in $H \setminus \left(\overline{B_1} \cup \overline{B_2}\right)$, one in $B_1 \setminus \overline{B_2}$, one in $B_2 \setminus \overline{B_1}$, and one in $B_1 \cap B_2$.

Remark 2.4 By Theorem 5.1 [17], we can replace (PS) condition by weaker (C) condition in Lemma 2.4.

Throughout this paper, we assume that

(J_1) $f_0 = \max_{k\in[1,T]} \limsup_{u\to 0} \left|\frac{f(k,u)}{u}\right| < \lambda_1$

(J_2) $\lim_{|u|\to\infty} \frac{f(k,u)}{u} = r$ for $k\in[1,T]$ where $r\in(0,+\infty)$ is a constant, or $r=+\infty$, $\upsilon>2$ and $C>0$ satisfy

$$\left|f(k,u)\right| \le C\left(1+|u|^{\upsilon-1}\right).$$

(J_3) (i) $\lim_{|u|\to\infty}\left[uf(k,u)-2F(k,u)\right] = -\infty, \forall k\in[1,T]$

or

(ii) $\lim_{|u|\to\infty}\left[uf(k,u)-2F(k,u)\right] = +\infty, \forall k\in[1,T]$.

where $F(k,u)=\int_0^u f(k,s)\mathrm{d}s$.

At last, we state our main results as following.

Theorem 2.1 Suppose (J_1) and (J_2) and $r>\lambda_2$. Then one has the following.

(i) If $r\in(\lambda_2,+\infty)$ is not an eigenvalue of (2.3), then (1.1) has at least three nontrivial solutions, one sign-changing, one positive and one negative.

(ii) If r is an eigenvalue of (2.3) and (J_3) holds, then the conclusion of (i) is true.

Theorem 2.2 If $\liminf_{|u|\to\infty} \frac{f(k,u)}{u} > \lambda_1$ and $\liminf_{u\to 0} \frac{f(k,u)}{u} < \lambda_1$ for all $k\in[1,T]$. Then (1.1) has at least two nontrivial solutions, one negative and one positive.

From Theorem 2.2, we can get

Corollary 2.3 Suppose $f(k,0)=0$ for any $k\in[1,T]$, we have:

(i) If $\liminf_{u\to-\infty} \frac{f(k,u)}{u} > \lambda_1$ and $\liminf_{u\to 0^-} \frac{f(k,u)}{u} < \lambda_1$ for any $k\in[1,T]$, then (1.1) has at least a negative solution.

(ii) If $\liminf_{u\to+\infty} \frac{f(k,u)}{u} > \lambda_1$ and $\liminf_{u\to 0^+} \frac{f(k,u)}{u} < \lambda_1$ for any $k\in[1,T]$, then (1.1) has at least a positive solution.

Our results improve previous work in the following way:

(1) [1] [2] [3] [4] considered Dirichlet boundary value problem, but it is unknown whether the solutions are sign-changing. While in this paper, the nonlinear term f can change sign.

(2) The nonlinearity f satisfies classical Ambrosett-Rabinowitz superlinear condition in [11] [12] [13] or locally Lipschitz continuity in [7] [8] [14], which are not used in our results.

3. Existence of Sign-Changing Solutions of (1.1)

In this section, we shall make use of Lemma 2.4 to complete the proof of Theo-

rem 2.2. Let convex cones $\Lambda=\{x\in H: x\geq 0\}$ and $-\Lambda=\{x\in H: x\leq 0\}$. The distance respecting to $\|\cdot\|_m$ in H is written by $dist_m$. For arbitrary $\varepsilon>0$, we denote

$$B_\varepsilon^+=\{x\in H: dist_m(x,\Lambda)<\varepsilon\},\quad B_\varepsilon^-=\{x\in H: dist_m(x,-\Lambda)<\varepsilon\}$$

then $B_\varepsilon^+, B_\varepsilon^-$ are open convex subsets on H with $B_\varepsilon^+\cap B_\varepsilon^-\neq\phi$. In addition, $H\setminus\left(B_\varepsilon^+\cup B_\varepsilon^-\right)$ contains only sign-changing functions.

Lemma 3.1 Suppose one of the following conditions holds.

(i) $r=+\infty$.

(ii) $r<+\infty$ is not an eigenvalue of (2.3), here r is defined by (J_2).

Then the functional I defined by (2.1) satisfies (PS) condition.

Proof. (i) Assume $r=+\infty$. Let $\{x_n\}\subset H$ be a (PS) sequence, *i.e.*, $I(x_n)$ is bounded and $I'(x_n)\to 0$ as $n\to\infty$. Since H is a finite dimensional Hilbert space, we only need to show that $\{x_n\}$ is bounded. If $r=+\infty$, choosing a constant $\gamma>0$, we have $F(k,u)\geq\lambda_T u^2-\gamma$ for all $(k,u)\in[1,T]\times R$. Then

$$\begin{aligned} I(x_n)&=\frac{1}{2}(Ax_n,x_n)-\sum_{k=1}^{T}F(k,x_n(k))\leq\frac{1}{2}\lambda_T\|x_n\|^2-\lambda_T\|x_n\|^2+T\gamma\\ &=-\frac{1}{2}\lambda_T\|x_n\|^2+T\gamma \end{aligned}\tag{3.1}$$

furthermore,

$$\|x_n\|^2\leq\frac{2T\gamma-2I(x_n)}{\lambda_T}.$$

Since $I(x_n)$ is bounded, we conclude that $\{x_n\}$ is a bounded sequence and (PS) condition is satisfied.

(ii) suppose $r<+\infty$ is not an eigenvalue of (2.3). We are now ready to prove that $\{x_n\}$ is bounded. Arguing by contradiction, we suppose there is a subsequence of $\{x_n\}$ with $\rho_n=\|x_n\|\to+\infty(n\to+\infty)$ and for each $k\in[1,T]$, either $\{x_n(k)\}$ is bounded or $x_n(k)\to+\infty$. Put $y_n=\frac{x_n}{\rho_n}$. Clearly, $\|y_n\|=1$. Then there have a subsequence of $\{y_n\}$ and $y\in H$ satisfying that $y_n\to y$ as $n\to\infty$. Write

$$d_n=\left(\frac{f(1,x_n(1))}{x_n(1)}y_n(1),\frac{f(2,x_n(2))}{x_n(2)}y_n(2),\cdots,\frac{f(T,x_n(T))}{x_n(T)}y_n(T)\right).$$

Since $\lim\limits_{|u|\to\infty}\frac{f(k,u)}{u}=r$ for all $k\in[1,T]$ and $I'(x_n)=x_n-K_m f_m x_n$, we get

$$\frac{I'(x_n)}{\rho_n}=y_n-\frac{1}{\rho_n}K_0 f_0 x_n=y_n-K_0 d_n\to y-K_0 ry.$$

Because of $\frac{I'(x_n)}{\rho_n}\to 0$ as $n\to\infty$, we have $y-K_0 ry\to 0$. In view of Lemma 2.2, we find that r is an eigenvalue of matrix A, which contradicts to the assumption. So $\{x_n\}$ is bounded and the proof is finished.

Lemma 3.2 I satisfies (C) condition under (J_3).

Proof . First assume (J_3) (i) be satisfied. There exists $M_1 > 0$, if $\{x_n\} \subset H$ be a sequence such that $I(x_n) \le M_1$ and $(1+\|x_n\|_m)\|I'(x_n)\|_m \le M_1$, there holds

$$\begin{aligned} &-3M_1 \le 2I(x_n) - (1+\|x_n\|_m)\|I'(x_n)\|_m \le 2I(x_n) - \langle I'(x_n), x_n\rangle_m \\ &= \sum_{k=1}^{T}\left[x_n(k) f(k, x_n(k)) - 2F(k, x_n(k))\right] \end{aligned} \tag{3.2}$$

Then we claim $\{x_n\}$ is bounded. Actually, if $\{x_n\}$ is unbounded, there possesses a subsequence of $\{x_n\}$ and some $k_0 \in [1,T]$ satisfying $|x_n(k_0)| \to +\infty\ (n \to \infty)$. According to ($J_3$) (i), we get

$$\begin{aligned} &\sum_{k=1}^{T}\left[x_n(k) f(k, x_n(k)) - 2F(k, x_n(k))\right] \\ &\le \left[x_n(k_0) f(k, x_n(k_0)) - 2F(k, x_n(k_0))\right] + (T-1)R \to -\infty \end{aligned}$$

and there has a positive constant $M_2 > 0$ such that $uf(k,u) - 2F(k,u) \le M_2$ for any $k \in [1,T]$ and $u \in R$. Therefore,

$$\begin{aligned} &\sum_{k=1}^{T}\left[x_n(k) f(k, x_n(k)) - 2F(k, x_n(k))\right] \\ &\le \left[x_n(k_0) f(k, x_n(k_0)) - 2F(k, x_n(k_0))\right] + (T-1)R \to -\infty \end{aligned}$$

which contradicts to (3.2). Then I satisfies (C) condition.

When (J_3) (ii) holds, we can prove I satisfies (C) condition in a similar way. Then Lemma 3.2 is verified.

Lemma 3.3 If (J_1) and (J_2) hold, there exist $m \ge 0$ and $\varepsilon_0 > 0$ such that for $0 < \varepsilon < \varepsilon_0$, we have

(i) if $x \in B_\varepsilon^-$ is a nontrivial critical point of I and $A_m(\partial B_\varepsilon^-) \subset B_\varepsilon^-$, then x is a negative solution of (1.1);

(ii) if $x \in B_\varepsilon^+$ is a nontrivial critical point of I and $A_m(\partial B_\varepsilon^+) \subset B_\varepsilon^+$, then x is a positive solution of (1.1).

Proof. (i) According to (J_1) and (J_2). For all $u \ne 0$ and $k \in [1,T]$, there exists $m \ge 0$ such that

$$u(f(k,u) + mu) > 0. \tag{3.3}$$

Let $y = A_m(x)$, $x^+ = \max\{x, 0\}$, $x^- = \max\{-x, 0\}$ for all $x \in H$. Since

$$\|x\|_m^2 = (\lambda + m)\|x\|^2,$$

it follows $\sqrt{\lambda_1 + m} \le \|x\|_m \le \sqrt{\lambda_T + m}$ and

$$\|x^+\| = \inf_{z \in -\Lambda}\|x - z\| \le \frac{1}{\sqrt{m+\lambda_1}} \inf_{z \in -\Lambda}\|x - z\|_m = \frac{1}{\sqrt{m+\lambda_1}} dist_m(x, -\Lambda). \tag{3.4}$$

By (J_1) and (J_2), there exist constants $\tau > 0$, $C > 0$ and $\upsilon > 2$ such that

$$|f(k,u) + mu| \le (m + \lambda_1 - \tau)|u| + C|u|^{\upsilon - 1}, \quad \forall (k,u) \in [1,T] \times R. \tag{3.5}$$

Choosing a positive constant D, since $x \in H$ is finite-dimensional, we have

$$|x|_{\upsilon} := \left(\sum_{k=1}^{T}|x(k)|^{\upsilon}\right)^{\frac{1}{\upsilon}} \le D\min\{\|x\|,\|x\|_m\}, \quad \forall x \in H. \tag{3.6}$$

It is obviously that $|x|_2 = \|x\|$. Moreover, $y^+ = y - y^-$ and $y^- \in -\Lambda$ imply

$$dist_m(y,-\Lambda) \le \|y - y^-\|_m = \|y^+\|.$$

Making use of (3.4)-(3.6), we get

$$\begin{aligned} dist_m(y,-\Lambda)\|y^+\|_m &\le \|y^+\|_m^2 \le \langle y, y^+\rangle_m = \langle A_m x, y^+\rangle_m = \langle K_m f_m x, y^+\rangle_m \\ &= \sum_{k=1}^{T}\left[f(k,x(k)) + mx(k)\right]y^+(k) \\ &\le \sum_{k=1}^{T}\left[f(k,x^+(k)) + mx^+(k)\right]y^+(k) \\ &\le (m+\lambda_1-\tau)\|x^+\|\|y^+\| + C|x^+|_{\upsilon}^{\upsilon-1}|y^+|_{\upsilon} \\ &\le \left(\frac{m+\lambda_1-\tau}{\sqrt{m+\lambda_1}}\|x^+\| + CD^{\upsilon}|x^+|^{\upsilon-1}\right)\|y^+\|_m \\ &\le \left(\frac{m+\lambda_1-\tau}{m+\lambda_1}dist_m(x,-\Lambda) + C_1\left(dist_m(x,-\Lambda)\right)^{\upsilon-1}\right)\|y^+\|_m, \end{aligned}$$

here $C_1 = \dfrac{CD^{\upsilon}}{\sqrt{(m+\lambda_1)^{\upsilon-1}}}$. Hence

$$dist_m(y,-\Lambda) \le \frac{m+\lambda_1-\tau}{m+\lambda_1}dist_m(x,-\Lambda) + C_1\left(dist_m(x,-\Lambda)\right)^{\upsilon-1}.$$

When $C_1\left(dist_m(x,-\Lambda)\right)^{\upsilon-2} = \dfrac{\tau}{2(m+\lambda_1)}$, there holds

Since $\dfrac{2(m+\lambda_1)-\tau}{2(m+\lambda_1)} < 1$, we obtain

$$A_m(\partial B_{\varepsilon}^-) \subset B_{\varepsilon}^-, \ \forall u \in B_{\varepsilon}^-.$$

If $x \in B_{\varepsilon}^-$ is a nontrivial critical point of I, it is clear that $I'(x) = x - A_m x = 0$. It follows from (3.7) that $x \in -\Lambda \setminus \{0\}$. Combining (3.3) and remark 2.1, we have $x(k) < 0$. Consequently, x is a negative solution of (1.1).

(ii) can be discussed similarly, we only need to change y^+ to y^- to prove (ii). For simplicity, we omit its proof.

Lemma 3.4 Suppose z_1, z_2 be eigenvectors corresponding to eigenvalues λ_1, λ_2 of (2.3) and $x \in H_2 = span\{z_1, z_2\}$. If $r > \lambda_2$, then $I(x) \to -\infty$ as $\|x\|_m \to +\infty$.

Proof. (1) If $r = +\infty$. From (3.1), we can see that $I(x) \to -\infty$ as $\|x\|_m \to +\infty$ for any $x \in H$.

(2) Assume $r \in (\lambda_2, +\infty)$. For $x \in H_2$, we have $x = \varepsilon_1 z_1 + \varepsilon_2 z_2$. In general, we can suppose $(z_1, z_2) = 0$. Thus $\|x\|^2 = \langle x, x\rangle = \varepsilon_1^2\|z_1\|^2 + \varepsilon_2^2\|z_2\|^2$ and there exists ε satisfying $0 < \varepsilon < \min\{r-\lambda_1, r-\lambda_2\}$. From $\lim\limits_{|x|\to\infty}\dfrac{f(k,x)}{x} = r$ for any $k \in [1,T]$

and $x \in R$, there exists $\varsigma > 0$ such that

$$F(k,x) \geq \frac{r-\varepsilon}{2}x^2 - \varsigma.$$

Then for $x \in H_2$, it follows

$$\begin{aligned} I(x) &= \frac{1}{2}(Ax,x) - \sum_{k=1}^{T} F(k,x(k)) \\ &\leq \frac{1}{2}\left(\lambda_1\varepsilon_1^2\|z_1\|^2 + \lambda_2\varepsilon_2^2\|z_2\|^2\right) - \frac{r-\varepsilon}{2}\|x\|^2 + T\varsigma \\ &= \frac{1}{2}(\lambda_1 - r + \varepsilon)\varepsilon_1^2\|z_1\|^2 + \frac{1}{2}(\lambda_2 - r + \varepsilon)\varepsilon_2^2\|z_2\|^2 + T\varsigma. \end{aligned}$$

Since $\lambda_1 - r + \varepsilon < 0$ and $\lambda_2 - r + \varepsilon < 0$, we find $I(x) \to -\infty$ as $\|x\|_m \to +\infty$. This completes the proof.

Now we are in the position to prove Theorem 2.1 by using Lemma 2.4.

Proof of Theorem 2.1 From (3.5), we get

$$F(k,x) + \frac{m}{2}|x|^2 \leq (m + \lambda_1 - \tau)\frac{1}{2}|x|^2 + \frac{C}{\upsilon}|x|^{\upsilon},$$

which combine with (3.6) gives that

$$\begin{aligned} I(x) &= \frac{1}{2}\langle x,x\rangle_m - \sum_{k=1}^{T}\left[F(k,x(k)) + \frac{m}{2}|x(k)|^2\right] \\ &\geq \frac{1}{2}\|x\|_m^2 - \frac{m+\lambda_1-\tau}{2}\|x\|^2 - \frac{C}{\upsilon}|x|_{\upsilon}^{\upsilon} \\ &= \frac{\tau}{2(m+\lambda_1)}\|x\|_m^2 - \frac{CD^{\upsilon}}{\upsilon}|x|_m^{\upsilon}. \end{aligned}$$

It follows from (3.4) that $\|x^{\pm}\| \leq \frac{1}{\sqrt{m+\lambda_1}} dist_m(x, \mp\Lambda) \leq \frac{1}{\sqrt{m+\lambda_1}}\varepsilon_0$ for any $x \in \overline{B_{\varepsilon}^+} \cap \overline{B_{\varepsilon}^-}$. Then there has $c_0 > -\infty$ such that $\inf\limits_{u \in B_{\varepsilon}^+ \cap B_{\varepsilon}^-} I(x) = c_0$. Moreover, in view of Lemma 3.4, we can choose $R > 2\varepsilon_0$ such that $I(x) < c_0 - 1$ for all $x \in H_2$ and $\|x\|_m = R$. To apply Lemma 2.4, we define a path $g : [0,1] \to H_2$ as

$$g(s) = R\frac{z_1\cos(\pi s) + z_2\sin(\pi s)}{\|z_1\cos(\pi s) + z_2\sin(\pi s)\|_m}.$$

By direct computation, we get

$$g(0) = R\frac{z_1}{\|z_1\|_m} \in B_{\varepsilon}^+ \setminus B_{\varepsilon}^-, \quad g(1) = -R\frac{z_1}{\|z_1\|_m} \in B_{\varepsilon}^- \setminus B_{\varepsilon}^+$$

and

$$\inf_{x \in B_{\varepsilon}^+ \cap B_{\varepsilon}^-} I(x) > \sup_{\tau \in [0,1]} I(g(\tau)).$$

Combining Lemmas 3.1, 3.3 and 2.4, we find there has a critical point in $H \setminus \left(\overline{B_{\varepsilon}^+} \cup \overline{B_{\varepsilon}^-}\right)$ corresponding to a sign-changing solution of (1.1). Moreover,

we also have a critical point in $B_\varepsilon^+ \setminus \overline{B_\varepsilon^-}\left(B_\varepsilon^- \setminus \overline{B_\varepsilon^+}\right)$ corresponding to a positive solution (a negative solution) of (1.1). The proof of (i) is completed.

Notice Lemma 3.2 and Remark 2.4, the proof of (ii) is analogous to (i) and we omit it.

4. Existence of Positive Solutions of (1.1)

In this section, we are now ready to prove existence of positive solutions of (1.1) using Lemma 2.3. Denote $x^+ = \max\{x,0\}$ and $x^- = \min\{x,0\}$. Assume $f(k,0)=0$ for all $k\in[1,T]$. To prove Theorem 1.2, we consider functionals

$$I_\pm(x) = \frac{1}{2}\langle x,x\rangle_0 - \sum_{k=1}^{T} F\left(k,x^\pm\right), \quad \forall x\in H.$$

It is easy to find that critical points of the function $I_+(I_-)$ correspond to positive solutions (negative solutions) of (1.1).

Lemma 4.1 If $\liminf_{|u|\to\infty} \frac{f(k,u)}{u} > \lambda_1$ for all $k\in[1,T]$, then I_+ and I_- satisfy (PS) condition.

Proof. Suppose $\{x_n\}\subset H$ be a sequence with $I_+(x_n)$ is bounded and $I_+'(x_n)\to 0$ as $n\to+\infty$. Denote $(f_+x)(k) = f\left(k,x^+(k)\right)$ for $k\in[1,T]$ and $x\in H$. In view of (2.6) and $x^- = \min\{x,0\}\le 0$, there holds

$$\left\|x_n^-\right\|_0^2 \le \left\langle x_n, x_n^-\right\rangle_0 \le \left\langle x_n - Kf_+x_n, x_n^-\right\rangle_0 = \left\langle I_+'(x_n), x_n^-\right\rangle_0 = o(1)\left\|x_n^-\right\|_0,$$

thus $x_n^-\to 0$ as $n\to\infty$. So we claim $\{x_n^+\}$ is bounded. We assume, by contradiction, that there has a subsequence of $\{x_n\}$ with $\rho_n = \left\|x_n^+\right\|_0 \to +\infty$ as $n\to\infty$. For each $k\in[1,T]$, either $\{x_n^+\}$ is bounded or $x_n^+(k)\to+\infty$. Put $y_n = \frac{x_n^+}{\rho_n}$. then $\|y_n\|_0 = 1$. Moreover, there has a subsequence of $\{y_n\}$ and $y\in E$ satisfying $y_n\to y$ as $n\to\infty$.

Denoting $z_1>0$, the eigenvector associated with λ_1, we obtain

$$\begin{aligned}\lambda_1\sum_{k=1}^{T} x_n(k)z_1(k) &= -\sum_{k=1}^{T}\Delta^2 x_n(k-1)z_1(k) = \sum_{k=1}^{T+1}\Delta x_n(k-1)z_1(k-1)\\ &= \left\langle x_n, z_1\right\rangle_0 = \left\langle Kf_+x_n, z_1\right\rangle_0 + I_+'(x_n)z_1\\ &= \sum_{k=1}^{T} f\left(k,x_n^+(k)\right)z_1 + \left\langle I_+'(x_n), z_1\right\rangle > 0.\end{aligned}$$

Dividing by ρ_n, it follows immediately that

$$\lambda_1\sum_{k=1}^{T} y_n(k)z_1(k) = \sum_{k=1}^{T}\frac{f\left(k,x_n^+(k)\right)}{x_n^+(k)}y_n(k)z_1(k) + o(1). \tag{4.1}$$

Since $\min_{k\in[1,T]}\liminf_{|u|\to\infty}\frac{f(k,u)}{u} > \lambda_1$ and $\|y_n\|_0 = 1$, then passing to the limit in (4.1), we get a contradiction. Hence, our claim is true. Since H is finite dimensional, the above argument means that $\{x_n\}$ has a convergent subsequence.

Consequently, I_+ satisfies (PS) condition.

Similarly, it is not difficult to know that I_- satisfies (PS) condition. Lemma 4.1 is proved.

Proof of Theorem 2.2 From $\max_{k\in[1,T]}\lim_{u\to 0}\dfrac{f(k,u)}{u}<\lambda_1$, there exist $\eta>0$ and $\delta>0$ such that

$$F(k,u)=\int_0^u f(k,s)\mathrm{d}s\le\frac{1}{2}(\lambda_1-\eta)|u|^2,\quad k\in[1,T],|u|\le\delta.$$

Now if we denote $B_\delta=\{x\in H:\|x\|_0<\delta\}$, then for $x\in\partial B_\delta$, there holds

$$I_+(x)=\frac{1}{2}\langle x,x\rangle_0-\sum_{k=1}^{T}F\left(k,x^+(k)\right)\ge\frac{1}{2}\|x\|_0^2-\frac{1}{2}(\lambda_1-\eta)\|x\|^2$$
$$\ge\frac{1}{2}\|x\|_0^2-\frac{\lambda_1-\eta}{2\lambda_1}\|x\|_0^2=\frac{1}{2\lambda_1}\eta\delta^2.$$

Because of $\min_{k\in[1,T]}\liminf_{u\to\infty}\dfrac{f(k,u)}{u}>\lambda_1$, there exists a constant $\xi>0$ such that

$$\min_{k\in[1,T]}\liminf_{u\to\infty}\frac{f(k,u)}{u}>\lambda_1+\xi.$$

Then we can choose a positive constant C_2 such that $F(k,u)=\dfrac{1}{2}(\lambda_1+\xi)u^2-C_2$ for all $(k,u)\in[1,T]\times R$. If υ is sufficiently large, we obtain

$$I_+(vz_1)\le\frac{v^2}{2}\|z_1\|_0^2-\frac{v^2}{2\lambda_1}(\lambda_1+\xi)\|z_1\|_0^2+C_2=-\frac{v^2\xi}{2\lambda_1}\|z_1\|_0^2+C_2<0.$$

In view of Lemma 2.3 and 4.1, we yield that there exists $\mu\in H$ such that $I_+'(\mu)=0$ and $I_+(\mu)\ge\dfrac{1}{2\lambda_1}\eta\delta^2>0$. Hence

$$\|\mu^-\|_0^2\le\langle\mu,\mu^-\rangle_0\le\langle\mu-Kf_+\mu,\mu^-\rangle_0=\langle I_+'(\mu),\mu^-\rangle_0.$$

Consequently, $\mu^-=0$. Thus $\mu=\mu^+\ge 0$.

If $\mu(k)=0$ for some $k\in[1,T]$, we find

$$-\mu(k+1)-\mu(k-1)=-\Delta^2\mu(k-1)=f(k,u(k))=0,$$

then $\mu(k\pm1)=0$. If $\mu(k)=0$ somewhere in $k\in[1,T]$, it vanishes identically. By $I_+'(\mu)\ge\rho>0$, we obtain $\mu(k)>0$ for $k\in[1,T]$. Therefore, μ is a positive solution of (1.1).

In a similar way as above, if we consider the case of I_-, a negative solution can be obtained. Then the proof of Theorem 2.2 is finished.

5. Applications

To illustrate Theorem 2.1 and Theorem 2.2, we will give two examples.

Example 5.1 Consider BVP

$$\begin{cases} -\Delta^2 x(k-1) = \dfrac{|x(k)|-n}{|x(k)|+1} mx(k), k \in [1,T] \\ x(0) = x(T+1) = 0 \end{cases} \tag{5.1}$$

where $m > 4\sin^2 \dfrac{\pi}{T+1}$, $0 < n < \dfrac{4\sin^2 \dfrac{\pi}{2(T+1)}}{m}$.

By direct calculation, we get

$$F(k,x) = \begin{cases} \dfrac{mx^2}{2} - m(n+1)\left(x - \ln(1+x)\right), & x \geq 0, \\ \dfrac{mx^2}{2} + m(n+1)\left(x + \ln(1-x)\right), & x < 0. \end{cases}$$

and $\lim\limits_{|x|\to\infty} \left[xf(k,x) - 2F(k,x) \right] = +\infty$ for all $k \in [1,T]$. According to (2.4), we obtain

$$\lambda_1 = 4\sin^2 \frac{\pi}{2(T+1)}, \lambda_2 = 4\sin^2 \frac{\pi}{T+1}.$$

In addition, $f_0 = \max\limits_{k\in[1,T]} \limsup\limits_{x\to 0} \left| \dfrac{f(k,x)}{x} \right| = mn < \lambda_1$ and

$\lim\limits_{|x|\to\infty} \dfrac{f(k,x)}{x} = r = m > \lambda_2$. From above argument, we find all conditions of Theorem 2.1 are satisfied, thus (5.1) has at least a sign-changing solution, a positive solution and a negative solution.

For a certain case, fix $T = 2$, here $m > 4\sin^2 \dfrac{\pi}{3} = 3$, $0 < n < \dfrac{1}{m}$, then we can choose $m = 4$, $n = \dfrac{1}{5}$. After not very complicated calculation, we find

$$\left(0, \frac{3}{5}, \frac{3}{5}, 0\right), \left(0, -\frac{3}{5}, -\frac{3}{5}, 0\right), \left(0, \frac{19}{5}, -\frac{19}{5}, 0\right), \left(0, -\frac{19}{5}, \frac{19}{5}, 0\right)$$

are positive solution, sign-changing solution, sign-changing solution and negative solution of (5.1), respectively.

Remark 5.1 From above example, we can get at least three nontrivial solutions of (1.1), one sign-changing, one positive and one negative if the nonlinearity f satisfy all the conditions of Theorem 2.1.

Example 5.2 Consider BVP

$$\begin{cases} -\Delta^2 x(k-1) = x^3(k) + ax(k), k \in [1,T] \\ x(0) = x(T+1) = 0 \end{cases} \tag{5.2}$$

here $a < 4\sin^2 \dfrac{\pi}{2(T+1)}$.

From (2.4), it is easy to see that $\lambda_1 = 4\sin^2 \dfrac{\pi}{2(T+1)}$. Moreover, $f(k,0) = 0$, $\limsup\limits_{x\to 0} \dfrac{f(k,x)}{x} = a < \lambda_1$ and $\limsup\limits_{|x|\to\infty} \dfrac{f(k,x)}{x} \geq \lambda_1$ for all $k \in [1,T]$. Therefore,

it follows from Theorem 2.2 that (5.2) has at least a positive solution and a negative solution.

In the case of $T=2$, because of $a<4\sin^2\frac{\pi}{6}=1$, we can choose $a=\frac{1}{2}$. After direct computation, we get that $\left(0,\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2},0\right)$ and $\left(0,-\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2},0\right)$ are positive solution and negative solution of (5.2), respectively.

Remark 5.2 From example 5.2, it is not difficult to know that if the nonlinearity f satisfy all the conditions of Theorem 2.2, we can obtain at least a positive solution and a negative solution of (1.1).

6. Conclusion

In this manuscript, some sufficient conditions on the existence of sign-changing solutions, positive solutions and negative solutions for a class of second-order nonlinear difference equations were established with Dirichlet boundary value problem by using invariant sets of descending flow and variational methods. Our results improve some existed ones in some literatures, because we not only establish some sufficient conditions on the existence of sign-changing solutions, but also we allow the nonlinearity f to dissatisfy Ambrosett-Rabinowitz type condition or locally Lipschitz continuity and to change sign.

Acknowledgements

This work was supported by Key Laboratory of Mathematics and Interdisciplinary Sciences of Guangdong Higher Education Institute. The authors would like to thank the reviewer for the valuable comments and suggestions, thanks.

References

[1] Agarwal, R.P., Perera, K. and O'Regan, D. (2004) Multiple Positive Solutions of Singular and Nonsingular Discrete Problems via Variational Methods. *NNonlinear Analysis: Theory, Methods & Applications*, **58**, 69-73. https://doi.org/10.1016/j.na.2003.11.012

[2] Zhang, G.Q., Zhang, W.G. and Liu, S.Y. (2007) Multiplicity Result for a Discrete Eigenvalue Problem with Discontinuous Nonlinearities. *Journal of Mathematical Analysis and Applications*, **328**, 1068-1074. https://doi.org/10.1016/j.jmaa.2006.05.077

[3] Zhang, G.D. and Sun, H.R. (2010) Existence of Multipe Solutions for a Class of Boundary Value Problem of Second Order Difference Equation. *Journal of Northwest Normal University*, **46**, 11-14.

[4] Luo, L.J. (2004) Existence of Positive Solutions to Second-Order Difference Equations Boundary Value Problem. *Journal of Guangzhou University*, **3**, 501-503. (in Chinese).

[5] Xu, X.A. (2004) Multiple Sign-Changing Solutions for Some M-Point Boundary Value Problem. *Electronic Journal of Differential Equations*, **89**, 281-286.

[6] Li, F.Y., Liang, Z.P., Zhang, Q. and Li, Y.H. (2007) On Sign-Changing Solutions for Nonlinear Operator Equations. *Journal of Mathematical Analysis and Applications*,

327, 1010-1028.

[7] Zhang, K.M. and Xie, X.J. (2009) Existence of Sign-Changing Solutions for Some Asymptotically Linear Three-Point Boundary Value Problems. *Nonlinear Analysis*, **70**, 2796-2805.

[8] Zhang, Z.T. and Perera, K. (2006) Sign-Changing Solutions of Kirchhoff Type Problems via Invariant Sets of Descending Flow. *Journal of Mathematical Analysis and Applications*, **317**, 456-463.

[9] Mao, A.M. and Zhang, Z.T. (2009) Sign-Changing and Multiple Solutions of Kirchhoff Type Problems. *Nonlinear Analysis*, **70**, 1275-1287.

[10] Sun, J.X. (1984) On Some Problems about Nonlinear Operators. PhD Thesis, Shandong University, Jinan.

[11] Liu, Z.L. and Sun, J.X. (2001) Invariant Sets of Descending Flow in Critical Point Theory with Applications to Nonlinear Differential Equations. *Journal of Differential Equations*, **172**, 257-299. https://doi.org/10.1006/jdeq.2000.3867

[12] Bartsch, T. and Liu, Z.L. (2004) On a Superlinear Elliptic p-Laplacian Equation. *Journal of Differential Equations*, **198**, 149-175.

[13] Bartsch, T., Liu, Z.L. and Weth, T. (2005) Nodal Solutions of a p-Laplacian Equation. *Proceedings of the London Mathematical Society*, **91**, 129-152. https://doi.org/10.1112/S0024611504015187

[14] Dancer, E.N. and Zhang, Z.T. (2000) Fucik Spectrum, Sign-Changing and Multiple Solutions for Semilinear Elliptic Boundary Value Problems with Resonance at Infinity. *Journal of Mathematical Analysis and Applications*, **250**, 449-464. https://doi.org/10.1006/jmaa.2000.6969

[15] He, T.S., Zhou, Y.W., Xu, Y.T. and Chen, C.Y. (2015) Sign-Changing Solutions for Discrete Second-Order Periodic Boundary Value Problems. *Bulletin of the Malaysian Mathematical Sciences Society*, **38**, 181-195. https://doi.org/10.1007/s40840-014-0012-1

[16] Rabinowitz, P.H. (1986) Minimax Methods in Critical Point Theory with Applications to Differential Equations. *CBMS Regional Conference Series in Mathematics*, *American Mathematical Society*, Providence, Vol. 65.

[17] Liu, X.Q. and Liu, J.Q. (2011) On a Boundary Value Problem in the Half-Space. *Journal of Differential Equations*, **250**, 2099-2142.

Journal of Applied Mathematics and Physics, 2017, 5, 2244-2261
http://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Topological Modelling of Deep Ulcerations in Patients with Ulcerative Colitis

Ian Morilla[1,2*], Mathieu Uzzan[2], Dominique Cazals-Hatem[2], Hatem Zaag[1], Eric Ogier-Denis[2], Gilles Wainrib[3], Xavier Tréton[2]

[1]Université Paris 13, Sorbonne Paris Cité, LAGA, CNRS (UMR 7539), Laboratoire d'excellence Inflamex, F-93430, Villetaneuse, France
[2]INSERM, UMRS1149, Team "Inflammation Intestinale", Research Centre of Inflammation, Paris, France
[3]Département d'Informatique, Equipe DATA, Ecole Normale Supérieure, Paris, France
Email: *morilla@math.univ-paris13.fr

How to cite this paper: Morilla, I., Uzzan, M., Cazals-Hatem, D., Zaag, H., Ogier-Denis, E., Wainrib, G. and Tréton, X. (2017) Topological Modelling of Deep Ulcerations in Patients with Ulcerative Colitis. *Journal of Applied Mathematics and Physics*, **5**, 2244-2261.
https://doi.org/10.4236/jamp.2017.511183

Received: September 7, 2017
Accepted: November 21, 2017
Published: November 24, 2017

Abstract

Aims: Steadily the clinicians of our team in inflammatory bowel disease encounter ulcerative colitis patients that develop deep ulcers during their treatment. Currently, these practitioners are only equipped with their grade of expertise in inflammatory domains to decide what new therapy maybe use in such cases. Encouraged by the limited knowledge of this frequent pathology, we seek to determine the molecular conditions underlying the recurrent formation of deep ulcerations in certain group of patients. ***Method***: The goal of this strategy is to expose differences between groups of patients based on similarities computed by *random walk graph kernels* and performing *functional inference* on those differences. ***Results***: We apply the methodology to a cohort of eleven miRNA microarrays of ulcerative colitis patients. Our results showed how the group of ulcerative colitis patients with presence of deep ulcers is topologically more similar (0.35) than ulcerative colitis patients (0.18) to control. Such topological constraint drove functional inference to complete the information that clinicians need. ***Conclusions***: Our analyses reveal highly interpretable in the guidance of practitioners to eventually correct initial therapies of ulcerative colitis patients that develop deep ulcers. The methodology can provide them with useful molecular hypotheses necessaries prior to make any decision on the newest course of the treatment.

Keywords

Ulcerative Colitis, Deep Ulcers, Fast Random Walk Graph Kernels, Conjugate Gradient Methods, Spectral Graph Theory

1. Introduction

Acute severe ulcerative colitis (ASC) is a multifaceted complication affecting

about 25% of ulcerative colitis (UC) patients nowadays. Such a complication is a chronic threatening state often requiring emergent colectomy in case of intensive medical treatment failure. Additionally, the presence of deep ulcers expose patients to serious episodes such as sepsis, toxic mega-colon, perforation or death [1] [2] [3]. Despite many efforts, the molecular conditions leading to ulcers formation are still not clear. As showed in following sections, our graph kernel analysis provides practitioners with an excellent medical tool to approach this serious episode of inflammatory disorder. In this sense, we infer plausible hypothesis that sheds light into such a pressing medical problem and fits previous experiments reported in the literature.

Graphs naturally model many types of structured data by means of nodes and edges. While nodes are representing general entities edges describe type of relations between such entities. On the other hand, machine learning methods applied to biomedical contexts [4] [5] concern about capturing relationships between structured entities. This tight coupling is of major interest in domains like medicine, where the seek of similarity between structures, here patients, is essential in preventing and fighting diseases. Kernel algorithms [6] provide an excellent framework to measure similarity $\left(\kappa\left(o,o'\right)\right)$ between objects o and o'. Notwithstanding, some few mathematical properties must be ensured first, *i.e.*, symmetry $\left(\kappa\left(o,o'\right)=\kappa\left(o',o\right)\right)$ and positive semi-definite (p.s.d.). Kernel methods may be used both to compare nodes within the same graph [7] and in inter-graph [8] [9] comparisons. The only constraint is its interpretability since we need to capture the pith of data encapsulated by the construction of a graph while we find ways suitable for the kernel evaluation. In this paper we evaluate in a novel scenario, inflammatory bowel disease, an extension of kernel methods [10] looking for topological similarity and combine functional context with the idea of performing medical inference in ulcerative colitis (UC). The paper maybe dissected in sections, namely: Section 2 portrays the inflammatory medical issue underlying this work; Section 3 gives us a reasonable landscape of the methods: spectral graph analysis, Conjugate Gradient Methods (CGs) to calibrate random walk graph kernel, and functional inference on our topological model; section 4 confirms our approach is valid when it is used in a real cohort of 11 patients having been diagnosed with acute severe ulcerative colitis; we provide our concluding remarks in Section 5.

2. Motivation: The Deep Ulcer Problem in ASC

Practitioners and scientists based at the "Centre de Recherche sur l'Inflammation" (INSERM, UMRS1149); Université Paris-Diderot Sorbonne have recently, conducted a primary pilot study targeted to determine why some patients having been diagnosed with ulcerative colitis, an idiotypic inflammatory bowel disease, develop a haemorrhagic mucosa with deep ulceration. Indeed, ulcerative colitis is characterised by superficial inflammatory damages in the colonic mucosa. Currently, there are no pathogenic factors identified to explain the occurrence of deep ulcers in severe form of UC, such as ASC. This newest complication of the

disease is an indicator of a poor response to medical therapy. Upon multiple medical assays as well as statistical approaches (*i.e.*, supervised hierarchical clustering, etc.) aiming at establishing predictive signatures to be used as diagnostic and prognostic; such phenomenon, apparently, seems to be "stochastic" within the treatment of ASC patients.

3. Material and Methods

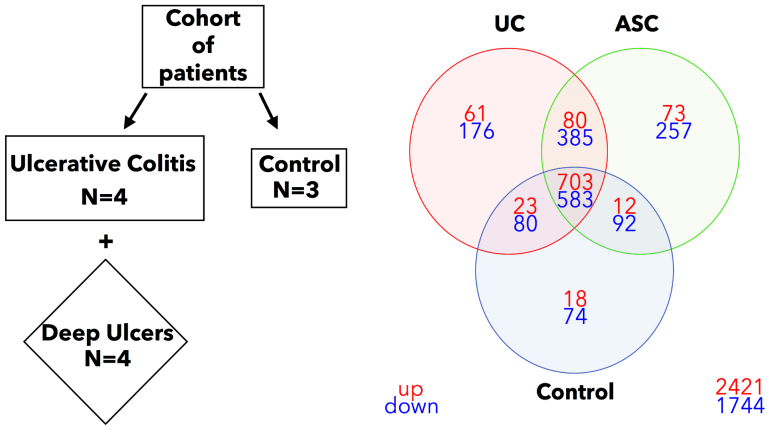
This section provides the reader with a summary description of the three constituent methods, *i.e.*, spectral graph theory, CGs in the efficiently computation of the graph kernel, and functional inference on topological models needed to understand the results showed in section 4.

3.1. Human Samples

All the biopsies analysed in the study were extracted from non-inflamed mucosa of the sigmoid colon. Paraffinised samples of colectomy were selected among three groups of patients: a first group consisting of four patients operated on UC in presence of deep ulcerations (ASC), what is a constituent marker of severity; a second sample made of three healthy subject with normal colonic mucosa and a last sample of four patients with refractory UC, *i.e.*, superficial inflammation without deep ulcers (**Figure 1(a)**). The extracted RNA derives from low inflammatory areas of the colon. MicroRNA (small non-coding RNA containing between 22 - 25 nucleotides) expression was measured by specific chip of microarray Affymetrix.

3.2. Differential miRNA Expression

Differential miRNA expression was performed using limma [11] by fitting a log-normal (LN) generalized linear model (GLM) that accounts for expression (mucosal) as well as group (UC/ASC).



(a) (b)

Figure 1. Human samples' scheme (a) and differential analysis of miRNA expression profiles per group of patients (b); UC, ASC and Control patients are highlighted in red, green and blue respectively.

3.3. Multi-Omic Graphs Integration

In this stage a cohort of eleven miRNA microarrays was used with the aim of co-integrating the differential miRNA expression profiles not present in the intersection in pairwise of UC patients (*i.e.*, UC, ASC and Control) and known human Protein-Protein Interaction (PPI, defined as miRNA-gene target product) from Genemania database [12]. Our approach is based on the assumption that genes with similar gene expression levels are translated into proteins that are more likely to interact. Recent works on gene expression and protein interaction data at genome-wide level expose such a conjecture: "Protein pairs encoded by co-expressed genes are much more likely to interact mutually than with any other type of proteins [13] [14]. Specifically, the rationale to transform the miRNA expression of a patient into a network is like this: We may want to represent a node in the graph for every protein encoded by a miRNA target gene provided its expression level was measured on this patient's microarray. We create an edge between two given proteins of this type if these proteins are reported as interacting by Genemania, and genes are up or down-regulated at the same time with respect to a provided measure tag (see previous subsection). Herein, no distinction is made between coding gene and protein.

3.4. Spectral Graph Properties

Briefly, we initially explore the geometric and algebraic behaviour of each co-integrated omic graph (UC, ASC and Control) by means of some few key spectral properties, namely: their spectra; *i.e.*, the eigenvalue for a given factor measures the variance in all the variables which is accounted for by that factor. Their algebraic connectivity calculated in the Laplacian matrix as its second smallest eigenvalue. Eigencentrality; *i.e.*, to weight the relative importance of a given i^{th} node in linking motifs within the graphs and defined as the i^{th} component of the eigenvector corresponding to the greatest eigenvalue; and their modularity by calculating the Fiedler's vector; *i.e.*, the vector corresponding to its algebraic connectivity [15] [16]. All the calculations were performed using MATLAB R2011a (maci64 architecture on a machine with a single 2.8 GHz processor and 8GB RAM distributed in two cores).

3.5. Fixing the Context: Reproducing Kernel Hilbert Space

Definition 1 *Lets* $M \in \mathbb{R}^{r\times c}$ *and* $M' \in \mathbb{R}^{p\times q}$ *be two real matrices, the Kronecker product* $M \otimes M' \in \mathbb{R}^{rp\times cq}$ *and column-stacking operator* $vec(M) \in \mathbb{R}^{rc}$ *are defined as*

$$M \otimes M' := \begin{bmatrix} M_{11}M' & M_{12}M' & \cdots & M_{1c}M' \\ M_{21}M' & M_{22}M' & \cdots & M_{2c}M' \\ \vdots & \vdots & & \vdots \\ M_{r1}B & M_{r2}M' & \cdots & M_{rc}M' \end{bmatrix},$$

$$vec(M) := \begin{bmatrix} M_{*1} \\ \vdots \\ M_{*c} \end{bmatrix},$$

where M_{*k} amounts the k^{th} column of M.

The Kronecker product and vec operator meet the following relationship (e.g., [17], Proposition 7.1.9):

$$vec(MNP) = \left(P^{\mathrm{T}} \otimes M\right) vec(N). \tag{1}$$

Another standard condition of the Kronecker product exploited in this work is ([17], Proposition 7.1.6):

$$\left(M \otimes M'\right)\left(N \otimes N'\right) = MN \otimes M'N'. \tag{2}$$

All these ideas are extendable to Reproducing Kernel Hilbert Spaces (RKHS). Let $\mathcal{H}$ be such a space, hence it is defined by a p.s.d. kernel $\kappa : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, where $\mathcal{Y}$ is a set of labels including the singular label ξ. $\mathcal{H}$ generates a feature map $\Phi : \mathcal{Y} \to \mathcal{H}$ satisfying $\kappa(y, y') = \langle \Phi(y), \Phi(y') \rangle_{\mathcal{H}}$ and mapping in $\mathcal{H}$ ξ to its zero element. We finally denote by $\Phi(Y)$ the matrix of a graph G associated to the feature map that enables lifting tensor algebra from $\mathcal{Y}$ to $\mathcal{H}$ [18].

3.6. General Setup of UC Graphs

Graphs G_s were constructed for each group of UC patients individually. These networks consisted of a set of n vertices $V = \{v_1, v_2, \cdots, v_n\}$ endowed with order and edges $E \subset V \times V$. The nature of the measured microRNA expression leads us to work on undirected graphs, *i.e.*, if $(v_p, v_q) \in E \Leftrightarrow (v_q, v_p) \in E$. Aditionally, $(v_p, v_p) \notin E$ for any p. Now, we define paths on those UC graphs as a sequence of indices $p_0, \cdots, p_l$ (l amounts path length) such that $v_{p_{s-1}} \sim v_{p_s}$, for all $1 \le s \le l$. Our graphs are robustly connected since a path can be traced in each direction between each pair of vertices of the graph. We also associate a weight $w_{pq} > 0$ to each edge (v_p, v_q) to capture the "strength" of an edge (v_p, v_q). Then $(v_p \nsim v_q)$ implies $w_{pq} = 0$ whereas for undirected weighted graphs we have $w_{pq} = w_{qp}$. Now, let $A := w_{qp} D^{-1}$ be the adjacency matrix[1] of our weighted graphs with D a diagonal matrix measuring the node degrees, that is, $D_{pp} = \sum_q w_{qp}$. Thus it may be used as transition matrix in a stochastic process since the sum of each of its columns is one. We transform a path on G_s into random by applying $P(p_{s+1} | p_1, \cdots, p_s) = A_{p_{s+1}, p_s}$ what generates sequences of vertices $v_{p_1}, v_{p_2}, v_{p_3}, \cdots$ proportionally linked to their weights in pairwise following the above probability. Hence, the probability of transition between any pair of vertex v_q and v_p through a path of length p can be induced by the expression $\left(A^p\right)_{pq}$. Finally, we say that two graphs $G = (V, E)$ and $G' = (V', E')$ are isomorphic ($G \cong G'$) if $(v_p, v_q) \in E$ iff $\left(g(v_p), g(v_q)\right) \in E'$, where $g : V \to V'$ is a bijection.

[1]In some others context this matrix might be differently defined, e.g., spectral graph theory.

3.7. Random Walk Graph Kernel

Henceforth, we note that all the definitions are generalised to the normalised case, whereas the edges are taken on a set with finite number of labels $\{1,2,\cdots,d\}$. In particular, we can take the induced RKHS $\mathcal{H}=\mathbb{R}^d$ endowed with the usual inner product.

Intuitive definition: Random walk graph kernel has been extensively reported in literature to classify and measure similarities of graphs [18] [19]. The rationale of this algorithm is as follows: The random walk kernel on graph counts the number of walks shared by a couple of graphs. Two walks are said to be shared if their lengths and label sequences are the same. Subsequently, the calculated number of shared walks enables to measure the similarity of the two graphs. To infer a formal definition of random walk graph kernel, we might want to present some basic concepts in direct product of graphs. The direct product of two graphs $G=\{V,E\}$ and $G'=\{V',E'\}$ is other graph, denoted by $G_\times=\{V_\times,E_\times\}$, where the node set $V_\times=\{(v_p,v_s')\,|\,v_p\in V,v_s'\in V'\}$, and the edge set $E_\times=\{((v_p,v_s'),(v_q,v_t'))\,|\,(v_p,v_q)\in E,(v_s',v_t')\in E'\}$. In particular, $G_\times$ can be associated to a weight matrix $W_\times=A\otimes A'$ (Definition 1) with non-zero entries provided the analogous edge is defined in the graph produced by the direct product. A random walk on the direct product graph $G_\times$ amounts the trace of random walks on G and G' at once. Let $c(d)$ and $c'(d')$ be the starting (stopping) probabilities of the random walks on G and G', respectively. Then, the number of shared walks of length l on the direct product graph $G_\times$ is calculated by $(d\otimes d')\left(A^T\otimes A'^T\right)^l(c\otimes c')$, where A and A' are the normalised adjacency matrices of G and G', respectively [20]. This definition enables the review of all the shared walks per each unique lengths. However, this sum might not be convergent. Thus, we introduce a non-negative coefficient of decay $\mu(l)$ to get rid of the longer walks.

Kernel definition Formally, the expression for the random walk kernel on graph is as follows:

$$k(G,G'):=\sum_{l=0}^{\infty}\mu(l)d_\times^TW_\times^lc_\times. \tag{3}$$

Hence, $c_\times:=c\otimes c'$ ($d_\times:=d\otimes d'$) is the starting (stopping) probability distribution associated to the graph produced by the direct product. Therefore, if the coefficients $\mu(l)$ assure the convergence of (3), then (3) is a valid p.s.d. kernel ([18], Theorem 3).

3.8. Conjugate Gradient Methods

We selected the conjugate gradient method for calculating the random walk kernel on our graphs since other methods such as the Sylvester or the spectral decomposition are not applicable for kernels on graphs in general [20].

The computation of a random walk kernel on graph with $\mu(l)=\lambda^l$ stands for inverting $(\mathbb{I}-\lambda W_\times)$, an $n^2\times n^2$ matrix if each graph G and G' have n vertices. Lets M and v be a matrix and a vector respectively, conjugate gradient (CG) method is used to solve systems as $Mx=v$ efficiently [20]. More general, since these methods are thought of symmetric p.s.d. matrices, CGs solve as well other linear systems efficiently. CG solvers improve their performances as the matrix has a small number of different eigenvalues, or is rank deficient. Remarkably, in cases where the matrix M is sparse the computation speed of matrix-vector products can be increased significantly [21].

The computation of the graph kernel (3) using CG maybe firstly described as the solution of the following linear system:

$$(\mathbb{I}-\lambda W_\times)x=c_\times, \tag{4}$$

for x, then we compute $d_\times^T x$. Next, it ought to contemplate proficient ways to solve (4) with the CG solver. We already know that W is a square matrix of size $n^2\times n^2$. The application of the CG method to a direct approach needs $O(n^4)$ iterations to multiply W by a vector y. However, if we exploit the above extended vec-MNP formula (1) into RKHS ([18], Lemma 12) with some new matrix $Y\in\mathbb{R}^{n\times n}$ with $y=vec(Y)$ and taking into account that in particular $W_\times=A\otimes A'$ (A and A' the normalised adjacency matrix for the graphs G and G' respectively), by ([18], Lemma 12) we can write

$$W_\times y=(A\otimes A')vec(Y)=vec(A'YA^T). \tag{5}$$

If $A\sim\Phi(\cdot)\in\mathbb{R}^d$ then we can compute the above multiplication of a matrix by a vector in time order of $O(dn^3)$. Furthermore, even more efficient computation of $A'YA^T$ is feasible provided that the matrices A and A' are sparse: Assuming that A and A' have $O(n)$ non-ξ entries, then computing (5) takes only $O(n^2)$ time.

Finally, note that the nearest Kronecker product [22] is not appropriate to approximate $W_\times$ since the number d of distinct labels in our labeled graph is not large enough.

3.9. Weisfeiler-Lehman Graph Kernels Cross-Validation

As validation of our results, we also propose to compare the random walk kernel on graph and the family of Weisfeiler-Lehman kernels. The later consists of proficient kernels to be used on graphs presenting discrete node labels. Such family is built on the Weisfeiler-Lehman test of isomorphism between graphs [23] and its valid 1-dimensional variant [24]. It captures topological and label information iteratively mapping the graph of reference onto a sequence of graphs with nodes displaying characteristic attributes. This catenation of graphs originating from the Weisfeiler-Lehman test can establish a family of kernels, including an adequate kernel to compare patterns taking subtree shape. Notice how the edges and length of such a sequence produce a final complexity in linear terms.

Definition 2 *Given the Weisfeiler-Lehman (WL) graph* $G_a=(V,E,l_a)$ *of height* a, *its sequence is denoted by*:

$$V=\{G_0,G_1,\cdots,G_h\}=\{(V,E,l_0),(V,E,l_1),\cdots,(V,E,l_h)\}, \tag{6}$$

where h counts iterations, and $\{G_0,\cdots,G_h\}$ and $\{G_0',\cdots,G_h'\}$ are respectively the sequences of G and G' associated to WL graphs.

Definition 3 *Provided the so-called base kernel* κ *is fixed, then the definition of Weisfeiler-Lehman kernel for* κ *is*

$$\kappa_{WL}^{(h)}(G,G')=\kappa(G_0,G_0')+\kappa(G_1,G_1')+\cdots+\kappa(G_h,G_h'), \tag{7}$$

where $G_0=G$ and $l_0=l$, the WL sequence up to height a of G.

Finally, $\kappa(h)$ is positive semidefinite if the base kernel κ is positive semidefinite [25], Theorem 3.

Definition 4 *Let* $\Gamma_k\subseteq\Gamma$ *be the set of node labels matching at least once in graphs* G *or* G' *at the end of the k-th iteration of the WL algorithm. We also fix* Γ_0 *as the set of original node labels of* G *and* G' *while* Γ_k *are pairwise disjoint. Then, we presume every* $\Gamma_k=\{\sigma_{k1},\cdots,\sigma_{k|\Gamma_k|}\}$ *is ordered. Define a map* $p_k:\{G,G'\}\times\Gamma_k\to\mathbb{N}$ *such that* $p_k(G,\sigma_{kl})$ *amounts the count of the letter* σ_{kl} *in a graph* G. *The* ***Weisfeiler-Lehman subtree kernel*** *on two graphs* G *and* G' *is as follows*:

$$\kappa_{WLsubtree}^{(h)}(G,G')=\left\langle\phi_{WLsubtree}^{(h)}(G),\phi_{WLsubtree}^{(h)}(G')\right\rangle, \tag{8}$$

where for G (resp. G')

$$\phi_{WLsubtree}^{(h)}(G)=\left(p_0(G,\sigma_{01}),\cdots,p_0\left(G,\sigma_{0|\Gamma_0|}\right),\cdots,p_h(G,\sigma_{h1}),\cdots,p_h\left(G,\sigma_{h|\Gamma_h|}\right)\right).$$

This algorithm basically seeks matching of vertex identifiers assuming that the corresponding subgraphs match.

Definition 5 *Provided a function* w *weighting the edges exits, we can described the corresponding base kernel* κ_E *by* $\sum_{e\in E}\sum_{e'\in E'}\delta(\alpha,\alpha')\delta(\beta,\beta')\kappa_w(w(e),w(e'))$, *where* δ *amounts Dirac kernel and* κ_w *is the similarity captured by a kernel between weights. Hence by 6, the* ***Weisfeiler-Lehman edge kernel*** *turns into*

$$\kappa_{WLedge}^{(h)}=\kappa_E(G_0,G_0')+\kappa_E(G_1,G_1')+\cdots+\kappa_E(G_h,G_h'),$$

where $\kappa_E=\langle\phi_E(G),\phi_E(G')\rangle$ and $\phi_E(G)$ is a vector of matching pairs (α,β), $\alpha,\beta\in\sum$, which amounts sorted final vertices of an edge in G.

Definition 6 *We also calculate the* ***shortest path version of the Weisfeiler-Lehman kernel****. Similarly, it is defined as*

$$\kappa_{WLshortestpath}^{(h)}=\kappa_{SP}(G_0,G_0')+\kappa_{SP}(G_1,G_1')+\cdots+\kappa_{SP}(G_h,G_h'),$$

where $\kappa_{SP}(G,G')=\langle\phi_{SP}(G),\phi_{SP}(G')\rangle$ and $\phi_{SP}(G)$ denotes a vector composed by the counts of matches for triplets (α,β,sp_l) in G/G', where $\alpha,\beta\in\sum$ are sorted final vertices of a shortest path and $sp_l\in\mathbb{N}_0$ is the shortest path length.

3.10. Inference on Random Walk Graph Kernels by Enrichment of Functional Annotations

So far, we described how to compare UC/ASC graphs, enabling the trace of the underlying similarity between them and their corresponding control samples by gene targets expression profiles from data. Now, we are interesting in performing inference on our topological model to characterise the genetic mechanisms of miRNA perturbations of gene graph in detail. In section 4, we discuss how inference schemes can be used on our estimated model to learn about downstream effects of miRNAs perturbations. We note that all of these inference schemes are based on enrichment analysis in functional annotations (calculation of Fisher's test [26] is performed to quantitatively capture the functional enrichment of genes according to their annotation terms) using the gene ontology database (GO) [27].

4. Results and discussion

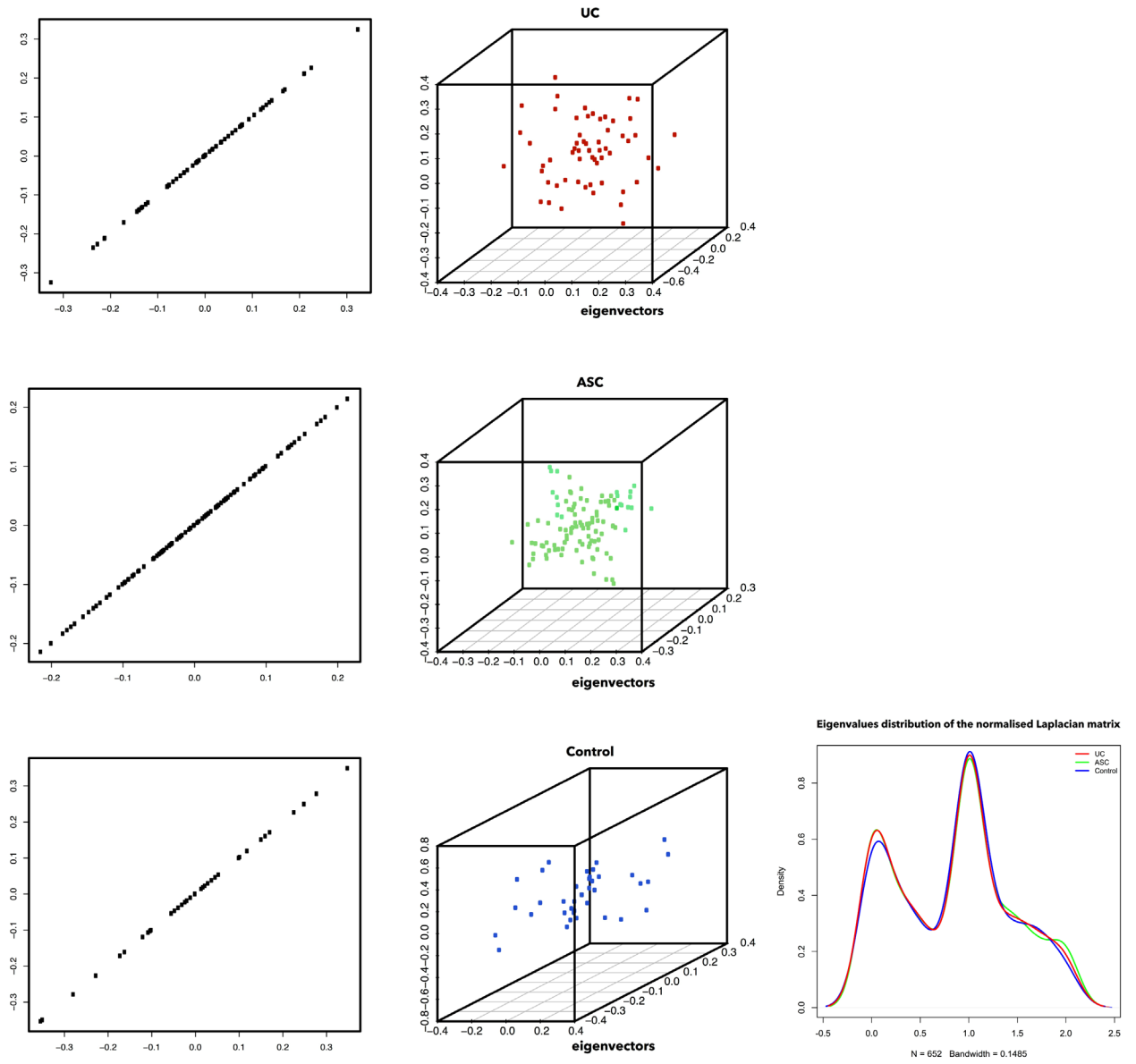
4.1. Data Integration and Spectral Behaviour between the UC Graphs

We analyse our sequence of graphs individually by comparing some algebraic characteristics.

As describe in section 3.3, we found that 2390 proteins (**Figure S1**) from Genemania [12] were reported by the gene expression levels of our miRNA microarrays (**Figure 1(b)**). The largest amount of those proteins (1071 for 330 miRNAs differentially expressed (see section 3.2) was identified in the ASC sample, whereas the UC patients sample matched in 804 (in 237 miRNAs differentially expressed); the remaining 515 (in 92 miRNAs differentially expressed) corresponded to the sample of control. These amounts seem to be consistent with the medical expectation of discovering, at a larger-scale, perturbed expression profiles involved in the pathways leading to deep ulcerations (ASC). Strikingly, the comparison of their spectra showed dissimilar conclusions; while the eigenvectors of ASC and Control patients exhibit similar patterns regarding UC patients (**Figure 2(a)**), the eigenvalue distributions of the three group of patients display the same Gaussian mixture models (**Figure 2(b)**). However, the algebraic connectivity in ASC and UC resembled each other with associated values of 21 and 18 what means almost twofold greater than the control group with a value of 10. No significant difference was detected among the remaining spectral parameters, *i.e.*, eigencentrality or simple modularity **Figure S1**. Although we enhance important algebraic and geometric characteristics of our graphs, it seems that no conclusions might be made regarding their similarities per group.

4.2. Topological Similarity between Pairs of UC Graphs by Random Walk Kernel

To measure topological similarity among our three groups of graphs, *i.e.*, UC, ASC and Control with a biological significant, we established a comparison



(a) (b)

Figure 2. Distribution of spectra per group of patients. Plots of the 2D and 3D eigenvector distributions of the laplacian matrix show how ASC (green) and Control (blue) patients exhibit similar behaviours as compared to UC (red) patients (a); However, the three groups of patients display the same type of Gaussian mixture models for their eigenvalue distributions (b).

between interacting and co-regulated groups of target genes per sample of patient. To this task a random walk kernel on graph is the appropriate selection, as for this graph a random walk amounts a set of target genes in which continuous genes by the walk side are co-expressed and interact. To efficiently compute the random walk, we made use of the CG methods using the parameter $\lambda = 0.001$ with convergence threshold set to 10^{-6}. In **Figure 3(a)** we contrast the scores of similarity measured by graph kernel computation of the conjugate gradient algorithm referred to UC patients modelled as labeled graphs with that of the direct sparse method. Our approach demonstrates how the group of ASC

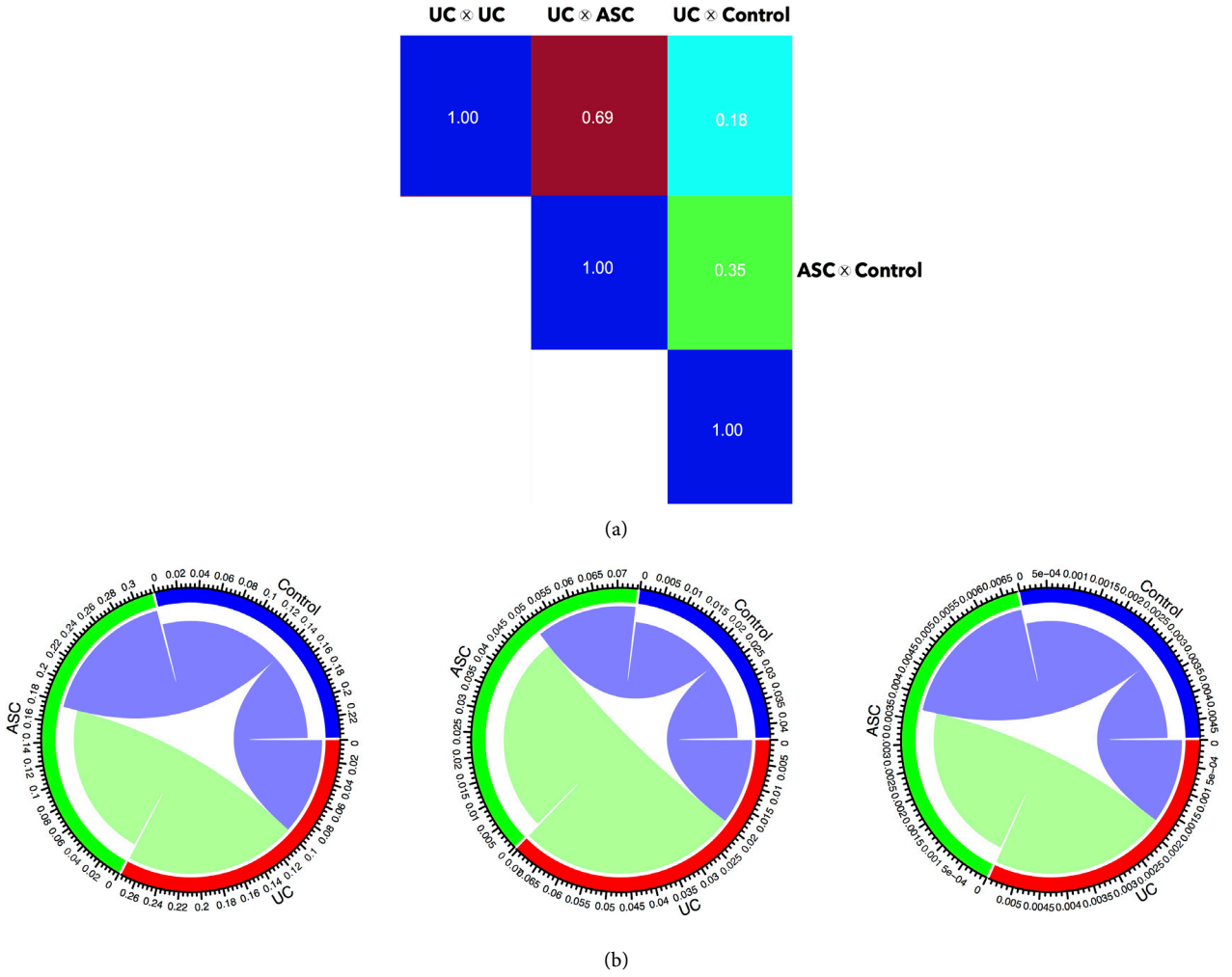


Figure 3. Scores of topological similarity between graphs of patients yielded by our random walk graph kernel ($\lambda = 0.01$ and tolerance set to 10^{-6} in its computation with conjugate gradient method). The UC and ASC groups are mutually similar the most; however the group of ASC patients resembles topologically better than UC to control (0.35/0.18) (a); Cross-validation of our results computed by the three instances of the general Weisfeiler-Lehman graph kernels, the Weisfeiler-Lehman subtree kernel, the Weisfeiler-Lehman edge kernel, and the Weisfeiler-Lehman shortest path kernel (b).

patients is topologically more similar to control patients (0.35 as normalised score $\in [0,1]$) than UC are (0.18). Here, closer to 1 means more similar graphs. We recall that the random walk kernel on graph measures the amount of walks shared by the couple of graphs involved in $G_{\times}$ (section 3.7). This topological relationship between ASC patients and their group of control is, although relatively unexpected, entirely plausible from a biological and thus medical point of view. Validation of similarities using the Weisfeiler-Lehman Graph Kernels The reliability of our results is also validated by comparing the performances of the random walk and Weisfeiler-Lehman graph kernels. The latter consists of a triplet of robust methods (see methods) in capturing topological and label information on graphs. These algorithms confirmed the same scheme described in our results, *i.e.*, ASC group is closer than UC patients to control group. Whereas a graphical visualisation of these data may

be displayed in Figure 3(b), the specific normalised (by all the possible paths on the graph) values of the pairwise comparison between patients' graphs are shown in the following Table 1.

4.3. Inference on the Topological Model: Malfunction of Lymphoid Structures Induces Deep Ulcers in UC Patients

We can perform inference on our topological model combining the similarity scores and functional enrichment analysis. Since ASC patients are topologically more similar than UC to Control (Figure 3 and Figure 5(a)), one natural idea is to explore the lack of or alternatively the low expression levels of miRNA-gene targets involved in enriched pathways from both ASC and control data with respect to UC patients. In the view of the enrichment analysis using GO (Table SI, Table SII and Table SIII) and the above inference constraint, the only enriched functional module fitting our topological model in the colon was that linked to lymphoid nodules (GO:0048541 with p-value $=2.45e^{-4}$ and q-value $=5.45e^{-2}$ associated to the Fisher Exact test). Such structures are the equivalent

Table 1. Weisfeiler-Lehman graph kernels' Validation.

Method/Graph Comparison	UC⊗ASC	UC⊗Control	ASC⊗Control
WL subtree	0.17	0.10	0.14
WL edge	0.05	0.01	0.02
WL shortest path	0.004	0.002	0.003

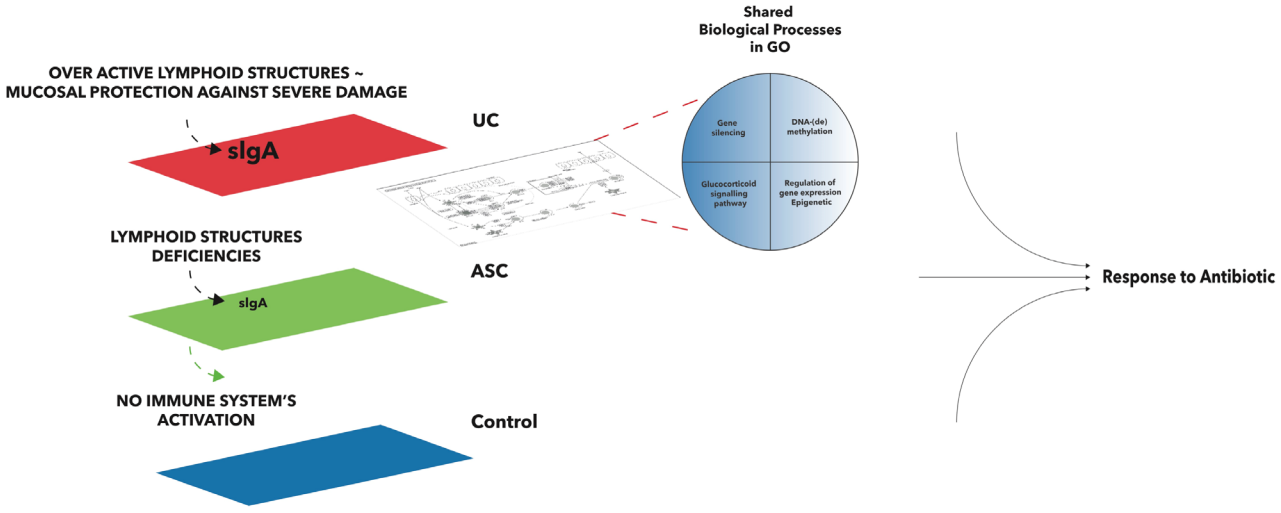


Figure 4. Inference derived from our topological model based on GO analysis of functional enrichment per group of patients. Enrichment in lymphoid nodules development is the only major difference between ASC group of patients. While the miRNA-target genes involved in lymphoid nodules pathways are over-expressed in the group of UC patients, these target genes are poorly under-expressed in ASC. This scenario resembles the molecular behaviour of Control patients (a); UC and ASC biological processes shared in GO database (b); GO biological process in common of the three group of patients after prospective drug-mediated treatment (c).

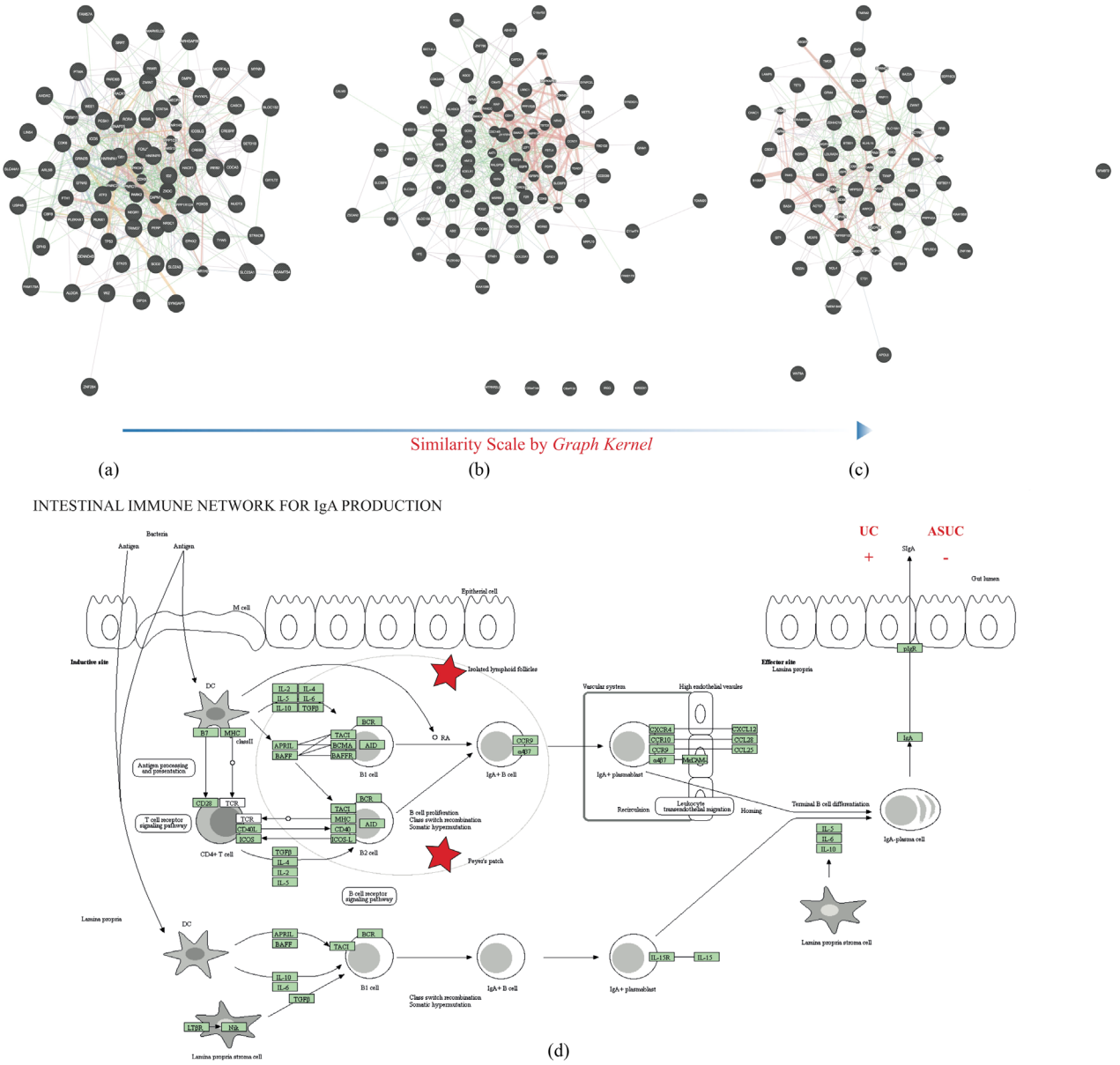


Figure 5. Description of the medical hypothesis-driven by our graph kernel analysis. Scales of similarity between group of patients, *i.e.*, ASC and Control becoming a topological constraint to be considered (a)-(c). Upon functional inference using GO, we deduce how the lack of production in sIgA/E for ASC patients (highlighted in green) prevent immune system's activation face bacteria's threat (d). This behaviour resembles the group of control maybe by a malfunction of lymphoid structures in the first line of activation in immune response. Edge colours in panels (a)-(c): purple, rose, blue and green amount to co-expression, physical interactions, co-localisation and genetic interactions respectively.

to the Peyer's patches (PPs) in the Ileum. This fact derives from the no detection of such enrichment in the production of immunoglobulins, *i.e.*, sIgA and sIgE, in ASC patients as compared to UC patients (**Figure 4**). Such a lack of production in sIgA is a consequence of the relative poorly enriched scores associated to the coding genes ID2 and STAT5, which control the intestinal immune network for sIgA production via negative regulation of class IgA/E class switching [28] and"on-off" recombination of immunoglobulin gene in developing pro-B cells

[29] (**Figure 5(d)**—starts highlighted in red). Now, we are equipped with enough information to infer the following hypothesis: There exists a very low production of immune globulin A (sIgA) within ASC patients occasioned by malfunction of lymphoid nodules. Indeed, there is no immune system's activation, whereas in UC patients we have over-expression of lymphoid nodules related pathways (**Figure 5(d)**).

The sIgA is an antibody—Y-shaped protein—that plays a critical role in immune function in the mucous membranes. This scenario matches the topological constraint yielded by our model between control and ASC patients. Furthermore, it is been already described how sIgA likely contains other propitious outcomes in overall immunity by means of a diminished inflammation in the digestive tract [30]. There is also evidence that sIgA (low/coding genes under-expression) secretion into body cavities in combination with malfunction of immune cells in PPs [31] [32] [33] is involved in allergic diseases (type 1 diabetes, Ulcerative Colitis/Crohn disease, hay fever or asthma). Thus, the formation of deep ulcers in some UC patients may be caused by the low production of sIgA as a consequence of lymphoid structures malfunction.

5. Conclusion

This paper first presented the urgent medical problem derived from the occurrence of deep ulcers during the therapy of patients with a severe chronic inflammation in the colon mucosa and how the efficiently computation of a Random walk graph kernel captures similarity between groups of these patients, namely: UC, ASC and Control. We adopt the extended linear algebra in an RKHS to overcome some issues of efficiency in kernels computations taking advantage of the shared structure intrinsic to these questions. The groups of patients were modelled as undirected labeled graphs based on the co-integration of target gene expression profiles and interaction. Thus, the nature of our data and the flexibility of conjugate gradient algorithm made of this method the most appropriate to compute geometrical random walks among other options such as spectral decomposition. We made use of models of sparsity, low effective rank, and Kronecker product to reduce the computational cost in the calculations and exploited specific forms of $W_{\times}$. While other methods of direct comparison to measure similarity like spectral properties are not conclusive; this approach reveals as much more interpretable. Indeed, our results demonstrate how the group of ASC patients topologically resembles Control better than UC patients do. In addition, we stress the reliability of our results by means of a robust triple validation. Albeit, an important caveat of our kernel approach concerns the possible values taken by the parameter λ in (3) which entirely relies on the range of $W_{\times}$ as weight matrix. We also show how the topological constraint imposed by the ASC and Control groups drives the analysis of enrichment in functional annotations enabling inference on our topological model. As a consequence, we are able to guide clinicians with a likely hypothesis regarding the low production of slgA and slgE in the ASC group to be conducted during patient's treatment.

Moreover, these results are being further validated by the clinicians and scientists of our team in the "Centre de Recherche sur l'Inflammation" as part of the future work based on this study. Specifically, we plan to performance immunofluorescence experiments, which would experimentally validate our results. we will also extend our analysis to a new cohort of patients applying improved versions of neighbour matching using deep learning models to capture similarities between graph of individual patients. Overall, this work provides practitioners with a useful and biologically meaningful tool to find similarities among patients profiles in a timely manner. Our approach allows them to avoid spending a large amount of time and effort on sweeping lots of experimental results to test eventual therapeutic hypotheses done by hand; therefore, the diagnosis efficiency and accuracy can be enhanced.

Acknowledgements

We acknowledge the financial support by Institut National de la Santé et de la Recherche Médicale (INSERM), Inserm-Transfert, Association Franois Aupetit (AFA), Université Diderot Paris 7, and the Investissements d'Avenir programme ANR-11-IDEX-0005-02 and 10-LABX-0017, Sorbonne Paris Cité, Laboratoire d'excellence INFLAMEX. IM would like to extend his thanks to Dr. Verónica G. Doblas for her invaluable discussions and ideas.

References

[1] Grainge, M.J., West, J. and Card, T.R. (2010) Role of Drug Transporters and Drug Accumulation in the Temporal Acquisition of Drug Resistance. *Lancet*, **375**, 657-663.

[2] McClements, D. and Probert, C. (2015) Managing Acute Severe Ulcerative Colitis in the Hosptialised Setting. *Frontline Gastroenterology*, **6**, 241-245.

[3] Wang, H., Vo, T., Hajar, A., Li, S., Chen, X., Parissenti, A.M., Brindley, D.N. and Wang, Z. (2014) Multiple Mechanisms Underlying Acquired Resistance to Taxanes in Selected Docetaxel-Resistant Mcf-7 Breast Cancer Cells. *BMC Cancer*, **14**, 37. https://doi.org/10.1186/1471-2407-14-37

[4] Moody, G. (2004) Digital Code of Life: How Bioinformatics is Revolutionizing Science, Medicine, and Business. Wiley, Hoboken, New Jersey.

[5] Gulshan, V., Peng, L., Coram, M., *et al.* (2016) Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, **316**, 2402-2410.

[6] Scholkopf, B. and Smola, A.J. (2002) Learning with Kernels. MIT Press, Cambridge.

[7] Kondor, R. and Lafferty, J.D. (2002) Diffusion Kernels on Graphs and Other Discrete Structures. *Proceedings of the International Conference on Machine Learning*, 315-322.

[8] Smola, A.J. and Kondor, R. (2003) Kernels and Regularization on Graphs. *Proceedings of the Annual Conference on Computational Learning Theory*, Lecture Notes in Computer Science, 144-158. https://doi.org/10.1007/978-3-540-45167-9_12

[9] Yanardag, P. and Vishwanathan, S.V.N. (2015) Deep Graph Kernels. In: *Proceedings of the* 21*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, 1365-1374.

https://doi.org/10.1145/2783258.2783417

[10] Roche-Lima, A. (2016) Implementation and Comparison of Kernel-Based Learning Methods to Predict Metabolic Networks. *Network Modeling Analysis in Health Informatics and Bioinformatics*, **5**, 26. https://doi.org/10.1007/s13721-016-0134-5

[11] Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. (2015) Limma Powers Differential Expression Analyses for Rnasequencing and Microarray Studies. *Nucleic Acids Research*, **43**, e47. https://doi.org/10.1093/nar/gkv007

[12] Warde-Farley, D., Donaldson, S.L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C.T., Maitland, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G.D. and Morris, Q. (2010) The Genemania Prediction Server: Biological Network Integration for Gene Prioritization and Predicting Gene Function. *Nucleic Acids Research*, **38**, 214-220.

[13] Fraser, H.B., Hirsh, A.E., Wall, D.P. and Eisen, M.B. (2004) Coevolution of Gene Expression among Interacting Proteins. *Proceedings of the National Academy of Science*, **24**, 9033-9038. https://doi.org/10.1073/pnas.0402591101

[14] Musungu, B.M., Bhatnagar, D., B, R.L., Payne, G.A., Brian, G.O., Fakhoury, A.M. and Geisler, M. (2016) A Network Approach of Gene Coexpression in the *Zea mays*/*Aspergillus avus* Pathosystem to Map Host/Pathogen Interaction Pathways. *Frontiers in Genetics*, **7**, 206. https://doi.org/10.3389/fgene.2016.00206

[15] Fiedler, M. (1973) Algebraic Connectivity of Graphs. *Czechoslovak Mathematical Journal*, **98**, Article ID: 298305.

[16] Brouwer, A. and Haemers, W.H. (2011) Spectral Graphs. Springer.

[17] Bernstein, D.S. (2005) Matrix Mathematics. Princeton University Press.

[18] Vishwanathan, S.V.N., Schraudolph, N.N., Kondor, R.I. and Borgwardt, K.M. (2010) Graph Kernels. *Journal of Machine Learning Research*, **11**, 1201-1242.

[19] Sugiyama, M. and Borgwardt, K. (2015) Halting in Random Walk Kernels. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M. and Garnett, R., Eds., *Advances in Neural Information Processing Systems* 28, Curran Associates, Inc., 1639-1647.

[20] Kang, U., Tong, H. and Sun, J. (2012) Fast Random Walk Graph Kernel. *Proceedings of the* 12*th SIAM International Conference on Data Mining SDM*, 828-838. https://doi.org/10.1137/1.9781611972825.71

[21] Nocedal, J. and Wright, S.J. (1999) Numerical Optimization. Springer Series in Operations Research. https://doi.org/10.1007/b98874

[22] Steeb, W.H. and Hardy, Y. (2011) Matrix Calculus and Kronecker Product: A Practical Approach to Linear and Multilinear Algebra. 2nd Edition, World Scientific Publishing Company. https://doi.org/10.1142/8030

[23] Weisfeiler and Lehman, A.A. (1968) A Reduction of a Graph to a Canonical Form and an Algebra Arising during This Reduction. *Nauchno-Technicheskaya Informatsia*, **9**.

[24] Babai, L. and Kucera, L. (1979) Canonical Labelling of Graphs in Linear Average Time. *Proceedings Symposium on Foundations of Computer Science*, 39-46. https://doi.org/10.1109/SFCS.1979.8

[25] Shervashidze, N., Schweitzer, P., van Leeuwen, E.J., Mehlhorn, K. and Borgwardt, K.M. (2011) Weisfeiler-Lehman Graph Kernels. *Journal of Machine Learning Research*, **12**, 2539-2561.

[26] Sprent, P. (1993) Applied Nonparametric Statistical Methods. 2nd Editon, Chapman and Hall, London.

[27] Consortium, G.O. (2017) Expansion of the Gene Ontology Knowledgebase and Resources. *Nucleic Acids Research*, **45**, D331-D338.

[28] Sugai, M., Gonda, H., Kusunoki, T., Kataki, T., Yokota, Y. and Shimizu, A. (2002) Essential Role of id2 in Negative Regulation of Ige Class Switching. *Nature Immunology*, **4**, 25-29.

[29] Malin, S., McManus, S., Cobaleda, C., Novatchkova, M., Delogu, A., Bouil-let, P., Strasser, A. and Busslinger, M. (2010) Role of stat5 in Controlling Cell Survival and Immunoglobulin Gene Recombination during Pro-b Cell Development. *Nature Immunology*, **11**, 171-179.

[30] Robinson, L.E. and Reeves, S. (2015) Review of Sigas Major Role as a Ffrst Line of Immune Defense and New Indications Regarding Inammation and Gut Health. *Epicor Science Report*, **25**, 25-29.

[31] Rai, T., Wu, X. and Shen, B. (2015) Frequency and Risk Factors of Low Immunoglobulin Levels in Patients with Inammatory Bowel Disease. *Gastroenterology Report*, **2**, 115-121.

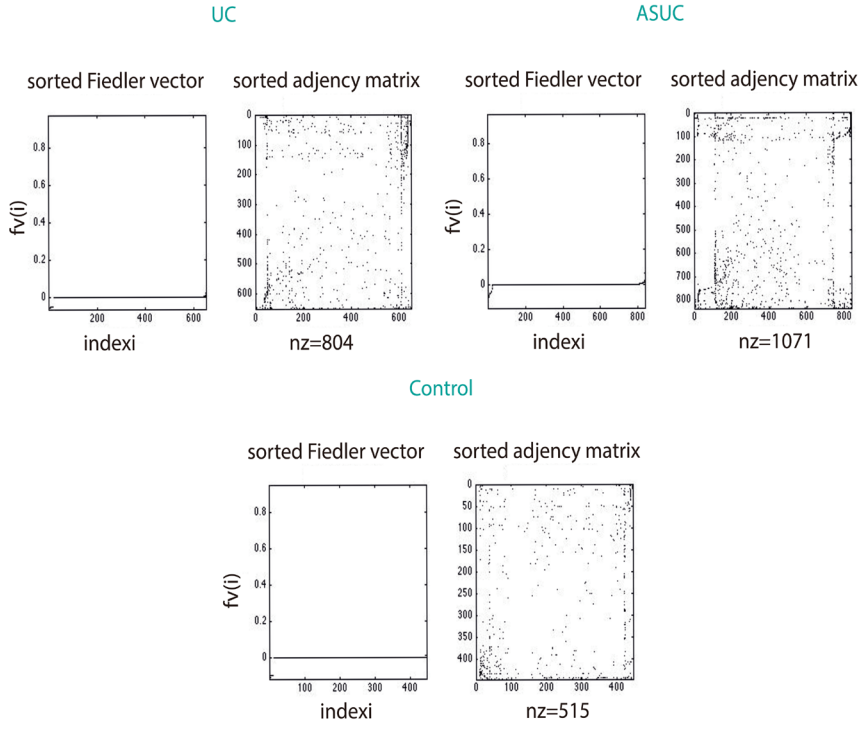
[32] Mulder, S.J. and Mulder-Bos, G.C. (2006) Most Probable Origin of Coeliac Disease Is Low Immune Globulin a in the Intestine Caused by Malfunction of Peyers Patches. *Medical Hypotheses*, **66**, 757-762.

[33] Kawakota, S., Tran, T.H., Maruya, M., Suzuki, K., Doi, Y., Tsutsui, Y., Kato, L.M. and Fagarasan, S. (2012) The Inhibitory Receptor pd-1 Regulates iga Selection and Bacterial Composition in the Gut. *Science*, **336**, 485-489.

Nomenclature

miRNA: micro-RNAs **ASC**: Acute Severe Ulcertative Colitis **UC**: Ulcerative Colitis κ: kernel application on graphs **PSD**: Positive Semi-definite Kernel **CGs**: Conjugate Gradient Methos **LN**: Log-Normal **GLM**: Generalized Linear Model **PPI**: Protein-Protein Interaction $\otimes$: Kronecker product of two matrices vec: column-stacking operator of a matrix RKHS: Reproducing Kernel Hilbert Spaces G: a set of ordered points generating a graph V: an ordered set of vertices E: set of edges of a graph G $\Phi(Y)$: matrix of a graph G w_{xx}: weight of an edge (x,x) $W_{\times}$: weight matrix associated to the Kronecker product of two matrices D: node degrees matrix A: adjacency matrix of a graph G $\mu(l)$: non-negative coefficient of decay for walks of length l WL: Weisfeiler-Lehman kernels κ_{WL}: Weisfeiler-Lehman kernel for κ Σ_{κ}: set of node labels matching at least once in a graph at the end of the k-ith Weisfeiler-Lehman iteration p_k: a map counting a specific node label in a graph $\kappa_{WLsubtree}$: Weisfeiler-Lehman subtree kernel on two graphs δ: Dirac kernel κ_{WLedge}: Weisfeiler-Lehman edge kernel for κ $\kappa_{WLshortestpath}$: Weisfeiler-Lehman shortest path kernel for κ **GO**: Gene Ontology database **PPs**: Peyer's Patches **sIgA/E**: immunoglobulins A/E **ID2**: Inhibitor Of DNA Binding 2 **STAT5**: Signal Transducer And Activator Of Transcription 5.

Figure S1



Tables. https://figshare.com/s/795ae25c8bf76ffb2489

Journal of Applied Mathematics and Physics, 2017, 5, 2262-2282
http://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

A Mathematical Model to Analyze Spread of Hemorrhagic Disease in White-Tailed Deer Population

Gerry Baygents*, Majid Bani-Yaghoub

Department of Mathematics and Statistics, University of Missouri Kansas City, Kansas City, MO, USA
Email: *baygentsg@umkc.edu

How to cite this paper: Baygents, G. and Bani-Yaghoub, M. (2017) A Mathematical Model to Analyze Spread of Hemorrhagic Disease in White-Tailed Deer Population. *Journal of Applied Mathematics and Physics*, **5**, 2262-2282.
https://doi.org/10.4236/jamp.2017.511184

Received: January 21, 2017
Accepted: November 26, 2017
Published: November 29, 2017

Abstract

Hemorrhagic disease (HD) is a fatal vector-borne disease that affects white-tailed deer and many other ruminants. A vector-borne disease model is proposed in the present work, which takes into account migrating effects of deer population using distributed delay terms. The model is employed to analyze the effects of deer migration on the HD spread. This is carried out in three steps. First, the conditions for existence and stability of the endemic and the disease free equilibria are established. Second, using the method of the Next Generation Matrix, the basic reproduction expression R_0 is derived from the model. Third, using the R_0 expression and its numerical simulations, it is illustrated that the severity of an HD outbreak is directly influenced by the migration rates of infected and susceptible deer (*i.e.*, d_I and d_S, respectively). For small values of d_S, the value of R_0 is increased with d_I, whereas R_0 decreases with d_I when d_S is large. Using the method of chain trick, the proposed model with distributed delay is reduced to a system of ordinary differential equations where the convergence of the system to endemic and diseases free equilibrium is numerically explored.

Keywords

Hemorrhagic Disease, Distributed Delay, Migration, Basic Reproduction Number

1. Introduction

Hemorrhagic disease (HD) is a fatal disease of white-tailed deer (*Odocoileus virginianus*). It is the collective term used for epizootic hemorrhagic disease and bluetongue disease (genus *Orbivirus*). These diseases have similar symptoms and

are frequently grouped together and referred to as HD. Symptoms include hemorrhaging, swelling due to fluid accumulation, sores, ulcers, sloughing of hooves, high fever, and loss of fear of humans [1] [2]. There are three different forms of HD (peracute, acute, and chronic) which dictate how long a deer will survive. Death can occur in as little as one to two weeks [3]. It is possible for a deer to survive, but it is rare. In addition to white-tailed deer, HD can be transmitted to other wild ruminants and domestic animals, most commonly hoof stock, but it rarely causes disease. The infection does not affect humans or non-ruminant animals [1]. The vector that spreads HD is small biting midge (*Culicoides Ceratopogondiae*). These midges are tiny, blood-sucking flies that are merely pests to humans, but they are the vectors in the spread of the disease in deer and livestock.

In the present work, we build a mathematical model to investigate the dynamics of HD. The amount of literature dedicated to the mathematical modeling of vector-borne diseases is extensive (See for example [4] [5] [6] [7]). The model by Nobel Prize winner Ronald Ross [4] is at the cornerstone of such models, and he used his model to investigate the spread of malaria. Over four decades later, George Macdonald developed it further [5]. In fact, there have been several extensions to the Ross-Macdonald model. For instance, Lou and Zhou [6] included advection and diffusion terms to take the spatial movements of individuals into account. Reaction-diffusion models have also been used for investigating dynamics of vector-borne diseases such as dengue fever [7] and Zika [8]. Using a deterministic modeling approach, the main objective of the present study is to have a better understanding of the possible effects of deer-midge interactions and deer migrations on HD dynamics in a deer population.

In recent years, more realistic models have been constructed which take into account dispersion time and host movements. A key article is the work by Neubert *et al.* [9], which argues that dispersion in Lotka-Volterra predator-prey models is unrealistic as individuals leaving an area (*i.e.*, a patch) immediately appear in another. In nature, an individual requires a finite amount of time to complete a trip from one patch to another or to complete a round trip leaving and returning to the same patch. During this time, the migrating individuals are not interacting with other predators or prey in this patch. Thus, Neubert and his co-authors [9] [10] demonstrate that models that incorporate explicit travel-time are often more stable.

Few models have been constructed to analyze the dynamics of HD in white-tailed deer populations and dairy farms. Park *et al.* [11] studied these dynamics by first fitting a statistical model to predict HD incidents as a function of seroprevalence (*i.e.*, the number of individuals in a population who tested positive for HD). Then, using ordinary differential equations (ODE), they formulated a mechanistic model to support the theory that there is a correlation between the number of HD cases and the number of deer in a population with

the virus. Their study suggests that the maximum number of cases occurs at intermediate levels of this seroprevalence. By constructing a realistic model, we will be able to analyze and simulate the dynamics of HD. A better understanding of HD dynamics gives epidemiologists and biologists the capacity to control and predict future epidemics in white-tailed deer populations. The present work is the first step toward realistic modeling of HD dynamics with a focus on migrating effects of white-tailed deer population.

The rest of this paper is organized as follows. In Section 2 we propose the vector-borne model of HD spread. This model takes into account the migration and immigration of deer from and into a single patch. In Section 3, we study the model through linearization, chain-trick method, and equilibrium analysis. We also calculate the R_0 expression and use it in Section 4 to numerically investigate the effects of the model parameters on outbreaks. Finally, in Section 5, we provide a discussion of results and outline the limitations of this study.

2. The Single-Patch Model

In the attempt to create a mathematical model of HD outbreak in a population of white-tailed deer, we make certain assumptions based on the ecology of deer and midge populations and the characteristics of HD. The deer (host) and midge (vector) populations are divided into susceptible and infected classes. At time t, there are $D_S(t)$ susceptible deer, $D_I(t)$ infected deer, $M_S(t)$ susceptible midges, and $M_I(t)$ infected midges. The total deer population at time t is $D_N(t) = D_S(t) + D_I(t)$, and the total midge population is $M_N(t) = M_S(t) + M_I(t)$. Susceptible deer become infected through bites of infected midges; susceptible midges become infected when they feed on the blood of an infected deer. As observed in the wild, deer will migrate (disperse) out of and back into a region (*i.e.*, a patch) due to seasonal variations, availability of food, or predators; midges, however, will not. They are weak fliers and typically disperse no more than about a mile from the site of larval development, with females flying farther than males [12]. Moreover, their flying activity is greatly reduced in windy conditions. They may fly as far as six miles or more, but this is very rare [13]. We therefore consider the following assumptions in the model construction:

1) All newborns are susceptible in both populations of deer and midges (*i.e.*, no inherited infection or vertical transmission is considered).

2) Susceptible deer become infected only by adequate contact with infected midges and cannot become infected via contact with an infected deer.

3) Once infected, a deer will die from the disease. (Note, in actuality, there are cases where a deer survives the infection, but it is rare.)

4) Individuals in both populations will die naturally by both density independent and density dependent factors.

5) By the law of mass action, we assume that infection transmission is proportional to the population densities of deer and midges.

6) Deer will frequently travel out of and into a geographic area (a patch), but

midges do not (as the amount of dispersal in midge populations is negligible).

A compartmental diagram of the proposed HD model is seen in Figure 1, and a summary of parameters and variables is given in Table 1. All parameters are assumed to be non-negative. Given the above-mentioned assumptions and the model diagram, the set of delayed differential equations representing the model is given by

$$\begin{aligned}
\frac{\mathrm{d}D_S}{\mathrm{d}t} &= \lambda_D D_N - \frac{\beta_D M_I D_S}{D_N} - \left(\mu_D + \rho + d_S + \mu_{2D} D_N\right) D_S \\
&\quad + d_S \int_0^\infty g(z)\mathrm{e}^{-\delta_S z} D_S(t-z)\mathrm{d}z \\
\frac{\mathrm{d}D_I}{\mathrm{d}t} &= \frac{\beta_D M_I D_S}{D_N} - \left(\mu_D + \rho + \gamma_D + d_I + \mu_{2D} D_N\right) D_I + d_I \int_0^\infty g(z)\mathrm{e}^{-\delta_I z} D_I(t-z)\mathrm{d}z \quad (1) \\
\frac{\mathrm{d}M_S}{\mathrm{d}t} &= \lambda_M M_N - \frac{\beta_M D_I M_S}{D_N} - \left(\mu_M + \sigma + \mu_{2M} M_N\right) M_S \\
\frac{\mathrm{d}M_I}{\mathrm{d}t} &= \frac{\beta_M D_I M_S}{D_N} - \left(\mu_M + \sigma + \mu_{2M} M_N\right) M_I
\end{aligned}$$

In absence of the disease, population growths of deer and midges are

Table 1. Summary of the variables and parameters used in the delayed HD model (1).

Symbol	Description
$D_S(t)$	Number of susceptible deer at time t
$D_I(t)$	Number of infected deer at time t
$D_N(t)$	Total deer population at time t
β_D	Infection rate (deer)
λ_D	Birth rate (deer)
ρ	Harvest rate
μ_D	Death rate (deer), density independent
μ_{2D}	Death rate (deer), density dependent
d_S	Net flux rate, susceptible deer
d_I	Net flux rate, infected deer
γ_D	Pathogenic induced death rate (deer)
δ_S	Probability of death per unit of time of a susceptible deer during migration
δ_I	Probability of death per unit of time of an infected deer during migration
$M_S(t)$	Number of susceptible midges at time t
$M_I(t)$	Number of infected midges at time t
$M_N(t)$	Total midge population at time t
β_M	Infection rate (midges)
λ_M	Birth rate (midges)
σ	Efficacy rate of midge control measures
μ_M	Death rate (midges), density independent
μ_{2M}	Death rate (midges), density dependent

Note: All variables and parameters are non-negative. The specific parameter values used in the analysis will be indicated in Table 2, Section 4.

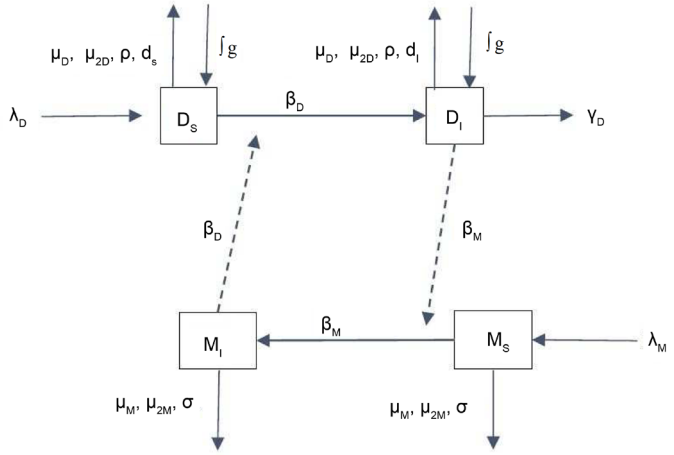


Figure 1. A compartmental diagram of the HD model (1) with population dispersal. Dashed lines represent the HD transmission between the vector and host. Deer migration into the patch is denoted by $\int g$ and migration out of the patch is denoted by d_S and d_I. See **Table 1** for a summary of the parameters and variables.

formulated with logistic growth models. These are the terms that include λ_D, λ_M, μ_{2D}, and μ_{2M} in model (1). Similar to [14], the carrying capacity for the deer population exists and must be positive. Hence, it is required that

$$H_1 : \lambda_D > \mu_D + \rho + d_S \tag{2}$$

and

$$H_2 : \lambda_D > \mu_D + \rho + \gamma_D + d_I. \tag{3}$$

Also, the carrying capacity for midges exists and is positive. Thus,

$$H_3 : \lambda_M > \mu_M + \sigma. \tag{4}$$

Individual deer immigrate from the patch at a constant per capita rate (d_S and d_I) and return z units of time after their departure. The integrals in the first two equations of model (1) are distributed delay terms representing the influx of susceptible and infected deer, respectively, from all points in time in the past up to and including the present time [15]. The function $g(z)$ in the integrals is a probability density function for the time it takes for a deer to disperse given that the deer survives the trip, and $g(z)\mathrm{d}z$ is the probability that a successfully dispersing deer departing at time t completes the trip between time $t+z$ and $t+z+\mathrm{d}z$. As $g(z)$ is a probability density function, it is normalized so that $\int_0^\infty g(z)\mathrm{d}z = 1$. The functions $\mathrm{e}^{-\delta_S z}$ and $\mathrm{e}^{-\delta_I z}$ in the integrals are the probabilities of a deer surviving a trip of duration z given constant probabilities per unit of time δ_S and δ_I for the mortality during travel of susceptible and infected deer, respectively. All deer migrating back into this single patch originated in the patch; in other words, there are no new deer entering the patching that originated from somewhere else. Hence, we are studying a herd of deer concentrated within a patch with the ability of migrating in and out of it.

3. Analysis of the Single-Patch Model

3.1. Linear Stability Analysis

In this section, we provide a formal procedure of linear stability analysis which leads to the characteristic equation and the stability conditions for the equilibrium solutions. Specifically, Disease Free Equilibrium (DFE) (*i.e.*, $D_I^* = 0$ and $M_I^* = 0$) and Endemic Equilibrium (EE) are the constant solutions of model (1). In epidemiology, a stable DFE is always desired whereas a stable EE can be of great concern. The first two equations of model (1) have an integral influx term that may be simplified by the following method. Letting

$$f^{[1]}\left(D_S, D_I, M_S, M_I\right) = \lambda_D D_N - \frac{\beta_D M_I D_S}{D_N} - \left(\mu_D + \rho + d_S + \mu_{2D} D_N\right) D_S, \quad (5)$$

we rewrite the first equation as

$$\frac{\mathrm{d}D_S}{\mathrm{d}t} = f^{[1]}\left(D_S, D_I, M_S, M_I\right) + d_S \int_0^\infty g(z)\mathrm{e}^{-\delta_S z} D_S(t-z)\mathrm{d}z. \quad (6)$$

Similarly, we rewrite the second equation as

$$\frac{\mathrm{d}D_I}{\mathrm{d}t} = f^{[2]}\left(D_S, D_I, M_S, M_I\right) + d_I \int_0^\infty g(z)\mathrm{e}^{-\delta_I z} D_I(t-z)\mathrm{d}z, \quad (7)$$

where

$$f^{[2]}\left(D_S, D_I, M_S, M_I\right) = \frac{\beta_D M_I D_S}{D_N} - \left(\mu_D + \rho + \gamma_D + d_I + \mu_{2D} D_N\right) D_I. \quad (8)$$

As the bottom two equations of model (1) have no integral term, we let $f^{[3]}$ and $f^{[4]}$ equal the right-hand side of the third and fourth equations in model (1), respectively. Let a solution $\left(D_S(t), D_I(t), M_S(t), M_I(t)\right)$ of model (1) nearby an equilibrium solution $E = \left(D_S^*, D_I^*, M_S^*, M_I^*\right)$ be in the form of

$$\begin{aligned} D_S(t) &= D_S^* + \tilde{D}_S(t), \quad D_I(t) = D_I^* + \tilde{D}_I(t), \\ M_S(t) &= M_S^* + \tilde{M}_S(t), \quad M_I(t) = M_I^* + \tilde{M}_I(t) \end{aligned} \quad (9)$$

for some $\tilde{D}_S(t)$, $\tilde{D}_I(t)$, $\tilde{M}_S(t)$, and $\tilde{M}_I(t)$. Using the Taylor expansion, we linearize the first equation in model (1) about equilibrium E by substituting (9) into (6) and dropping the nonlinear terms. Thus, the first equation of model (1) is linearized as follows.

$$\begin{aligned} \frac{\mathrm{d}D_S}{\mathrm{d}t} = \frac{\mathrm{d}\tilde{D}_S}{\mathrm{d}t} &= f^{[1]}\left(D_S, D_I, M_S, M_I\right) + d_S \int_0^\infty g(z)\mathrm{e}^{-\delta_S z} D_S(t-z)\mathrm{d}z \\ &= f^{[1]}\left(D_S^*, D_I^*, M_S^*, M_I^*\right) + \frac{\partial f^{[1]}}{\partial D_S}(E)\cdot\tilde{D}_S + \frac{\partial f^{[1]}}{\partial D_I}(E)\cdot\tilde{D}_I + \frac{\partial f^{[1]}}{\partial M_S}(E)\cdot\tilde{M}_S \\ &\quad + \frac{\partial f^{[1]}}{\partial M_I}(E)\cdot\tilde{M}_I + d_S \int_0^\infty g(z)\mathrm{e}^{-\delta_S z}\left(D_S^* + \tilde{D}_S(t-z)\right)\mathrm{d}z \\ &= f^{[1]}\left(D_S^*, D_I^*, M_S^*, M_I^*\right) + \frac{\partial f^{[1]}}{\partial D_S}(E)\cdot\tilde{D}_S + \frac{\partial f^{[1]}}{\partial D_I}(E)\cdot\tilde{D}_I + \frac{\partial f^{[1]}}{\partial M_S}(E)\cdot\tilde{M}_S \\ &\quad + \frac{\partial f^{[1]}}{\partial M_I}(E)\cdot\tilde{M}_I + d_S D_S^* \int_0^\infty g(z)\mathrm{e}^{-\delta_S z}\mathrm{d}z + d_S \int_0^\infty g(z)\mathrm{e}^{-\delta_S z}\tilde{D}_S(t-z)\mathrm{d}z. \end{aligned} \quad (10)$$

We know that equilibrium E satisfies the first equation of model (1), hence

$$f^{[1]}\left(D_S^*, D_I^*, M_S^*, M_I^*\right) + d_S D_S^* \int_0^\infty g(z)\mathrm{e}^{-\delta_S z}\mathrm{d}z = 0, \tag{11}$$

and thus

$$d_S D_S^* \int_0^\infty g(z)\mathrm{e}^{-\delta_S z}\mathrm{d}z = -f_1\left(D_S^*, D_I^*, M_S^*, M_I^*\right). \tag{12}$$

Substituting (12) into (10) yields

$$\begin{aligned}\frac{\mathrm{d}\tilde{D}_S}{\mathrm{d}t} &= \frac{\partial f^{[1]}}{\partial D_S}(E)\cdot\tilde{D}_S + \frac{\partial f^{[1]}}{\partial D_I}(E)\cdot\tilde{D}_I + \frac{\partial f^{[1]}}{\partial M_S}(E)\cdot\tilde{M}_S + \frac{\partial f^{[1]}}{\partial M_I}(E)\cdot\tilde{M}_I \\ &\quad + d_S\int_0^\infty g(z)\mathrm{e}^{-\delta_S z}\tilde{D}_S(t-z)\mathrm{d}z.\end{aligned} \tag{13}$$

Applying the same procedure to equation (7), we get that the second equation of model (1) is linearized by

$$\begin{aligned}\frac{\mathrm{d}\tilde{D}_I}{\mathrm{d}t} &= \frac{\partial f^{[2]}}{\partial D_S}(E)\cdot\tilde{D}_S + \frac{\partial f^{[2]}}{\partial D_I}(E)\cdot\tilde{D}_I + \frac{\partial f^{[2]}}{\partial M_S}(E)\cdot\tilde{M}_S + \frac{\partial f^{[2]}}{\partial M_I}(E)\cdot\tilde{M}_I \\ &\quad + d_I\int_0^\infty g(z)\mathrm{e}^{-\delta_I z}\tilde{D}_I(t-z)\mathrm{d}z.\end{aligned} \tag{14}$$

Using Equations (9)-(14), model (1) is linearized about equilibrium E and takes the form

$$Y'(t) = AY(t), \tag{15}$$

where $Y(t) = \left[\tilde{D}_S(t), \tilde{D}_I(t), \tilde{M}_S(t), \tilde{M}_I(t)\right]^{\mathrm{T}}$ and A is the Jacobian matrix evaluated at E. However, the specific form of matrix A cannot be extracted due to the presence of the integral terms in (13) and (14). To bypass this issue, we use the Fundamental Theorem of linear systems of differential equations [16] and look for exponential solutions of the form

$$\begin{bmatrix}\tilde{D}_S(t)\\ \tilde{D}_I(t)\\ \tilde{M}_S(t)\\ \tilde{M}_I(t)\end{bmatrix} = \begin{bmatrix}r_1\\ r_2\\ r_3\\ r_4\end{bmatrix}\mathrm{e}^{\lambda t} = R\mathrm{e}^{\lambda t}. \tag{16}$$

We also let $\tilde{g}$ be the (one-sided) Laplace transform of the travel-time distribution $g(z)$. That is,

$$\tilde{g}(x) \equiv \int_0^\infty g(z)\mathrm{e}^{-xz}\mathrm{d}z. \tag{17}$$

We have the following Lemma.

Lemma 1 *The Laplace transform $\tilde{g}$ is a positive, decreasing function that is bounded above by* 1 *for all non-negative values of x.*

Proof. Let $g(z)$ be a probability density function as described above. Because the function e^{-xz} is positive for all real x and fixed z, $\mathrm{e}^{-xz} = 1$ when $x = 0$, and e^{-xz} decreases for all $x > 0$. Therefore, it must be the case that $0 < g(z)\mathrm{e}^{-xz} \le 1$ and $g(z)\mathrm{e}^{-xz}$ decreases for all non-negative x. Thus, $\tilde{g}(x) \equiv \int_0^\infty g(z)\mathrm{e}^{-xz}\mathrm{d}z$ is a positive decreasing function bounded above by 1.

By substituting (16) into (15) and simplifying the terms, we get the specific form of matrix A, and 15 is rewritten as

$$\begin{bmatrix} \dfrac{\partial f^{[1]}(E)}{\partial D_S} + d_S\tilde{g}(\lambda+\delta_S) - \lambda & \dfrac{\partial f^{[1]}(E)}{\partial D_I} & \dfrac{\partial f^{[1]}(E)}{\partial M_S} & \dfrac{\partial f^{[1]}(E)}{\partial M_I} \\ \dfrac{\partial f^{[2]}(E)}{\partial D_S} & \dfrac{\partial f^{[2]}(E)}{\partial D_I} + d_I\tilde{g}(\lambda+\delta_I) - \lambda & \dfrac{\partial f^{[2]}(E)}{\partial M_S} & \dfrac{\partial f^{[2]}(E)}{\partial M_I} \\ \dfrac{\partial f^{[3]}(E)}{\partial D_S} & \dfrac{\partial f^{[3]}(E)}{\partial D_I} & \dfrac{\partial f^{[3]}(E)}{\partial M_S} - \lambda & \dfrac{\partial f^{[3]}(E)}{\partial M_I} \\ \dfrac{\partial f^{[4]}(E)}{\partial D_S} & \dfrac{\partial f^{[4]}(E)}{\partial D_I} & \dfrac{\partial f^{[4]}(E)}{\partial M_S} & \dfrac{\partial f^{[4]}(E)}{\partial M_I} - \lambda \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \quad (18)$$

The linear system in (18) has a nontrivial solution if and only if the determinant of the matrix is zero. This leads to the characteristic equation corresponding to model (1) linearized about E. Before deriving the characteristic equation, we prove the existence of DFE.

Proposition 1 *The disease free equilibrium of model* (1) *exists if and only if* $\lambda_D > \mu_D + \rho + d_S\left(1-\tilde{g}\left(\delta_S\right)\right)$ *and* $\lambda_M > \mu_M + \sigma$ *are satisfied.*

Proof. Noting that $D_I^* = 0$, $D_N = D_S^*$, and $\dfrac{\mathrm{d}D_S}{\mathrm{d}t} = 0$ at the DFE, the first equation in model (1) gives us $D_S^* = \dfrac{\lambda_D - (\mu_D + \rho) - d_S\left(1-\tilde{g}\left(\delta_S\right)\right)}{\mu_{2D}}$. Similarly, $M_I^* = 0$ and $M_N = M_S^*$, and the third equation of model (1) gives rise to $M_S^* = \dfrac{\lambda_M - (\mu_M + \sigma)}{\mu_{2M}}$. As $D_S^* > 0$ and $M_S^* > 0$ by parameter assumptions, the disease free equilibrium exists.

Remark 1 *The inequalities* (2) *and* (4) *and Lemma* 1 *imply that the conditions of Proposition* 1 *are always satisfied. Hence, the DFE always exists and it is given by*

$$\begin{aligned} D_S^* &= \frac{\lambda_D - (\mu_D + \rho) - d_S\left(1-\tilde{g}\left(\delta_S\right)\right)}{\mu_{2D}} \\ D_I^* &= 0 \\ M_S^* &= \frac{\lambda_M - (\mu_M + \sigma)}{\mu_{2M}} \\ M_I^* &= 0 \end{aligned} \quad (19)$$

By linearizing model (1) about the DFE, we get the characteristic equation

$$det\left(J(\lambda)\right) = 0, \quad (20)$$

where $J(\lambda)$ is the matrix in (18) evaluated at $E = DFE$, and it simplifies to

$$J(\lambda) = \begin{bmatrix} J_1(\lambda) & \lambda_D - \mu_{2D}D_S^* & 0 & -\beta_D \\ 0 & J_2(\lambda) & 0 & \beta_D \\ 0 & -\dfrac{\beta_M M_S^*}{D_S^*} & J_3(\lambda) & \lambda_M - \mu_{2M}M_S^* \\ 0 & \dfrac{\beta_M M_S^*}{D_S^*} & 0 & J_4(\lambda) \end{bmatrix} \quad (21)$$

such that

$$J_1(\lambda) = \lambda_D - \mu_D - \rho - d_S - 2\mu_{2D} D_S^* + d_S \tilde{g}(\lambda + \delta_S) - \lambda, \tag{22}$$

$$J_2(\lambda) = -\mu_D - \rho - \gamma_D - d_I - \mu_{2D} D_S^* + d_I \tilde{g}(\lambda + \delta_I) - \lambda, \tag{23}$$

$$J_3(\lambda) = \lambda_M - \mu_M - \sigma - 2\mu_{2M} M_S^* - \lambda, \tag{24}$$

and

$$J_4(\lambda) = -\mu_M - \sigma - \mu_{2M} M_S^* - \lambda. \tag{25}$$

Hence, the characteristic equation (20) is rewritten

$$J_1(\lambda) J_3(\lambda) \left[J_2(\lambda) J_4(\lambda) - \frac{\beta_D \beta_M M_S^*}{D_S^*} \right] = 0. \tag{26}$$

Since $J_1(\lambda)$ and $J_2(\lambda)$ are not polynomials, the Routh-Hurwitz criteria [17] is not applicable for determining stability. However, with a specific form of $g(z)$, we may compute the roots of the characteristic equation and determine the necessary and sufficient conditions for the stability of the DFE.

3.2. Basic Reproduction Number

The basic reproduction number R_0 is defined as the expected number of secondary infections produced by a single case of an infection introduced to a completely susceptible population [18]. When $R_0 > 1$, the infection will spread as the number of infected individuals increases. When $R_0 < 1$, the infection will die out in the long run. Thus, we seek conditions and parameter values so that $R_0 < 1$.

The magnitude of R_0 determines the severity of infection. Larger values of $R_0 > 1$ lead to faster disease spread, whereas smaller values of $R_0 < 1$ lead to the disease dying out more rapidly. Using the Next Generation Matrix (NGM) approach [19] [20], the expression for R_0 can be derived. Specifically, the next generation matrix is given by $K = FV^{-1}$, and the spectral radius of K is equal to R_0. The elements of matrix F, using the extended definition of the matrix F [21], represent new infections, where the entry (i, j) of F represents the rate at which secondary individuals appear in class i per individual of type j. The elements of matrix V are the transition of infections.

In order to calculate the R_0 expression, we make some simplifying assumptions in our model. In particular, we assume the integral terms in the first and second equations of model (1) are simplified to

$$\int_0^\infty g(z) \mathrm{e}^{-\delta_S z} D_S(t - z) \mathrm{d}z = \tilde{g}(\delta_S) D_S(t) \tag{27}$$

and

$$\int_0^\infty g(z) \mathrm{e}^{-\delta_I z} D_I(t - z) \mathrm{d}z = \tilde{g}(\delta_I) D_I(t) \tag{28}$$

respectively.

Remark 2 *The assumptions in* (27) *and* (28) *result in a positive outflow of deer out of the patch. The first equation of model* (1) *contains the expression*

$-d_S D_S(t) + d_S \int_0^\infty g(z) e^{-\delta_S z} D_S(t-z)\,dz$. *Using* (27), *this simplifies to* $d_S(\tilde{g}(\delta_S)-1)D_S(t)$ *which is negative by the above Lemma. In other words, there are more susceptible deer leaving the patch than entering it. The same is true for the infected deer as concluded from the second equation of model* (1) *and assumption* (28).

Using the assumptions in (27) and (28), we get that

$$F = \begin{bmatrix} d_I \tilde{g}(\delta_I) & \beta_D \\ \dfrac{\beta_M M_S^*}{D_S^*} & 0 \end{bmatrix}, \tag{29}$$

$$V = \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix}, \tag{30}$$

and

$$FV^{-1} = \begin{bmatrix} \dfrac{d_I \tilde{g}(\delta_I)}{V_1} & \dfrac{\beta_D}{V_2} \\ \dfrac{\beta_M M_S^*}{V_1 D_S^*} & 0 \end{bmatrix}, \tag{31}$$

where

$$V_1 = \mu_D + \rho + \gamma_D + d_I + \mu_{2D} D_S^* \tag{32}$$

and

$$V_2 = \mu_M + \sigma + \mu_{2M} M_S^*. \tag{33}$$

As mentioned earlier, the basic reproduction number R_0 is the spectral radius of FV^{-1}. Since FV^{-1} is a positive definite matrix, R_0 is equal to the largest eigenvalue of FV^{-1}. After simplifying, the expression for R_0 can be written as

$$R_0 = \frac{1}{2}\left(R_0^{[1]} + \sqrt{\left(R_0^{[1]}\right)^2 + 4R_0^{[2]}} \right), \tag{34}$$

where

$$R_0^{[1]} = \frac{d_I \tilde{g}(\delta_I)}{V_1}, \tag{35}$$

representing the contribution of deer migration to disease outbreaks, and

$$R_0^{[2]} = \frac{\beta_D \beta_M M_S^*}{V_1 V_2 D_S^*}, \tag{36}$$

representing the effects of the deer-midge interactions on disease outbreaks. Therefore, the migration effects of infected deer and the effects of deer-midge interactions within the patch on HD outbreaks can be studied separately.

1) **Pure migration effects** ($R_0^{[2]} = 0$). This occurs when either β_D or β_M is zero, and thus there is no transmission of the disease between the midges and the deer (or vice-versa) within the patch. Using Equation (34), $R_0^{[2]} = 0$ implies

$R_0 = R_0^{[1]}$. In reality, this can effectively occur when the midge population in the patch is negligible. It can be seen that $R_0^{[1]}$ is a concave down increasing function of d_I. Thus, the flux rate of infected deer d_I may increase $R_0^{[1]}$. From Equation (32), we get that $\lim_{d_I \to \infty} R_0^{[1]} = \tilde{g}(\delta_I)$. Using Lemma 1, $\tilde{g}(\delta_I) \le 1$. Therefore, d_I alone cannot cause an outbreak even though it increases the $R_0^{[1]}$ value. In fact, using Equations (32) and (35), it can be easily shown that $R_0^{[1]} < 1$ for all parameter values of the model. Hence, assumptions (27) and (28) are underestimating the migration effects of deer population on disease outbreak.

2) **Residential effects** ($R_0^{[1]} = 0$). This occurs when $d_I = 0$, which means that infected deer have limited mobility and cannot leave or enter the patch due to illness. In this case, $R_0^{[1]} = 0$ implies $R_0 = \sqrt{R_0^{[2]}}$. In this case, an epidemic may be prevented if $R_0^{[2]} < 1$. This, in fact, may be possible as the harvest rate, ρ, is a part of the expression of $R_0^{[2]}$. On the other hand, small values of V_2 (*i.e.*, low mortality of midges) may result in an outbreak.

The following proposition indicates the effects of parameter values on R_0 in general.

Proposition 2 *The basic reproduction number* R_0 *is defined in Equation* (34) *and it has the following properties:*

1) R_0 is an increasing function of δ_S and d_S.

2) R_0 is a decreasing function of δ_I.

3) R_0 is an increasing function of d_I if d_S or the product $\beta_D \beta_M$ is sufficiently small.

4) R_0 is a decreasing function of d_I if d_S or the product $\beta_D \beta_M$ is sufficiently large.

Proof. Part (*i*): As shown below, the partial derivative of R_0 with respect to δ_S is positive.

$$\frac{\partial R_0}{\partial \delta_S} = \frac{-\beta_D \beta_M d_S M_S^* \tilde{g}'(\delta_S)}{\mu_{2D} V_1 V_2 (D_S^*) \sqrt{\left(R_0^{[1]}\right)^2 + 4R_0^{[2]}}} > 0. \tag{37}$$

Note that $\tilde{g}'(\delta_S) < 0$ because $\tilde{g}(\delta_S)$ is a decreasing function (See Lemma 1). Similarly, the partial derivative of R_0 with respect to d_S is positive.

$$\frac{\partial R_0}{\partial d_S} = \frac{d_I \tilde{g}(\delta_I)\left(1 - \tilde{g}(\delta_S)\right)}{2V_1^2} + \frac{1}{\sqrt{\left(R_0^{[1]}\right)^2 + 4R_0^{[2]}}} \left[\frac{\left(d_I \tilde{g}(\delta_I)\right)^2 \left(1 - \tilde{g}(\delta_S)\right)}{2V_1^3} + \frac{\beta_D \beta_M M_S^*}{\mu_{2D} V_2 \left(V_1 D_S^*\right)^2 \left(1 - \tilde{g}(\delta_S)\right)} \left(\mu_{2D} D_S^* + V_1\right) \right] > 0. \tag{38}$$

Part (*ii*): The partial derivative of R_0 with respect to δ_I is negative.

$$\frac{\partial R_0}{\partial \delta_I} = \frac{d_I \tilde{g}'(\delta_I)}{2} \left(\frac{1}{V_1} + \frac{d_I \tilde{g}(\delta_I)}{V_1^2 \sqrt{\left(R_0^{[1]}\right)^2 + 4R_0^{[2]}}} \right) < 0. \tag{39}$$

To prove statements (iii) and (iv), note that the partial derivative of R_0 with respect to d_I is given by

$$\frac{\partial R_0}{\partial d_I}=\frac{\tilde{g}(\delta_I)(V_1-d_I)}{2V_1^2}+\frac{1}{V_1^2\sqrt{\left(R_0^{[1]}\right)^2+4R_0^{[2]}}}\left[\frac{d_I(V_1-d_I)\left(\tilde{g}(\delta_I)\right)^2}{2V_1}-\frac{\beta_D\beta_M M_S^*}{V_2D_S^*}\right] \quad (40)$$

Also note that $V_1-d_I=\mu_D+\rho+\gamma_D+\mu_{2D}D_S^*>0$. The expression

$$\frac{d_I(V_1-d_I)\left(\tilde{g}(\delta_I)\right)^2}{2V_1}-\frac{\beta_D\beta_M M_S^*}{V_2D_S^*}>0 \quad (41)$$

is equivalent to

$$d_IV_2D_S^*(V_1-d_I)\left(\tilde{g}(\delta_I)\right)^2-2V_1\beta_D\beta_M M_S^*>0. \quad (42)$$

Recall that $D_S^*=\dfrac{\lambda_D-(\mu_D+\rho)-d_S(1-\tilde{g}(\delta_S))}{\mu_{2D}}$. When d_S is sufficiently small, $d_IV_2D_S^*(V_1-d_I)\left(\tilde{g}(\delta_I)\right)^2$ will be sufficiently large and the inequality holds. When $\beta_D\beta_M$ is sufficiently small, $2V_1\beta_D\beta_M M_S^*$ will be sufficiently small and the inequality holds. Thus $\dfrac{\partial R_0}{\partial d_I}>0$. Similarly, when either d_S or the product $\beta_D\beta_M$ is sufficiently large, $\dfrac{\partial R_0}{\partial d_I}<0$.

Remark 3 *Proposition* 2 *implies that the flux rate* d_I *of infected deer can have two opposing effects based on the value of* d_S *or the product* $\beta_D\beta_M$. *Because the directional behavior of* R_0 *changes due to the value of these, there must be critical values* ($d_S^{[c]}$ *and* $(\beta_D\beta_M)^{[c]}$) *such that* R_0 *is an increasing function of* d_I *when* d_S *or* $\beta_D\beta_M$ *are below either of the critical values and* R_0 *is a decreasing function of* d_I *when* d_S *or* $\beta_D\beta_M$ *are above either of them.*

The following Lemma is associated with the structure of the R_0 expression in equation (34).

Lemma 2 *For* $a,b\geq 0$, $a+b<1$ *if and only if* $\dfrac{1}{2}\left(a+\sqrt{a^2+4b}\right)<1$.

Proof. ($\Rightarrow$) If $a+b<1$, then $b<1-a$. Also, as $0\leq a<1$, $|a-2|=2-a$. Thus,

$$\begin{aligned}\frac{1}{2}\left(a+\sqrt{a^2+4b}\right)&<\frac{1}{2}\left(a+\sqrt{a^2+4(1-a)}\right)=\frac{1}{2}\left(a+\sqrt{a^2-4a+4}\right)\\&=\frac{1}{2}\left(a+|a-2|\right)=\frac{1}{2}(a+2-a)=1\end{aligned} \quad (43)$$

($\Leftarrow$)

$$\begin{aligned}&\frac{1}{2}\left(a+\sqrt{a^2+4b}\right)<1\\&a+\sqrt{a^2+4b}<2\\&\sqrt{a^2+4b}<2-a\end{aligned} \quad (44)$$

$$a^2 + 4b < 4 - 4a + a^2$$
$$4a + 4b < 4$$
$$a + b < 1$$

Remark 4 *Let* $a = R_0^{[1]}$ *and* $b = R_0^{[2]}$. *Using Lemma* 2, *we get that* $R_0 < 1$ *is equivalent to* $R_0^{[1]} + R_0^{[2]} < 1$. *As indicated in* [22] [23], *the expression* $R_0^{[1]} + R_0^{[2]}$ *is known as a Type-Reduction number which can be more accurate than* R_0 *to calculate the minimum disease eradication efforts.*

Proposition 3 *Under the assumptions* (27) *and* (28), *the DFE of model* (1) *is locally asymptotically stable if and only if* $R_0^{[1]} + R_0^{[2]} < 1$ *or, equivalently,* $R_0 < 1$.

Proof. ($\Leftarrow$) We determine stability conditions at the DFE by using the Jacobian of the system of equations. The DFE is locally asymptotically stable if the real parts of all eigenvalues of the Jacobian matrix are negative as explained in Section 3.1. Using assumptions (27) and (28), the Jacobian matrix evaluated at the DFE is given by:

$$A = \begin{bmatrix} A_1 & \lambda_D - \mu_{2D} D_S^* & 0 & -\beta_D \\ 0 & A_2 & 0 & \beta_D \\ 0 & -\dfrac{\beta_M M_S^*}{D_S^*} & A_3 & \lambda_M - \mu_{2M} M_S^* \\ 0 & \dfrac{\beta_M M_S^*}{D_S^*} & 0 & A_4 \end{bmatrix}, \tag{45}$$

where $A_1 = \lambda_D - \mu_D - \rho - \left(d_S\left(1 - \tilde{g}\left(\delta_S\right)\right) + 2\mu_{2D} D_S^*\right)$, $A_2 = d_I \tilde{g}\left(\delta_I\right) - V_1$, $A_3 = \lambda_M - \mu_M - \sigma - s\mu_{2M} M_S^*$, and $A_4 = -V_2$. The characteristic equation of this matrix, using Λ for the eigenvalues, is

$$f\left(\Lambda\right) = \left(A_1 - \Lambda\right)\left(A_3 - \Lambda\right)\left[\left(A_2 - \Lambda\right)\left(A_4 - \Lambda\right) - \frac{\beta_D \beta_M M_S^*}{D_S^*}\right]. \tag{46}$$

For the first eigenvalue A_1, we note that since the DFE must satisfy $D_S' = 0$, we can show that $\lambda_D = \mu_D + \rho + d_S + \mu_{2D} D_S^* - d_S \tilde{g}\left(\delta_S\right) = V_1 - d_S \tilde{g}\left(\delta_S\right)$. Therefore,

$$\begin{aligned} & A_1 \lambda_D - \mu_D - \rho - \left(d_S\left(1 - \tilde{g}\left(\delta_S\right)\right) + 2\mu_{2D} D_S^*\right) \\ & = \left(\mu_D + \rho + d_S + \mu_{2D} D_S^* - d_S \tilde{g}\left(\delta_S\right)\right) - \mu_D - \rho - \left(d_S\left(1 - \tilde{g}\left(\delta_S\right)\right) + 2\mu_{2D} D_S^*\right) \\ & = -\mu_{2D} D_S^* < 0 \end{aligned} \tag{47}$$

Similarly, for the second eigenvalue, given that the DFE must satisfy $M_S' = 0$, we can show $\lambda_M = \mu_M + \sigma + \mu_{2M} M_S^*$, and thus
$A_3 = \lambda_M - \mu_M - \sigma - 2\mu_{2M} M_S^* = -\mu_{2M} M_S^* < 0$.

For the remaining two eigenvalues, we rewrite the part of the characteristic equation in brackets as

$$\Lambda^2 - \left(A_2 + A_4\right)\Lambda + A_2 A_4 - \frac{\beta_D \beta_M M_S^*}{D_S^*} = 0. \tag{48}$$

This is a quadratic of the form $\Lambda^2 + b\Lambda + c$. According to the Routh-Hurwitz

criteria [17], the roots of a quadratic will have negative real parts if the linear coefficient and the constant term are positive. The linear coefficient is $-\left(A_2+A_4\right)$ and is positive as shown below.

$$\begin{aligned} A_2+A_4 &= d_I\tilde{g}\left(\delta_I\right)-V_1-V_2 \\ &= d_I\tilde{g}\left(\delta_I\right)-\mu_D-\rho-\gamma_D-d_I-\mu_{2D}D_S^*-V_2 \\ &= -d_I\left(1-\tilde{g}\left(\delta_I\right)\right)-\mu_D-\rho-\gamma_D-d_I-\mu_{2D}D_S^*-V_2<0 \end{aligned} \tag{49}$$

If $R_0^{[1]}+R_0^{[2]}<1$, then

$$\begin{aligned} &\frac{d_I\tilde{g}\left(\delta_I\right)}{V_1}+\frac{\beta_D\beta_M M_S^*}{V_1V_2D_S^*}<1 \\ &d_I\tilde{g}\left(\delta_I\right)V_2+\frac{\beta_D\beta_M M_S^*}{D_S^*}<V_1V_2 \\ &V_1V_2-d_I\tilde{g}\left(\delta_I\right)V_2-\frac{\beta_D\beta_M M_S^*}{D_S^*}>0 \end{aligned} \tag{50}$$

Hence, the constant term of the characteristic equation

$$\begin{aligned} A_2A_4-\frac{\beta_D\beta_M M_S^*}{D_S^*} &= -\left(d_I\tilde{g}\left(\delta_I\right)-V_1\right)V_2-\frac{\beta_D\beta_M M_S^*}{D_S^*} \\ &= V_1V_2-d_I\tilde{g}\left(\delta_I\right)V_2-\frac{\beta_D\beta_M M_S^*}{D_S^*}>0 \end{aligned} \tag{51}$$

Therefore, both roots of the quadratic (*i.e.* the two eigenvalues) must have negative real parts. Thus, under the given conditions, the system is stable at DFE.

($\Rightarrow$) If the DFE of model (1) is locally asymptotically stable, then by Theorem 8.12. *iii* of [24], the real parts of all eigenvalues of the Jacobian matrix A are negative. By (50), this occurs when $V_1V_2-d_I\tilde{g}\left(\delta_I\right)V_2-\frac{\beta_D\beta_M M_S^*}{D_S^*}>0$ which is the same as $R_0^{[1]}+R_0^{[2]}<1$.

We must now prove the existence of an endemic equilibrium solution in the proposed model. However, this is difficult as two of the variables, D_S and D_I, are contained within the integral dispersion terms. Therefore, we utilize a technique called the chain trick [15] to reduce model (1) to an ODE model.

3.3. Reduction to ODE Model

Using the chain trick method [15], we can rewrite the first two equations as

$$\begin{aligned} \frac{\mathrm{d}D_S}{\mathrm{d}t} &= \lambda_D D_N-\frac{\beta_D M_I D_S}{D_N}-\left(\left(\mu_D+\rho\right)+d_S+\mu_{2D}D_N\right)D_S+\bar{D}_S \\ \frac{\mathrm{d}D_I}{\mathrm{d}t} &= \frac{\beta_D M_I D_S}{D_N}-\left(\left(\mu_D+\rho\right)+\gamma_D+d_I+\mu_{2D}D_N\right)D_I+\bar{D}_I \end{aligned} \tag{52}$$

where

$$\bar{D}_S = d_S\int_0^\infty g\left(z\right)\mathrm{e}^{-\delta_S z}D_S\left(t-z\right)\mathrm{d}z \tag{53}$$

and

$$\bar{D}_I = d_I \int_0^\infty g(z) \mathrm{e}^{-\delta_I z} D_I(t-z)\,\mathrm{d}z. \tag{54}$$

These quantities are treated as new model variables, so we may now differentiate both of them and amend the existing set of equations.

In time delay models, there are two distributions that are commonly used. The first is a uniform distribution with mean τ given by

$$g(u) = \begin{cases} \dfrac{1}{\tau\rho}, & \text{for } \tau\left(1-\dfrac{\rho}{2}\right) \le u \le \tau\left(1+\dfrac{\rho}{2}\right) \\ 0, & \text{elsewhere.} \end{cases} \tag{55}$$

The second is the gamma distribution given by

$$g(u) = \frac{u^{p-1}\alpha^p \mathrm{e}^{-\alpha u}}{\Gamma(p)}, \tag{56}$$

where $\alpha, p \ge 0$ are parameters which determine the shape of the distribution and the mean of the distribution is p/α. In the case when $p=1$, the result is the exponential distribution, $g(z) = \alpha \mathrm{e}^{-\alpha z}$. Using (56) with $p=1$, the expression for $\dfrac{\mathrm{d}\bar{D}_S}{\mathrm{d}t}$ is computed to be:

$$\begin{aligned} \bar{D}_S &= \int_0^\infty g(z) \mathrm{e}^{-\delta_S z} D_S(t-z)\,\mathrm{d}z \\ &= \int_{-\infty}^t g(t-u) \mathrm{e}^{-\delta_S (t-u)} D_S(u)\,\mathrm{d}u \\ &= \mathrm{e}^{-\delta_S t} \int_{-\infty}^t g(t-u) \mathrm{e}^{-\delta_S u} D_S(u)\,\mathrm{d}u \\ &= \mathrm{e}^{-\delta_S t} \int_{-\infty}^t \alpha \mathrm{e}^{-\alpha(t-u)} \mathrm{e}^{-\delta_S u} D_S(u)\,\mathrm{d}u \\ &= \alpha \mathrm{e}^{-(\delta_S+\alpha)t} \int_{-\infty}^t \mathrm{e}^{(\delta_S+\alpha)u} D_S(u)\,\mathrm{d}u \end{aligned} \tag{57}$$

Thus, by the product rule for differentiation,

$$\begin{aligned} \frac{\mathrm{d}\bar{D}_S}{\mathrm{d}t} &= \alpha\left(-(\delta_S+\alpha)\right)\mathrm{e}^{-(\delta_S+\alpha)t} \int_{-\infty}^t \mathrm{e}^{(\delta_S+\alpha)u} D_S(u)\,\mathrm{d}u + d_S \alpha \mathrm{e}^{-(\delta_S+\alpha)t} \mathrm{e}^{(\delta_S+\alpha)u} D_S(t) \\ &= -(\delta_S+\alpha)\bar{D}_S + \alpha D_S \end{aligned} \tag{58}$$

The simplification is the same for $\bar{D}_I$, and so the delayed model in (1) is reduced to the ODE model formulated by

$$\begin{aligned} \frac{\mathrm{d}D_S}{\mathrm{d}t} &= \lambda_D D_N - \frac{\beta_D M_I D_S}{D_N} - (\mu_D + \rho + d_S + \mu_{2D} D_N) D_S + d_S \bar{D}_S \\ \frac{\mathrm{d}D_I}{\mathrm{d}t} &= \frac{\beta_D M_I D_S}{D_N} - (\mu_D + \rho + \gamma_D + d_I + \mu_{2D} D_N) D_I + d_I \bar{D}_I \\ \frac{\mathrm{d}M_S}{\mathrm{d}t} &= \lambda_M M_N - \frac{\beta_M D_I M_S}{D_N} - (\mu_M + \sigma + \mu_{2M} M_N) M_S \\ \frac{\mathrm{d}M_I}{\mathrm{d}t} &= \frac{\beta_M D_I M_S}{D_N} - (\mu_M + \sigma + \mu_{2M} M_N) M_I \\ \frac{\mathrm{d}\bar{D}_S}{\mathrm{d}t} &= -(\delta_S+\alpha)\bar{D}_S + \alpha D_S, \frac{\mathrm{d}\bar{D}_I}{\mathrm{d}t} = -(\delta_I+\alpha)\bar{D}_I + \alpha D_I \end{aligned} \tag{59}$$

The disease free equilibrium (DFE) is computed to be

$$
\begin{aligned}
&D_S^* \frac{\alpha+(\delta_S+\alpha)(\lambda_D-(\mu_D+\rho)-d_S)}{(\delta_S+\alpha)\mu_{2D}}\\
&D_I^*=0\\
&M_S^*=\frac{\lambda_M-(\mu_M+\sigma)}{\mu_{2M}}\\
&M_I^*=0\\
&\bar{D}_S^*=\frac{\alpha\left[d_S\alpha+(\delta_D+\alpha)(\lambda_D-(\mu_D+\rho)-d_S)\right]}{(\delta_S+\alpha)^2\mu_{2D}}\\
&\bar{D}_I^*=0
\end{aligned}
\tag{60}
$$

In the next section, we provide the numerical simulations of the ODE model (59) and the R_0 expression (34).

4. Numerical Simulations

Using Matlab 9.1, we generated the surface plots of R_0 values based on the model parameters $R_0^{[1]}$ and $R_0^{[2]}$ (See Figure 2). As proven in Proposition 2,

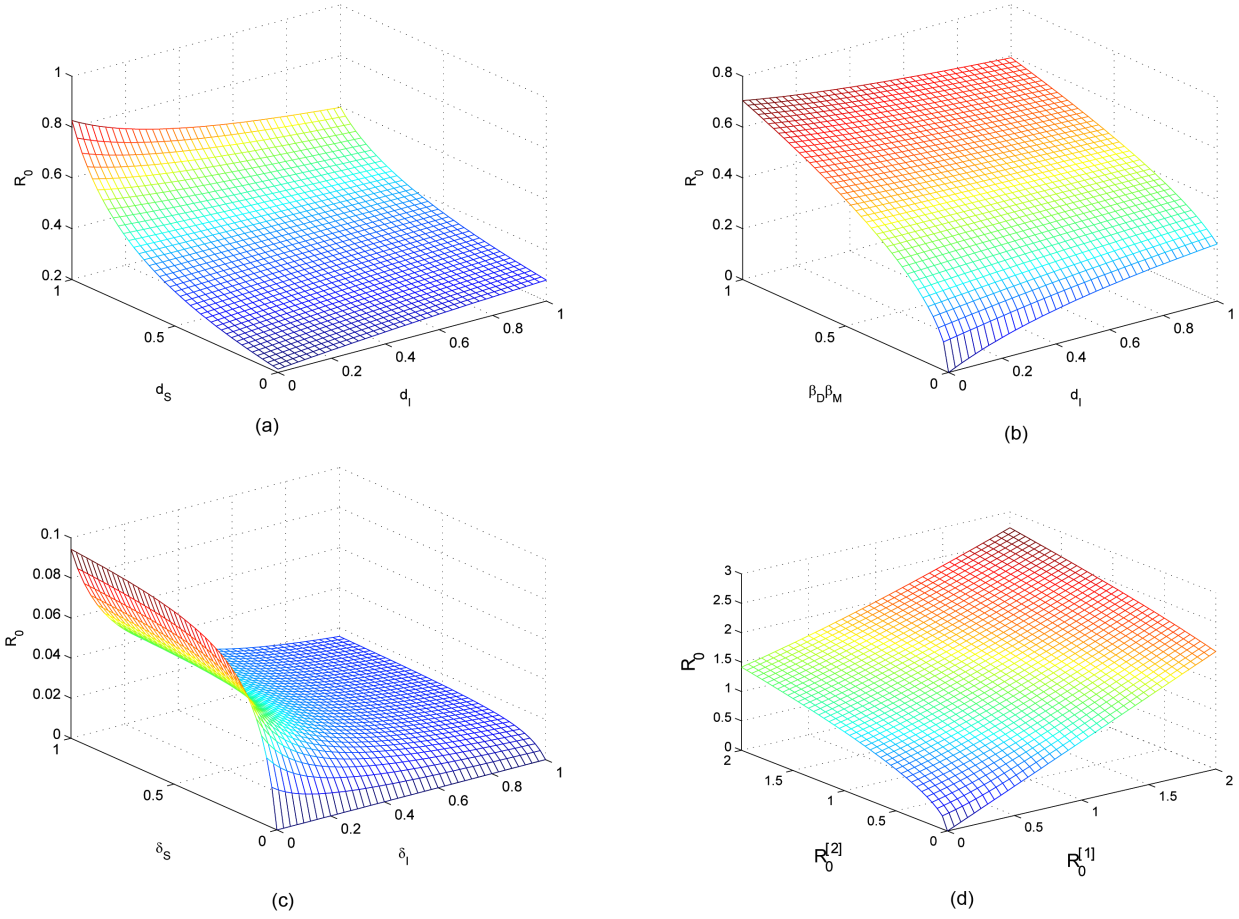


Figure 2. Numerical simulations of R_0 as a function of the selected model parameters. (a) R_0 values increase with d_I provided d_S values are small. When d_S values are large, R_0 decreases with d_I; (b) R_0 increases both with $\beta_D\beta_M$ and d_I; (c) R_0 increases with δ_S and decreases with δ_I; (d) R_0 increases linearly with $R_0^{[1]}$ and increases parabolically with $R_0^{[2]}$.

Figure 2(a) shows that R_0 is an increasing function with respect to d_S. The influx of additional, susceptible deer into a patch leads to an increased number of potential interactions with infected midges and thus an increase in the number of infections overall. Figure 2(c) shows that R_0 is an increasing function with respect to δ_S and a decreasing function with respect to δ_I. Figure 2(a) and Figure 2(b) demonstrate the behavior of R_0 with respect to the influx of infected deer, d_I. For smaller values of d_S or $\beta_D\beta_M$, R_0 is an increasing function with respect to d_I; for larger values of d_S or $\beta_D\beta_M$, it is a decreasing function with respect to d_I. Thus, there must be a critical value ($d_S^{[c]}$ or $(\beta_D\beta_M)^{[c]}$) where the behavior changes.

If we consider R_0 as a function of the deer-midge interactions, then R_0 is essentially a linear function of $R_0^{[1]}$ and a function of the square root of $R_0^{[2]}$. The graph of R_0 would be increasing and concave down with respect to an increase in $R_0^{[2]}$ (See Figure 2(d)). This is consistent with what we would expect to happen. As the amount of interaction increases, so does the number of potential new infections with a greater chance of an outbreak occurring. Plus, as a greater proportion of the deer population becomes infected, the rate of increase of R_0 must decrease as the number of uninfected deer will consequently drop.

We also demonstrate numerically that the solutions of model (1) converge to the endemic equilibrium if $R_0 > 1$ and achieves a disease free equilibrium if $R_0 < 1$. To do this, a MATLAB code was written utilizing the ODE45 solver, and the results were verified against the computed R_0 value for a given set of parameters. At time $t = 0$, we have the following initial values: $D_S(0) = 30$, $D_I(0) = 10$, $M_S(0) = 20$, $M_I(0) = 5$, $\bar{D}_S(0) = 10$, and $\bar{D}_I(0) = 1$. See Table 2 for the specific parameter values used for the numerical simulations.

Figure 3(a) and Figure 3(c) show the long-term behavior of the four classes of deer populations-total susceptible, total infected, susceptible influx, and infected influx-plotted on the same graph, while Figure 3(b) and Figure 3(d) show the long-term behavior of the susceptible and infected midge populations.

Table 2. Parameter values used in model simulation and the calculated R_0 values.

Parameter	Value when $R_0 = 0.40$	Value when $R_0 = 2.19$	Parameter	Value when $R_0 = 0.40$	Value when $R_0 = 2.19$
β_D	0.2	1.1	γ_D	0.35	0.1
λ_D	0.9	0.9	β_M	0.2	1.6
ρ	0.2	0.2	λ_M	0.9	0.9
μ_D	0.1	0.1	σ	0.2	0.2
μ_{2D}	0.2	0.5	μ_M	0.05	0.05
d_S	0.3	0.1	μ_{2M}	0.05	0.05
d_I	0.3	0.42			

Note: The R_0 values are consistent with the numerical simulations shown in Figure 3. Similar results were obtained using different sets of parameter values.

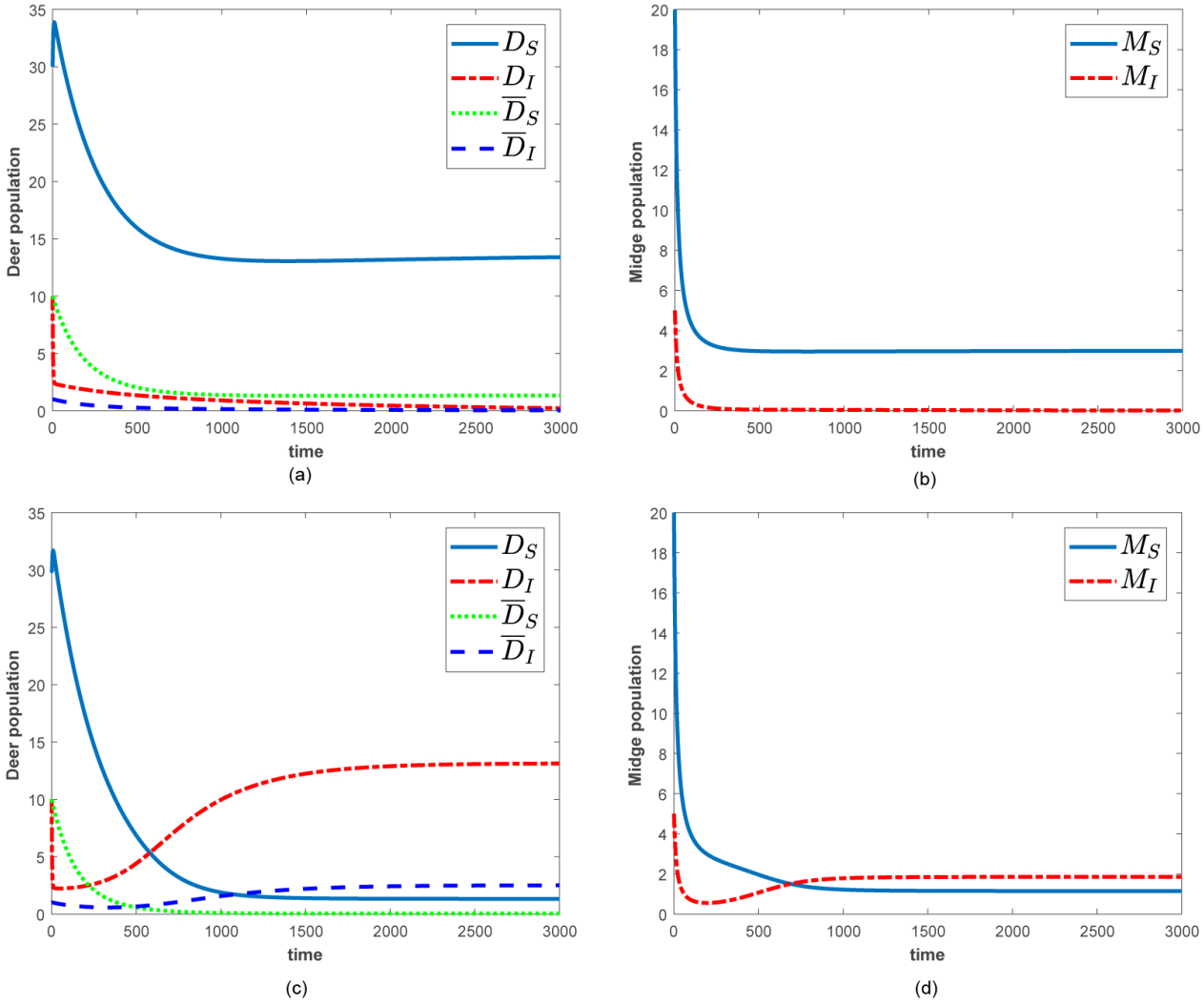


Figure 3. (a) (b) When the basic reproduction number $R_0 < 1$, the system stabilizes to its disease free equilibrium and the number of infected deer, the number of dispersing infected deer, and the number of infected midges tends to zero as t increases; (c) (d) When the basic reproduction number $R_0 > 1$, the system stabilizes to its endemic equilibrium. See Table 2 for the specific values used and the corresponding values of R_0.

Figure 3(a) and Figure 3(b) indicate that when $R_0 < 1$, the system will stabilize to its disease free equilibrium. Figure 3(c) and Figure 3(d) show that when $R_0 > 1$, the system will stabilize to an endemic equilibrium. These outcomes are robust for large sets of initial values and parameter values.

5. Discussion

In this paper, we have developed a distributed delay model for transmission dynamics of HD in a deer population. Though mathematical models for disease and HD specifically are established, we chose to focus on how the dynamics are affected by the dispersion (migration) of deer specifically and how the basic reproduction number is affected by these dispersion rates (*i.e.*, d_S and d_I). The results show that there are critical values for the interaction parameters

$(\beta_D\beta_M)^{[c]}$ and rates of susceptible deer dispersion $d_S^{[c]}$. Hence, possible outbreaks could be avoided by controlling how and where these deer move.

One of the primary limitations of this study is the lack of actual parameter values. Although the qualitative behavior of model (1) remains fairly distinctive, (*i.e.*, convergence to DFE or EE) for large sets of parameter values, many of the values were chosen randomly. It is our goal to estimate some of the parameter values using data from the Missouri Department of Conservation concerning the prevalence of HD in Missouri's white-tailed deer. Nevertheless, the graphs presented in **Figure 2** and **Figure 3** show consistent tendencies in the behavior in the model. We also have not considered behavior in a multi-patch system, where migrating individuals leave one patch and eventually enter a neighboring patch, nor did we consider a delay in the traveling time. Holt [25] and Weisser *et al.* [26] extended their results to a system of multiple patches joined through a pool of dispersing individuals. Moreover, the proposed model (1) does not include the effect of predators on the population of white-tailed deer. As a prey species, deer are linked with local predators. In Missouri, the coyote is one such predator. Some coyote predator studies have been done, but these are admittedly outdated. However, deer make up a portion of a coyote's diet and that large increases or decreases in predator populations may influence deer mortality rates [28]. Finally, our model assumed only one vector for the transmission of HD. With the species richness of the *Culicoides* genus, we may reasonably expect more and different interaction rates and different levels of control efficacy [27]. We also note that weather has an effect on both the midge population and the life cycle of the HD virus [2] [29]. Midge populations thrive in damper areas, and in 2012, there was an above average amount of rain in the late winter/early spring, filling ponds and other water bodies in Missouri [28]. In addition, record warm temperatures in that spring and summer may cause midges to become more active sooner than normal [28]. Next, the high temperatures caused water sources to dry up, and not only did the resulting mud flats become ideal breeding areas for subsequent generations of midges, but also caused deer to visit water sources more frequently due to lower water content in the plants they ate as part of their diet. These same high temperatures also cause female midges to lay more eggs, and Wittmann *et al.* also revealed that higher temperatures decrease the extrinsic incubation period of the HD virus within the midges [30]. Thus, the virus develops faster and allows a midge to infect more deer during its life span. None of these factors have been considered in the model (1). Instead, the main focus has been on migration effects of deer population on overall HD dynamics within a patch.

6. Conclusion

The above mentioned limitations demand model extensions to study the effectiveness of control and preventive strategies. Deer species are important members of the ecosystem as they feed on brush and grass in a given area and

keep them in check. In conclusion, the present work is the first step towards inclusion of migration effects of deer population modeling of HD dynamics. The R_0 expression provides insights into the effects of deer movement on the spread of disease.

References

[1] Flinn, E. and Sumners, J. (2013a) Breaking Down the Hemorrhagic Disease Outbreak. *Missouri Conservationist*, **74**, 24-29.

[2] Sleeman, J.M., Howell, J.E., Knox, W.M. and Stenger, P.J. (2009) Incidence of Hemorrhagic Disease in White-Tailed Deer Is Associated with Winter and Summer Climactic Conditions. *EcoHealth*, **6**, 11-15. https://doi.org/10.1007/s10393-009-0220-6

[3] Foster, N.M., Breckon, R.D., Luedke, A.J., Jones, R.H. and Metcalf, H.E. (1977) Transmission of Two Strains of Epizootic Hemorrhagic Disease Virus in Deer by *Culicoides variipennis*. *Journal of Wildlife Diseases*, **13**, 9-16. https://doi.org/10.7589/0090-3558-13.1.9

[4] Ross, R. (1911) The Prevention of Malaria. John Murray, London.

[5] Macdonald, G. (1957) The Epidemiology and Control of Malaria. Oxford University Press, London.

[6] Lou, Y. and Zhao, X.-Q. (2009) The Periodic Ross-Macdonald Model with Diffusion and Advection. *Applicable Analysis*, **89**, 1067-1089. https://doi.org/10.1080/00036810903437804

[7] Wang, W. and Zhao, X.-Q. (2011) A Nonlocal and Time-Delayed Reaction Diffusion Model of Dengue Transmission. *SIAM Journal of Applied Mathematics*, **71**, 147-168. https://doi.org/10.1137/090775890

[8] Fitzgibbon, W.E., Morgan, J.J. and Webb, G.B. (2017) An Outbreak Vector-Host Epidemic Model with Spatial Structure: The 2015-2016 Zika Outbreak in Rio De Janeiro. *Theoretical Biology and Medical Modeling*, **14**, 7. https://doi.org/10.1186/s12976-017-0051-z

[9] Neubert, M.G., Klepac, P. and Van Den Driessche, P. (2001) Stabilizing Dispersal Delays in Predator-Prey Metapopulations Models. *Theoretical Population Biology*, **61**, 339-347. https://doi.org/10.1006/tpbi.2002.1578

[10] Klepac, P., Neuber, M.G. and Van Den Driessche, P. (2007) Dispersal Delays, Predator-Prey Stability, and the Paradox of Enrichment. *Theoretical Population Biology*, **7**, 436-444. https://doi.org/10.1016/j.tpb.2007.02.002

[11] Park, A.W., Magori, K., White, B.A. and Stallknecht, D.E. (2013) When More Transmission Equals Less Disease: Reconciling the Disconnect between Disease Hotspots and Parasite Transmission. *PLoS ONE*, **8**, e61501. https://doi.org/10.1371/journal.pone.0061501

[12] Purdue University, Extension E-250-W. Biting Midges: Biology and Public Health Risk. https://extension.entm.purdue.edu/publichealth/insects/bitingmidge.html

[13] Sedda, L., Brown, H.E., Purse, B.V., Burgin, L., Gloster, J. and Rogers, D.J. (2012) A New Algorithm Quantifies the Roles of Wind and Midge Flight Activity in the Bluetongue Epizootic in Northwest Europe. *Proceedings: Biological Sciences*, **279**, 235462. http://www.jstor.org.proxy.library.umkc.edu/stable/41549546 https://doi.org/10.1098/rspb.2011.2555

[14] Ngwa, G. and Shu, W. (1999) A Mathematical Model for Endemic Malaria with Va-

riable Human and Mosquito Populations. United Nations Educational Scientific and Cultural Organization and International Atomic Energy Agency, The Abdus Salam International Centre for Theoretical Physics, Miramare-Trieste.

[15] Kuang, Y. (1993) Delay Differential Equations with Applications in Population Dynamics. Academic Press, Inc., San Diego.

[16] Perko, L. (2001) Differential Equations and Dynamical Systems. 3rd Edition, Springer, New York. https://doi.org/10.1007/978-1-4613-0003-8

[17] Routh, E.J. (1877) A Treatise on the Stability of a Given State of Motion: Particularly Steady Motion. Macmillan, London.

[18] Dietz, K. (1993) The Estimation of the Basic Reproduction Number for Infectious Diseases. *Statistical Methods in Medical Research*, **2**, 23-41. https://doi.org/10.1177/096228029300200103

[19] Bani-Yaghoub, M., Gautam, R., Ivanek, R., van den Driessche, P. and Shuai, Z. (2012) Reproduction Numbers for Infections with Free-Living Pathogens Growing in the Environment. *Journal of Biological Dynamics*, **6**, 923-940. https://doi.org/10.1080/17513758.2012.693206

[20] Van den Driessche, P. and Watmough, J. (2002) Reproduction Numbers and Sub-Threshold Endemic Equilibria for Compartmental Models of Disease Transmission. *Mathematical Biosciences*, **180**, 29-48. https://doi.org/10.1016/S0025-5564(02)00108-6

[21] Hurford, A., Cownden, D. and Day, T. (2010) Next-Generation Tools for Evolutionary Invasion Analyses. *Journal of the Royal Society Interface*, **7**, 561-571. https://doi.org/10.1098/rsif.2009.0448

[22] Heesterbeek, J.A.P. and Dietz, K. (1996) The Concept of R0 in Epidemic Theory. *Statistica Neerlandica*, **50**, 89-110. https://doi.org/10.1111/j.1467-9574.1996.tb01482.x

[23] Heesterbeek, J.A.P. and Roberts, M.G. (2007) The Type-Reproduction Number T in Models for Infectious Disease Control. *Mathematical Biosciences*, **206**, 3-10. https://doi.org/10.1016/j.mbs.2004.10.013

[24] Jordan, D.W. and Smith, P. (1999) Nonlinear Ordinary Differential Equations: An Introduction to Dynamical Systems. Oxford University Press, Oxford.

[25] Holt, R.D. (1984) Spatial Heterogeneity, Indirect Interaction, and the Coexistence of Prey Species. *American Naturalist*, **124**, 377-406. https://doi.org/10.1086/284280

[26] Weisser, W.W., Jansen, V.A.A. and Hassell, M.P. (1997) The Effects of a Pool of Dispersers on Host-Parasitoid Systems. *Journal of Theoretical Biology*, **189**, 413-425. https://doi.org/10.1006/jtbi.1997.0529

[27] Park, A.W., Cleveland, C.A., Dallas, T.A. and Corn, J.L. (2016) Vector Species Richness Increases Haemorrhagic Disease Prevalence through Functional Diversity Modulating the Duration of Seasonal Transmission. *Parasitology*, **143**, 874-879. https://doi.org/10.1017/S0031182015000578

[28] Flinn, E. and Sumners, J. (2013b) State of the States Deer Herd. *Missouri Conservationist*, **74**, 24-29.

[29] Mellor, P.S., Boorman, J. and Baylis, M. (2000) Culicoides Biting Midges: Their Role as Arbovirus Vectors. *Annual Review of Entomology*, **45**, 307-340. https://doi.org/10.1146/annurev.ento.45.1.307

[30] Wittmann, E.J., Mellor, P.S. and Baylis, M. (2002) Effect of Temperature on the Transmission of Orbiviruses by the Biting Midge, *Culicoides sonorensis*. *Medical and Veterinary Entomology*, **16**, 147-156. https://doi.org/10.1046/j.1365-2915.2002.00357.x

Scientific Research Publishing

Journal of Applied Mathematics and Physics, 2017, 5, 2283-2290
http://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Bianchi Type-V Cosmological Models for Perfect Fluid with Time-Varying Gravitational and Cosmological Constant

Mohammed Aman Ullah, Mohammad Amjad Hossain, Mohammad Moksud Alam

Department of Mathematics, University of Chittagong, Chittagong, Bangladesh
Email: *aman@cu.ac.bd, mah62235@gmail.com, moksud.math@cu.ac.bd

How to cite this paper: Ullah, M.A., Hossain, M.A. and Alam, M.M. (2017) Bianchi Type-V Cosmological Models for Perfect Fluid with Time-Varying Gravitational and Cosmological Constant. *Journal of Applied Mathematics and Physics*, **5**, 2283-2290.
https://doi.org/10.4236/jamp.2017.511185

Received: September 30, 2017
Accepted: November 26, 2017
Published: November 29, 2017

Abstract

Einstein's field equations with variable gravitational and cosmological constants are considered in presence of perfect fluid for locally-rotationally-symmetric (LRS) Bianchi type-V space-time discussion in context of the particle creation. We present new shear free solutions for both absence and presence of particle creation. The solution describes the particle and entropy generation in the anisotropic cosmological models. We observe that time variation of gravitational and cosmological constant is needed for particle creation phenomena. Moreover, we obtained the particle production rate $\Gamma(t)$ for this model and discussed in detail.

Keywords

LRS Bianchi Type-V, Perfect Fluid, Gravitational Constant, Cosmological Constant, Particle Creation

1. Introduction

Einstein field equation in general relativity [1] and cosmology contains two parameters: Newton's gravitational constant G and the cosmological constant Λ. A number of authors have considered the cosmological model with the cosmological constant Λ as a function of cosmic time. An important role of cosmological constant with the relation $\Lambda \propto t^{-2}$ studied by Berman *et al.* [2] and Berman [3] [4]. The time variation of the gravitational constant G was first proposed by Dirac [5] and extensively discussed in the literature by Weinberg [6]. Lau [7], Lau and Prokhovnik [8] proposed generalized field equations with time-dependent G and Λ, since then many authors have investigated cosmological models with variables G and Λ. Arbab [9] has discussed a viscous model

with variable G and Λ claiming that energy is conserved. Singh *et al.* [10] discussed a number of classes of solutions to Einstein's field equations with variable G and Λ, and bulk viscosity for a flat Robertson-Walker universe in the framework of general relativity.

The Bianchi cosmologies which are spatially homogeneous and anisotropic play an important in theoretical cosmology and have been much studied since 1960s. Coley [11] has investigated Bianchi type-V spatially homogeneous imperfect fluid cosmological models which contain both viscosity and heat flow. Coley and Hoogen [12] have studied a locally-rotationally-symmetric (LRS) Bianchi type-V metric for imperfect fluid source with both viscosity and heat conduction. Singh and Beesham [13] have presented LRS Bianchi type-V cosmological models in the presence of perfect fluid with heat conduction. Singh [14] has extended the work to LRS Bianchi type-V cosmological models and obtained solutions of the field equations in general relativity. A number of studies in cosmological model are performed using the gravitational constant and cosmological constant with cosmic time variable in the context of isotropic perfect fluid [9] [15] [16] and anisotropic [17] [18] [19] [20].

Recently, particle creation and in the absence of particle creation in cosmology and its phenomenological description are discussed in detail by Singh [21]. Previously, it was also attracted a lot of interests shown in [22] [23].

In this paper, we have studied the evolution and dynamics of a perfect fluid LRS Bianchi type-V models with variable G and Λ which describe the particle creation. We also try to present the exact solutions of Einstein's field equations in the case of particle creation and in the absence of particle creation.

2. Field Equations

We consider a LRS Bianchi type-V space time with the metric [11]

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + A^2(t)\mathrm{d}x^2 + \mathrm{e}^{2x}B^2(t)\left(\mathrm{d}y^2 + \mathrm{d}z^2\right) \tag{1}$$

where $A(t)$ and $B(t)$ are the cosmic scale functions. The Einstein's field equations with time-dependent G and Λ are given by Abdel-Rahaman [15].

$$R_{ij} - \frac{1}{2}g_{ij}R = -8\pi G(t)T_{ij} + \Lambda(t)g_{ij} \tag{2}$$

where the energy-momentum tensor T_{ij} is that of a perfect fluid.

For perfect fluid, T_{ij} is given by

$$T_{ij} = (p+\rho)u_i u_j + pg_{ij} \tag{3}$$

where ρ is the matter density, p is the thermodynamics pressure and u_i is the four-velocity vector of the fluid satisfying $u_i u^i = -1$. In a co-moving coordinate system, the field Equations (2), for the metric (1), in case of (3), read as [18]:

$$2\frac{\ddot{B}}{B} + \frac{\dot{B}^2}{B^2} - \frac{1}{A^2} = -8\pi G(t)p + \Lambda(t), \tag{4}$$

$$\frac{\ddot{A}}{A}+\frac{\ddot{B}}{B}+\frac{\dot{A}}{A}\frac{\dot{B}}{B}-\frac{1}{A^2}=-8\pi G(t)\,p+\Lambda(t), \tag{5}$$

$$2\frac{\dot{A}}{A}\frac{\dot{B}}{B}+\frac{\dot{B}^2}{B^2}-\frac{3}{A^2}=8\pi G(t)\rho+\Lambda(t), \tag{6}$$

$$\frac{\dot{A}}{A}-\frac{\dot{B}}{B}=0 \tag{7}$$

where dot denotes the ordinary derivative with respect to the cosmic time t and double dot stands for second derivative w.r. to the same.

The vanishing of the covariant divergence of the Einstein tensor leads to the following useful equation

$$\dot{\rho}+(\rho+p)\left(\frac{\dot{A}}{A}+2\frac{\dot{B}}{B}\right)=-\left(\frac{\dot{G}}{G}\rho+\frac{\dot{\Lambda}}{8\pi G}\right) \tag{8}$$

The average scale factor R for the LRS Bianchi type-V model is defined as

$$R=\left(AB^2\right)^{\frac{1}{3}} \tag{9}$$

The generalized mean Hubble parameter H can be defined as in [24]

$$H=\frac{\dot{a}}{a}=\frac{1}{3}\left(\frac{\dot{A}}{A}+2\frac{\dot{B}}{B}\right) \tag{10}$$

The physical quantities of observation of interest in cosmology and the expansion scalar θ, the average anisotropy parameter A_p and the shear scalar σ^2, are defined as [21]

$$\theta=u^i_{;i}=\left(\frac{\dot{A}}{A}+2\frac{\dot{B}}{B}\right) \tag{11}$$

$$A_p=\frac{1}{3}\sum_{i=1}^{3}\left(\frac{H_i-H}{H}\right)^2 \tag{12}$$

$$\sigma^2=\frac{1}{2}\sigma_{ij}\sigma^{ij}=\frac{1}{3}\left(\frac{\dot{A}}{A}-\frac{\dot{B}}{B}\right)^2 \tag{13}$$

The particle content of the early universe is formed from a non-interacting comoving relativistic fluid having a particle number density $n(t)$ and obeying the equation of state of the form

$$n=n_0\left(\frac{\rho}{\rho_0}\right)^{\frac{1}{1+\alpha}} \tag{14}$$

$$p=\alpha\rho \tag{15}$$

where $n_0\geq 0$ and $\rho_0\geq 0$ are constants and $0\leq\alpha\leq 1$. Equation (8) can be written, using Equations (14)-(15), in the form of a particle balance equation

$$\dot{n}+3Hn=\Gamma(t)n \tag{16}$$

where $\Gamma(t)$ is the particle production rate given by

$$\Gamma(t) = -\frac{1}{(1+\alpha)G(t)}\left[\frac{\dot{\Lambda}(t)}{8\pi\rho(t)} + \dot{G}(t)\right] \tag{17}$$

For particle creation, it is required that Equation (17) satisfies $\Gamma(t) \geq 0$. This implies that the space-time can produce matter, while the reverse case is thermodynamically forbidden. Which leads to

$$\frac{\dot{\Lambda}(t)}{8\pi\rho(t)} + \dot{G}(t) \leq 0 \tag{18}$$

The entropy S generated during the particle creation is given by

$$T\frac{\mathrm{d}S}{\mathrm{d}t} = -\frac{1}{G(t)}\left[\frac{\dot{\Lambda}(t)}{8\pi\rho(t)} + \dot{G}(t)\right]\rho(t)R^3 \tag{19}$$

which can be written as

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \frac{(1+\alpha)\rho(t)R^3}{T}\Gamma(t) \tag{20}$$

In a cosmological fluid where the density and pressure are functions of the temperature only, *i.e.* $\rho = \rho(T)$ and $p = p(T)$, the entropy is given by [6]

$$S = \frac{(\rho + p)a^3}{T} = \frac{(1+\alpha)\rho a^3}{T} \tag{21}$$

The total entropy of the cosmological fluid function of the particle production rate is given by

$$S = S_0 \mathrm{e}^{\int \Gamma(t)\mathrm{d}t} \tag{22}$$

where $S_0 \geq 0$ is a constant of integration. In the case of no particle production, *i.e.* $\Gamma(t) = 0$, we have the usual particle conservation law of the standard cosmology, *i.e.*

$$\dot{\rho} + 3(1+\alpha)H_p = 0 \tag{23}$$

$$\dot{\Lambda}(t) + 8\pi\rho\dot{G}(t) = 0 \tag{24}$$

For the case of entropy $S(t) = S_0 = \text{constant}$, and thus the change of entropy is zero.

3. Solution of Field Equations

From Equation (7), we have

$$B = kA \tag{25}$$

where k is constant of integration.

From Equations (4) and (5), we have

$$\frac{\dot{B}^2}{B^2} = \frac{\dot{A}^2}{A^2} \tag{26}$$

Using the relations (25) and (26), Equations (5) and (6) yields

$$2\frac{\dot{A}^2}{A^2} - 2\frac{\ddot{A}}{A} - \frac{2}{A^2} = 8\pi G(1+\alpha)\rho \tag{27}$$

Again, from the Equations (5) and (6), we have

$$2\Lambda + 8\pi G(t)(p-\rho) = 2\frac{\ddot{A}}{A} + 4\frac{\dot{A}^2}{A^2} - \frac{4}{A^2} \tag{28}$$

As described in [21], we have a system of five Equations (4)-(8) with six unknowns, namely A, B, G, ρ, Λ and p. To find the exact solution of the field equations, we use the power law equation. The particle creation is described due to the variation of G and Λ as shown in [18]. We obtain the solution for ρ, G and Λ in the following two cases:

3.1. In the Absence of Particle Creation:

Consider the left part of Equation (8) equal to zero *i.e.*

$$\dot{\rho} + (\rho + p)\left(\frac{\dot{A}}{A} + 2\frac{\dot{B}}{B}\right) = 0$$

Using the Equations (25) and (15), the above equation can be written as

$$i.e.\ \frac{\dot{\rho}}{\rho} + 3(1+\alpha)\frac{\dot{A}}{A} = 0$$

$$i.e.\ \rho = \frac{k_1}{A^{3(1+\alpha)}} \tag{29}$$

where k_1 is integrating constant. Now, using Equation (29), Equation (27) reduces to

$$G(t) = \frac{A^{3\alpha+1}}{4\pi(1+\alpha)k_1}\left\{\dot{A}^2 - A\ddot{A} - 1\right\} \tag{30}$$

Again, using the Equations (25) and (26), Equations (5) + (6) give

$$\Lambda = \frac{2}{1+\alpha}\cdot\frac{\ddot{A}}{A} + \left(\frac{3\alpha+1}{1+\alpha}\right)\cdot\frac{\dot{A}^2}{A^2} - \left(\frac{3+\alpha}{1+\alpha}\right)\cdot\frac{1}{A^2} \tag{31}$$

Also, from the right side of Equation (8) gives

$$\frac{\dot{G}}{G}\rho + \frac{\dot{\Lambda}}{8\pi G} = 0$$

$$i.e.\ A\ddot{A} - \dot{A}^2 + \frac{3+\alpha}{3\alpha+1} = 0$$

$$i.e.\ A = \frac{\beta}{2\gamma}\left(e^{\gamma t} - e^{-\gamma t}\right) \tag{32}$$

where β and γ are constants. So,

$$B = \frac{k\beta}{2\gamma}\left(e^{\gamma t} - e^{-\gamma t}\right) \tag{33}$$

and $R^3 = AB^2$ which gives

$$R = \frac{k\beta}{2\gamma}\left(e^{\gamma t} - e^{-\gamma t}\right) \tag{34}$$

The Hubble parameter and deceleration parameter are respectively given by

$$H=\gamma\left(\frac{1+e^{-2\gamma t}}{1-e^{-2\gamma t}}\right),\quad q=-\left(\frac{1-e^{-2\gamma t}}{1+e^{-2\gamma t}}\right)^2=-\tanh^2(\gamma t) \tag{35}$$

The energy density, gravitational and cosmological constants and expansion scalar are given by

$$\rho=\frac{k(2\gamma)^{3(1+\alpha)}}{\left\{\beta\left(e^{\gamma t}-e^{-\gamma t}\right)\right\}^{3(1+\alpha)}} \tag{36}$$

$$G(t)=\frac{\beta^2-1}{4\pi k_1(1+\alpha)}\cdot\left(\frac{\beta}{2\gamma}\right)^{3\alpha+1}\left(e^{\gamma t}-e^{-\gamma t}\right)^{3\alpha+1} \tag{37}$$

$$\Lambda(t)=3\gamma^2\left(\frac{e^{\gamma t}+e^{-\gamma t}}{e^{\gamma t}-e^{-\gamma t}}\right)^2-\frac{3+\alpha}{1+\alpha}\cdot\frac{4\gamma^2}{\beta^2}\left(e^{\gamma t}-e^{-\gamma t}\right)^{-2} \tag{38}$$

$$\theta=3\gamma\frac{\left(e^{\gamma t}+e^{-\gamma t}\right)}{\left(e^{\gamma t}-e^{-\gamma t}\right)}=3\gamma\coth(\gamma t) \tag{39}$$

The above discussion, it is clear that the Hubble's parameter, energy density, cosmological constant, deceleration parameter and expansion scalar are decreasing as functions of time. But only gravitational constant is increasing as functions of time.

Some authors [5] [16] considered that G the cosmological constant is a decreasing parameter as function of time, while some other [9] [25] are taken G as increasing as function of time.

3.2. In the Case of Particle Creation

Let us consider, $A=at^m$ and $G=bt^r$, where a,b,m and r are constants.

Now, we have from Equation (27)

$$\rho=\frac{1}{8\pi(1+\alpha)}\left\{\frac{2m}{b}\cdot\frac{1}{t^{r+2}}-\frac{2}{a^2b}\cdot\frac{1}{t^{2m+r}}\right\} \tag{40}$$

$$\Lambda(t)=\frac{3(1+\alpha)m^2-2m}{1+\alpha}\cdot\frac{1}{t^2}-\frac{3\alpha+1}{a^2(1+\alpha)}\cdot\frac{1}{t^{2m}} \tag{41}$$

which satisfies the Equation (8).

The particle production rate is given by (17)

$$\Gamma(t)=\frac{1}{1+\alpha}\left[\left\{\frac{6a^2(1+\alpha)m^2t^{2m+1}-\left[4m+2m(3\alpha+1)t^3\right]}{2ma^2t^{2m+2}-2t^4}\right\}+\frac{r}{t}\right] \tag{42}$$

The other physical quantities are

$$H=3\frac{m}{t} \tag{43}$$

$$\theta=u^i_{;i}=3\frac{m}{t} \tag{44}$$

The physical quantities such as energy density, cosmological constant, Hub-

ble's parameter and expansion scalar are decreasing with the increasing of the time and tend to zero when time tends to infinity. The particle creation decreasing when time increasing and tends to zero when time tends to infinity. The mass within the co-moving volume $V \propto R^3$ is given by $nM_p = \rho R^3$, where n is the number density of the particles and M_p the particle mass.

4. Conclusion

We have obtained solution for two cases: viz absence of particle creation and for particle creation of Einstein's field equation of a locally-rotationally-symmetric Bianchi type-V universe with cosmological constant and gravitational constant as a cosmic time. The physical quantities are realistic *i.e.* they behave physically for cosmological cases. The particle production rate $\Gamma(t)$ decreasing when time is increasing but when time tends to zero $\Gamma(t)$ tends to infinity which is also physical because of Big bang theory.

Acknowledgements

Authors would like to thank Professor Dr. Mohammad Ashraful Islam, Department of Mathematics, University of Chittagong, Chittagong, Bangladesh for his valuable suggestion and kind help during this work. Authors also thank the reviewers for their valuable comments and suggestions, which are really helpful in revising this paper.

References

[1] Einstein, A. (1917, 1919) English translation: The Principle of Relativity (Methuen, 1923, Reprinted by Dover, 1924). *Sitz. Ber. Preuss. Acad. Wiss.,* 177 and 191.

[2] Berman, M.S., Som, M.M. and Gomide, F.M. (1989) Brans-Dicke static universes. *General Relativity and Gravitation*, **21**, 287-292. https://doi.org/10.1007/BF00764101

[3] Berman, M.S. (1990a) Static Universe in a Modified Brans-Dicke Cosmology. *International Journal of Theoretical Physics*, **29**, 567-570. https://doi.org/10.1007/BF00672031

[4] Berman, M.S. (1990b) Kantowski-Sachs Cosmological Models with Constant Deceleration Parameter. *Nuovo Cimento B,* **105**, 239-242. https://doi.org/10.1007/BF02723079

[5] Dirac, P.A.M. (1937) The Cosmological Constants. *Nature*, **139**, 323. https://doi.org/10.1038/139323a0

[6] Weinberg, S. (1989) The Cosmological Constant Problem. *Reviews of Modern Physics*, **61**, 1-23. https://doi.org/10.1103/RevModPhys.61.1

[7] Lau, Y.K. (1985) The Large Number Hypothesis and Einstein's Theory of Gravitation. *Australian Journal of Physics,* **38**, 547. https://doi.org/10.1071/PH850547

[8] Lau, Y.K. and Prokhovnik, S.J. (1986) The Large Numbers Hypothesis and a Relativistic Theory of Gravitation. *Australian Journal of Physics*, **39**, 339-346. https://doi.org/10.1071/PH860339

[9] Arbab, A.I. (1997) Cosmological Models with Variable Cosmological and Gravitational "Constants" and Bulk Viscous Models. *General Relativity and Gravitation*,

29, 61-74. https://doi.org/10.1023/A:1010252130608

[10] Singh, C.P., Kumar, S. and Pradhan, A. (2007) Early Viscous Universe with Variable Gravitational and Cosmological "Constants". *Class. Quantum Gravity*, **24**, 455-474. https://doi.org/10.1088/0264-9381/24/2/011

[11] Coley, A.A. (1990) Bianchi-V Imperfect Fluid Cosmology. *General Relativity and Gravitation*, **22**, 3-18. https://doi.org/10.1007/BF00769241

[12] Coley, A.A. and Hoogen, R.J. (1994) Qualitative Analysis of Diagonal Bianchi Type-V Imperfect Fluid Cosmological Models. *Journal of Mathematical Physics*, **35**, 4117-4144. https://doi.org/10.1063/1.530845

[13] Singh, C.P. and Beesham, A. (2009) Locally-Rotationally-Symmetric Bianchi Type-V Cosmology with Heat Flow. *Pramana—Journal of Physics*, **73**, 793-798. https://doi.org/10.1007/s12043-009-0147-z

[14] Singh, C.P. (2009) Locally-Rotationally-Symmetric Bianchi Type-V Cosmology in General Relativity. *Pramana—Journal of Physics*, **72**, 429-443. https://doi.org/10.1007/s12043-009-0038-3

[15] Abdel-Rahaman, A.M.M. (1990) A Critical Density Cosmological Model with Varying Gravitational and Cosmological "Constants". *General Relativity and Gravitation*, **22**, 655-663. https://doi.org/10.1007/BF00755985

[16] Beesham, A. (1986) The Cosmological Constant (Λ) as a Possible Primordial Link to Einstein's Theory of Gravity, the Properties of Hadronic Matter and the Problem of Creation. *Nuovo Cimento B*, **96**, 17-20. https://doi.org/10.1007/BF02725574

[17] Chakrabarty, I. and Pradhan, A. (2001) LRS Bianchi I Models with Time-Varying Gravitational and Cosmological Constants. *Gravitation and Cosmology*, **7**, 55-57.

[18] Pradhan, A. and Yadav, V.K. (2002) Bulk Viscous Anisotropic Cosmological Models with Variable G and Λ. *International Journal of Modern Physics D*, **11**, 893-912. https://doi.org/10.1142/S0218271802002050

[19] Pradhan, A., Singh, A.K. and Otarod, O. (2007) FRW Universe with Variable G and Λ-Terms. *Romanian Journal of Physics*, **52**, 445.

[20] Ram, S., Zeyauddin, M. and Singh, C.P. (2009) Bianchi Type-V Cosmological Models with Perfect Fluid and Heat Flow in Saez-Ballester Theory. *Pramana—Journal of Physics*, **72**, 415-427. https://doi.org/10.1007/s12043-009-0037-4

[21] Singh, C.P. (2011) Cosmological Models with Time-Varying Gravitational and Cosmological "Constants". *Astrophysics and Space Science*, **331**, 337-342. https://doi.org/10.1007/s10509-010-0439-2

[22] Harko, T. and Mak, M.K. (1999) Particle Creation in Cosmological Models with Varying Gravitational and Cosmological "Constants". *General Relativity and Gravitation*, **31**, 849-862. https://doi.org/10.1023/A:1026634204476

[23] Harko, T. and Mak, M.K. (1999) Particle Creation in Varying Speed of Light Cosmological Models. *Classical and Quantum Gravity*, **16**, 2741. https://doi.org/10.1088/0264-9381/16/8/312

[24] Grøn, Ø. (1985) Expansion Isotropization during the Inflationary Era. *Physical Review D*, **32**, 2522. https://doi.org/10.1103/PhysRevD.32.2522

[25] Kalligas, D., Wesson, P. and Everitt, C.W.F. (1992) Flat FRW Models with Variable G and Λ. *General Relativity and Gravitation*, **24**, 351-357. https://doi.org/10.1007/BF00760411

Scientific Research Publishing

Journal of Applied Mathematics and Physics, 2017, 5, 2291-2301
http://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Application of Improved Artificial Bee Colony Algorithm in Urban Vegetable Distribution Route Optimization

Zhenzhen Zhang, Lianhua Wang

Beijing Wuzi University, Beijing, China
Email: 739336961@qq.com, lianhuawang@sina.com

How to cite this paper: Zhang, Z.Z. and Wang, L.H. (2017) Application of Improved Artificial Bee Colony Algorithm in Urban Vegetable Distribution Route Optimization. *Journal of Applied Mathematics and Physics*, **5**, 2291-2301.
https://doi.org/10.4236/jamp.2017.511186

Received: October 19, 2017
Accepted: November 26, 2017
Published: November 29, 2017

Abstract

According to the characteristics and requirements of urban vegetable logistics and distribution, the optimization model is established to achieve the minimum distribution cost of distribution center. The algorithm of artificial bee colony is improved, and the algorithm based on MATLAB software is designed to solve the model successfully. At the same time, combined with the actual case, the two algorithms are compared to verify the effectiveness of the improved artificial bee colony algorithm in the optimization of urban vegetable distribution path.

Keywords

Urban Vegetable, Vehicle Routing, Optimized Artificial Bee Colony Algorithm, Path Optimization

1. Introduction

City vegetable logistics is engaged in transportation, warehousing and other series of vegetable logistics activities within the city limits, it is one of the logistics, business flow and information flow, according to different customers on the delivery time and quantity of vegetables and other requirements, to provide personalized service for customers and distribution [1]. As the characteristics of the city itself, making the city vegetable logistics and distribution with a wide range of customers, the number of customer points, customer points between the distribution distance is relatively close to the characteristics of urban vegetable logistics and distribution generally use the car for distribution, not only because the special truck has the advantages of strong maneuverability, flexible response

[2], and these advantages can also meet the customer a small number of goods, the number of purchases, delivery to the door of the consumer characteristics. In addition, the urban vegetable logistics and distribution of urban residents have a great impact on daily life. It is necessary to construct a reasonable urban vegetable logistics and distribution system not only to meet the requirements of timeliness, convenience, distribution environment and high distribution efficiency, but also should meet the characteristics of urban customers, distribution flow and other characteristics [3].

Urban vegetable logistics distribution vehicle routing problem can be described as: there are a number of customers need the distribution center to deliver a certain amount of goods, the specific location of each customer point is known. The distribution center with the delivery task has the same type of distribution vehicle, and the quantity meets the demand, all the delivery vehicles have the same carrying capacity of [4]. The vehicles that carry the goods depart from the distribution center, along the planned route, the goods to the customer on the route of the hands of the final return to the distribution center. There are certain conditions in the distribution process constraints, each customer can only be a distribution vehicle responsible for the distribution of the task, the time of delivery of the goods within the time required by the customer, otherwise given the appropriate delivery vehicle must be punished [5] [6]. In this paper, the establishment of the model first considers the requirements of the customer point to the delivery time, and then considers the optimization model with the total cost as the goal, which is based on the load and distance constraints of the vehicle. Using the improved artificial bee colony algorithm, using MATLAB software programming solution, combined with the case, to verify the feasibility and effectiveness of improving the artificial bee colony algorithm.

2. Establishment of Model

2.1. Assumptions

As a result of the actual distribution, we will encounter a variety of practical problems, from a different point of view, to solve the problem achieved by the target results are not the same, in order to study and establish the issue of urban vegetable distribution mathematical model [7]. Here make the following assumptions: 1) The starting point of each delivery vehicle for the distribution center, and ultimately return to the distribution center. 2) The total delivery amount of each distribution route is less than the maximum load of the vehicle. 3) The distance traveled by each delivery vehicle shall not exceed its distance. 4) Each customer can only be completed by a car distribution, the cost of each vehicle start and travel unit distance is known.

2.2. Model Establishment

In order to facilitate the description of the constructed model, the following symbols are defined:

(a) $S=\{i \mid i=0,1,2,\cdots,n\}$: collection of a single distribution center and all of its customers, $i=0$ is the distribution center, $i=1,2,\cdots,n$ is customer who provides delivery services by the distribution center;

(b) $K=\{k \mid k=0,1,2,\cdots,m\}$: a set of vehicles that can be used by the distribution center;

(c) Q: the maximum load of a vehicle, in which the vehicle is the same type of vehicle;

(d) q_i: the customer needs, that is, distribution center should be provided for the customer point of delivery, $i=1,2,\cdots,n$;

(e) D: the maximum distance of the vehicle is the same type of vehicle, so the maximum distance is the same;

(f) d_{ij}: the distance from the point i to the point j, $i, j=0,1,2,\cdots,n$;

(g) t_{ij}: the time is used from point i to point j, $i, j=0,1,2,\cdots,n$;

(h) C_0 is the fixed cost of the vehicle, the starting cost required for each vehicle to be delivered, C_1 is the distance traveled by the vehicle, the cost per unit length of time per vehicle;

(m) t_i: the time at which the vehicle serves the customer, $i=0,1,2,\cdots,n$;

(p) S_{ik}: the time period during which the vehicle provides services to the customer, S_{ike} and S_{ikl} are the starting and ending time of vehicles serving customers respectively, $i=0,1,2,\cdots,n$;

(r) ET_i, LT_i: the customer service time window start and end time, during this time to provide delivery services is the highest customer satisfaction, $i=0,1,2,\cdots,n$;

(v) $P_i(S_{ik})$: the penalty cost caused by the service of the customer, that is, the punitive cost caused by the failure to provide service at the time of customer satisfaction.

(w) x_{ijk}: whether or not the vehicle provides service from the customer to the customer, $i, j=0,1,2,\cdots,n$, $k=1,2,\cdots,m$;

(q) y_{ik}: whether the vehicle provides the customer with the delivery, $i=1,2,\cdots,n$, $k=1,2,\cdots,m$,

The decision variables for the problem are as follows:

$$x_{ijk}=\begin{cases}1 & \text{the vehicle is served from point } i \text{ to point } j \\ 0 & \text{otherwise}\end{cases} \quad i, j=0,1,2,\cdots,n; k=1,2,\cdots,m$$

$$y_{ik}=\begin{cases}1 & \text{the task of customer } i \text{ is done by vehicle } k \\ 0 & \text{otherwise}\end{cases} \quad i=1,2,\cdots,n; k=1,2,\cdots,m$$

In this question, the penalty coefficient is introduced and the opportunity cost is expressed as the unit weight per unit time. When the distribution vehicle fails to provide service in the customer satisfaction period, it will be punished in certain proportion. When the actual delivery time is earlier than the service time window, the penalty coefficient is α, when the actual delivery time is later than the service time window, the penalty coefficient is β. $P_i(S_{ik})$ formula of penalty cost is as follows:

$$P_i(S_{ik}) = \begin{cases} \alpha q_i (ET_i - S_{ike}) & S_{ike} < ET_i \\ 0 & ET_i \le S_{ike}, S_{ikl} \le LT_i \\ \beta q_i (S_{ikl} - LT_i) & S_{ikl} > LT_i \end{cases} \quad i = 0,1,2,\cdots,n \quad (2\text{-}1)$$

When the cost of cargo damage is calculated, it is assumed that the damage rate of the vegetables is related to the time of the vegetables in the prescribed low-temperature transportation environment, and the deterioration rate of the vegetables is constant and the deterioration rate of the vegetables is constant and the metamorphic function [8] as follows:

$$Q_i' = Q_i C e^{-\delta t} \quad (2\text{-}2)$$

Among them, Q_i the product is the quality of the goods in good condition, t is the product experience the logistics time, δ is the product of the sensitivity coefficient of time, C for the product at a constant temperature deterioration of a constant speed change. In the metamorphic function, the product is more sensitive to time, the δ value is relatively small, on the contrary, the value is larger.

The cost of the entire distribution process is:

$$H = \sum_{i=1}^{n} q_i \left(1 - C e^{-\delta S_{ike}}\right) p \quad (2\text{-}3)$$

p is the unit of the value of the loss of vegetable products.

The VRPSTW studied in this paper will be time cost, with the lowest total cost as the optimization target. The objective function is shown in (2-4).

$$\min Z = \sum_{k=1}^{m}\sum_{i=0}^{n}\sum_{j=0}^{n} C_1 d_{ij} x_{ijk} + \sum_{i=1}^{n} P_i(S_{ik}) + C_0 \sum_{i=0}^{n}\sum_{k=1}^{m} x_{i0k} + \sum_{i=1}^{n} q_i \left(1 - C e^{-\delta S_{ike}}\right) p \quad (2\text{-}4)$$

The constraints are:

$$\sum_{i=1}^{n} q_i y_{ik} \le Q, \quad k = 1,2,\cdots,m \quad (2\text{-}5)$$

$$\sum_{k=1}^{m} y_{ik} = 1, \quad i = 1,2,\cdots,n \quad (2\text{-}6)$$

$$\sum_{k=1}^{m} y_{0k} = m \quad (2\text{-}7)$$

$$\sum_{j=1}^{n} x_{0jk} = 1, \quad k = 1,2,\cdots,m \quad (2\text{-}8)$$

$$\sum_{i=1}^{n} x_{i0k} = 1, \quad k = 1,2,\cdots,m \quad (2\text{-}9)$$

$$\sum_{i=0}^{n} x_{ijk} = y_{jk}, \quad j = 1,2,\cdots,n; k = 1,2,\cdots,m \quad (2\text{-}10)$$

$$\sum_{j=0}^{n} x_{ijk} = y_{ik}, \quad i = 0,1,\cdots,n; k = 1,2,\cdots,m \quad (2\text{-}11)$$

$$\sum_{i=0}^{n}\sum_{j=0}^{n} d_{ij} x_{ijk} \le D, \quad k = 1,2,\cdots,m \quad (2\text{-}12)$$

$$x_{ijk} = 0,1,\ \ i, j = 0,1,\cdots,n; k = 1,2,\cdots,m \tag{2-13}$$

$$y_{ik} = 0,1,\ \ i = 0,1,\cdots,n; k = 1,2,\cdots,m \tag{2-14}$$

$$S_{ike} + t_i x_{ijk} + t_{ij} x_{ijk} - M\left(1 - x_{ijk}\right) \le S_{jke},\ i = 0,1,\cdots,n;\ j = 0,1,\cdots,n; k = 1,2,\cdots,m \tag{2-15}$$

$$S_{ikl} + t_{ij} x_{ijk} - M\left(1 - x_{ijk}\right) \le S_{jke}, i = 0,1,\cdots,n;\ j = 0,1,\cdots,n; k = 1,2,\cdots,m \tag{2-16}$$

$$S_{ikl} = S_{ike} + t_i y_{ik},\ \ i = 0,1,\cdots,n; k = 1,2,\cdots,m \tag{2-17}$$

$$S_{ike} \le M y_{ik},\ \ i = 0,1,\cdots,n; k = 1,2,\cdots,m \tag{2-18}$$

$$S_{ike} \ge 0, S_{ikl} \ge S_{ike},\ \ i = 0,1,\cdots,n; k = 1,2,\cdots,m \tag{2-19}$$

$$S_{0ke} = 0, S_{0kl} = 0,\ \ k = 1,2,\cdots,m \tag{2-20}$$

The description of the constraint is as follows: Equation (2-5) indicates that the total demand for all customers per vehicle service does not exceed the capacity limit of the vehicle. (2-6) (2-8) means that each customer point can only be delivered by one car. Type (2-7) means that each vehicle from the distribution center. (2-9) means that each car finally returns to the distribution center. (2-10) means that each car if the arrival point must be service. (2-11) means that if the vehicle is customer service, the vehicle will leave the point after the task is completed. Equation (2-12) indicates that the total travel distance of each vehicle must not exceed its travel distance limit. (2-15) and (2-16) show the time relationship in which the vehicle arrives at two customers in their distribution route. Equation (2-17) indicates the relationship between the starting and ending time of the vehicle for customer service. (2-18) means that if the vehicle does not provide delivery to the customer, the vehicle will not reach the customer. Equation (2-19) represents the time limit for the start and end of the service time for the customer service. Equation (2-20) indicates that the departure time of the vehicle from the distribution center is zero.

3. Improved Artificial Bee Colony Algorithm

Human beings based on the nature of bees to collect honey activities, summed up the artificial anthropocentric algorithm theory and to describe. The take bee (the lead bee) to go out looking for honey, and then by jumping dancing and another worker bees in the form of probability to share the source of food information, follow the bee and reconnaissance bee first wait, until the bees bring back the food source information, then choose to follow the honey or find the food near the source to find nectar. From the above description can be learned to adopt bees, follow the bee and reconnaissance bee be able to achieve the identity of the conversion.

3.1. Parameter Initialization

According to the constructed path optimization model, the number of customers to remove the distribution center is $n - 1$. The number of vehicles participating in the distribution center is defined by m, and Q is the number of vehicles

in each distribution vehicle. In the actual urban vegetable distribution vehicle routing problem, q_i represents customer demand. The expression formula for setting N before the initialization phase is as follows:

$$N = m - \left(\left| \frac{\sum_{i=1}^{n} q_i}{q} \right| + 1 \right) \tag{3-1}$$

Population size $N = 50$; $N/2$ indicates the number of food sources is also the number of honey mining bee; single maximum limit = 20; maximum number of iterations MaxCycle = 500.

Solution of the initialization function [solu] = Initial (num), this method is used to generate the problem of the initial solution. Method of production: by looking for the nearest customer service point, one by one distribution, and meet the constraints.

The solution space is defined by $V[N][n][n]$, and the initial feasible solution number is denoted by N.

$$v[i][j][0] = -1, \ i \in N, j \in n \tag{3-2}$$

3.2. Evaluation Method of Profitability

Based on the optimization model of the urban vegetable distribution VRP problem created by the objective function Formula (2-4), we can see that the actual problem to achieve is to minimize the total distribution costs. Artificial colony algorithm uses the reciprocal of the minimum distribution cost as the fitness function, so that the high cost corresponds to the low fitness and the small food source income value.

3.3. Population Update

The algorithm at the beginning of operation, there is not much difference in the probability of each food source is worker to find the number of iterations, a gradual increase in worker, not only to expand the search probability of each food source, and previously set good pheromone is gradually cut, resulting in not being left in to find a food source on early in the process of iterative pheromone is nearly equal to zero. In addition, in the posterior iteration of the algorithm, the algorithm will generate local extreme value because of the large amount of information gathered on the path. The update formula is as follows:

$$Q_{ij}(N+1) = \begin{cases} \rho^{1+\varepsilon(N)} Q_{ij}(N) + \Delta Q_{ij}(N), Q \lhd Q_{\max} \\ \rho^{1-\varepsilon(N)} Q_{ij}(N) + \Delta Q_{ij}(N), Q \rhd Q_{\min} \end{cases} \tag{3-3}$$

Formula: $\varepsilon(N) = N/\tau$, τ is a constant. According to this formula can be achieved with the change of the pheromone solution to adjust the use of such methods can improve the efficiency of the ABC algorithm in the global scope, and can effectively prevent the local extreme situation.

When the algorithm is in different iterative stages, the algorithm can not only avoid the local optimization but also accelerate the convergence speed. The set-

ting of the constant is as shown in Equation (3-4).

$$\delta = \begin{cases} \delta_1, & 0 \le N < N_1 \\ \delta_2, & N_1 \le N < N_2 \\ \delta_3, & N_2 \le N < NC_{\max} \end{cases} \tag{3-4}$$

3.4. Nectar Selection

Follow the bee according to the nectar fitness value corresponding to the selected probability to choose the appropriate honey to go honey, the formula is as follows:

$$P_i = \frac{fit_i}{\sum_{i=1}^{N} fit_i} \tag{3-5}$$

Follow the bee according to the nectar fitness value corresponding to the selected probability to choose the appropriate honey to go honey, the formula such as follow the bee to a certain probability in the poor near the nectar to find a new source. Compare the current nectar and the previous nectar fitness value, if the former is greater than or equal to the latter, then use the current nectar to replace the previous nectar, otherwise, remain unchanged.

3.5. Population Elimination

When the bees (or follow bees) continuous limit failed to search better circulation nectar, this solution is determined (*i.e.* the nectar will fall into the local optimal solution). Bees (or follow bees) to give up the nectar, and into the investigation bee, continue to search for new nectar, the search formula is:

$$x_{ij} = x_{\min j} + rand[0,1]\left(x_{\max j} - x_{\min j}\right) \tag{3-6}$$

4. Case Analysis

The distribution center of an enterprise provides distribution service for 15 customer points. The distribution center is represented by the serial number 0, and the serial number 1 - 15 represents 15 clients respectively. The enterprise uses the same type of freight car for 15 customer point distribution. Among them, the maximum carrying capacity of each car is 5 tons, the vehicle fixed cost is 100, the consumption time cost of 250 vehicles, earlier than the time window to the customer point penalty cost is 250 per ton per hour, the vehicle later arrived time windows penalty cost customers per ton of 300 per hour. The distribution center and customer information statistics are shown in **Table 1**. It is necessary to arrange vehicles to complete the distribution task, so that the total cost of the distribution cost and the penalty cost is the least, so as to find the reasonable distribution route for the distribution center. Data processing was carried out by Matlab2010 software.

Distribution center and customer distribution status are shown in **Figure 1**.

Table 1. Distribution center and customer information.

number	X	Y	q_i	ET_i	LT_i	S_{ik}
0	4	0	0	0	230	0
1	6.6	7.6	0.6	146	166	20
2	7.8	4.1	1.2	61	81	20
3	3.6	2.2	1.1	76	96	20
4	3.8	2.8	0.8	53	73	20
5	5	1.5	1.2	67	87	20
6	0.5	3.8	0.9	116	136	20
7	5.3	7.9	0.8	90	110	20
8	0	2	0.6	135	155	20
9	1	2.5	1	117	137	20
10	2.5	2.3	1.2	114	134	20
11	7.1	0.5	0.7	148	168	20
12	4.1	6.8	0.9	92	112	20
13	6.4	9	1	84	104	20
14	3.8	5.1	0.8	102	122	20
15	10.5	11.5	0.7	128	148	20

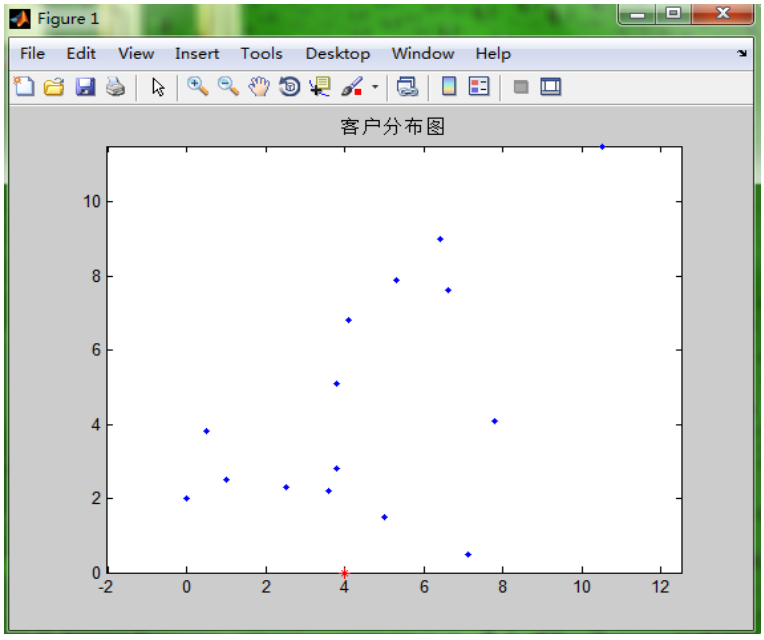


Figure 1. Distribution center and customer distribution.

4.1. Case Solving and Analysis

The number of leading bees accounted for 50%, followed by bees accounted for 10%, reconnaissance bees accounted for 40%, $A = 1.5$, $B = 2$, the number of iterations $Nmax = 200$, of which 50 times before the iteration, Delta 1 = 100, the

middle 51 - 150 times, Delta 2 = 50, 50 times after the iteration, Delta 3 = 200. The distribution route diagram and convergence process of artificial bee colony algorithm are shown in Figure 2, and the distribution route diagram and convergence process of artificial bee colony algorithm are improved, as shown in Figure 3.

4.2. Comparison of Different Algorithms

In order to facilitate the comparison, the paper analyzes the data of the same enterprise and the customer point of the same group. On the basis of the same distance, traffic volume and distance limit, the optimization process of vehicle routing problem under soft time window constraint is simulated, The optimal target value of the algorithm is the total cost of distribution, and the specific data of the number of iterations of the total distribution vehicle convergence are shown in Table 2.

Comparison of the case data can be seen, the same customer restrictions on the distance, distance, volume, when the time constraint is not too strict, is the vehicle routing problem with soft time windows, an improved artificial bee colony algorithm in general distribution vehicle problems and artificial bee colony algorithm is the same, but the distribution of the total cost savings of 1091.78;

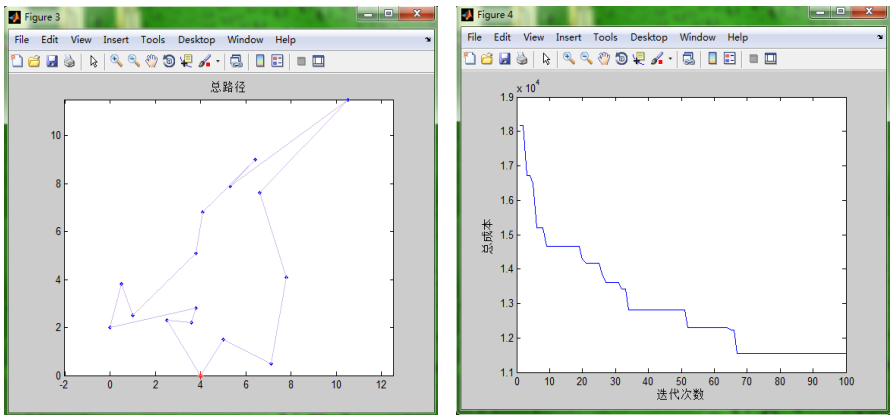


Figure 2. Distribution route and convergence process for algorithm.

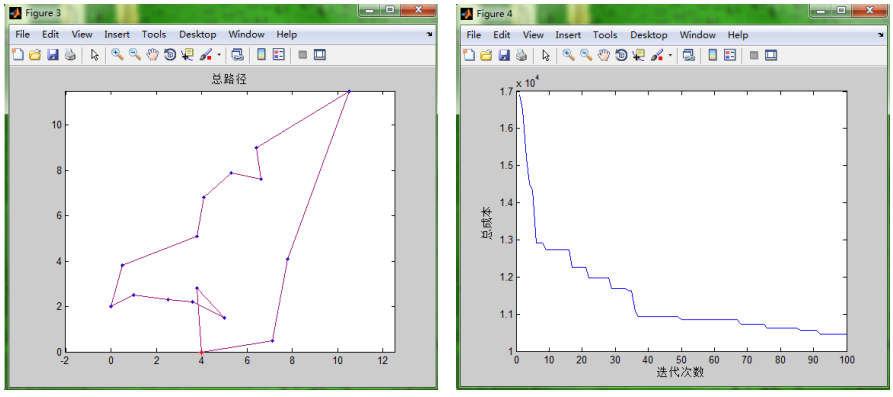


Figure 3. Improved route and convergence process for algorithm.

Table 2. Result comparison.

case	cost	vehicle	Iteration times
1	11547.11	1	68
2	10455.33	1	50

and convergence the speed of the improved artificial bee colony algorithm is faster. The main reason is the analysis algorithm, improved artificial bee colony algorithm with variable neighborhood search update strategy, the best individual search direction is better than stronger, compared with the traditional artificial bee colony algorithm, optimize the operation speed and quality optimization.

5. Conclusion

Logistics and distribution activities is an important part of the logistics system to achieve the optimization of vehicle routing that can directly optimize the logistics and distribution activities, not only can improve the economic efficiency of enterprises, but also to help achieve the logistics management of scientific. Vehicle routing problem as a combinatorial optimization problem, has a strong theory, application value, in the field of logistics and distribution has been widely used [9]. Although the degree of attention to the research of vehicle routing problem is growing, but the expansion of the basic distribution routing problem with time constraint is not deep enough, but has yet to find the solution more quickly and accurately, so there are many scholars pay close attention in the limited time to find the most satisfactory solution to the problem of [10]. On the basis of the previous scholars' research, this paper studies the vehicle routing problem with time constraints, takes into account the influence of time on the distribution path, analyzes the different criteria of the customer's time requirements, but the problems studied in this paper do not consider the actual distribution. The impact of road conditions on the speed of the road, in the next step can also be added to the peak period, the impact of non-peak time on the speed of the introduction of real-time changes in speed to study.

References

[1] Huang, H. and Zhang, Z.X. (2010) Research Status and Prospect of Vehicle Routing Problem. *Logistics technology*, **10**, 21-24.

[2] Nourossana, S. and Erfani, H. (2012) Bee Colony System: Preciseness and Speed in Discrete Optimization. *International Journal on Artificial Intelligence Tools*, **21**, 1250006-1250016. https://doi.org/10.1142/S0218213011000474

[3] Lan, H. and He, Q.F. (2015) Optimization of Cold Chain Logistics Distribution Route Considering Road Access. *Journal of Dalian Maritime University*, **41**, 67-74.

[4] Bi, X.J. (2012) Improved Artificial Bee Colony Algorithm. *Journal of Harbin Engineering University*, **33**, 117-123.

[5] Bao, W.W. and Liu, T. (2012) A Survey of Artificial Bee Colony Algorithm. *Shanxi Electronic Technology*, **2**, 90-92.

[6] Ozturk, C., Hancer, E. and Karaboga, D. (2015) Dynamic Clustering with Improved Binary Artificial Bee Colony Algorithm. *Applied Soft Computing*, **28**, 69-80. https://doi.org/10.1016/j.asoc.2014.11.040

[7] Shao, K. Research and Application of Vehicle Routing Problem Based on Artificial Bee Colony Algorithm. Master Thesis, Wuhan University of Technology, Wuhan.

[8] Wang, Q. Study on Logistics Distribution Location and Transportation Route Optimization of Cold Chain Food with Time Window. Master Thesis, Changan University, Xi'an.

[9] Yang, J. and Ma, L. (2010) Application of Bee Colony Optimization Algorithm in Vehicle Routing Problem. *Computer Engineering and Applications*, **46**, 214-216.

[10] Yu, X.D. and Lian, L. (2016) Application of Artificial Bee Colony Algorithm in Vehicle Routing Problem with Single Time Windows. *Science and Technology Innovation and Application*, **19**, 62-62.

Journal of Applied Mathematics and Physics, 2017, 5, 2302-2310
http://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Modification of Even-A Nuclear Mass Formula

Jingyi Zhang

Department of Physics, University of Shanghai for Science and Technology, Shanghai, China
Email: jingchyi@sina.cn

How to cite this paper: Zhang, J.Y. (2017) Modification of Even-A Nuclear Mass Formula. *Journal of Applied Mathematics and Physics*, **5**, 2302-2310.
https://doi.org/10.4236/jamp.2017.511187

Received: October 24, 2017
Accepted: November 26, 2017
Published: November 29, 2017

Abstract

In this paper we obtain an empirical mass formula of even-A nuclei based on residual proton-neutron interactions. The root-mean-squared deviation (RMSD) from experimental data is at an accuracy of about 150 Kev. While for heavy nuclei, we give another formula that fits the experimental data better (RMSD ≈ 119 Kev). We have successfully described the experimental data of nuclear masses and predicted some unknown masses (like ^{200}Ir not involved in AME2003, the deviation of our predicted masses from the value in AME2012 is only about 82 keV). The predictive power of our formula is more competitive than other mass models.

Keywords

Residual Proton-Neutron Interactions, Nuclear Masses, Binding Energies

1. Introduction

The study of nuclear masses and energy levels has always been one of the most challenging frontiers in the field of nuclear physics. There are two types to describe and understand the nuclear masses, one of which is global relations, and the other is local. Some global nuclear mass models such as Weizäscker model [1], Duflo-Zuker model [2], the finite range droplet model [3], a recent macroscopic-microscopic mass formula [4] [5] [6] etc., successfully produce the measured masses with accuracy at the level of 300 - 600 Kev. However, the global mass models require more physics and more information about nuclear force to get better description of the nuclear masses. On the other hand, the local mass relations, such as the isobaric multiplet mass equation (IMME), the Garvey-Kelson (GK) relations, which use the predicted nuclear masses and the residual proton-neutron interactions to evaluate the mass. It is found that the local mass relations are just approximately satisfied in known masses, so it has a good potential to predict the unknown masses.

In this paper, our purpose is to obtain a residual proton-neutron interactions formula of even-A nuclei from those of neighboring nuclei. In Section II we introduce the residual proton-neutron interactions and obtain our formula based on the proton-neutron interactions between the last proton and the last neutron. Then we introduce two modifications to improve our formula. The RMSD from experimental data is about 150 Kev. And for heavy nuclei, we obtain another formula fits with the experimental data even more precise. With our further refinement of heavy nuclei, the RMSD gets even smaller to about 120 Kev. In Section III we successfully predict some unknown masses. The result shows that the predict power of our formula is competitive with others. In Section IV we discuss and summarize the results of this paper.

2. The Residual Proton-Neutron Interactions

The residual proton-neutron interaction plays an important role in the evolution of collective, deformation and phase transition [7] [8] [9] [10], so it has attracted many attentions [11]-[17]. The proton-neutron interactions between the last i protons and j neutrons is given by

$$V_{ip-jn}(Z,N)=B(Z,N)+B(Z-i,N-j)-B(Z,N-j)-B(Z-i,N). \quad (1)$$

The famous formula GKL and GKT were derived from the neutron-proton interactions between the last neutron and proton [18] [19]. The relationship between Garvey-Kelson quality is a semi empirical relationship between 6 adjacent nuclear mass. If the interaction between neighboring nuclei changes slowly in the local range, it can be completely counteracted by the addition and subtraction of many adjacent nuclei. Garvey-Kelson mass relationship has two common relationships:

$$\begin{aligned} &M(N,Z+1)+M(N-1,Z-1)+M(N+1,Z)\\ &-M(N,Z-1)-M(N-1,Z)-M(N+1,Z+1)=0, \end{aligned} \quad (2)$$

$$\begin{aligned} &M(N,Z-1)+M(N-1,Z+1)+M(N+1,Z)\\ &-M(N,Z+1)-M(N-1,Z)-M(N+1,Z-1)=0, \end{aligned} \quad (3)$$

where $M(N,Z)$ denotes the mass of a nucleus with neutron number N and proton number Z. Equation (2) is called the longitudinal Garvey-Kelson relation (GKL), and Equation (3) the transverse (GKT).

In this section, we use the residual proton neutron interactions between the last proton and the last neutron to form our formula. According to the Equation (1), it is easy to obtain that the residual proton-neutron interactions between the last proton and the last neutron is defined as

$$\begin{aligned} V_{1p-1n}(Z,N)&=B(Z,N)+B(Z-1,N-1)-B(Z,N-1)-B(Z-1,N)\\ &=M(Z,N)+M(Z-1,N-1)-M(Z,N-1)-M(Z-1,N) \end{aligned} \quad (4)$$

The Garvey-Kelson mass relations require six nuclei, but our formula requires only four. So our formula involves less number of nuclei, its predictions in iterative extrapolations is the more reliable, and its deviations are smaller in the

extrapolation process.

In recent years, many papers tried to find formulas to describe and evaluate the nuclear masses, but many of them have a large RMSD. In this work, we focus on the even-A nuclei, through the study on the neighboring nuclei with the database in AME2012 [20].

For the residual nuclear proton-neutron interactions which $A \geq 42$, we calculate the δV_{1p-1n} as shown in **Figure 1**. Based on that, we empirically obtained the residual proton-neutron interactions formula of even-A nuclei. The formula is as follows:

$$\begin{aligned}\overline{\delta V_{1p-1n}} &= B(Z,N+1)+B(Z-1,N)-B(Z,N)-B(Z-1,N+1)\\ &\cong \frac{515.6}{A^2}+\frac{62.78}{A}+0.1079\ \text{keV}\end{aligned} \tag{5}$$

$\overline{\delta V_{1p-1n}}$ is the average values of δV_{1p-1n} for nuclei with the same mass number A.

We find that the average binding energy of our predicted mass agrees well with the specific binding energy curve. We successfully describe and predict some even-A nuclear masses by using these equations and some known experimental nuclear masses in AME2012 for calculation of δV_{1p-1n}.

It can be seen from the **Figure 1** that the interaction of proton-neutron is

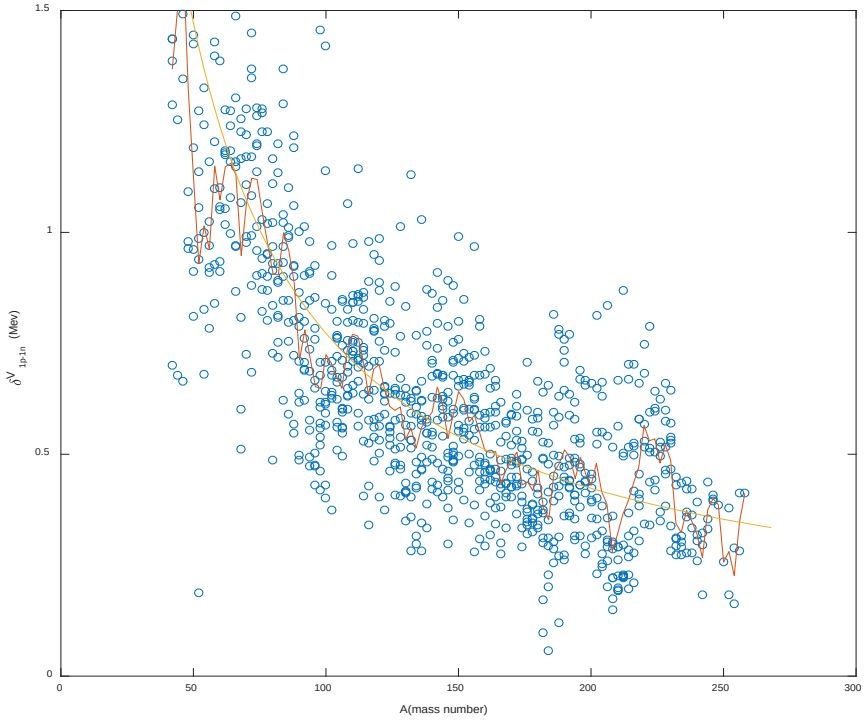


Figure 1. Circles show that the residual proton-neutron interactions δV_{1p-1n}.The curve is plotted by using the average values of δV_{1p-1n} for nuclei with the same mass number A, expressed as $\overline{\delta V_{1p-1n}}$. The smoothed curve are plotted in terms of equation $\overline{\delta V_{1p-1n}}(A) = \frac{515.6}{A^2}+\frac{62.78}{A}+0.1079\ \text{keV}$ for even-A nuclei with $A \geq 42$.

more stable in the heavy nuclei region than in the light nuclei region.

In order to better describe the quality of the nucleus, we will improve the above formula with some amendments, donated by δV_{1p-1n}^{cal} as the final improvement results [4] [5] [6]. The first is called the Coulomb correction, denoted by Δ_C:

$$\Delta_C(Z,N) \approx a_C\left(-\frac{4}{9}Z^{4/3}A^{-7/3} - \frac{2}{3}ZA^{-4/3} + \frac{4}{9}Z^2A^{-7/3} + \frac{4}{9}Z^{1/3}A^{-4/3}\right),$$

the second is called the symmetry energy correction, denoted by Δ_{sym}:

$$\Delta_{sym}(Z,N) = a_{sym}\frac{1}{A\left(2+|IA|\right)^3} + b_{sym}A^{-1},$$

where $I=(N-Z)/A$ and $a_C=10.51$, $a_{sym}=20126$, $b_{sym}=-61.25$ as parameters [17] [21].

The revised $\delta V_{1p-1n}(Z,N)$ is as follows:

$$\delta V_{1p-1n}^{cal}(Z,N) = \overline{\delta V_{1p-1n}} - \Delta_C(Z,N) - \Delta_{sym}(Z,N). \tag{6}$$

The improvement of these two corrections on our predicted δV_{1p-1n} is about 5 keV. Although the two contributions are small, but with more understanding of the symmetry energy of the nucleus, we believe that these contributes will become more important in the future.

In order to describe the nuclear mass obtained by our theory vividly, we compare the average RMSD of the nuclear mass with the experimental data to represent the difference, and the formula is as follows:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(M_i^{exp} - M_i^{cal}\right)^2}.$$

The RMSD is about 150 Kev. In **Figure 2** we show deviations (in units of keV) between our calculated δV_{1p-1n}^{cal} by applying Equations (6) and those experimental data of binding energies compiled in AME2012 [20]. It can be seen that the RMSDs of these δV_{1p-1n} decrease with A. The description is better in the medium mass nucleus and heavy nucleus.

As early as 1960s, the nuclear structure theory predicts the existence of a number of new elements in the long life near the proton number $Z=114$ and neutron number $N=184$ (*i.e.* island of super heavy nuclei) and the island of super heavy nuclear plays an important role in the entire nuclear physics field. So for the heavy nuclei, we obtain another formula to describe the mass and it fits more closely with the experimental data. And in order to achieve better result, the different parameters are given between even-even nuclei and odd-odd nuclei, the formula is as follows:

$$V_{1p-1n}(A) = \frac{a}{A^2} + \frac{b}{A} + c. \tag{7}$$

Parameter	*a*	*b*	*c*
Even-even	−9464	146.3	−0.06435
Odd-odd	46000	−324.1	0.9124

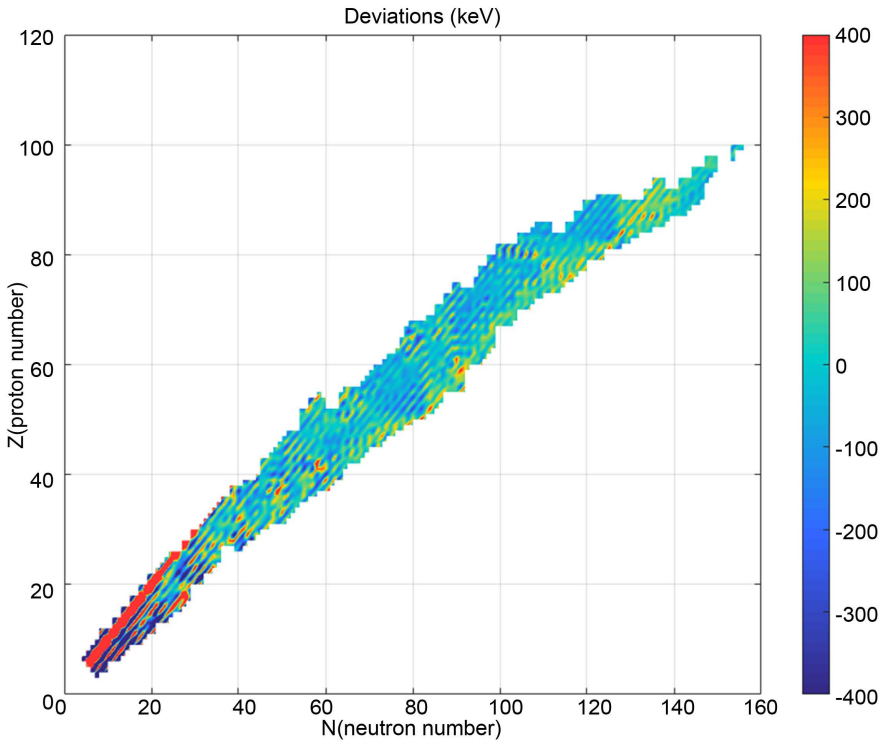


Figure 2. (Color online) Deviations (in units of keV) of our calculated δV_{1p-1n}^{cal} by using Equations (6) with respect to those extracted from experimental binding energies [Equation (4)], for the nuclei with $A \geq 16$.

When we use the Equation (6) to describe the nuclear masses, the RMSD is about 150 Kev, but if we try the Equation (7) where $A > 200$, the RMSD is 119 Kev, it shows that our formula of heavy nuclei is more accurate.

Figure 3 displays the difference between the experimental values and calculated values, we compare it with Ref [21], one can see that our result is better.

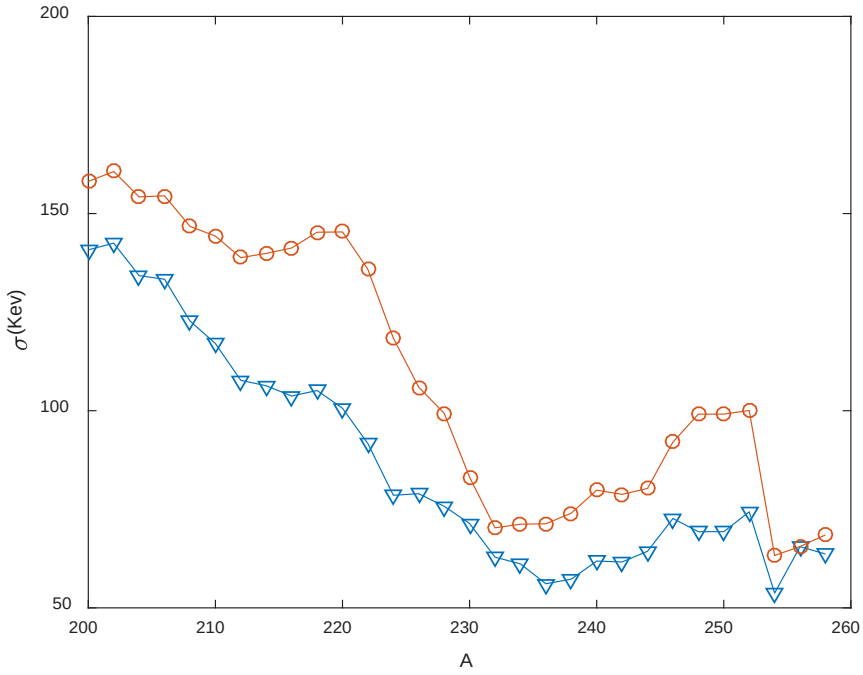
3. Mass Predictions

Through above study, we find our formula has a good performance in describing the nuclear masses. In this section, we use our formula and the residual proton-neutron interaction to predict the nuclear mass not obtained in the experiment. Based on the Equation (4), we can obtain

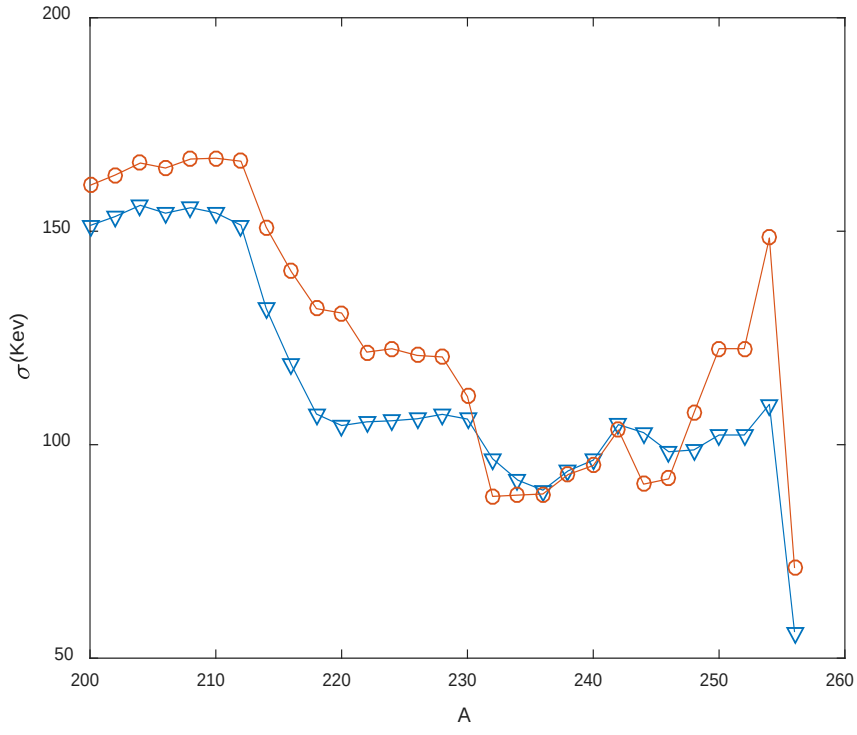
$$M(Z,N) = M(Z-1,N) + M(Z,N-1) - M(Z-1,N-1) + \overline{\delta V_{1p-1n}}(A).$$

The unknown mass $M(Z,N)$ is predicted by using the three nuclei masses around it and the $\delta V_{1p-1n}(Z,N)$ we empirical obtained.

Now let's focus on a few examples of our predictions. **Table 1** shows mass excess of some nuclei are not predictive in ame 2003 or ame 2012 databases. These unknown masses are important not only in the context of astrophysics, but also in the nuclear structure. Interestingly, our predicted values show good in comparison with the experimental results. For ^{182}Lu, the deviation of our predicted masses from the value in AME2012 is only ~63 keV. Three additional



(a)



(b)

Figure 3. Shows the RMSDs of even-A nuclei. (a) represents the odd-odd nuclei; (b) represents the even-even nuclei. We obtain the even-A nuclear masses from some experimentally known nuclear masses and the residual proton-neutron interactions formula. Comparing calculated values with the AME2012 databases obtain the RMSDs. The triangles are plotted by using the RMSDs of our calculated values. The circles are plotted by using the formula in Ref [21].

Table 1. Mass excess of some mass nuclei with us and predicted results in the AME2003 database and the AME2012 databsae. (keV).

Nucleus	AME2003	AME2012	M^{pred}
^{52}Ni	−22,650	−23,470	−23,187
^{74}Sr	−40,700	−40,830	−40,952
^{86}As	−59,150	−58,962	−58,316
^{98}Kr	−44,800	−44,310	−44,555
^{126}Pr	−60,260	−60,320	−60,573
^{148}Tm	−39,270	−38,765	−38,713
^{164}Re	−27,640	−27,523	−27,422
^{182}Lu	−41,880	−41,880	−41,817
^{190}At	null	null	10,290
^{200}Ir	null	−21,611	−21,693
^{202}Pt	−22,600	−22,692	−22,592
^{224}Np	null	31,876	31,793
^{232}Am	43,400	43,268	43,376
^{272}Mt	133,890	133,582	133,671
^{286}Ed	168,120	169,725	169,700

nuclei are ^{202}Pt, ^{232}Am and ^{286}Ed, the differences between our predicted values and those in AME2012 are approximately 100 keV. It seems our formula shows a great accuracy and can be used predict nuclear masses.

4. Discussion and Conclusions

In this paper, we obtain the residual proton-neutron interactions formula to describe and predict the mass of even-A nuclei. In order to improve the accuracy of the δV_{1p-1n}, we use the average value of the δV_{1p-1n} (denoted as $\overline{\delta V_{1p-1n}}$ modification) and introduce two modifications.

For further understanding of the super heavy nuclei, we use another formula to describe the δV_{1p-1n}, and its results fit the experiment data more accurate, one can see that the RMSD decreases considerably.

Then we investigate the predictive power of these new formulas by numerical experiments. They are competitive with other local mass relations. The deviation of predicted results from experimental values is less compared with other models.

Based on results so far, our method of studying the neighboring nuclei has a good performance. We can predict other unknown masses by using our empirical formula to provide useful reference points for experimental physics.

Acknowledgements

The author would like to thank G.Y.Gao for reading and commenting of this

paper.

References

[1] Von Weizsäcker, C.F. (1935) Zur Theorie der Kernmassen. *Zeitschrift für Physik*, **96**, 431. https://doi.org/10.1007/BF01337700

[2] Duflo J. and Zuker, A.P. (1995) Microscopic Mass Formulas. *Physical Review C,* **52**, R23. https://doi.org/10.1103/PhysRevC.52.R23

[3] Möller P., Myers W. D., Sagawa, H. and Yoshida, S. (2012) New Finite-Range Droplet Mass Model and Equation-of-State Parameters. *Physical Review Letters,* **108**, 052501. https://doi.org/10.1103/PhysRevLett.108.052501

[4] Wang N., Liang Z. Y., Liu, M. and Wu, X. Z. (2010) Mirror Nuclei Constraint in Nuclear Mass Formula. *Physical Review C,* **82**, 044304. https://doi.org/10.1103/PhysRevC.82.044304

[5] Wang N., Liu, M. and Wu, X. Z. (2010) Modification of Nuclear Mass Formula by Considering Isospin Effects. *Physical Review C,* **81**, 044322. https://doi.org/10.1103/PhysRevC.81.044322

[6] Mendoza-Temis, J., Hirsch, J. G. and Zuker, A. P. (2010) The Anatomy of the Simplest Duflo–Zuker Mass Formula. *Nuclear Physics A*, **843**, 14. https://doi.org/10.1016/j.nuclphysa.2010.05.055

[7] De Shalit A. and Goldhaber, M. (1953) Mixed Configurations in Nuclei. *Physical Review Journals Archive*, **92**, 1211. https://doi.org/10.1103/PhysRev.92.1211

[8] Federman P. and Pittel, S. (1977) Towards a Unified Microscopic Description of Nuclear Deformation. *Physics Letters B,* **69**, 385. https://doi.org/10.1016/0370-2693(77)90825-5

[9] Casten R. F. and Zamfir, N. V. J. (1996) The Evolution of Nuclear Structure: The N_pN_n Scheme and Related Correlations. *Journal of Physics G: Nuclear and Particle Physics*, **22**, 1521. https://doi.org/10.1088/0954-3899/22/11/002

[10] Talmi, I. (1962) Effective Interactions and Coupling Schemes in Nuclei. *Reviews of Modern Physics*, **34**, 704. https://doi.org/10.1103/RevModPhys.34.704

[11] Brenner, D.S., Wesselborg, C., Casten, R.F., Warner, D.D. and Zhang, J.Y. (1990) Empirical p-n Interactions: Global Trends, Configuration Sensitivity and N=Z Enhancements. *Physics Letters B,* **243**, 1. https://doi.org/10.1016/0370-2693(90)90945-3

[12] Mouze, G., Bidegainberry, S., Rocaboy, A. and Ythier, C. (1993) The Neutron-Proton Interaction Energy of the Valence Nucleons. *Il Nuovo Cimento A* (1965-1970), **106**, 885.

[13] Gao, Z.C., Chen, Y.S. and Meng, J. (2001) Garvey-Kelson Mass Relations and n-p Interaction. *Chinese Physics Letters*, **18**, 1186. https://doi.org/10.1088/0256-307X/18/9/310

[14] Cakirli R. B., Brenner D. S., Casten, R. F. and Millman, E. A. (2005) Proton-Neutron Interactions and the New Atomic Masses. *Physical Review Letters,* **94**, 092501. https://doi.org/10.1103/PhysRevLett.94.092501

[15] Cakirli R. B. and Casten, R. F. (2006) Direct Empirical Correlation between Proton-Neutron Interaction Strengths and the Growth of Collectivity in Nuclei. *Physical Review Letters,* **96**, 132501. https://doi.org/10.1103/PhysRevLett.96.132501

[16] Breitenfeldt M. et al., (2010) Approaching the N = 82 Shell Closure with Mass Measurements of Ag and Cd Isotopes. *Physical Review C,* **81**, 034313. https://doi.org/10.1103/PhysRevC.81.034313

[17] Jiao, B.B. (2017) Description and Prediction of Even-A Nuclear Masses Based on Residual Proton-Neutron Interactions. https://arxiv.org/abs/1706.08686

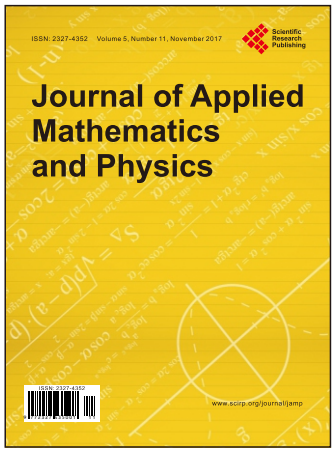
[18] Garvey, G. T. and Kelson, I. (1966) New Nuclidic Mass Relationship. *Physical Review Letters,* **16**, 197. https://doi.org/10.1103/PhysRevLett.16.197

[19] Garvey G. T., Gerace W. J., Jaffe R. L., Talmi, I. and Kelson, I. (1969) Set of Nuclear-Mass Relations and a Resultant Mass Table. *Reviews of Modern Physics*, **41**, S1. https://doi.org/10.1103/RevModPhys.41.S1

[20] Wang, M., Audi, G., Wapstra, A.H., *et al.* (2012) The Ame2012 Atomic Mass Evaluation. *Chinese Physics C*, **36**, 1603.

[21] Fu, G.J., Lei, Y., Jiang, H., *et al.* (2011) Description and Evaluation of Nuclear Masses Based on Residual Proton-Neutron Interactions. *Physical Review C,* **84**, 034311. https://doi.org/10.1103/PhysRevC.84.034311

Call for Papers



Journal of Applied Mathematics and Physics

ISSN Print: 2327-4352 ISSN Online: 2327-4379
http://www.scirp.org/journal/jamp

Journal of Applied Mathematics and Physics is an international journal dedicated to the latest advancement of applied mathematics and physics. The goal of this journal is to provide a platform for researchers and scientists all over the world to promote, share, and discuss various new issues and developments in different areas of applied mathematics and physics. We aim to publish high quality research articles in terms of originality, depth and relevance of content, and particularly welcome contributions of interdisciplinary research on applied mathematics, physics and engineering.

Subject Coverage

The journal publishes original papers including but not limited to the following fields:

- Applications of Systems
- Applied Mathematics
- Applied Non-Linear Physics
- Applied Optics
- Applied Solid State Physics
- Biophysics
- Computational Physics
- Condensed Matter Physics
- Control Theory
- Cryptography
- Differentiable Dynamical Systems
- Engineering and Industrial Physics
- Experimental Mathematics
- Fluid Mechanics
- Fuzzy Optimization
- Geophysics
- Integrable Systems
- Laser Physics
- Methodological Advances
- Multi-Objective Optimization
- Nanoscale Physics
- Nonlinear Partial Differential Equations
- Non-Linear Physics
- Nuclear Physics
- Numerical Computation
- Optical Physics
- Portfolio Selection
- Riemannian Manifolds
- Scientific Computing
- Set-Valued Analysis
- Soliton Theory
- Space Physics
- Symbolic Computation
- Topological Dynamic Systems
- Variational Inequality
- Vector Optimization

We are also interested in: 1) Short reports—2-5 page papers where an author can either present an idea with theoretical background but has not yet completed the research needed for a complete paper or preliminary data; 2) Book reviews—Comments and critiques.

Notes for Intending Authors

The journal publishes the highest quality original full articles, communications, notes, reviews, special issues and books, covering both the experimental and theoretical aspects including but not limited to the above materials, techniques and studies. Papers are acceptable provided they report important findings, novel insights or useful techniques within the scope of the journal. All manuscript must be prepared in English, and are subjected to a rigorous and fair peer-review process. Accepted papers will immediately appear online followed by prints in hard copy.

Website and E-Mail

http://www.scirp.org/journal/jamp **E-mail: jamp@scirp.org**

What is SCIRP?

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

What is Open Access?

All original research papers published by SCIRP are made freely and permanently accessible online immediately upon publication. To be able to provide open access journals, SCIRP defrays operation costs from authors and subscription charges only for its printed version. Open access publishing allows an immediate, worldwide, barrier-free, open access to the full text of research papers, which is in the best interests of the scientific community.

- High visibility for maximum global exposure with open access publishing model
- Rigorous peer review of research papers
- Prompt faster publication with less cost
- Guaranteed targeted, multidisciplinary audience

Scientific Research Publishing

Website: http://www.scirp.org

Subscription: sub@scirp.org

Advertisement: service@scirp.org