

Computational Methods in Applied Sciences 38

Nikos D. Lagaros  
Manolis Papadrakakis *Editors*

# Engineering and Applied Sciences Optimization

Dedicated to the Memory of Professor  
M.G. Karlaftis



 Springer

# **Computational Methods in Applied Sciences**

Volume 38

**Series editor**

Eugenio Oñate, Barcelona, Spain

More information about this series at <http://www.springer.com/series/6899>

Nikos D. Lagaros · Manolis Papadrakakis  
Editors

# Engineering and Applied Sciences Optimization

Dedicated to the Memory  
of Professor M.G. Karlaftis

 Springer



*Editors*

Nikos D. Lagaros  
Institute of Structural Analysis and  
Antiseismic Research  
National Technical University of Athens  
Athens  
Greece

Manolis Papadrakakis  
Institute of Structural Analysis and  
Antiseismic Research  
National Technical University of Athens  
Athens  
Greece

ISSN 1871-3033

Computational Methods in Applied Sciences

ISBN 978-3-319-18319-0

ISBN 978-3-319-18320-6 (eBook)

DOI 10.1007/978-3-319-18320-6

Library of Congress Control Number: 2015938311

Springer Cham Heidelberg New York Dordrecht London

© Springer International Publishing Switzerland 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media  
([www.springer.com](http://www.springer.com))

# Memoriam

## In Memoriam to a Great Scientist, An Excellent Educator and a Beloved Friend

Professor Matthew G. Karlaftis  
21 December 1969–4 June 2014



Asklepieion, Kos Island, 3rd of June 2014. Professor Matthew G. Karlaftis (second from the *right*) and his friends Iordanis, Christos and Nikos (*left to right*).

It is always very difficult to bid farewell a dear friend, a mentor, a part of your life. Professor Matthew G. Karlaftis (Matt) met an untimely death at the age of 45, while at the OPT-i 2014 conference. Matt was the heart of the OPT-i conference; he envisaged it to be a leading scientific event and a unique stage for sharing and exchanging research ideas in the field of optimization and its applications in Engineering and Applied Sciences. This volume is by all means the “capping stone” of that vision. Eleni and I had the privilege of being Matt’s first Ph.D. students, closest research associates and academic colleagues, since his first steps at the NTUA in the early 2000s. We were extremely fortunate to share Matt’s academic vision and his passion for research and development, his teaching charisma and, foremost, his friendship and support. For us, Matt was, before everything else, family.

Professor Karlaftis was born in Athens, in 1969. He received his B.Sc. and M.Sc. degrees in Civil Engineering from the University of Miami and his Ph.D. in Transportation Engineering from Purdue University, in 1996. After completion of his graduate studies, he had the appealing option of applying for a faculty in a top US University. Nonetheless, he decided to return to Greece and, following the completion of mandatory military service, to seek for a faculty position in a Greek University. While he rarely expressed it, Matt was very fond of his origins and national heritage and, as such, returning to his home land seemed the logical option. He became a faculty member at the National Technical University of Athens in 2001, and, tragically, he was about to officially advance to a Full Professor position, shortly after his death. Despite the strict national legislation and procedures regulating promotion and tenure of faculty members, he managed to become Full Professor in less than 14 years. In parallel to his academic activities, Matt served as the vice-president of the Athens Urban Transport Organization, during the critical period of the 2004 Athens Olympics and as the elected President of the Hellenic Institute of Transportation Engineers.

We could spend numerous pages talking about Matt and his academic achievements; a leading scholar, a charismatic teacher, a prolific writer, a pioneer in the field of quantitative methods and their applications in Engineering were indicative aspects of his academic profile. With over 120 publications in peer-reviewed journals, authorship of six books, several book chapters, and the scientific leadership of over 45 research projects in almost 15 years, Matt was an academic “rock star” in the areas of transportation and traffic planning, civil infrastructure design, and management. His editorial involvement was also impressive; he was the editor in chief for Transportation Research Part C, European editor of ASCE’s Journal of Transportation Engineering, Associate Editor of ASCE’s Journal of Infrastructure Systems, and an editorial board member for eight other journals. During his short career, he received a number of prestigious awards, including the Fulbright Scholar Grant (2006–2007), the ASCE Walter L. Huber Civil Engineering Research Prize (2005), the TRB ABJ80 Best Paper Award (2009), and the ASCE State-of-the-Art Paper Award (2011). His international bestselling book on transportation econometrics and statistics is a standard text for many scholars and students worldwide.

Matt was a gifted scholar. He had this rare ability of analyzing complex problems in a simple, yet, elegant and robust way. This, along with his extensive knowledge of quantitative methods, as well as transportation planning and engineering topics, allowed him to successfully apply advanced optimization, statistical and econometric models in difficult engineering problems. But, above all his virtues and accomplishments, was his exceptional ability to reach to his students. Matt was an excellent teacher and mentor for many undergraduate and graduate students, among which are five faculty members in Greece and abroad. Most importantly though, Matt was a fascinating person, a man of honor, an open-hearted, devoted, and trustworthy friend. We will miss you dearly Matt.

Konstantinos L. Kepaptsoglou, Ph.D.  
Lecturer  
National Technical University of Athens

Eleni I. Vlahogianni, Ph.D.  
Assistant Professor  
National Technical University of Athens

# Preface

## **Engineering and Applied Sciences Optimization: Dedicated to the memory of Professor M.G. Karlaftis**

This volume is published to commemorate the life and memory of Prof. Matthew G. Karlaftis. Numerous memorial events have been organized and acclamations have been written about Matt since his untimely passing on June 4, 2014, a few hours before his opening speech at the First International Conference on Engineering and Applied Sciences Optimization (OPT-i) that we were co-organizing. He was a very special person who will be long remembered as a great scientist and educator as well as a beloved friend.

The subject areas of the volume ranges from Structural Optimization, Logistics, Transportation, Traffic and Telecommunication Networks to Operational Research, Metaheuristics, Multidisciplinary and Multiphysics Design Optimization, etc. The chapters which appear in this volume are selected studies presented at OPT-I and works written by his friends and former colleagues and students; all in the area of optimization that Matt loved and was so quantitatively driven. All contributions reflect the warmth and genuine friendship which Matt enjoyed from his associates and show how much his scientific contribution has been appreciated. He will be greatly missed and we hope that this volume will be proven as a suitable memorial to his life and achievements.

The volume consists of 25 chapters which are grouped into three categories, in the first category, the chapters deal with optimization studies related to logistics, transportation and traffic and telecommunication networks; in the second, various works are presented where metaheuristic optimization methods are used for solving various engineering problems and in the third, structural optimization and operational research problems are solved.

**First Part:** In the work of Roncoli et al., it is described a novel approach for defining optimal strategies in motorway traffic flow control, considering that a portion of vehicles are equipped with vehicle automation and communication systems; an optimization problem, formulated as a convex quadratic programming problem, was

developed with the purpose of minimizing traffic congestion. Qian et al. make a comprehensive use of the large-scale taxi trip data and present a three-fold study on urban dynamics pattern in NYC. First, the spatiotemporal pattern of urban activities are examined from trip dynamics by aggregating pick-up and drop-off locations; second, they explore the inherent similarities among taxi trips and reveal the underlying connections among detached places using two-step clustering algorithms. Paz et al. proposed a methodology aiming to calibrate microscopic traffic flow simulation models, which was found to be capable to calibrate simultaneously all the calibration parameters as well as demand patterns for any type of network. *Gkiotsalitis and Stathopoulos* investigated the importance of big-data in improving the organizational efficiency of physical meetings among multiple travelers in urban environments. In particular, they examined the state-of-the-art on capturing travelers' patterns based on their data traces and the expected gains from leveraging user-generated data for optimizing leisure travel. In the work of Cruciol et al. it is introduced the application of the methods of data mining to get the knowledge from air traffic big-data in management processes. The proposed approach uses a Bayesian network for data analysis to reduce the costs of flight delay. *Papathanasopoulou and Antoniou* enhance the capabilities of an existing data-driven approach while it is further validated using another training dataset; in addition, the methodology is enriched and an improved methodological framework is suggested for the optimization of car-following models. Vlahogianni presents a detailed review of the unique opportunities provided by ITS and big data and discuss the emerging approaches for transportation modeling; furthermore, the challenges and emerging opportunities posed for researchers working with such approaches are also discussed.

**Second Part:** Yang presents the key features of nature-inspired metaheuristic algorithms by analyzing their diversity and adaptation, exploration and exploitation, attractions and diffusion mechanisms. The author also highlights the success and challenges concerning swarm intelligence, parameter tuning, and parameter control as well as some open problems. Saka et al. used five optimum design algorithms for cold-formed steel frames made of thin-walled sections using the recent metaheuristic techniques. The algorithms considered are firefly, cuckoo search, artificial bee colony with levy flight, biogeography-based optimization, and teaching-learning-based optimization algorithms. Mehmani et al. present a new model management technique to be incorporated into population-based heuristic optimization; according to this technique different computational models are selected adaptively in order to be used during optimization, with the overall objective to result in optimal designs with high fidelity function estimates at a reasonable computational expense. Simos presents a general algorithmic method for constructing  $2^q$ -level design matrices in order to explore and optimize response surfaces where the predictor variables are each at  $2^q$  equally spaced levels, by utilizing a genetic algorithm. Hosseini et al. present a new optimization technique named as mesh adaptive direct search (MADS) that is used to solve optimal steady-state performance of power systems. MADS is utilized to determine the optimal settings of control variables, such as generator voltages and transformer taps for optimal reactive power and voltage control of IEEE 30-bus system.

**Last Part:** Murakami et al. propose a new optimization procedure including a variable adaptive step length for shear buildings with hysteretic dampers when subjected to a set of design earthquake ground motions under a constraint on total cost. The response sensitivity of buildings including hysteretic dampers is high and a devised algorithm of adaptive step-length is useful to obtain a smooth and reliable response sensitivity. Nigdeli and Bekdaş present an optimization methodology for tuning of tuned mass dampers on structures subjected to seismic loading for two different objectives, such as reducing the displacement of first story and absolute acceleration of top story of the structure. Alexandersen and Lazarov present a methodology for tailoring macroscale response by topology optimizing micro-structural details, where the microscale and macroscale response are completely coupled by treating the full model. Giannakoglou et al. present adjoint methods for the computation of the first- and higher-order derivatives of objective functions used in optimization problems governed by the Navier–Stokes equations in aero/hydrodynamics. Gogarty and Pasini present a 2D hierarchical topology optimization scheme aiming to design a cellular scaffold that optimally reconciles bone resorption and permeability, two antagonist objectives of bone tissue scaffolds. Gandomi et al. study the method of evolutionary boundary constraint handling that is very easy to implement and very effective. In particular, they intended to improve the optimization results by means of evolutionary boundary constraint handling scheme on slope stability optimization problem. Talgorn et al. present different formulations for the surrogate problem considered at each search step of the mesh adaptive direct search algorithm using a surrogate management framework. Bekas et al. aim to couple the problem of structural optimization of building frames, with that of the optimization of design options for their energy efficiency. Bekdaş and Nigdeli in their work iteratively search to find the flexural moment capacity of columns under axial loading. Waycaster et al. propose a framework for understanding the types of interactions that may take place and their effect on design optimization formulation by means of game theory. These effects were considered as an economic uncertainty that arises due to limited information about interactions between stakeholders. Antoni and Giannessi present a new approach for handling bilevel multi-objective problems. The advantage of this new approach consists of the following characteristics, for solving the upper level, it does not require to know explicitly the lower level. Georgioudakis et al. integrated the extended finite element into a shape design optimization framework aiming to improve the service life of structural components subject to fatigue. Charmpis and Dimitriou developed an optimal budget allocation framework and stress-tested for the optimal scheduling of a bridges upgrading program. A suitable test case is developed for performing in-depth analysis that takes into consideration the most important features involved.

The editors of the volume would like to express their deepest gratitude to all the contributors for their most valuable support during the preparation of this volume, for their time and effort devoted to the completion of their contributions, and for their expert help in the reviewing process. We are also grateful to all the colleagues who, although they did not contribute chapters to the volume, were kind enough to offer their expert help during the reviewing process. Finally, we would also like to

thank all the personnel of Springer Publishers, especially Johanna F.A. Pot (Editorial Assistant in Engineering) and Nathalie Jacobs (Senior Publishing Editor in Engineering), for their most valuable continuous support with the publication of this volume.

March 2015

Nikos D. Lagaros  
Manolis Papadrakakis



# Contents

<b>Motorway Traffic Flow Optimisation in Presence of Vehicle Automation and Communication Systems . . . . .</b>	<b>1</b>
Claudio Roncoli, Markos Papageorgiou and Ioannis Papamichail	
<b>Characterizing Urban Dynamics Using Large Scale Taxicab Data . . . .</b>	<b>17</b>
Xinwu Qian, Xianyuan Zhan and Satish V. Ukkusuri	
<b>Holistic Calibration of Microscopic Traffic Flow Models: Methodology and Real World Application Studies. . . . .</b>	<b>33</b>
Alexander Paz, Victor Molano and Javier Sanchez-Medina	
<b>Optimizing Leisure Travel: Is BigData Ready to Improve the Joint Leisure Activities Efficiency? . . . . .</b>	<b>53</b>
K. Gkiotsalitis and A. Stathopoulos	
<b>Air Traffic Flow Management Data Mining and Analysis for In-flight Cost Optimization. . . . .</b>	<b>73</b>
Leonardo L.B.V. Cruciol, Li Weigang, John-Paul Clarke and Leihong Li	
<b>Simulation Optimization of Car-Following Models Using Flexible Techniques . . . . .</b>	<b>87</b>
Vasileia Papathanasopoulou and Constantinos Antoniou	
<b>Computational Intelligence and Optimization for Transportation Big Data: Challenges and Opportunities . . . . .</b>	<b>107</b>
Eleni I. Vlahogianni	
<b>Nature-Inspired Algorithms: Success and Challenges. . . . .</b>	<b>129</b>
Xin-She Yang	

**Comparative Study on Recent Metaheuristic Algorithms in Design Optimization of Cold-Formed Steel Structures . . . . .** 145  
M.P. Saka, S. Carbas, I. Aydogdu, A. Akin and Z.W. Geem

**Adaptive Switching of Variable-Fidelity Models in Population-Based Optimization . . . . .** 175  
Ali Mehmani, Souma Chowdhury, Weiyang Tong and Achille Messac

**Genetic Algorithms for the Construction of 2<sup>2</sup> and 2<sup>3</sup>-Level Response Surface Designs . . . . .** 207  
Dimitris E. Simos

**Reactive Power and Voltage Control Based on Mesh Adaptive Direct Search Algorithm . . . . .** 217  
Seyyed Soheil Sadat Hosseini, Amir H. Gandomi, Alireza Nemati and Seyed Hamidreza Sadat Hosseini

**Optimal Placement of Hysteretic Dampers via Adaptive Sensitivity-Smoothing Algorithm . . . . .** 233  
Yu Murakami, Katsuya Noshi, Kohei Fujita, Masaaki Tsuji and Izuru Takewaki

**Design of Tuned Mass Dampers via Harmony Search for Different Optimization Objectives of Structures . . . . .** 249  
Sinan Melih Nigdeli and Gebrail Bekdas

**Tailoring Macroscale Response of Mechanical and Heat Transfer Systems by Topology Optimization of Microstructural Details . . . . .** 267  
Joe Alexandersen and Boyan Stefanov Lazarov

**Aerodynamic Shape Optimization Using “Turbulent” Adjoint And Robust Design in Fluid Mechanics . . . . .** 289  
Kyriakos C. Giannakoglou, Dimitrios I. Papadimitriou, Evangelos M. Papoutsis-Kiachagias and Ioannis S. Kavvadias

**Hierarchical Topology Optimization for Bone Tissue Scaffold: Preliminary Results on the Design of a Fracture Fixation Plate . . . . .** 311  
Emily Gogarty and Damiano Pasini

**Boundary Constraint Handling Affection on Slope Stability Analysis . . . . .** 341  
Amir H. Gandomi, Ali R. Kashani and Mehdi Mousavi

**Blackbox Optimization in Engineering Design: Adaptive Statistical Surrogates and Direct Search Algorithms** . . . . . 359  
Bastien Talgorn, Le Digabel Sébastien and Michael Kokkolaras

**Life Cycle Analysis and Optimization of a Steel Building** . . . . . 385  
G.K. Bekas, D.N. Kaziolas and G.E. Stavroulakis

**Optimization of Reinforced Concrete Columns Subjected to Uniaxial Loading** . . . . . 399  
Gebrail Bekdaş and Sinan Melih Nigdeli

**The Effect of Stakeholder Interactions on Design Decisions** . . . . . 413  
Garrett Waycaster, Christian Bes, Volodymyr Bilotkach, Christian Gogu, Raphael Haftka and Nam-Ho Kim

**A Fixed Point Approach to Bi-level Multi-objective Problems** . . . . . 427  
Carla Antoni and Franco Giannessi

**Reliability-Based Shape Design Optimization of Structures Subjected to Fatigue** . . . . . 451  
Manolis Georgioudakis, Nikos D. Lagaros and Manolis Papadrakakis

**A Stress-Test of Alternative Formulations and Algorithmic Configurations for the Binary Combinatorial Optimization of Bridges Rehabilitation Selection** . . . . . 489  
Dimos C. Charmpis and Loukas Dimitriou

# Motorway Traffic Flow Optimisation in Presence of Vehicle Automation and Communication Systems

Claudio Roncoli, Markos Papageorgiou and Ioannis Papamichail

**Abstract** This paper describes a novel approach for defining optimal strategies in motorway traffic flow control, considering that a portion of vehicles are equipped with Vehicle Automation and Communication Systems (VACS). An optimisation problem, formulated as a convex Quadratic Programming (QP) problem, is developed with the purpose of minimising traffic congestion. The proposed problem is based on a first-order macroscopic traffic flow model able to capture the lane changing and the capacity drop phenomena. An application example demonstrates the achievable improvements if the vehicles travelling on the motorway are influenced by the control actions computed as a solution of the optimisation problem.

## 1 Introduction

The mitigation of traffic congestion on motorway systems is a useful but complex task that could generate significant economical and environmental advantages for the modern society. As a matter of fact, motorways, particularly in and around metropolitan areas, suffer from congestion for long periods during every day and, ironically, the major congestion and related infrastructure degradation appear during the period of maximum traffic demand. Despite the huge improvements achieved in Information Technology during the last decades, a smart and widespread application of these technologies to alleviate traffic congestion is still not fully achieved.

On the other hand, there has been an enormous interdisciplinary effort by the automotive industry as well as by numerous research institutions around the world

---

C. Roncoli (✉) · M. Papageorgiou · I. Papamichail  
Dynamic Systems and Simulation Laboratory, Technical University of Crete,  
73100 Chania, Greece  
e-mail: croncoli@dssl.tuc.gr

M. Papageorgiou  
e-mail: markos@dssl.tuc.gr

I. Papamichail  
e-mail: ipapa@dssl.tuc.gr

to plan, develop, test, and start deploying a variety of Vehicle Automation and Communication Systems (VACS) that are expected to revolutionise the features and capabilities of individual vehicles within the next decades. Several research works were realised in the past, sometimes foreseeing future scenarios where self-driving vehicles are part of a completely connected road infrastructure. The seminal work [13] introduced the concept of highly automated Intelligent Vehicle Highway System (Smart-IVHS), introducing a possible hierarchical control structure with the purpose of increasing highway capacity and safety. The authors defined simple policies for prescribing and regulating lane-changing policies and desired speeds in an interconnected control system. Further studies exploited the concept of Automated Highway System (AHS), defining a set of layers and developing control strategies for each one of them. In this context, an interesting work was presented in [9], where the authors analysed specifically the link-layer control problem and proposed a control law for the stabilisation of traffic conditions. It must be highlighted that the concept of platooning (i.e. the organisation of vehicles into closely spaced groups) is often considered as a good approach, capable of increasing the motorway capacity and reducing instability. Another interesting research work is described in [6]; the authors defined a model based on linear programming for assigning traffic to lanes. The AHS was modelled as a static trip-based multi-commodity network, in which the objective was to maximise the total outflow subject to predetermined O/D patterns.

It is a common opinion that an extensive use of VACS will cause an improvement of traffic conditions, however a lot of effort is required in order to define models and strategies that could generate the expected enhancement. In this paper, it is assumed that the use of VACS permits to exploit new control actions, allowing to achieve higher improvement of traffic conditions.

The paper is written according to the following structure: in Sect. 2, the proposed traffic flow model is described, whereas Sect. 3 presents the formulation of the optimisation problem. In Sect. 4, the proposed problem is applied to an example network, stating the obtained improvements and highlighting some aspects to be considered for practical purposes. Section 5 concludes the paper and proposes possible future extensions.

## 2 A Traffic Flow Model for Multiple-Lane Motorways

### 2.1 Multiple-Lane Traffic Flow Models

The motorway traffic flow models that are commonly studied in literature (e.g., the Cell Transmission Model-CTM [2] and METANET [11]) take into account aggregate dynamics for all the lanes of each modelled motorway stretch. This simplification is reasonable for most of the control purposes since the control actions normally involve all the lanes together. However, having vehicles equipped with intelligent devices creates the possibility of defining more customised control strategies (e.g. assigning

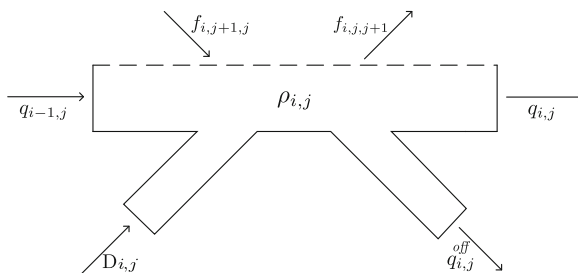
to each equipped vehicle corresponding tasks to be performed). For this reason, the proposed model is defined considering the lanes of the motorway network as different entities, characterised by their own state and control variables while developing the dynamic equations.

Only a few works on multiple-lane motorways have been carried out in past research. In the first main work [4], it is assumed that lane densities on a multi-lane highway oscillate around an equilibrium density; the authors developed a methodology to attenuate the disturbances and tried to increase the stability of the system. That work inspired the authors of [10], that proposed three models for capturing the lane changing behaviour. The first model is a continuum model based on the assumption that vehicles change lanes according to the difference of the deviations of their densities from equilibrium values. The second model extends the first one, taking also into account acceleration and inertia effects, obtaining a second-order model. A third extension is also proposed, considering also the street width. However, these models were formulated without applying any discretisation scheme. In the more recent work [7], the authors exploited the kinematic wave (KW) theory, proposing a multi-lane KW-based model as a first module of a more complex model that considers also moving blockages treated as particles characterised by bounded acceleration rates; lane changings are assigned according to the difference of mean speed between two adjacent lanes.

## 2.2 Model Formulation

The multiple-lane motorway is represented introducing the indices  $j = 1, \dots, J$  for lanes and  $i = 1, \dots, I$  for segments. The simulation time  $t = kT$  is defined considering the discrete time step  $T$  and the simulation index  $k = 1, \dots, K$ , where  $K$  defines the simulation horizon. The motorway is spatially subdivided introducing the segment-lane entities (see Fig. 1), characterised by the following variables:

- the density  $\rho_{i,j}(k)$  [veh/km], i.e. the number of vehicles in the segment  $i$ , lane  $j$ , at time step  $k$ , divided by the segment length  $L_i$ ;



**Fig. 1** The segment-lane variables used in the model formulation

- the longitudinal flow  $q_{i,j}(k)$  [veh/h], i.e. the traffic volume leaving segment  $i$  and entering segment  $i + 1$  during time interval  $(k, k + 1]$ , remaining in lane  $j$ ;
- the lateral flow  $f_{i,j,\bar{j}}(k)$  [veh/h] ( $\bar{j} = j \pm 1$ ), i.e. the traffic volume moving from lane  $j$  to lane  $\bar{j}$  (vehicles changing lane remain in the same segment during the current time interval); and
- the demand flow  $D_{i,j}(k)$  [veh/h], i.e. the flow entering from the on-ramp located at segment  $i$ , lane  $j$ , during the time interval  $(k, k + 1]$ .

The off-ramp flow is determined as a percentage of the total flow passing through all the lanes of the segment, defined by the given turning rates  $\gamma_{i,j}(k)$ :

$$q_{i,j}^{off}(k) = \gamma_{i,j}(k) \sum_{j=1}^J q_{i,j}(k). \quad (1)$$

The following conservation equation is introduced, defining the dynamics of traffic density  $\rho_{i,j}(k)$ :

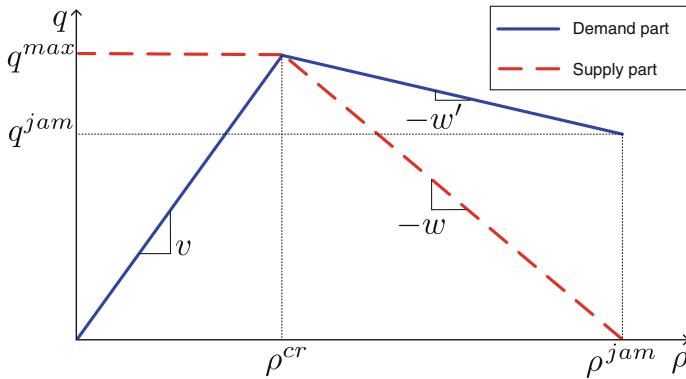
$$\begin{aligned} \rho_{i,j}(k+1) = \rho_{i,j}(k) + \frac{T}{L_i} & \left[ q_{i-1,j}(k) + D_{i,j}(k) - q_{i,j}(k) - q_{i,j}^{off}(k) \right. \\ & \left. + f_{i,j+1,j}(k) + f_{i,j-1,j}(k) - f_{i,j,j-1}(k) - f_{i,j,j+1}(k) \right]. \quad (2) \end{aligned}$$

In order to ensure numerical stability, the time step  $T$  must be selected so as to respect the Courant-Friedrichs-Lewy (CFL) condition [1]:

$$T \leq \min_{i,j} \frac{L_i}{v_{i,j}^{free}} \quad (3)$$

where  $v_{i,j}^{free}$  is the free speed defined for segment  $i$ , lane  $j$ .

The next modelling issue to address is the specification of bounds for the longitudinal flow. The starting basis for this is the well-known CTM [2, 3], that nevertheless does not take into account the capacity drop phenomenon, i.e. the reduction of discharge flow once a congestion is formed. The reasons for this phenomenon are not exactly known, however it seems to be caused by the limited acceleration of vehicles while exiting a congested area. In second-order models, such as METANET [11], the capacity drop is generated by the equations describing the spatiotemporal evolution of speed. This option is not available for first-order LWR models, and, in order to overcome this shortcoming, several attempts have been made. The chosen approach is based on [8], where the problem is addressed by imposing an upper bound to the acceleration depending on the traffic phase, distinguishing between LWR and maximum acceleration. The proposed modelling approach, that is represented only by (piecewise) linear equations, is thus represented by a modification of the demand part of the Fundamental Diagram (FD) in the following way: in case of congestion ( $\rho_{i,j}(k) > \rho_{i,j}^{cr}$ , where  $\rho_{i,j}^{cr}$  is the critical density), the demand flow is linearly



**Fig. 2** Graphical representation of the proposed Fundamental Diagram

decreased according to a fixed slope  $-w'$ . This leads to a flow  $q_{i,j}^{jam}$  that is allowed to leave a segment in a completely congested state ( $\rho_{i,j}(k) = \rho_{i,j}^{jam}$ ). A graphical representation of the proposed FD is depicted in Fig. 2. A more extensive description of the macroscopic model can be found in [14].

### 3 The Optimisation Problem

The model described in Sect. 2 is exploited for the formulation of an optimisation problem with the aim of improving the motorway conditions by reducing traffic congestion. It is supposed that the following control actions are utilised:

- Ramp Metering (RM) is currently applied on many motorways (see e.g. [12]) and does not necessarily require any additional equipment to be performed.
- Mainstream Traffic Flow Control (MTFC) via Variable Speed Limits (VSL): it is assumed that the exiting flows (and consequently the speeds) are controlled for each segment-lane; thus all equipped vehicles travelling on a segment-lane will receive and apply the respective speed as a speed limit. For a sufficient penetration of equipped vehicles, this will result in the observance of the speed limit by non-equipped vehicles as well.
- Lane-Changing Control: the optimal lateral flows are computed for each segment-lane, but the implementation of this control action is more cumbersome and uncertain than the previous two, unless all vehicles are under full guidance by the control center; in this latter case, it is not difficult to implement the control action by sending lane-changing orders to an appropriate number of vehicles. In all other cases, an intermediate algorithm should decide on the number and ID of equipped vehicles that should receive a lane-changing advice, taking into account the compliance rate and the spontaneous lane-changings; the latter may be reduced by involv-



ing additional “keep-lane” advice to other vehicles. These issues are currently in course of investigation and development.

Since RM actions are applied, the following variables are added considering the creation of queues at on-ramps:

- the queue length  $w_{i,j}(k)$  [veh], i.e. the number of vehicles queuing at on-ramp located in segment  $i$ , lane  $j$ , at time step  $k$ ; each queue is characterised by a maximum length  $w_{i,j}^{max}$ ;
- the on-ramp flow  $r_{i,j}(k)$  [veh/h], as the flow entering the network, leaving the queue generated in segment  $i$ , lane  $j$ , during the time interval  $(k, k + 1]$ ; this variable replaces the demand flow  $D_{i,j}(k)$  in (2);
- the extra-queue length  $W_{i,j}(k)$  [veh], that represents an additional state variable considering vehicles that cannot enter the queue because it has reached its maximum length; the introduction of this variable permits to avoid the infeasibility of the optimisation problem in scenarios with very high demand; and
- the flow  $d_{i,j}(k)$  [veh/h], i.e. the demand flow that is capable to enter the real queue; therefore in case the maximum size is not reached, it results  $d_{i,j}(k) = D_{i,j}(k)$ .

The problem is formalised as a convex Quadratic Program (QP), characterised by a convex quadratic cost function and uniquely linear constraints, allowing its application also for large networks.

$$\begin{aligned}
 \min_{\rho, w, W, q, r, f} Z = & T \sum_{k=1}^K \sum_{i=1}^I \sum_{j=1}^J [L_i \rho_{i,j}(k) + w_{i,j}(k)] \\
 & + M \sum_{k=1}^K \sum_{i=1}^I \sum_{j=1}^J W_{i,j}(k) \\
 & + \sum_{k=1}^K \sum_{i=1}^I \sum_{j=1}^J [\beta_{i,j,j-1} f_{i,j,j-1}(k) + \beta_{i,j,j+1} f_{i,j,j+1}(k)] \\
 & + \lambda^r \sum_{k=2}^K \sum_{i=1}^I \sum_{j=1}^J [r_{i,j}(k) - r_{i,j}(k-1)]^2 \\
 & + \lambda^f \sum_{k=2}^K \sum_{i=1}^I \left\{ \sum_{j=2}^J [f_{i,j,j-1}(k) - f_{i,j,j-1}(k-1)]^2 \right. \\
 & \quad \left. + \sum_{j=1}^{J-1} [f_{i,j,j+1}(k) - f_{i,j,j+1}(k-1)]^2 \right\} \\
 & + \lambda^{st} \sum_{k=2}^K \sum_{i=1}^I \sum_{j=1}^J \frac{\left[ q_{i,j}(k) - q_{i,j}(k-1) + v^{free} [\rho_{i,j}(k) + \rho_{i,j}(k-1)] \right]^2}{(\rho_{i,j}^{cr})^2} \\
 & + \lambda^{sl} \sum_{k=1}^K \sum_{i=2}^I \sum_{j=1}^J \frac{\left[ q_{i,j}(k) - q_{i-1,j}(k) + v^{free} [\rho_{i,j}(k) + \rho_{i-1,j}(k)] \right]^2}{(\rho_{i,j}^{cr})^2} \tag{4}
 \end{aligned}$$

Subject to:

$$\rho_{i,j}(k+1) = \rho_{i,j}(k) + \frac{T}{L_i} \left[ q_{i-1,j}(k) + r_{i,j}(k) - q_{i,j}(k) - q_{i,j}^{off}(k) + f_{i,j+1,j}(k) + f_{i,j-1,j}(k) - f_{i,j,j-1}(k) - f_{i,j,j+1}(k) \right] \quad (5)$$

$$w_{i,j}(k+1) = w_{i,j}(k) + T[d_{i,j}(k) - r_{i,j}(k)] \quad (6)$$

$$W_{i,j}(k+1) = W_{i,j}(k) + T[D_{i,j}(k) - d_{i,j}(k)] \quad (7)$$

$$q_{i,j}(k) \leq v_{i,j}^{free} \rho_{i,j}(k) \quad (8)$$

$$q_{i,j}(k) \leq \frac{v_{i,j}^{free} \rho_{i,j}^{cr} - q_{i,j}^{jam}}{\rho_{i,j}^{jam} - \rho_{i,j}^{cr}} \rho_{i,j}(k) + \frac{\rho_{i,j}^{cr} (v_{i,j}^{free} \rho_{i,j}^{jam} - \rho_{i,j}^{jam})}{\rho_{i,j}^{jam} - \rho_{i,j}^{cr}} \quad (9)$$

$$q_{i,j}(k) \leq v_{i+1,j}^{free} \rho_{i+1,j}^{cr} \quad (10)$$

$$q_{i,j}(k) \leq -\frac{v_{i+1,j}^{free} \rho_{i+1,j}^{cr}}{\rho_{i+1,j}^{jam} - \rho_{i+1,j}^{cr}} \rho_{i+1,j}(k) + \frac{v_{i+1,j}^{free} \rho_{i+1,j}^{cr} \rho_{i+1,j}^{jam}}{\rho_{i+1,j}^{jam} - \rho_{i+1,j}^{cr}} \quad (11)$$

$$[f_{i,j,j-1}(k) + f_{i,j,j+1}(k)] \leq \frac{L_i}{T} \rho_{i,j}(k) \quad (12)$$

$$[f_{i,j-1,j}(k) + f_{i,j+1,j}(k)] \leq \frac{L_i}{T} [\rho_{i,j}^{jam} - \rho_{i,j}(k)] \quad (13)$$

$$f_{i,j,j-1}(k) \leq f^{max}, \quad f_{i,j,j+1}(k) \leq f^{max} \quad (14)$$

$$\rho_{i,j}(k) \leq \rho_{i,j}^{jam}, \quad w_{i,j}(k) \leq w_{i,j}^{max}, \quad r_{i,j}(k) \leq r_{i,j}^{max} \quad (15)$$

The cost function (4) to be minimised is composed by various terms:

- the first and most important one is the Total Time Spent (TTS), that considers the overall time spent by vehicles both travelling and queuing at the on-ramps;
- the other linear terms are penalty terms defined with the purpose of reducing extra queues and lateral flows; and
- the quadratic terms are introduced in order to penalise time and space oscillations in the control values: ramp outflow, lateral movements, and the speed values; as a matter of fact, the last two terms represent a linearisation of the non-linear constraints that consider speed variation; these oscillations are penalised with respect to time and to space.

The first set of constraints represents the dynamics for the densities (5), that derives from (2), however replacing the external demand with the on-ramp flow; for the queues generated at on-ramps because of the RM actions (6); and for the extra-queues (7).

The following set of constraints corresponds to the FD described in Sect. 2, resulting in two equations for the demand term (8, 9) and two equations for the supply term (10, 11); it is important to highlight that, having the possibility of controlling the

flow (and indirectly the speed) of vehicles, the constraints could simply be described by linear inequalities that represent upper-bounds for the segment outflow.

The third set of constraints is related to the lateral flows. In this case, they appear only in the form of upper-bounds, representing the available vehicles (12) and the available space (13), allowing the optimiser to assign the best lateral flow.

The last set of constraints contains the upper-bounds for lateral flows (14), densities, on-ramp queues, and ramp flows (15).

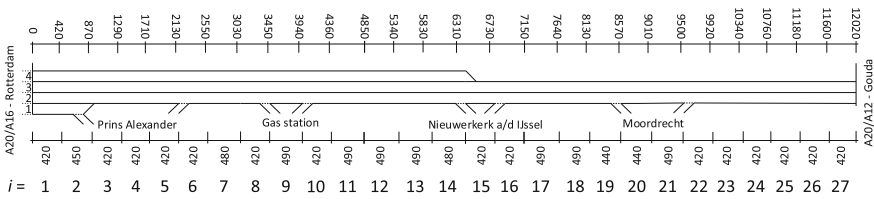
## 4 Application Example

In order to evaluate and illustrate the potential improvements that could be obtained by applying the described methodology, a stretch of the motorway A20 from Rotterdam to Gouda, the Netherlands, taken from [17], is used. The topological characteristics of this network (lane-drops, on-ramps and off-ramps) make it a very interesting test-bed for evaluating the results of the proposed optimisation problem.

The stretch, about 12 km in length, is subdivided into 27 segments of 450 m in average, as shown in Fig. 3. The time step is set to  $T = 15$  s. The lanes are numbered  $j = 1, \dots, 4$  from the inner lane (close to the roadside) to the outer lane (close to the road median).

It is supposed that both the densities and ramp queues are initialised to 0 at the beginning of the simulation. All links have the same characteristic values: the critical density is set to  $\rho_{i,j}^{cr} = 22$  veh/km, the jam density is set to  $\rho_{i,j}^{jam} = 180$  veh/km, the maximum speed is  $v_{i,j}^{max} = 100$  km/h, and the maximum flow at jam density is  $q_{i,j}^{jam} = 1467.4$  veh/h (obtained by setting a slope  $w' = w/3$ ). The exit rates at off-ramps are set as follows:  $\gamma_{2,1}(k) = 0.2$ ,  $\gamma_{8,2}(k) = 0.0085$ ,  $\gamma_{14,2}(k) = 0.3$ , and  $\gamma_{19,2}(k) = 0.2$ , for all  $k = 1, \dots, K$ .

A significant aspect is the tuning of the cost function weights: once a proper value of  $M$  is determined in order to avoid extra-queues (in this case,  $M = 10$ ), the tuning procedure is focused on keeping virtually the same TTS while, at the same time, trying to obtain reduced lateral flows and smooth control actions. For the linear penalty term related to lateral flow, an important aspect is also related to the locations of these control actions. In locations where strong lateral flow actions are



**Fig. 3** The A20 motorway stretch from Rotterdam to Gouda used to test the proposed approach

**Table 1** Comparison of computation time with respect to the optimisation horizon (and consequently the size of the optimisation problem)

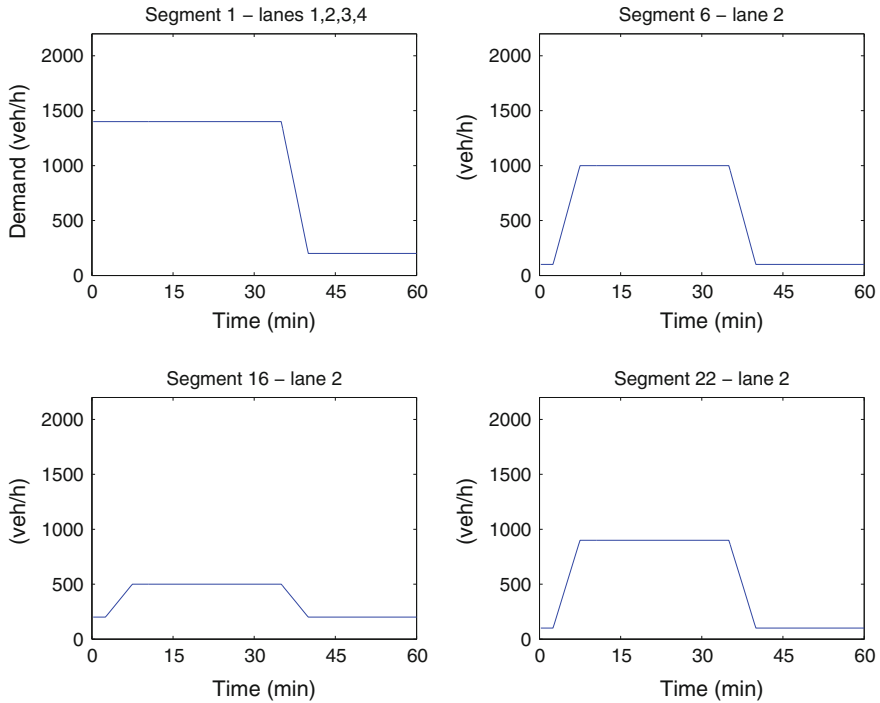
Optimisation horizon (min)	Number of variables	Number of equalities	Number of inequalities	Computation time (s)
60	206820	77760	308880	~20
45	154980	58320	96660	~13
30	103340	38880	64260	~7

expected (e.g. at lane-drops and on-ramps), vehicles are encouraged to change lane in the segment immediately upstream by setting the weight  $\beta_{i,j,\bar{j}} = 0$ ; in all other segments, the values are set to  $\beta_{i,j,\bar{j}} = 0.01$ . As a last step, the weigh parameters of quadratic terms were tuned, obtaining the following values:  $\lambda_f = 10^{-5}$ ,  $\lambda_r = 10^{-7}$ ,  $\lambda_{st} = 10^{-5}$ , and  $\lambda_{sl} = 10^{-6}$ .

For the resolution of the optimisation problem the solver Gurobi [5] has been utilised, choosing a barrier method algorithm for QP solving. Despite the considerable size of the optimisation problem that is obtained even for small networks, the solution is achieved in a reasonable time. In fact, as it is shown in Table 1, setting an optimisation horizon of 45 min or less, the solution is achieved in a computation time that is smaller than the simulation step, making this approach feasible also for real-time control, e.g. using this problem as a module in a Model Predictive Control (MPC) framework, as proposed in [16].

In order to highlight the computed control actions, the following example is examined considering an optimisation horizon of 60 minutes. The utilised demand profile is shown in Fig. 4, where the reduced entering flows during the last 20 min represent a cool down period that will ensure the dissolution of any congestion at the end of the simulation.

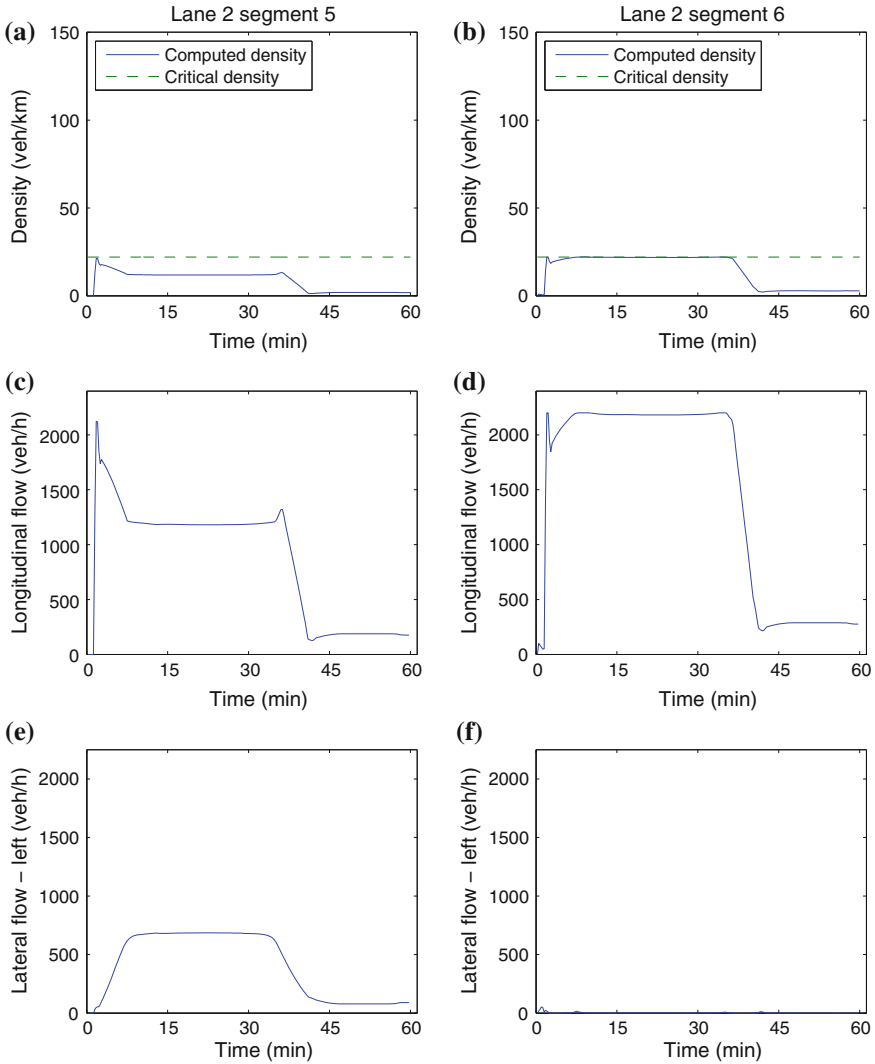
It should be highlighted that, thanks to the realistic modelling of the motorway traffic and the inherent intelligence of optimisation algorithms, the optimisation problem solution leads automatically to a plethora of complex, interrelated, and highly efficient control actions that have to be unveiled via careful observation, analysis, and interpretation of the obtained solution. A guiding principle while analysing the obtained optimal results is the attempt to maximise the flows at bottlenecks, i.e. at on-ramp merge segments and at the lane-drop segment. Bottleneck flow would be reduced if the corresponding merge segment density becomes overcritical, leading to reduced outflow due to capacity drop according to Fig. 2; maximum bottleneck flow is achieved if the density of the merge segment is maintained at its critical value. If the overall upstream and on-ramp demand can be accommodated via appropriate lane assignments, corresponding lane changing orders are produced to avoid any overcritical densities. On the other hand, if the demand exceeds the bottleneck capacity, this may call for holding back traffic at the on-ramp, via appropriately timed and sized RM actions; and, if the on-ramp storage is not sufficient, additional MTFC actions may be needed to also hold back traffic upstream of the bottleneck. In other words, unavoidable (due to high demand) queueing, congestion, and delays need to



**Fig. 4** Demand profiles at on-ramps; “Gas station” is omitted due to the very low entering flow (a maximum value of 30 veh/h)

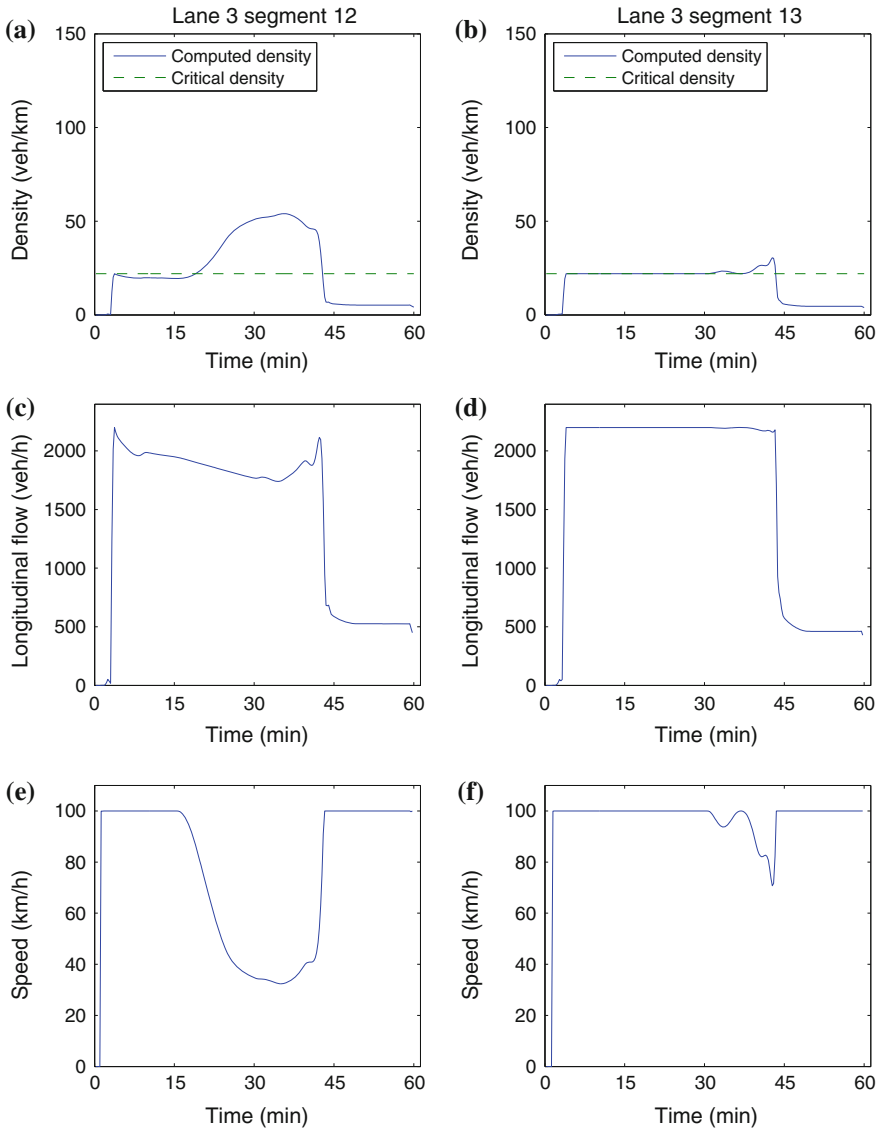
be placed intelligently in space and time, so as to maximise throughput and hence efficiency of the traffic system. The following specific actions could be identified in the optimal control solution:

- Whenever the all-lanes segment capacity is sufficient to accommodate the flow entering at the on-ramp, the required space at the merge segment is created by assigning inward lateral flow actions just upstream of the on-ramp merge area. This happens at on-ramp “Prins Alexander”, as depicted in Fig. 5. It is interesting to highlight that these lane-changing actions lead to achieve the maximum flow at merging segment 6 (that includes also the flow entering from the on-ramp “Prins Alexander”). Since the lateral movements are performed in the upstream segment, it is also worthwhile to point out that the phenomenon of having vehicles entering the motorway and changing lane in the same segment is avoided, as it is visible in Fig. 5f.
- In the lane-drop area of segment 14, the space for vehicles that have to change lane is created through some MTFC actions in the upstream segments, as shown in Fig. 6, thus avoiding an excessive increase of density and excessive vehicle movements at the segment of the lane-drop.



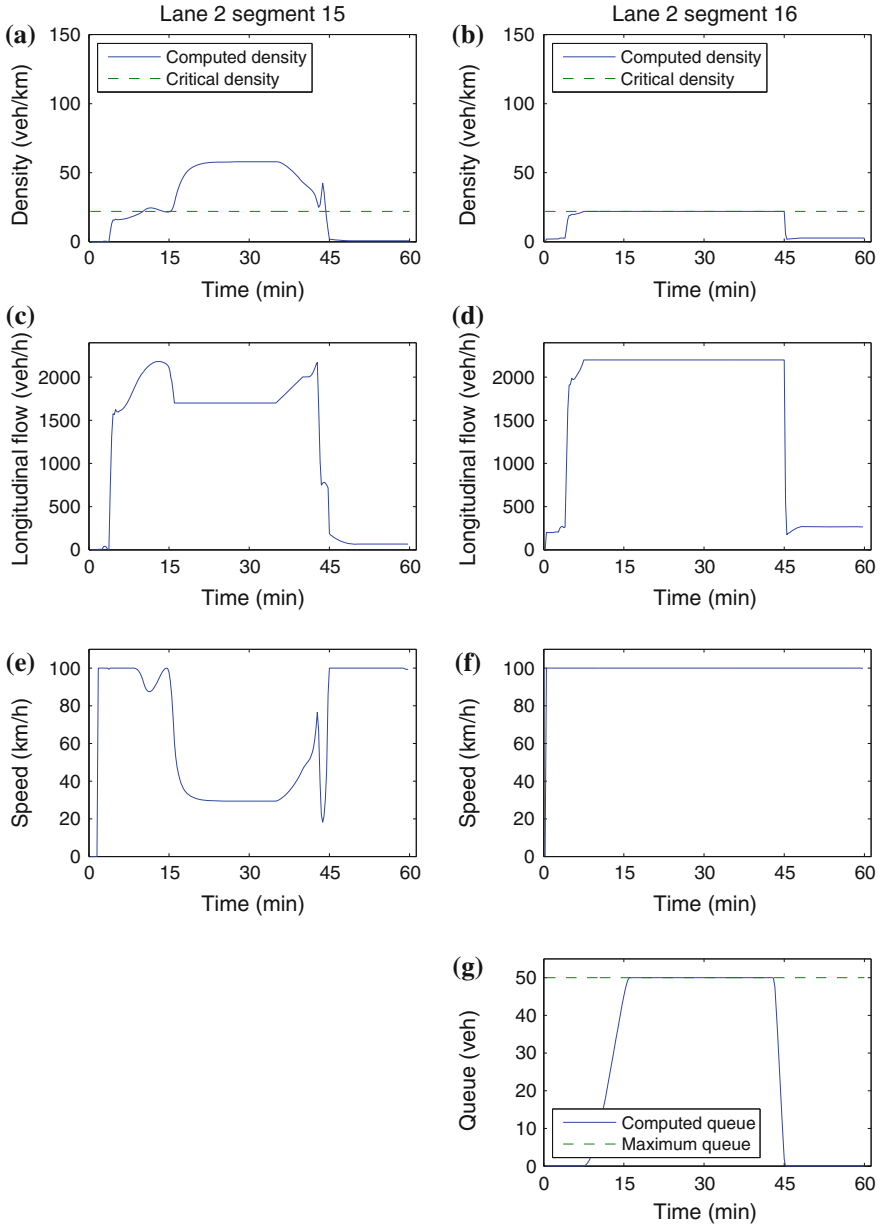
**Fig. 5** A potential congestion forming at the merge area “Prins Alexander” (segment 6) is avoided by creating some “space” in lane 2 (shoulder lane) for traffic entering from the on-ramp, as can be seen from the density plot (a) and the corresponding flow (c). This is achieved by assigning lateral flow from lane 2 to lane 3 (e); in contrast, no lane-changes are ordered at the merge segment (f). After vehicles have entered from the on-ramp, the flow reaches its capacity value (d), and the density is at the critical value (b)

- In case the demand flow is higher than the segment overall capacity, RM and VSL actions are jointly applied, allowing to maintain capacity flow and to avoid any speed breakdown. These actions appear both at the on-ramp “Nieuwerkerk a/d IJssel” (see Fig. 7) and “Moordrecht” (see Fig. 8).



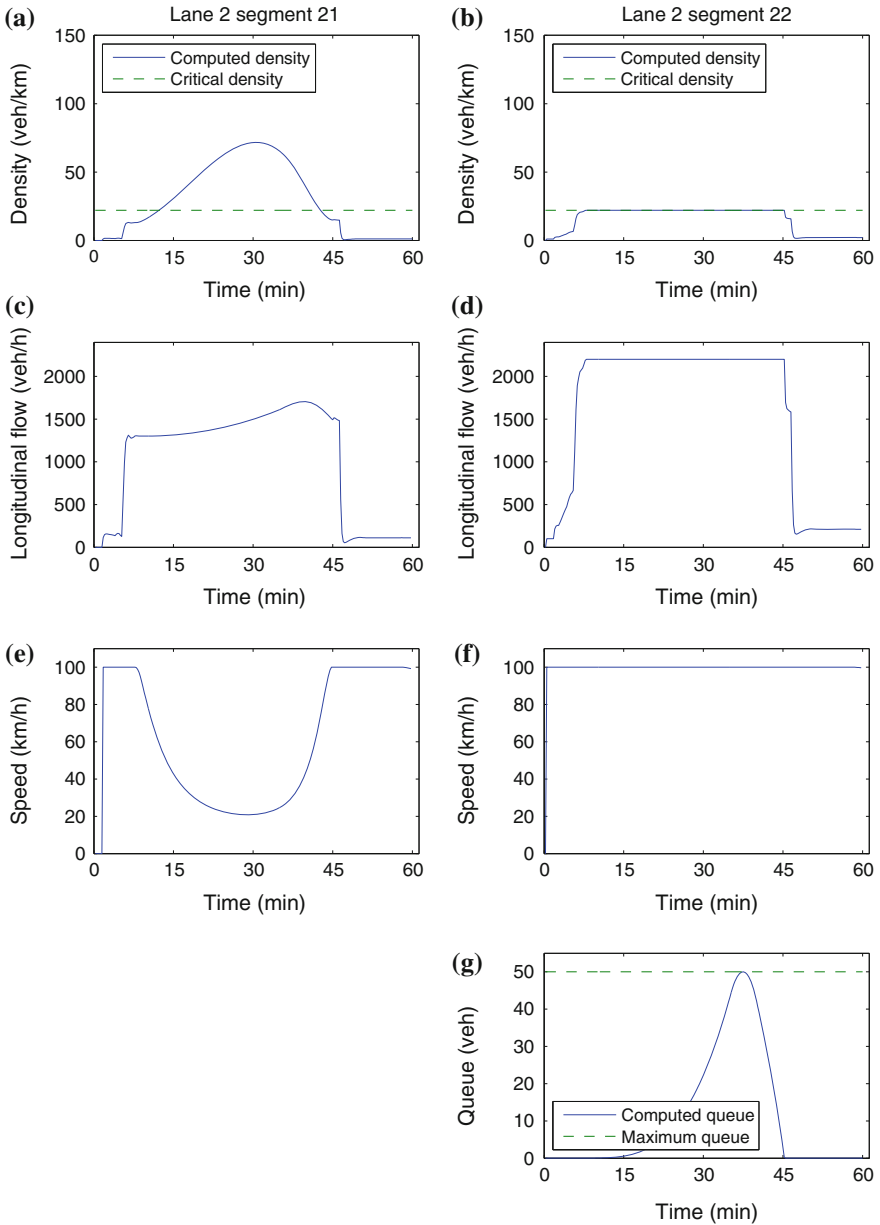
**Fig. 6** At the lane-drop area (segment 14), the space necessary for vehicles moving from lane 4 to lane 3, is created performing VSL actions in the upstream segment (e); this leads to an increase of density (a) but also to the reduction of the corresponding flow (c); these actions permit to achieve the critical density (b), and, consequently, the capacity flow (d) at segment 13

It is interesting to point out that MTFC actions are performed only in the lanes which face a capacity problem due to merging, i.e. lane 2 for all on-ramps and lane 3 for the lane-drop location; whereas in the other lanes the flow (and consequently the speed) remains constantly at the maximum value. In addition, some minor VSL actions are

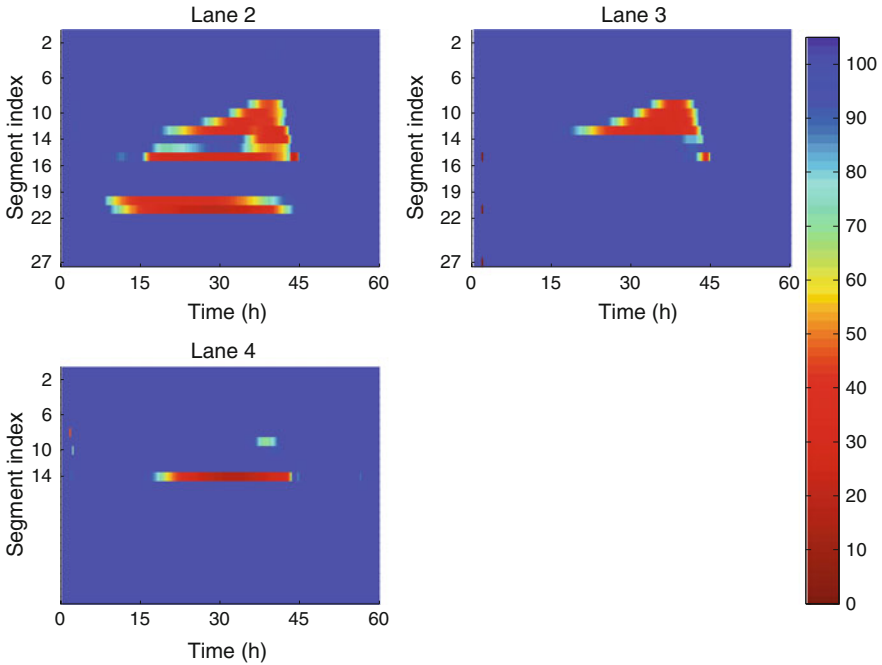


**Fig. 7** At the “Nieuwerkerk a/d IJssel” on-ramp, strong RM actions are performed leading to maximum ramp queue length over an extended period of time (g). As soon as the ramp queue reaches its maximum value (at  $t = 15$  min), MTFC via VSL (e) is activated in the upstream segment to limit the mainstream slow that enters the merge area (c). This leads to an upstream congestion (a), but enables to obtain the critical density (b) and the capacity flow (d) at the merge area, while the speed continues to be high (f)





**Fig. 8** Similarly to the case illustrated in Fig. 7, also at the “Moorrecht” on-ramp, a combined use of RM (g) and MTFC via VSL (c, e) is performed, achieving again flow maximisation (d) and critical density (b) at segment 22



**Fig. 9** The contour plots show the decrease of speed due to MTFC actions in the proximity of on-ramps, appearing only during the period of higher demand

taken in order to help vehicles that are changing lanes in proximity of the on- and off-ramps, as well as because of the lane drop. Space-time contour plots of speeds are displayed in Fig. 9. The visible congestion areas are unavoidable due to high demand (exceeding capacity), but are placed optimally (in space and time) for maximum throughput.

Interested readers may consult [15] for optimal control results obtained with the same methodology, but for a different (real) motorway infrastructure that features partly different phenomena and control actions.

## 5 Conclusions

The paper describes a novel multiple-lane traffic flow model on which an optimisation problem is based. The model includes some simplifications that have allowed to obtain a QP with only linear constraints. The low required computation time makes this methodology suitable for real-time applications, as well as usable in a hierarchical control approach. The exploitation of this work as the optimisation module in a MPC scheme is the subject of ongoing research activities [16]. However, in order to make

possible to implement this strategy also in case of mixed traffic (that contains both vehicles equipped with VACS and traditional ones), the design of an appropriate hierarchical control structure seems to be necessary.

**Acknowledgments** The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 321132, project TRAMAN21.

The authors would like to thank Prof. Bart van Arem and his group for their support in providing information related to the network used in the application example.

## References

1. Courant R, Friedrichs K, Lewy H (1928) Über die partiellen Differenzengleichungen der mathematischen Physik. *Math Ann* 100(1):32–74 Dec
2. Daganzo CF (1994) The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transp Res B Methodol* 28(4):269–287 Aug
3. Daganzo CF (1995) The cell transmission model, part II: network traffic. *Transp Res B Methodol* 29(2):79–93 Apr
4. Gazis DC, Herman R, Weiss GH (1962) Density oscillations between lanes of a multilane highway. *Oper Res* 10:658–667
5. Gurobi Optimization Inc (2013) Gurobi optimizer reference manual, Houston
6. Hall RW, Lotspeich D (1996) Optimized lane assignment on an automated highway. *Transp Res C Emer Technol* 4(4):211–229 Aug
7. Laval JA, Daganzo CF (2006) Lane-changing in traffic streams. *Transp Res B Methodol* 40(3):251–264 Mar
8. Lebacque J (2003) Two-phase bounded-acceleration traffic flow model: analytical solutions and applications. *Transp Res Rec* 1852(1):220–230 Jan
9. Li PY, Horowitz R, Alvarez L, Frankel J, Robertson AM (1997) An automated highway system link layer controller for traffic flow stabilization. *Transp Res C Emer Technol* 5(1):11–37 Feb
10. Michalopoulos PG, Beskos DE, Yamauchi Y (1984) Multilane traffic flow dynamics: some macroscopic considerations. *Transp Res B Methodol* 18(4–5):377–395 Aug
11. Papageorgiou M, Messmer A (1990) METANET: a macroscopic simulation program for motorway networks. *Traffic Eng Control* 31(9):466–470
12. Papamichail I, Papageorgiou M, Vong V, Gaffney J (2010) Heuristic ramp-metering coordination strategy implemented at monash freeway, Australia. *Transp Res Rec J Trans Res Board* 2178:10–20
13. Rao B, Varaiya P (1994) Roadside intelligence for flow control in an intelligent vehicle and highway system. *Transp Res C: Emerg Technol* 2(1):49–72 Mar
14. Roncoli C, Papageorgiou M, Papamichail I (2014) Traffic flow optimisation in presence of vehicle automation and communication systems—Part I: a first-order multi-lane model for motorway traffic. *Emerging Technologies*, Submitted to Transportation Research Part C
15. Roncoli C, Papageorgiou M, Papamichail I (2014) Traffic flow optimisation in presence of vehicle automation and communication systems—Part II: Optimal control for multi-lane motorways. *Emerging Technologies*, Submitted to Transportation Research Part C
16. Roncoli C, Papamichail I, Papageorgiou M (2014) Model predictive control for multi-lane motorways in presence of VACS. In: 17th International IEEE conference on intelligent transportation systems (ITSC2014)
17. Schakel W, van Arem B (2013) Improving traffic flow efficiency by In-car advice on lane, speed, and headway. *Transportation Research Board 92nd Annual Meeting*

# Characterizing Urban Dynamics Using Large Scale Taxicab Data

Xinwu Qian, Xianyuan Zhan and Satish V. Ukkusuri

**Abstract** Understanding urban dynamics is of fundamental importance for the efficient operation and sustainable development of large cities. In this paper, we present a comprehensive study on characterizing urban dynamics using the large scale taxi data in New York City. The pick-up and drop-off locations are firstly analyzed separately to reveal the general trip pattern across the city and the existence of unbalanced trips. The inherent similarities among taxi trips are further investigated using the two-step clustering algorithm. It builds up the relationship among detached areas in terms of land use types, travel distances and departure time. Moreover, human mobility pattern are inferred from the taxi trip displacements and is found to follow two stages: an exponential distribution with short trips and a truncated power law distribution for longer trips. The result indicates that the taxi trip may not fully represent human mobility and is heavily affected by trip expenses and the urban form and geography.

## 1 Introduction

The rapid urbanization process gives birth to megacities cities such as Tokyo, Shanghai and New York City (NYC). Not only are megacities big in terms of population density, they also bring up unprecedented opportunities and challenges. With large population density and tremendous human activities, one critical challenge is how to manage the giant urban system efficiently and sustainably. Urban dynamics represents the spatiotemporal principles followed by urban functioning evolvments [21].

---

X. Qian · X. Zhan · S.V. Ukkusuri (✉)  
Purdue University, 550 Stadium Mall Dr, West Lafayette, IN, USA  
e-mail: qian39@purdue.edu

X. Zhan  
e-mail: zhanxianyuan@purdue.edu

S.V. Ukkusuri  
e-mail: sukkusur@purdue.edu

Understanding urban dynamics helps to capture the pulse of urban activities, which undoubtedly provides a huge step forward to address the problem.

In the past few decades, efforts have been made in modeling and simulating urban dynamics using data from transportation systems [1, 2, 8, 9]. However, the inherent complexity such as random behavior and the impact of geographical boundary can hardly be described properly using mathematical models. In the era of big data, we are widely exposed to various data sources and data-driven methods start to gain popularities. Compared with traditional data collected from surveys and questionnaires, the pervasive computing devices are able to collect abundant data in an efficient and accurate manner. Moreover, the digital footprints from mobile sensors such as GPS device and cellular mobile provides an opportunity to learn in-depth fundamentals of human mobility. Several pioneering studies have implemented various data sources to reveal urban activity participation and individual mobility patterns [4, 7, 18, 19]. A case study in Milan discovered the urban spatiotemporal variations of activity intensity [18]. The intensity of activity locations is further used to locate hot spots and identify city structure by analyzing spatiotemporal signatures of Erlang data, which is a measurement of network bandwidth [19]. Gonzalez et al. revealed a highly regulated human mobility pattern [7] from 100,000 mobile phone users trajectories, and Calabrese et al. established a multivariate regression model to predict daily human mobility [4]. Hasan et al. [10] examined both aggregate and individual activity patterns from social media check-in data. Brockmann et al. [3] studied the distance of human travel from the distribution of bank notes. They found that distance distribution follows a power law and can be well approximated using continuous-time random walk.

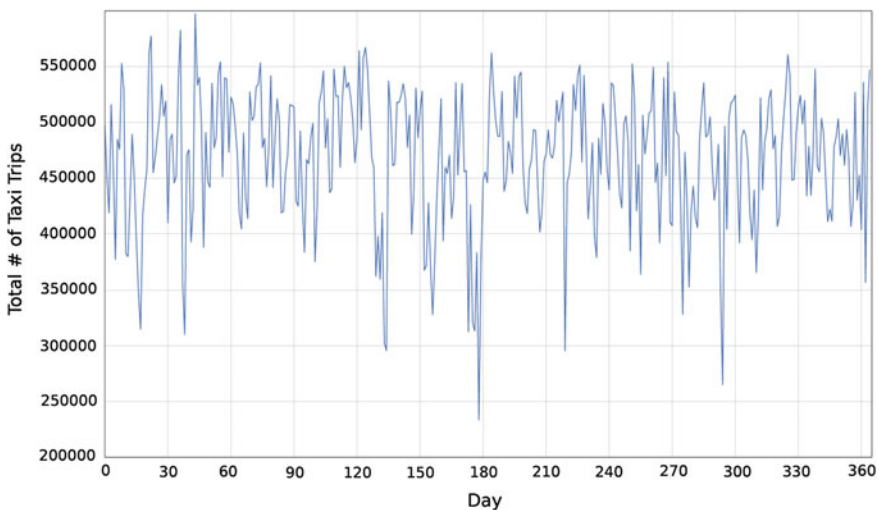
In large cities, public transportation is the direct carrier of urban life and taxicab is an indispensable component of it. As of 2007, 10% of total passenger volume are served by 18,000 taxicabs in Hong Kong [23]. By the end of 2012, 55,000 of taxis transport 1.5 million passengers daily [15]. Equipped with GPS devices, the taxi trip data enjoys the merit of sufficient temporal and spatial coverage due to the large passenger volume and 24-7 operation hours. Therefore, it is an advantageous data source for urban studies and has already received great attentions from researches. The pick-up and drop-off locations are processed with data mining and clustering algorithms to reveal urban activity patterns such as hotspots information for taxi drivers [5, 12, 24] and land use inference [14, 16]. However, urban dynamics are merely understood on the surface if pick-up/drop-off locations are only analyzed separately. Taxicab provides door-to-door service and is often used as a non-stop transportation tool. Therefore, the joint analysis of pick-up and drop-off location builds up a direct bridge between origin and destination and can aggregately reveal the underlying connections among detached urban places. Moreover, since each taxi trip is a form of human movement, the taxi trip data is also analyzed to disclose the uniformity of human mobility in large cities [11, 13, 17]. However, result discrepancies result are observed among very limited works. As a special case of human mobility, the displacement of taxi trips is restrained by the trip expenses and more importantly, the functionality structure of a city. Therefore the relationship between taxi trips and human mobility requires further investigation. In this paper, we make

a comprehensive use of the large scale taxi trip data and present the study on urban dynamics pattern in NYC from three aspects. First, the spatiotemporal pattern of urban activities is examined from trip dynamics by aggregating pick-up and drop-off locations. Secondly, we explore the inherent similarities among taxi trips and reveal the underlying connections among detached places using two-step clustering algorithms. In the end, we investigate the relationship between the taxi trips and uniformity of human mobility.

The rest of the paper is organized as follows. Section 2 gives an overview of the data and Sect. 3 analyzes the demand pattern of overall study area and several hot spots. Then the similarity among different trips is captured and the mobility pattern of taxicabs is presented. Conclusions and limitations are discussed in the final part.

## 2 Data

The taxi trip data used in this research is collected by New York City Taxi & Limousine Commission (NYCTLC) from December, 2008 to January, 2010. About 300,000 to 500,000 daily trips are recorded during the time and an overview of the annual trip distribution in 2009 is given in Fig. 1. A repeated and stable pattern is observed for weekly trips over the year and drastic drops are detected on holidays such as Thanks Giving and Christmas. Approximately 300,000 to 500,000 daily trips are recorded during the study period.



**Fig. 1** Annual daily distribution of taxi trips

**Table 1** Taxi data statistics

Date	Number of trips recorded	Number of trips after cleaning
10.5.2009	431,828	428,553
10.6.2009	467,649	464,273
10.7.2009	492,914	488,895
10.8.2009	517,079	512,781
10.9.2009	536,039	531,965
10.10.2009	532,179	528,032
10.11.2009	454,573	451,059

The dataset contains complete trip information, including the pick-up and drop-off timestamps and locations, the number of passengers onboard, the travel distance and the trip expense. Detailed trip trajectories are not available due to privacy concerns.

In addition to taxi data, census tract geography and land use information are also introduced in the analysis. The census tracts are extracted from the census tract area file provide in TRANSCAD. There are 2,211 census tracts within the study area, which cover Manhattan, Bronx, Queens, Brooklyn, Long Island, and a small portion of New Jersey. The land use map implemented in the study is obtained from New York City Department of City Planning (NYCDCP), which divides the city into three fundamental zoning districts: commercial (C), residential (R) and manufacturing (M). The three types are further categorized from low density to high density.

The taxi trip data from October 5th–11th are processed for further analysis, where no major social events were recorded during the period. The statistics of the one week data is presented in Table 1. Erroneous trip records are firstly removed, such as trips with zero travel distance or fare less than the initial price. Then all pick-up and drop-off locations are coupled with geography map to eliminate trips outside the study area. Finally, the remaining trips in the dataset are viewed as qualified and tagged with the overlaid census tract ID and land use type.

### 3 Trip Dynamics

#### 3.1 Overall Pattern

In this section, patterns of urban activity participation are examined from the arrival and departure dynamics of taxi trips. NYC is one of the busiest cities in the world. Around 5.7 million passengers moving around the city during the study period, generating more than 3.4 million taxi trips. The pick-up and drop-off location of all trips are aggregated at the census tract level based on the geographical coordinates and the overall geographical distributions of taxi trips are visualized in Fig. 2.

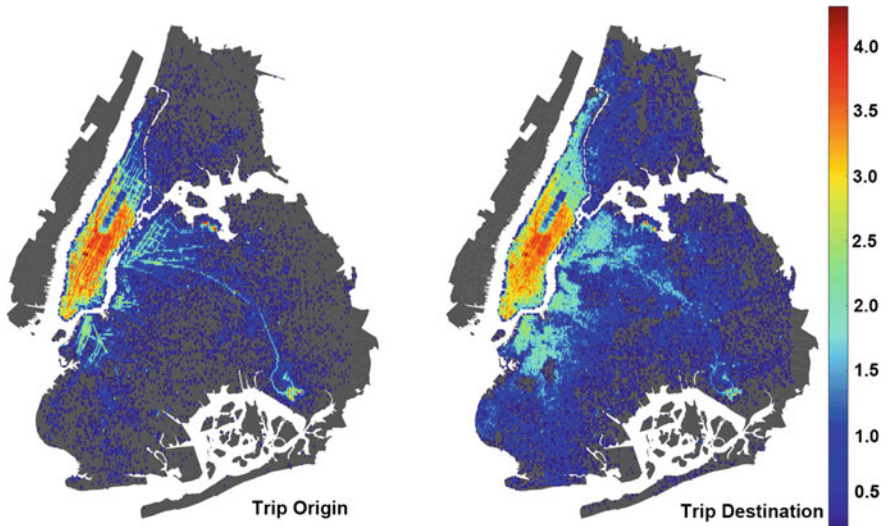


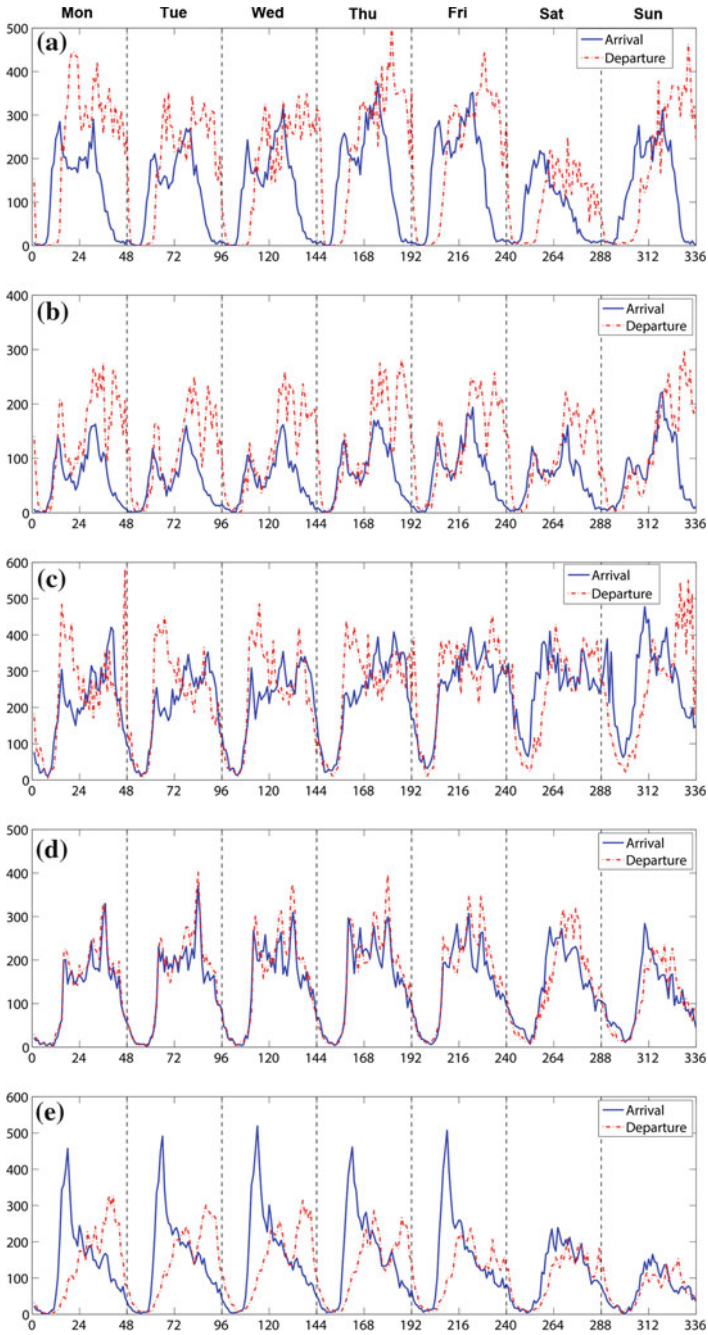
Fig. 2 Aggregated weekly density plot

The most appealing observation is that both trip origins and trip destinations exhibit highly centralized distribution towards Manhattan area. The result is not surprising since Manhattan serves as the business center of NYC. The number of trips decreases significantly with the increase of the distance to the city center, which reflects the typical sprawl of urban forms. While most places far from Manhattan have very low amount of trips, patterns at LaGuardia airport (LGA) and John F. Kennedy international airport (JFK) are entirely different. Approximately 90% of total trips are associated with Manhattan area. While majority of the trips congregate at midtown Manhattan and lower Manhattan, the upper Manhattan area is apparently less preferred by both passengers and drivers.

### 3.2 Hot Spots

Hot spots refer to the most frequent visited places in a city and usually have great activity intensity. The analysis of hotspots dynamics helps to understand the urban functionality in depth. By ranking total trip frequencies, most popular places are identified and five specific tracts are selected which cover the LGA, JFK, Penn Station, Central Park and the Fifth Avenue (the segment between 49th street and 56th street). Each individual hotspot has indispensable functionality including transportation terminals (with different purposes), recreational place and commercial area. To analyze the dynamics at hotspots, the temporal patterns across the week are plotted in Fig. 3. Penn Station and Madison Square Garden locate in the census tract where





**Fig. 3** Weekly trip pattern at hotspots x-axis is the time horizon and y-axis represents the number of trips **a** LaGuardia Airport, **b** JFK Airport, **c** Pen Station, **d** Central Park, **e** 5th Avenue

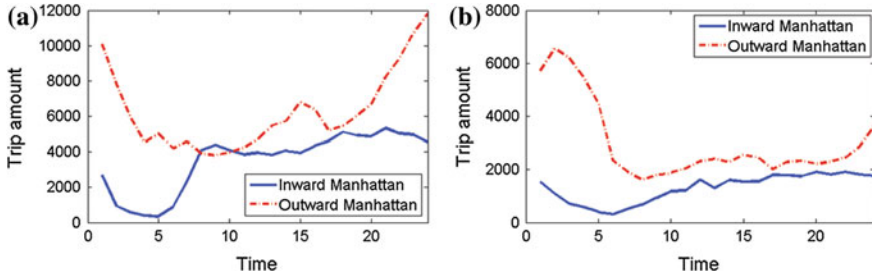
the greatest number of taxi trips are generated. Penn Station is not only the terminal for Amtrak trains, it also serves as the connection station for multiple subway lines. According to the morning arrival (trip origin) and evening departure (trip destination) peaks in weekdays, taxicab is very likely to function as the last and first mile transportation. Over the weekend, most arrivals and departures take place within daytime and at night. The pattern coincides with the functionality of Madison Square Garden, which is an entertainment place and is surrounded by many hotels.

Trip patterns at airports are distinct from that in central part. For both airports, while arrival curves are comparatively smooth, departure curves are observed to be noisy due to the periodical entry of flights. Besides, the intrinsic differences between the two airports are also disclosed from trip dynamics. Due to the effect of travel distance, the trip amount at JFK is significant lower than that at LGA. Secondly, since LGA are mainly used for domestic flights, the apparent morning peaks for flight arrivals during weekdays and the drop of trip amount on weekends. Moreover, as an airport mainly for international flights, JFK has more arrivals in the afternoon and the pattern is surprisingly consistent over the week. The result suggests that, during the week, the trip purpose is stable for international flights but varying significantly for domestic flights.

The Central Park is a recreational place. It occupies a larger area compared with other census tracts which contributes to its trip frequency. The comparison of trip dynamics between at the Central Park and at the Fifth Avenue perfectly interprets the functionality of corresponding land use attributes. The Fifth Avenue is a remarkable business street at midtown Manhattan and morning taxi arrival and evening taxi departure peaks are unsurprisingly retrieved. Reversely, due to the large portion of residential areas around the Central Park, most departures take place in the morning and majority of taxi arrivals are observed during evening rush hours.

### ***3.3 Unbalanced Taxi Trip***

Except for being able to capture activity dynamics at hotspots, the data also carries implicit yet significant insights such as the existence of unbalanced taxi trip. The number of taxi trips is closely associated with the level of economic development and the variation of urban functionality. Due to concerns such as trip margins and safety issues, taxi drivers usually have their preferred destinations, which eventually leads to the geographical discrimination. For example, taxi drivers may be unwilling to make trips to destinations where it is hardly possible to find potential passengers. The second type of unbalanced taxi trips is usually caused by sudden fluctuations in passenger demand. While the supply of taxis is fixed, the influx of commuters during peak hours makes it extremely hard to hail a vacant taxi.



**Fig. 4** Inward/outward-manhattan unbalanced trips **a** Weekday, **b** Weekend

From the overall spatial distribution, we observe a tremendous centrality of taxi trips at the developed Manhattan area. The great trip density suggests the easiness of finding passengers in Manhattan and stickiness of drivers to Manhattan area. As a result, we start looking into phenomenon and plot the temporal distributions for trips inwards and outwards Manhattan in Fig. 4. During daytime, the overall pattern for inward and outward trips turns out to be stable and balanced. However, when time goes to late night, we surprisingly witness an enormous gap: the highest amount of outward trips and the lowest amount of inward trips take place simultaneously. People may stay at Manhattan very late for entertainments and relaxations, while buses and metros having a reduced accessibility at the time. As taxi becomes very popular at a late time, drivers may refuse to leave Manhattan as they have to run the risk of returning empty. Hence, the unbalanced trip pattern implies the existence of geographical discrimination and a reduced level of service for taxi industry.

In order to reveal the unbalanced condition inbound Manhattan, we extract only weekday trips and spatial distributions of trip origins and destinations are presented in Fig. 5. Three typical time intervals are selected which cover off-peaks and morning and evening rush hours. Both morning peak and evening peak display eminent differences between trip origins and destinations and their patterns appear to be symmetric. Moreover, trips are found to be unbalanced with notable geographic characteristics. The northeastern part of midtown Manhattan is a large residential area and the midtown is mainly covered by commercial floors. As a result, most taxi trips inflow into midtown during morning peak and dissipate from the center area in the evening.

The existence of unbalanced taxi trips suggests an imminent need of designing policies to mitigate negative impacts. An additional fee can be charged or a subsidy can be assigned for trips outward Manhattan only after midnight as taxi drives are less likely to leave Manhattan at that time. Moreover, since morning and evening trips have distinct origins and destinations, the shuttle service following the direction of human migration should be to be effective. It can narrow the demand-supply gap of taxi service and reduce congestion at the same time.

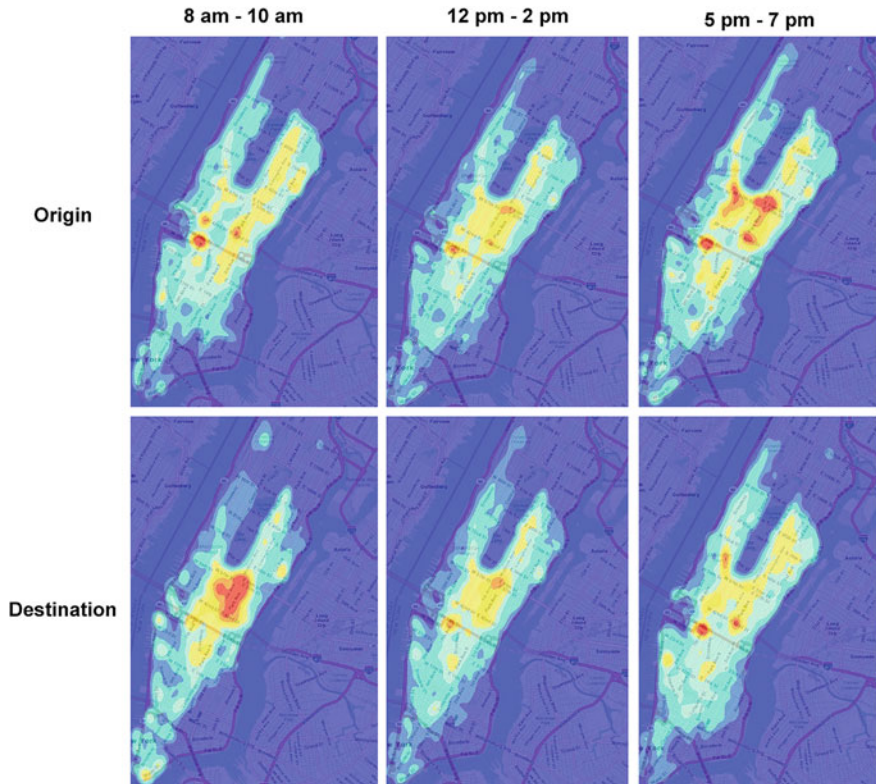


Fig. 5 Trip density plot inbound manhattan density increases from *blue* to *red*

## 4 Trip Classification

### 4.1 Clustering Algorithm

It is recognized that dynamics of trip origins and destinations are largely influenced by the geographical location, land use pattern and functionality of a particular place. Moreover, unlike other public transportation modes, the door-to-door service of taxicab builds up the straightforward connection between trip origin and destination. Therefore, how different urban areas are related can be understood by exploring the inherent similarities of taxi trips. Clustering algorithms are widely used to classify individual cases in large database into homogeneous groups. Considering spatial and temporal characteristics of taxi trips, each piece of taxi trip  $x_i$  can be represented as an eight dimensional tuple which takes the form:

$$x_i = (lat_i^o, long_i^o, lat_i^d, long_i^d, p_i^o, p_i^d, d_i, t_i) \quad (1)$$

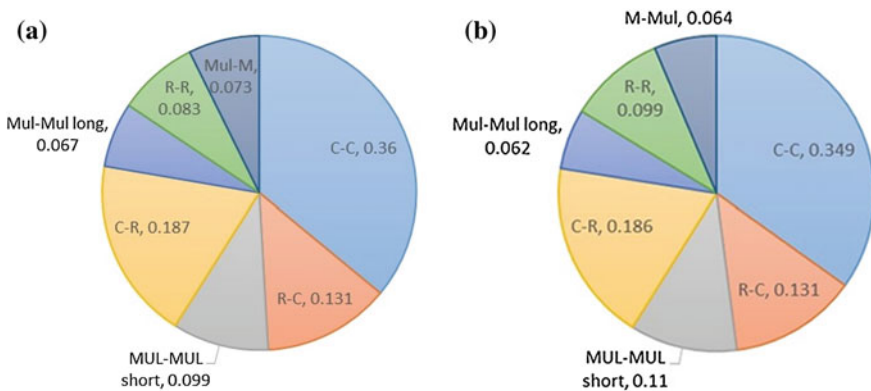
Where  $o, d$  represent the trip origin and destination respectively,  $lat$  and  $long$  are the latitude and longitude of trip locations,  $p$  refers to the land use attribute,  $d$  is the trip distance and  $t$  stands for the trip starting time. The clustering problem cannot be tackled by popular approaches such as k-means and DBSACN due to the presence of categorical variables (land use attribute).

Alternatively, the two-step clustering algorithm [6] is implemented to address the mixed variable clustering problem following two stages. The first stage is a pre-clustering approach which uses a sequential clustering method to generate initial sub-clusters. The second stage uses the agglomerative hierarchical approach which processes the sub-clusters from in the first stage recursively. The number of clusters is determined automatically by comparing BIC values. For interested readers, the detailed description for each step of the algorithm can be referred to SPSS manual [20].

### 4.2 Clustering Result

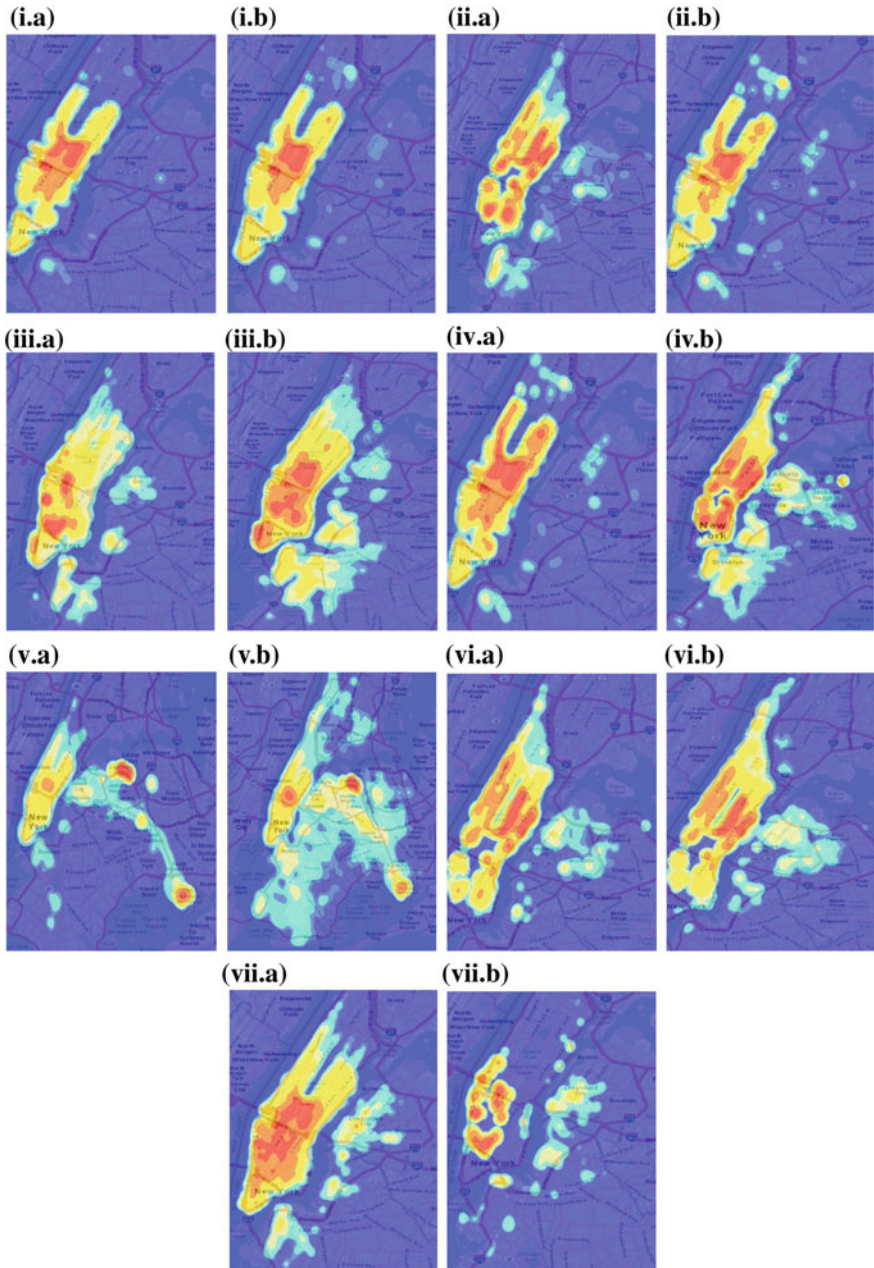
An overview of the clustering result is presented in Fig. 6. For both weekday and weekend taxi trips, the exactly same configuration with 7 distinct trip groups is obtained. Moreover, the percentage for the same cluster is pretty close. We name each cluster by its land use feature accordingly, including C-C, R-C, C-R, R-R, Mul (Mixed land use type)-Mul-S (short trip distance), Mul-Mul-L (long trip distance), and Mul-M trips. To better understand the characteristics of each cluster, the spatial distributions of trip origins and destinations in each cluster on weekdays are visualized in Fig. 7.

In general, C-C trips contribute to over one-third (36.0% for weekday and 34.9% for weekend) of the total taxi trips in NYC. Further, there are another 30% of trips that



**Fig. 6** Clustering result C-Commercial, R-Residential, M-Manufacturing, Mul-Mixture of the three Short/Long-Short/Long travel distance **a** Weekday, **b** Weekend





**Fig. 7** Spatial density plot of cluster origins and destinations **a** for trip origin and **b** for destination; i: commercial to commercial; ii: residential to commercial; iii: mixed to mixed with short travel distance; iv: commercial to residential; v: mixed to mixed with long travel distance; vi: residential to residential; vii: mixed to manufacturing; density increases from *blue* to *red*

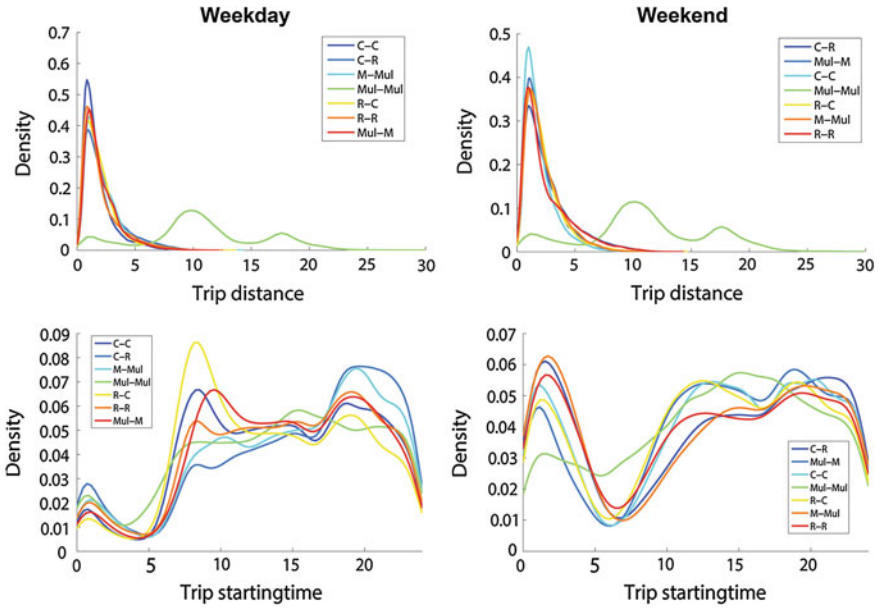


Fig. 8 Travel distance and trip starting time distribution for 7 clusters

are associated with commercial area (with either origin or destination in commercial area). This suggests the significant impact of land use pattern, especially commercial floors, on the amount of taxi trips. More specifically, commercial areas where trip originated from and arrived at cover the entire midtown and lower Manhattan. As a result, it is believed that most activities and functionalities of the city are concentrated in these places. Viewing the distribution of residential related trips, one can tell that there are considerable amount of people living on the peripheral area and they are connected to the city center by taxicab.

We also plot distributions of travel distance and trip starting time as important attributes for each cluster in Fig. 8. Apparently, the distance distribution suggests that taxi trips are heavily used for short-range travel, especially for trips less than 5 miles. Such pattern is mainly determined by the urban structure of NYC, as majority activities and functional places are agglomerated in a small area. While the distance distribution is stable over the week, there are prominent discrepancies observed for trip starting time between weekday and weekend. Firstly, all clusters except C-R and Mul-Mul-L trips have morning and evening peaks, reflecting that taxicabs are heavily used for work commuting in urban areas. Secondly, the temporal pattern of most urban activity is shifted from daytime to late night, as the trip intensity remains at a high level until 3 am.

Though taxi trips are mostly commercial and residential related, we observe that the Mul-Mul-L group is a very special type of taxi trips with unique characteristics. Based on the trip location distribution, these trips connect midtown Manhattan, LGA

and JFK to the rest of NYC. While all other clusters have very short trip distance, the mean travel distance of the group is approximately 11 miles. Two peaks are revealed from the distance distribution, which locate at 10 miles and 17 miles. The two points are matched with the travel distance from Manhattan to LGA and JFK respectively. As a result, the exclusive pattern is largely caused by the urban forms, as airports are usually far from the city center but with very high passenger volumes. The group of trip should be treated separately during urban studies as it is heavily biased from the general mobility pattern of taxi trips.

## 5 Taxi Mobility

Individual mobility pattern have been realized barely random. Several studies using data from the movement of an online game [22], the dispersal of bank notes [3], as well as trajectories from cellular data [7] have found highly regulated pattern in human movement. And the human movement is observed to follow a heavy-tailed plot under logarithmic scale and can be well approximated by scaling law. With human beings as the main participants, the taxi trips are results of human movement in an urban context as well. Hence, we try to reveal the taxi mobility and examine the relationship with individual mobility.

To uncover the taxi mobility, we first plot the distribution of travel distance under logarithmic scale in Fig. 9a. From the observation, the distribution of travel distance can be divided into parts: an ascending ranges from 0 to 0.8 mile, and then gradually descending as trip distance increases. Two minor peaks around 10 miles and 20 miles in the distribution are mainly caused by trips to LGA and JFK airport. The interference of airport trips has been discussed in previous section. We remove the trips to and from the two airports as they have specific purposes and unique characteristics. A refined distribution is generated in Fig. 9b.

Trips with distance less than 0.8 mile take 16.89% of total trips. As very short trips within walking radius, these trips differ from the general pattern of taxi mobility on a decision making process of whether to take taxis. The first part of the trips can be approximated with distribution:

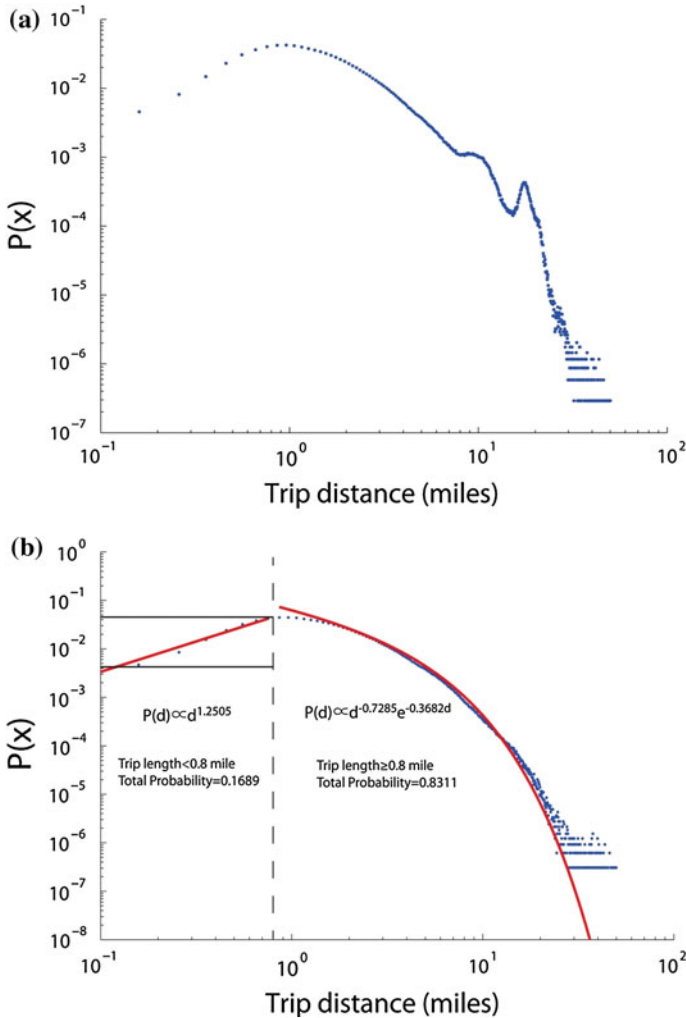
$$P(d) \propto d^\beta \quad (2)$$

Where exponent  $\beta = 1.2505$ .

The distribution resembles a power-law like distribution (straight line under logarithmic scale), however, the exponent takes a positive value. As mentioned earlier this phenomenon captures model choice process in whether take a taxi. And it is intuitive that with the increase in distance, the probability of taking a taxi also increases until attaining its maximum around 0.8 miles.

The refined second part is used to capture urban mobility features of taxi trips. The trips greater than 0.8 mile contribute 83.11% of total trips. It is found that the





**Fig. 9** Taxi trip distance distribution. **a** Distance distribution of all trips. **b** Distance of all but airport trips

distribution of taxi trip distance is well approximated by a power-law with exponential cut-off (also known as truncated power-law):

$$P(d) \propto d^{-\alpha} e^{-\lambda d} \tag{3}$$

With exponent  $\alpha = 0.7285$  and  $\lambda = 0.3682$ . The distribution is found to be heavy-tailed. Unlike the power-law distribution of human movement reported (8, 18, 19), the taxi trip distance distribution has a faster probability decay in the tail part (the effect of the exponential cut-off term). This indicates that the unique effects of urban

environment on the distribution of taxi trip distance. Since the underlying size of urban area limits the distance of taxi trip, very long trips (e.g. >30 miles) are less likely to happen, and the scale-free property of a typical power-law distribution fails. It is notable that as taxi trips are important component of urban human movement, the trip distance distribution reflects a unique perspective of human mobility. That is, the taxi mobility pattern reveals the hidden role of urban geographical boundaries in limiting urban human movement.

## 6 Conclusion and Future Work

In this paper, we exploit New York taxi trip data and comprehensively explore underlying patterns of urban taxi trips. We first look at the general level of demand and find out the spatial and temporal patterns for the most popular places. A potential unbalanced trip pattern is further discussed. Next, we use the two-step clustering algorithm to figure out the intrinsic taxi trip classes. Differences are discussed based on land use, travel distance and starting time distributions. In the end, taxi trip mobility is analyzed from the overall travel distance distribution.

Taxi data has been proved to be an efficient tool to understand urban dynamics and several interesting insights are raised in our paper. Unbalanced trips are common in taxi industry and should be carefully investigated to improve the level of service. Airport trips is a special part of taxi trips and differ from regular taxi trip patterns. Land use has significant impact on taxi trip types, and different types of taxi trips are able to uncover the structure of a city. Moreover, we find that the mobility of taxi trips are restricted by the urban geographical boundaries.

However, the paper also has several limitations. The current paper is primarily focused on exploring patterns. The following study will build a model from the patterns discovered to account for human movement within urban context. Moreover, more information such as social economics can be combined into the data analysis to provide more insights. Furthermore, it would be interesting to develop a methodology to infer urban land use type from taxi patterns. Also, attentions can be paid on extracting travel information from taxi dynamics and provide feedbacks to users.

## References

1. Batty M, Xie Y (1994) From cells to cities. *Environ Plan B* 21:31
2. Batty M, Xie Y, Sun Z et al (1999) Modeling urban dynamics through GIS-based cellular automata. *Comput Environ Urban Syst* 23:205–233
3. Brockmann D, Hufnagel L, Geisel T et al (2006) The scaling laws of human travel. *Nature* 439:462–465
4. Calabrese F, Diao M, Di Lorenzo G et al (2013) Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. *Transp Res Part C Emerg Technol* 26:301–313

5. Chang H, Tai Y, Chen H et al (2008) iTaxi: context-aware taxi demand hotspots prediction using ontology and data mining approaches. In: Proceedings of 13th conference artificial intelligence and applications (TAAI 2008)
6. Chiu T, Fang D, Chen J et al (2001) A robust and scalable clustering algorithm for mixed type attributes in large database environment. In: Proceedings of seventh ACM SIGKDD international conference on knowledge discovery and data mining—(KDD 01) 263–268
7. Gonzalez MC, Hidalgo CA, Barabasi A-L et al (2008) Understanding individual human mobility patterns. *Nature* 453:779–782
8. Giuliano G (2004) Land use impacts of transportation investments: highway and transit. In: Hanson S and Giuliano G (eds) *Geography of urban transportation*, 3rd edn. Guilford Press, New York
9. Harris B (1985) Urban simulation models in regional science. *J Reg Sci* 25:545–567
10. Hasan S, Zhan X, Ukkusuri SV et al (2013) Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In: Proceedings of 2nd ACM SIGKDD international work. *Urban computing*, p 6
11. Jiang B, Yin J, Zhao S et al (2009) Characterizing the human mobility pattern in a large street network. *Phys Rev E* 80:21136
12. Li B, Zhang D, Sun L et al (2011) Hunting or waiting? discovering passenger-finding strategies from a large-scale real-world taxi dataset. In: *IEEE international conference on pervasive computing and communications workshops (PERCOM Workshops)*, pp 63–68
13. Liang X, Zheng X, Lv W et al (2012) The scaling of human mobility by taxis is exponential. *Phys A Stat Mech Its Appl* 391:2135–2144
14. Liu Y, Wang F, Xiao Y et al (2012) Urban land uses and traffic 'source-sink areas': evidence from GPS-enabled taxi data in Shanghai. *Landsc Urban Plan* 106:73–87
15. NYCTLC(2012) New York city taxi and limousine commission (2012) Annual Report
16. Pan G, Qi G, Wu Z et al (2013) Land-use classification using taxi GPS traces. *IEEE Trans Intell Transp Syst* 14:113–123
17. Peng C, Jin X, Wong K-C et al (2012) Collective human mobility pattern from taxi trips in urban area. *PLoS One* 7:e34487
18. Ratti C, Pulselli RM, Williams S et al (2006) Mobile landscapes: using location data from cell phones for urban analysis. *Environ Plan B Plan Des* 33:727–748
19. Reades J, Calabrese F (2007) Cellular census: explorations in urban data collection. *Pervasive Comput IEEE* 6:30–38
20. SPSS INC (2001) The SPSS twoStep cluster component: a scalable component to segment your customers more effectively
21. Sun L, Chen C, Zhang D et al (2013) Understanding urban dynamics from Taxi GPS traces. *Creat Pers Soc Urban Aware Through Pervasive Comput*, p 299
22. Szell M, Sinatra R, Petri G et al (2012) Understanding mobility in a social petri dish. *Sci Rep* 2:457
23. Yang H, Fung CS, Wong KI et al (2010) Nonlinear pricing of taxi services. *Transp Res Part A Policy Pract* 44:337–348
24. Yuan J, Zheng Y, Zhang L et al (2011) Where to find my next passenger. In: Proceedings of 13th international conference on ubiquitous computing, pp 109–118

# Holistic Calibration of Microscopic Traffic Flow Models: Methodology and Real World Application Studies

Alexander Paz, Victor Molano and Javier Sanchez-Medina

**Abstract** This study proposes and applies a methodology to calibrate microscopic traffic flow simulation models. The proposed methodology has the capability to calibrate simultaneously all the calibration parameters as well as demand patterns for any type of network. Parameters considered include global and local as well as driver behaviour and vehicle performance parameters. Demand patterns, in terms of turning volumes, are included in the calibration framework. Multiple performance measures involving link counts and speeds are used to formulate and solve the proposed calibration problem. In addition, multiple time periods were considered. A Simultaneous Perturbation Stochastic Approximation (SPSA) algorithm is used to search for the vector of the model's parameters that minimizes the difference between actual and simulated network states. (Punzo V, Ciuffo B, Montanino M *Transp Res Rec J Transp Res Board* 2315(1):11–24 2012, Punzo et al. [1]) commented on the uncertainties present in many calibration methodologies. The motivation to consider simultaneously all model parameters is to reduce that uncertainties to a minimum, by leaving to the experience of the engineers as little parameter tuning as possible. The effects of changing the values of the parameters are taken into consideration to adjust them slightly and simultaneously. This results in a small number of evaluations of the objective function. Three networks were calibrated with excellent results. The first network was an arterial network with link counts and speeds used as performance measurements for calibration. The second network included a combination of freeway ramps and arterials, with link counts used as performance measurements.

---

A. Paz · V. Molano  
Department of Civil and Environmental Engineering, University of Nevada,  
Las Vegas, USA  
e-mail: apaz@unlv.edu

V. Molano  
e-mail: victor.hugo.molano@gmail.com

J. Sanchez-Medina (✉)  
CICEI, Department of Computer Science, University of Las Palmas de  
Gran Canaria (ULPGC), Las Palmas, Spain  
e-mail: javier.sanchez@ulpgc.es; javier.sanchez.medina@ieeee.org

The third network was an arterial network, with time-dependent link counts and speed used as performance measurements. The experimental results illustrate the effectiveness and validity of this proposed methodology. The same set of calibration parameters was used in all experiments.

## 1 Introduction

Micro-simulation models provide tremendous capabilities to model, at a high level of resolution, complex systems in a broad range of fields, including economy, sociology, physics, chemistry, and engineering [2].

In the context of vehicular traffic systems, microscopic traffic flow models enable the modelling of many aspects of the actual system, including the manoeuvres of individual vehicles and their interactions, the various types and characteristics of facilities, and the vast number of control settings. These capabilities are associated with a large number of modelling parameters that typically need to be tailored for each vehicular system. For example, driver behaviour includes parameters associated with car following, lane-changing manoeuvres, and gap acceptance.

In, Punzo et al. reflect on the uncertainties present in many of the current car-following based traffic flow simulation calibration methodologies. It is a fact that the accuracy of a model and the validity of its results are highly dependent on the correctness of the chosen parameters [3–9].

Punzo et al. [1] discussed uncertainties present in many of the existing methodologies for the calibration of car-following-based traffic flow simulation models. It is clear that the accuracy of a model and the validity of its results are highly dependent on the correctness of the chosen parameters [3–9].

Hence, it is important to consider all these model parameters simultaneously with the aim to capture their intricate interactions, thereby seeking convergence and stability of the solutions.

In [10] we drafted a method for the simultaneous calibration of all of the parameters of a CORSIM model. In the present work we have sharpen, extended and applied that methodology to three different big test cases with excellent results: (i) Pyramid Highway, in Reno, Nevada, USA; (ii) Interstate-75 in Miami, Florida, USA; and (iii) a Network of McTrans Sample Data Sets.

This study proposes a methodology to calibrate simultaneously all model parameters and demand patterns based on link counts and speeds. In addition, multiple performance measures were used, demand patterns were not pre-calibrated, and multiple time periods were explicitly considered with target performance values for each period. That is, the proposed methodology implements a Simultaneous Perturbation Stochastic Approximation (SPSA) algorithm to determine an adequate set for all model parameters and turning volumes for multiple time periods using multiple performance measures. Even though there is a significant body of literature around the proposed problem context, to the best of the authors knowledge, no study has considered simultaneously all the aspects listed in this paragraph and included in our

implementation and experimental framework. The state-of-the-art is summarized in the following subsection.

The SPSA was chosen based on its computationally efficiency and ability to handle large numbers of parameters [11–18]. Only two traffic flow simulation evaluations per iteration of the SPSA are required to update all model parameters. Running a low number of traffic flow simulations represents important savings in terms of time and other resources. However, the SPSA algorithm performs better when the initial model parameters relatively close to the optimal solution.

Comparative studies between SPSA and other algorithms could be found in the literature [11, 12, 18]. In addition, the SPSA algorithm has been used to calibrate and optimize various transportation applications [13, 19, 20].

The rest of this paper is organized as follows: We do a brief literature review in the next subsection. We expose the proposed methodology in Sect. 2. Then we share the experiments performed alongside with the corresponding results in Sect. 3. Finally we put together some concluding remarks in section .

## 1.1 State of the Art

A broad number of optimization algorithms, ranging from genetic algorithms to finite difference stochastic approximation, have been used to determine an adequate set of model parameters for a particular traffic system [3, 4, 6, 21, 22].

For example, the sequential simplex algorithm was used to calibrate parameters for car-following, acceleration/deceleration, and lane-changing behaviour [6]. However, only a subset of parameters was considered, maybe because of the lack of enough computing power in 2002. Moreover, parameters associated with infrastructure and vehicle performance were not considered. The algorithm provided adequate results under congested conditions. However, under low-congestion conditions, manual calibration provided better results [6].

In [23] they calibrate the VISSIM model of the NGSIM corridor, using a quite limited optimization technique, exploring only the limits. They calculate a number of restrictions for some parameters and accept values only if they satisfy all the restrictions. Additionally, they are only tuned to a specific period of the day.

In a recent study, [24], Markov Chain Monte Carlo (MCMC) method using Bayesian estimation theory. Only five parameters of a linear car following model [25] are calibrated.

Genetic Algorithms (GA) has been extensively used to calibrate traffic simulation parameters. In [26] the use a simple GA to calibrate the parameters of a CORSIM [27] based simulation of a 5.8 km expressway in Singapore. In [28], a freeway segment in California was used as a test example to attempt the optimization of two PARAMICS calibration parameters.

In both cases, the results proved limited success reducing discrepancies between real word and simulations.

Genetic Algorithms were used for the calibration of global and local capacity and occupancy parameters [20, 29]. A sequential approach was used to update global and local parameters.

In [30] a Genetic Algorithm was used to calibrate a small subset of all the PARAMICS [31] parameters.

In [32] a Multiobjective version of the Non-dominated Sorting Genetic Algorithm (NSGA-II, [33]) was applied to solve the multi-objective optimization task of parameter calibration. Results are modest and they were optimizing or calibrating a very few, only five of VISSIM's [34].

In [35] five PARAMICS [31] parameters were optimized for a larger model of down town Toronto, Canada. They tested three different GA approaches but they finally did not obtained significant improvements in the accuracy of the model.

In [8] they tuned 11 CORSIM [36] parameters of a 22.4 km segment of Interstate 10 in Houston, Texas. The authors used a Genetic Algorithm to perform an automated calibration of these parameters. Their results were remarkable, including a sensitivity analysis. As happens for every GA approach to traffic simulation calibration, there were a few set-up parameters in the Genetic Algorithm that must be carefully selected, because the quality of results is very dependent on them. There was no computing performance information provided for such work, which should be a very interesting element for comparison with SPSA-based approaches, likely to be faster, more suited to real world on-line applications.

In [37] yet another GA based PARAMICS parameter calibration was proposed. The authors only calibrate 5 parameters that needed to be initialized at "default values". In addition, there were eight additional configuration parameters that need to be tuned for the Genetic Algorithm to obtain better performance. This parameter adjustment required significant trial-and-error and experience by the researcher.

Regarding specifically SPSA algorithms we have selected a few interesting and related studies. In [13] Lee used SPSA algorithms to calibrate model parameters using distributions to generate input for various stages. The calibration capabilities of GA and SPSA algorithms were shown to be similar in [20]; however, SPSA algorithms were less computationally expensive.

In [38], the authors proposed a SPSA algorithm for the calibration of a simulation model of the Massachusetts Bay Transportation Authority (MBTA) Red Line. The authors used a generic simulator, SimMETRO. The effort involved a multiple objective function and simultaneous parameter calibration. It is important to notice that the simulation of one Metro line involves less parameters compared to a vehicular traffic system. This makes the problem more computationally affordable and less complex.

In they proposed a rail simulation SPSA based parameters calibration for the test case of the Massachusetts Bay Transportation Authority (MBTA) Red Line, using a generic simulator they called SimMETRO. Even when it is not exactly the same problem to solve than in our case, this is a remarkable application of multiple objective simultaneous parameter calibration. It is also true, though, that a one Metro line simulation has not as many calibration parameters as a vehicular simulation like

CORSIM may include, making the problem more computationally affordable and also less complex.

Another very interesting application of a SPSA algorithm to Intelligent Transportation Systems was published in [39]. A dynamical emission model was optimized to estimate aggregate emission patterns for traffic fleets so as to predict local air conditions.

SPSA and Finite Difference Stochastic Approximation algorithms have been proposed for the calibration of time depending Origin-Destination matrices. For example, in [11] driver behaviour parameters were pre-calibrated considering various time intervals. Other important performance measures, such as speed, were not considered.

In [40], a SPSA algorithm is used for the simultaneous adjustment of a dynamic traffic O-D matrix using traffic counts and speeds. However, the author states that some parameters must be tuned by hand to get close to the desired solutions. Hence, the proposed approach is infeasible for a large amount of calibration parameters as it requires significant user involvement and experience.

$$Min. NRMS = \frac{1}{\sqrt{N}} \times \sum_{t=1}^T \left( W \times \sqrt{\sum_{i=1}^N \left( \frac{V_i - \hat{V}(\theta)_i}{V_i} \right)^2} + (1 - W) \times \sqrt{\sum_{i=1}^N \left( \frac{S_i - \hat{S}(\theta)_i}{S_i} \right)^2} \right)$$

Subject to:

$$Lower\ bound \leq \theta \leq Upper\ bound$$

(1)

Ben-Akiva et al. worked on the calibration of a dynamic traffic O-D matrix [41] for a large network in Beijing. The SPSA algorithm was used given its capability to address noise. The significant work conducted using the SPSA algorithm to perform related research motivated its use in the proposed study.

are

## 2 Methodology

### 2.1 Formulation of the Calibration Problem

The calibration problem for all model parameters,  $\theta$ , is formulated using a mathematical programming approach. The analysis period is divided into a number T of discrete time periods. The objective function, normalized root mean square (NRMS), as denoted by Eq. 1, is the sum over all calibration time-periods of the average of the sum over all links I of the root square of the square of the normalized differences between actual and simulated link counts and speeds. The normalization enables the consideration of multiple performance measures, in this case, link counts and speeds. In our experimental set-up, the initial parameters for a model are selected as



the default values used in CORSIM models. The calibration problem is formulated as shown in Eq. 1, where:

- $V_i$  = actual link counts for link i
- $\tilde{V}(\theta)_i$  = simulated link counts for link i
- $S_i$  = actual speeds for link i
- $\tilde{S}(\theta)_i$  = simulated speeds for link i
- N = total number of links in the model
- T = total number of time periods t
- W = weight used to assign more or less value to counts or speeds

$$g_k \theta_k = \frac{y(\theta_k + c_k \Delta_k) - y(\theta_k - c_k \Delta_k)}{2c_k} [\Delta_{k1}^{-1}, \Delta_{k2}^{-1}, \Delta_{k3}^{-1}, \dots, \Delta_{kp}^{-1}]^T \quad (2)$$

## 2.2 Calibration Criteria

The calibration criteria for this study were based on guidelines from the Federal Highway Administration. The difference between actual and simulated link counts should be less than 5% for all links; and, the GEH statistic, in Eq. 3, should be less than 5 for at least 85% of the links [27].

$$GEH = \sqrt{\frac{2(V_i - \tilde{V}(\theta)_i)^2}{V_i + \tilde{V}(\theta)_i}} \quad (3)$$

$V_i$  = actual link counts at the link i.  
 $\tilde{V}(\theta)_i$  = simulated link counts at the link i.

## 2.3 Simultaneous Perturbation Stochastic Approximation Algorithm

The SPSA algorithm is an iterative approach that uses gradient estimations of the objective function to determine an optimal solution. Details of its implementation are provided by Spall [15–18]. In each iteration of SPSA, the vector of model parameters is updated using Eq. 4; where:

$$\theta_{k+1} = \theta_k - a_k g_k \theta_k \quad (4)$$

- $\theta_{k+1}$  = vector of updated parameters at iteration k + 1
- $\theta_k$  = vector of initial parameters at iteration k + 1
- $a_k$  = gain coefficient at iteration k + 1 calculated using Eq. 5
- $g_k \theta_k$  = estimated gradient at iteration k + 1.

$$a_k = \frac{a}{(k + 1 + A)^\alpha} \quad (5)$$

where  $a$ ,  $A$ , and  $\alpha$  are empirical non-negative coefficients. These coefficients affect the convergence of the SPSA algorithm. The simultaneous perturbation and gradient estimate are represented by  $g_k \theta_k$ , and is calculated using Eq. 2.

Here,  $c_k$  is calculated using Eq. 6 where  $c$  and  $\gamma$  are empirical non negative coefficients.

$$c_k = \frac{c}{(k + 1)^\gamma} \quad (6)$$

where,  $c = 2.7598$  and  $\gamma = 0.1666$ .

The elements in the random perturbation vector are Bernoulli-distributed, with a probability of one-half for each of the two possible outcomes (Eq. 7).

$$\Delta k = [\Delta_{k1}^{-1}, \Delta_{k2}^{-1}, \Delta_{k3}^{-1}, \dots, \Delta_{kp}^{-1}]^T \quad (7)$$

The SPSA algorithm is implemented using the following steps [18]:

- Step 1: Set counter  $k$  equal to zero. Initialization of coefficients for the gain function  $a$ ,  $A$ , and  $\alpha$  and calibration parameters  $\theta_0$ .
- Step 2: Generation of the random perturbation vector  $\Delta_k$ .
- Step 3: Evaluation of the objective function plus and minus the perturbation.
- Step 4: Evaluation of the gradient approximation  $g_k \theta_k$ .
- Step 5: Update the vector of calibration parameters using Eq. 4 along with the corresponding constraints denoted by Eq. 3.
- Step 6: Check for stopping criteria. If criteria is achieved, stop; otherwise, set counter  $k = k + 1$  and repeat Steps 1–6.
- Convergence is achieved when all the criteria in Table 1 is satisfied or the maximum number of iterations is reached.

## 2.4 Stopping Criteria

Stopping criteria is reached when the inequality in Eq. (4) is satisfied or a user pre-specified maximum number of iterations is reached. At convergence, the calibration criteria are expected to be satisfied or a significantly better model is obtained.

$$\frac{\sum_{k=n+1}^k \sqrt{(NRMS_{AV} - NRMS_k)^2}}{n} < \rho \quad (8)$$

where,

- $NRMS_{AV}$  = average NRMS of the last  $n$  iterations
- $NRMS_k$  = NRMS at  $k$  iteration
- $k$  = iteration counter
- $n$  = pre-specified integer = 10, and
- $\rho$  = pre-specified convergence condition = 0.015.

## 3 Experiments and Results

### 3.1 *Micro-simulation Model*

The proposed methodology was tested using CORSIM, a tool that integrates two different models to represent a complete traffic system, FRESIM for freeways and NETSIM for surface streets [36, 42]. The Traffic Analysis Toolbox Volume IV: Guidelines for Applying CORSIM Micro-simulation Modelling Software [5] describes a procedure for the calibration of micro-simulation traffic flow models, with a focus on CORSIM. The suggested procedure in these guidelines uses three sequential and iterative steps, including the calibration of (i) capacity at key bottlenecks, (ii) traffic volumes, and (iii) system performance. However, the guidelines do not suggest any particular methodology to perform the calibration in an efficient and effective manner. For example, issues associated with convergence and stability of the solutions are not discussed. Nevertheless, alternative studies have proposed and developed practical procedures to accelerate the calibration process, which typically is time consuming [43]. However, stability and convergence still are issues.

### 3.2 *Calibration Parameters for CORSIM Models*

The calibration of CORSIM models can involve Driver Behaviour and Vehicle Performance parameters [36, 42]. These parameters can be defined exclusively for surface streets or freeways or both models simultaneously. In addition, the resolution of these parameters can be global or link-based defined. This study considered all types of parameters and levels of resolution. In addition, parameters related to demand patterns were included. Table 1 shows all the different parameters used for the calibration of CORSIM models. Several studies have conducted sensitivity analysis for the calibration of CORSIM models [8]. These studies have showed that the maximum non-emergency deceleration rate, for example, does not affect the outcomes of a specific FRESIM model. However, the specific vehicle distributions improve the accuracy of the model [8]. Driver behaviour parameters were found to affect the time to breakdown and the flow on ramps. Flow related parameters showed low effects.

**Table 1** Calibration parameters for NETSIM and FRESIM models

NETSIM model surface streets		
Driver behaviour	Vehicle performance	Demand patterns
<ul style="list-style-type: none"> <li>• Queue discharge headway</li> <li>• Start-up lost time</li> <li>• Distribution of free-flow speed by driver type</li> <li>• Mean duration of parking manoeuvres</li> <li>• Lane change parameters</li> <li>• Maximum left and right turning speeds</li> <li>• Probability of joining spillback</li> <li>• Probability of left turn jumpers and laggings</li> <li>• Gap acceptance at stop signs</li> <li>• Gap acceptance for left and right turns</li> <li>• Pedestrian delays</li> <li>• Driver familiarity with their path</li> </ul>	<ul style="list-style-type: none"> <li>• Speed and acceleration characteristics</li> <li>• Fleet distribution and passenger occupancy</li> </ul>	<ul style="list-style-type: none"> <li>• Surface street turn movements</li> </ul>
FRESIM model-freeways		
<ul style="list-style-type: none"> <li>• Mean start-up delay at ramp meters</li> <li>• Distribution of free flow speed by driver type</li> <li>• Incident rubbernecking factor</li> <li>• Car-following sensitivity factor</li> <li>• Lane change gap acceptance parameters</li> <li>• Parameters that affect the number of discretionary lane changes</li> </ul>	<ul style="list-style-type: none"> <li>• Speed and acceleration characteristics</li> <li>• Fleet distribution and passenger occupancy</li> <li>• Maximum deceleration</li> </ul>	<ul style="list-style-type: none"> <li>• Freeway turn movements</li> </ul>

The calibration parameters have different effects for specific networks and conditions. The interaction between these parameters is very complex and might vary from model to model. As a starting point, the proposed methodology uses a set of default CORSIM values for the parameters listed in Table 1. This decreases the effort during the selection of the calibration parameters and set-up. During calibration, the value of the selected parameters is adjusted while constraining their boundaries in order to avoid unrealistic values.

### 3.3 Experimental Set-Up and Results

Three experiments were designed to test the capabilities of the proposed methodology to calibrate simultaneously, using vehicle counts and speeds. A software tool

was developed to implement the proposed calibration methodology. The tool was developed using a basic layered architecture where each layer handles a group of related functions. A Graphical User Interface (GUI) provides access to the entire software capabilities. The entire software was developed in Java; it includes more than 5,000 lines of code.

### System Specifications

- Operative System: Windows Server, Standard Edition, 2007, Service Pack 2 64Bit
- System: Intel Xeon CPU E7450 2.4 GHz (4 processors)
- Ram memory: 32 GB

### First Experiment: Pyramid Highway in Reno, Nevada, USA

In this experiment a CORSIM model for a portion of the Pyramid Highway in Reno, Nevada, was calibrated. This portion of highway is located between Milepost 1.673 and 5.131. This calibration focused on speeds and link counts for the entire simulation. The weight factor in the objective function was set to 0.7. This value is constant for the first two experiments because link counts were obtained using more accurate data collection methods compared to speeds. The model included 126 arterial links, and no freeways were included. Link counts and speeds were only available for 45 of these links. Coefficients for the SPSA algorithm were selected using guidelines from the literature [18]). These values affected the convergence of the algorithm. The time required for calibration was 25.5 min.

Figure 1a shows a Google map of the Pyramid Highway. Figure 1b illustrates the corresponding CORSIM model. Figure 2 illustrates how the objective function was minimized. The noisy trajectory was a consequence of the stochastic perturbation applied to all calibration parameters to obtain the gradient approximation at each iteration. The characteristics of the traffic model made the function noisier due to rounding. The NRSM was 0.042 before calibration and 0.010 after calibration. The calibration process stopped around the 80th iteration, when a stable region was found.

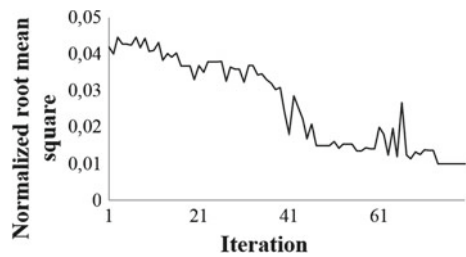
Figure 3a shows the actual and simulated counts and speeds before calibration. These values present poor initial conditions, especially for the volumes over 1500 vehicles per hour (vph). Figure 3b shows the actual and simulated counts and speeds after calibration. The proposed methodology is able to reduce the gap between actual and simulated counts. The results illustrate larger improvements for the large counts. Figure 3a clearly shows that links with counts over 1500 vph were improved, while the values with good initial conditions were slightly modified.

As illustrated in Fig. 3a, simulated speeds are far from actual speeds. The simulation model underestimates many speed values. After calibration (Fig. 3a), the speeds were improved for 23 of the links. The rest of the speeds were kept close to the initial values with a variation less than 1 mile per hour (mph). This can be associated to the relative large value of the weight assigned to the counts in the objective function ( $W = 0.7$ ). In addition, the experimental results show that link counts are more sensitive than speeds to changes in the calibration parameters. The GEH statistics for the models before and after calibration are shown in Table 2. This statistic is included in our analysis because it is recommended by the Traffic Analysis Tool-box [5]. It is



**Fig. 1** Pyramid highway, Reno, Nevada, USA (a) google map and CORSIM model (b) for the first experiment

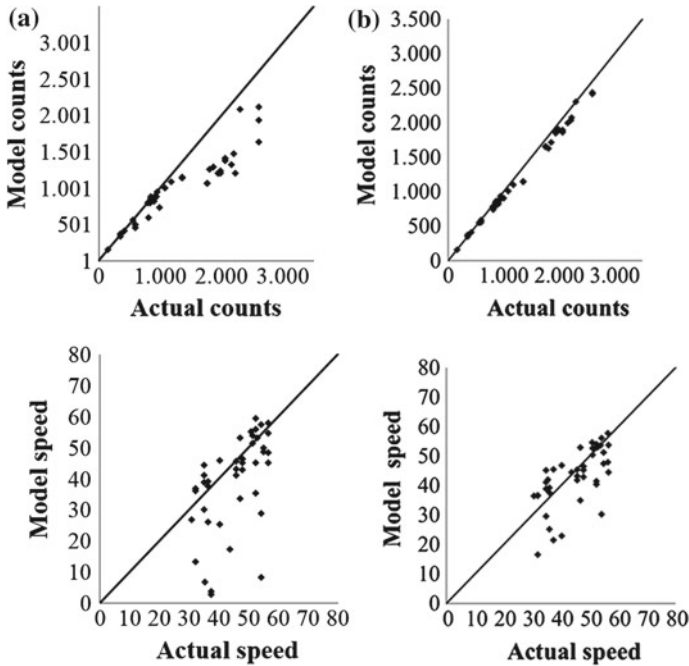
**Fig. 2** Objective function for the first experiment



clear that the calibration model significantly improves the GEH statistic. All the links reach a GEH statistic less or equal to 5, thereby satisfying the calibration criteria. The results show that the three calibration criteria are satisfied. In general, the proposed methodology was able to improve significantly the model outcomes.

Table 2 summarizes the calibration results for the first experiment. The total difference between actual and simulated link counts is 6% for all links in the network.

A sensitivity analysis was conducted using the Pyramid Highway model. With  $W = 0.5$  and  $W = 1.0$  the difference between simulated and link counts increased significantly.



**Fig. 3** Actual versus simulated counts and speeds before (a) and after (b) calibration, for the first experiment

**Table 2** Summary of calibration results for the first experiment

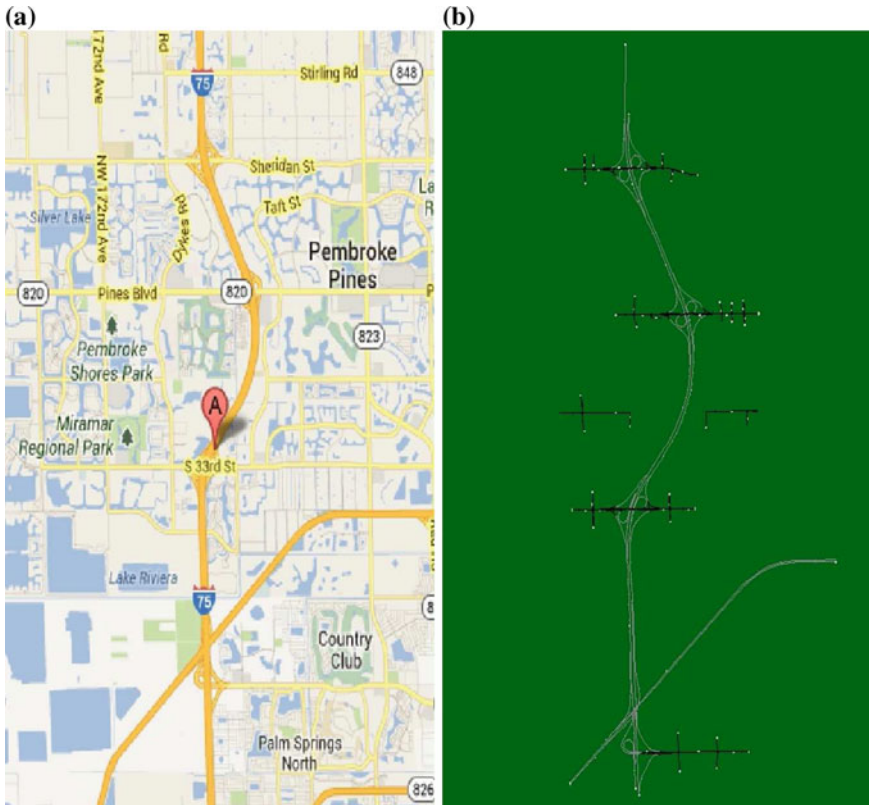
	NRMS	Total link counts	GEH
Before calib.	0.042	45,359	< 5 for 74 % of the cases
After calib.	0.010	55,882	< 5 for 100 % of the cases
Actual		59,610	

Second Experiment: I-75 in Miami, Florida, USA

In this experiment, a portion of I-75 in Miami, Florida was calibrated. A total of 375 freeway ramps and 334 arterial links were included in the model. Data was available for 353 freeway ramps and 59 arterial links for a morning peak period of one hour. The coefficients of the SPSA algorithm were the same as those used in the first experiment. All the calibration parameters in the network were included as well as the turning volumes for freeways and arterials. The weight factor in the objective function was set to 0.7. The time required for calibration was 125 min.

Figure 4a shows the Google map of I-75 highway in Miami, Florida, USA. Figure 4b illustrates the corresponding CORSIM model.

Figure 5 illustrates the trajectory of the objective function for this experiment. The NRMS goes from 0.270 to 0.245.



**Fig. 4** I-75 in Miami, Florida, USA (a) google map and CORSIM model (b), for the second experiment

**Fig. 5** Objective function for the second experiment

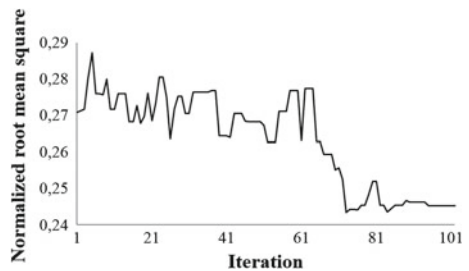
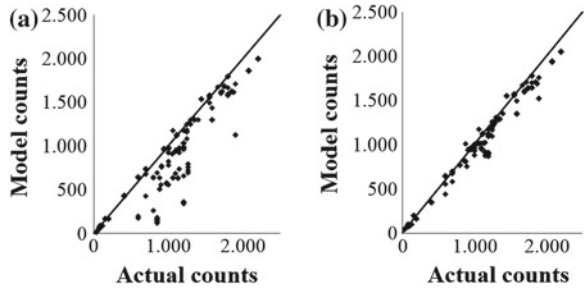


Figure 6a illustrates the link counts for the ramp segments in the model before calibration. Figure 6b shows the link counts for the ramps after calibration. These results clearly show that the calibration process significantly reduces the difference between actual and simulated link counts. It is clear that the calibration model significantly improves the GEH statistic. 99.6% of the links reach a GEH statistic less or equal to 5, thereby satisfying the calibration criteria.



**Fig. 6** Links counts before (a) and after (b) calibration for freeway ramps in the network (second experiment)



**Fig. 7** Links counts before (a) and after (b) calibration for arterials in the network (second experiment)

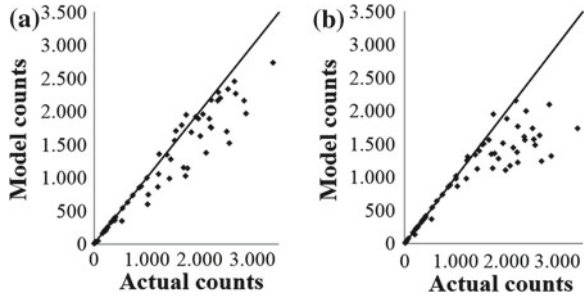


Figure 7a illustrates the link counts for the arterials before calibration. Figure 7b shows the link counts for the ramps after calibration. These results show that there is significant improvement for links with large link counts. The calibration model significantly improves the GEH statistic. Seventy-six percent (76%) of the freeway ramp links reach a GEH statistic less or equal to 5.

Figures 6 and 7 together show that the calibration methodology provides better results for freeway ramps than for arterials. This could be a consequence of having more data available for freeway ramps than for arterials, thereby giving more weight to the ramps.

Table 3 shows the ‘before’ and ‘after’ GEH statistics. As illustrated, the calibration improves the statistics, especially for the highest GEHs. However, some GEH values need to be improved because they are over 5.

**Table 3** Summary of calibration results for the second experiment

		Total link counts (vph)	GEH
Freeway	Before calib.	234,928.2	< 5 for 86 % of the cases
	After calib.	257,454.1	< 5 for 99.6 % of the cases
	Actual	271,908	
Arterials	Before calib.	61,097	< 5 for 66 % of the cases
	After calib.	68,927	< 5 for 76 % of the cases
	Actual	80,524	

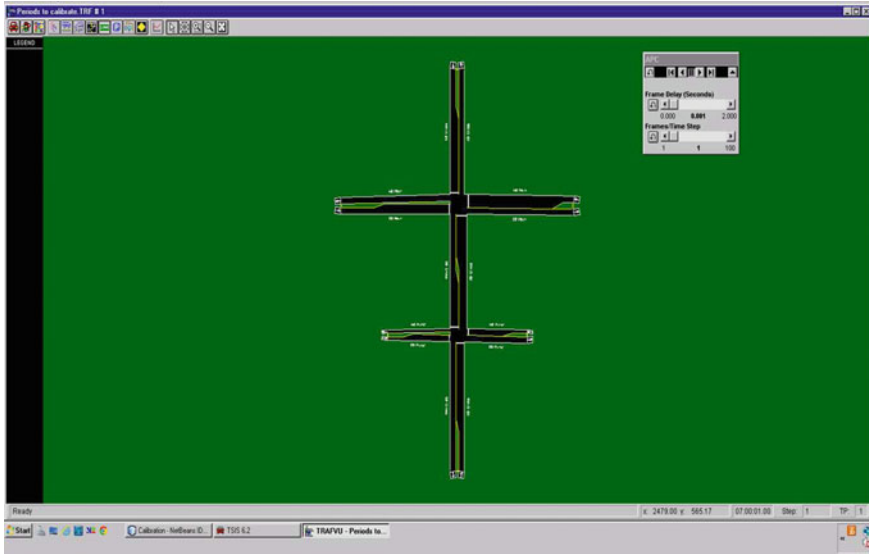


Fig. 8 CORSIM Model for the third experiment: network from McTrans sample datasets

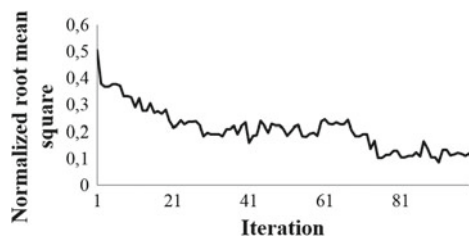
### Third Experiment: Network from McTrans Sample Datasets

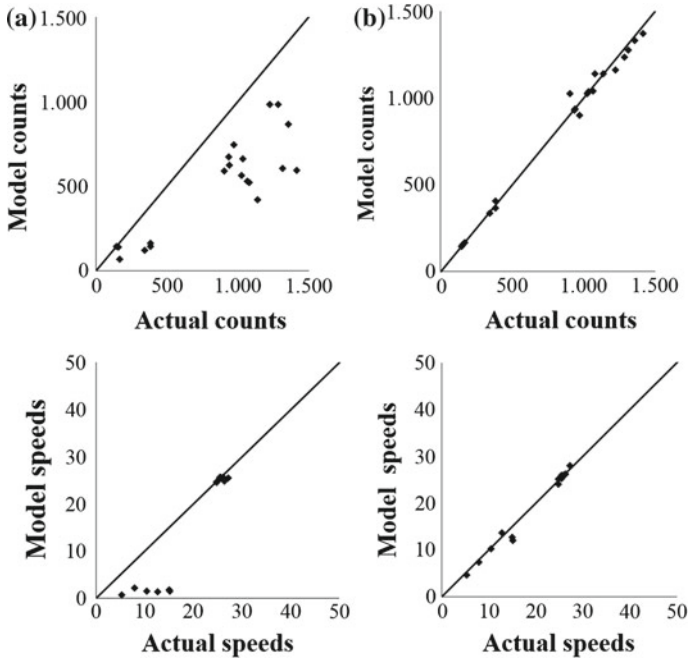
In this experiment, a network with arterials from McTrans official web page was calibrated. A total of 20 arterial links were included in the model. Data was available for all arterial links. Figure 8 shows the CORSIM model for this experiment. The time required for calibration was 10 min.

The total simulation time was 1 h divided in 4 time periods  $t$  of 15 min each ( $T = 4$ ). In this experiment, all parameters for all links for all four time periods were updated. The coefficients of the SPSA algorithm were the same as those used in the previous experiments. All the calibration parameters in the network as well as the turning volumes were included. The weight factor in the objective function was set to 0.7.

Figure 9 illustrates the trajectory of the objective function corresponding to the third experiment. The initial NRMS value is 0.51, while the minimum obtained after 100 iterations of the optimization algorithm is 0.09.

Fig. 9 Objective function for the third experiment





**Fig. 10** Actual versus simulated counts and speeds before (a) and after (b) calibration for time period 1, (third experiment)

Figure 10 illustrates the link counts and speeds before and after the calibration results for all links in the network for the first time period of the simulation. These results clearly show that the calibration process significantly reduces the difference between actual and simulated link counts and speeds.

Similar to Fig. 10, Table 4 shows the summary of link counts and speeds for all links in the network for the second, third, and fourth simulation time period, respectively. The calibrated results are significantly closer to the actual values, relative to the ‘before calibration’ results. In addition, all links have a GEH statistic below the threshold limit of 5 for all time periods. Speeds were improved for most links especially for values less than 20 mph.

In this experiment, optimal parameters for the model were determined in order to reproduce time-dependent link counts and speeds. The calibrated parameters took a single value during the entire simulation process; that is, they were not time-dependent. In contrast, the link counts and speeds were time-dependent. These results illustrate the ability of the proposed calibration methodology to adjust model parameters so as to calibrate the time-dependent link counts and speeds.

The summary of the results are showed in Table 4.

**Table 4** Summary of the calibration results for the third experiment

Goalkeeper	GK	Total link counts (vph)	GEH
Time period 1	Before calib.	10,126	< 5 for 10 % of the cases
	After calib.	17,136	< 5 for 100 % of the cases
	Actual	17,276	
Time period 2	Before calib.	13,498	< 5 for 10 % of the cases
	After calib.	22,625	< 5 for 100 % of the cases
	Actual	22,891	
Time period 3	Before calib.	10,502	5 for 0 % of the cases
	After calib.	17,820	< 5 for 100 % of the cases
	Actual	18,767	
Time period 4	Before calib.	10,533	< 5 for 0 % of the cases
	After calib.	17,939	< 5 for 95 % of the cases
	Actual	19,013	

## 4 Conclusions

This study proposed a methodology for the calibration of micro-simulation traffic flow models. The design and implementation of this methodology seeks to enable the calibration of generalized models. The proposed calibration methodology was developed independent of characteristics for any particular microscopic traffic flow simulation model. It minimizes the difference between actual and simulated time dependent link counts and speeds by considering all model parameters and turning volumes simultaneously.

The methodology used the Simultaneous Perturbation Stochastic Approximation (SPSA) algorithm to determine the calibrated set of model parameters. Previous studies have proposed the use of the SPSA algorithm for the calibration of vehicular traffic systems; however, few parameters were considered, and the calibration typically was based on a single performance measure, usually link counts. During the experiments developed, the proposed algorithm always reached convergence and stability.

The proposed methodology was tested using CORSIM models. However, there is nothing preventing the implementation of the proposed methodology for the calibration of other models. Three different vehicular traffic systems were calibrated, taking into consideration all their model parameters by using various performance measures, including link counts and speeds. The first experiment included arterials, using as performance measures link counts and speeds. The second system included both arterials and freeways. Considering arterials and freeways represented a significant challenge because two different models with different parameters needed to be considered simultaneously. The third experiment included time-dependent link counts and speeds for four time periods during this experiment; in addition, global, individual, and time-dependent parameters were considered. Further analysis was

required to determine the weight factor,  $W$ . This value was set constant because link counts were obtained using more accurate data collection methods compared to speeds. Information about the data collection and data quality can be used to set the weight factor.

The experimental results illustrated the effectiveness of the proposed methodology. The three vehicular traffic systems used in this study were successfully calibrated; specifically, the calibration criteria were satisfied after the calibration was performed. The results from the first and third experiment showed that speeds were improved after the calibration. The quality of the second vehicular traffic system improved significantly. However, further sensitivity analysis of the parameters used by the SPSA algorithm is required to achieve better results and satisfy the calibration criteria. These parameters were chosen using sensitivity analysis. A pattern to find optimal values for the SPSA parameters was not found. Further, as the number of parameters required for calibration increases, the complexity of the optimization problem also increases as well as the complexity to determine the set of required optimization coefficients.

The same set of calibration parameters was used in all the experiments. Therefore, any effort during parameter selection has been reduced. The results were improved for the entire model. All calibrated parameters were within reasonable boundaries. Similarly, no irregularities were observed using the graphical user interface. The calibration software developed in this study can be downloaded, along with a user's guide and examples, using this link: <http://faculty.unlv.edu/apaz/files/CalibrationToolDemo.zip>. Hence, the reviewers can replicate the results from this study.

The calibration tool developed as part of this study used an optimization algorithm that required a set of coefficients to find the appropriate set of CORSIM model parameters. A time-consuming sensitivity analysis of these coefficients was required to achieve desired results.

A bi-level optimization framework is required to enable the simultaneous calibration of traffic flow and SPSA parameters. The first level of the bi-level framework represents the existing calibration tool developed as part of the existing project, whose objective was the calibration of CORSIM models under saturated conditions. Here, and Simultaneous Perturbation Stochastic Approximation (SPSA) optimization algorithm was used to determine the appropriate calibration parameters. The second level of the proposed bi-level framework corresponds to future research, whose objective is to automate the sensitivity analysis that is required to find the right set of optimization coefficients for the SPSA algorithm. A parallel paper currently under review describes the proposed bi-level framework.

**Acknowledgments** The authors thank the Nevada Department of Transportation for sponsoring this project. The authors also appreciate Carlos Gaviria's and Cristian Arteaga's help and support during the implementation stage of this study.

## References

1. Punzo V, Ciuffo B, Montanino M (2012) Can results of car-following model calibration based on trajectory data be trusted? *Transp Res Rec J Transp Res Board* 2315(1):11–24
2. Anderson RE, Hicks C (2011) Highlights of contemporary microsimulation. *Soc Sci Comput Rev* 29(1):3–8
3. Breški D, Cvitanić D, Lovrić I (2006) Sensitivity analysis of the corsim simulation model parameters. *Gradevinar* 58(7):539–548
4. Brockfeld E, Kühne RD, Wagner P (2005) Calibration and validation of microscopic models of traffic flow. *Transp Res Rec J Transp Res Board* 1934:179–187
5. Holm P, Tomich D, Sloboden J, Lowrance C (2007) Traffic analysis toolbox volume iv: guidelines for applying corsim microsimulation modeling software. (no. fhwa-hop-07-079). it industries
6. Kim KO, Rilett L (2003) Simplex-based calibration of traffic microsimulation models with intelligent transportation systems data. *Transp Res Rec J Transp Res Board* 1855(1):80–89
7. Kondyli A, Soria I, Duret A, Elefteriadou L (2012) Sensitivity analysis of corsim with respect to the process of freeway flow breakdown at bottleneck locations. *Simul Model Pract Theory* 22:197–206
8. Schultz GG, Rilett L (2004) Analysis of distribution and calibration of car-following sensitivity parameters in microscopic traffic simulation models. *Transp Res Rec J Transp Res Board* 1876(1):41–51
9. Schultz GG, Rilett LR (2005) Calibration of distributions of commercial motor vehicles in corsim. *Transp Res Rec J Transp Res Board* 1934(1):246–255
10. Paz A, Molano V, Gaviria C (2012) Calibration of corsim models considering all model parameters simultaneously. In: 15th international IEEE conference on intelligent transportation systems (ITSC), pp 1417–1422. doi:[10.1109/ITSC.2012.6338841](https://doi.org/10.1109/ITSC.2012.6338841)
11. Balakrishna R, Antoniou C, Ben-Akiva M, Koutsopoulos HN, Wen Y (2007) Calibration of microscopic traffic simulation models: methods and application. *Transp Res Rec J Transp Res Board* 1999(1):198–207
12. Chin DC (1997) Comparative study of stochastic algorithms for system optimization based on gradient approximations. *IEEE Trans Syst Man Cybern B Cybern* 27(2):244–249
13. Lee JB (2008) Calibration of traffic simulation models using simultaneous perturbation stochastic approximation (spsa) method extended through bayesian sampling methodology. ProQuest, Ann Arbor
14. Maryak JL, Spall JC (2005) Simultaneous perturbation optimization for efficient image restoration. *IEEE Trans Aerosp Electron Syst* 41(1):356–361
15. Spall JC (1995) Stochastic version of second-order (Newton-Raphson) optimization using only function measurements. In: Proceedings of simulation conference, pp 347–352
16. Spall JC (1998) Implementation of the simultaneous perturbation algorithm for stochastic optimization. *IEEE Trans Aerosp Electronic Syst* 34(3):817–823
17. Spall JC (1998) An overview of the simultaneous perturbation method for efficient optimization. *Johns Hopkins APL Tech Dig* 19(4):482–492
18. Spall JC (2003) Introduction to stochastic search and optimization: estimation, simulation, and control. Wiley, New York
19. Lee JB, Ozbay K (2009) New calibration methodology for microscopic traffic simulation using enhanced simultaneous perturbation stochastic approximation approach. *Transp Res Rec J Transp Res Board* 2124(1):233–240
20. Ma J, Dong H (2007) Calibration of microsimulation with heuristic optimization methods. *Transp Res Rec J Transp Res Board* 1999(1):208–217
21. Cunha AL, Bessa JR JE, Setti JR (2009) Genetic algorithm for the calibration of vehicle performance models of microscopic traffic simulators. In: Progress in artificial intelligence, Springer, pp 3–14

22. Toledo T, Ben-Akiva ME, Darda D, Jha M, Koutsopoulos HN (2004) Calibration of microscopic traffic simulation models with aggregate data. *Transp Res Rec J Transp Res Board* 1876(1):10–19
23. Henclewood D, Suh W, Rodgers M, Hunter M, Fujimoto R (2012) A case for real-time calibration of data-driven microscopic traffic simulation tools. In: *Simulation conference (WSC), proceedings of the 2012 winter*, pp 1–12. doi:10.1109/WSC.2012.6465294
24. Rahman M (2013) Application of parameter estimation and calibration method for car-following models
25. Bando M, Hasebe K, Nakayama A, Shibata A, Sugiyama Y (1995) Dynamical model of traffic congestion and numerical simulation. *Phys Rev E* 51(2):1035
26. Cheu RL, Jin X, Ng KC, Ng YL, Srinivasan D (1998) Calibration of FRESIM for Singapore expressway using genetic algorithm. *J Transp Eng* 124(6):526–535
27. Holm P, Tomich D, Sloboden J, Lowrance C (2007) Traffic analysis toolbox volume iv: guidelines for applying corsim microsimulation modeling software. Technical Report. [http://ops.fhwa.dot.gov/trafficanalysisitools/tat\\_vol4/index.htm](http://ops.fhwa.dot.gov/trafficanalysisitools/tat_vol4/index.htm)
28. Lee DH, Yang X, Chandrasekar P (2001) Parameter calibration for paramics using genetic algorithm. In: *80th annual meeting of the transportation research board*, Washington, DC
29. Jha M, Gopalan G, Garms A, Mahanti BP, Toledo T, Ben-Akiva ME (2004) Development and calibration of a large-scale microscopic traffic simulation model. *Transp Res Rec J Transp Res Board* 1876(1):121–131
30. Li Z, Liu H, Li J (2010) A calibration and validation procedure for microscopic simulation model. In: *13th international ieee conference on intelligent transportation systems (ITSC)*, pp 563–568. doi:10.1109/ITSC.2010.5625018
31. Quadstone PARAMICS V4.2 (2003) Modeller reference manual. Quadstone Ltd, Edinburgh
32. Weinan H, Jian S (2009) A nsga-ii based parameter calibration algorithm for traffic microsimulation model. In: *International conference on measuring technology and mechatronics automation, 2009. ICMTMA '09*, vol 3, pp 436–439. doi:10.1109/ICMTMA.2009.437
33. Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Trans on Evol Comput* 6(2):182–197
34. VISSIM 5.10 (2011) User manual. PTV, Karlsruhe, Germany
35. Ma T, Abdulhai B (2002) Genetic algorithm-based optimization approach and generic tool for calibrating traffic microscopic simulation parameters. *Transp Res Rec J Transp Res Board* 1800(1):6–15
36. Center M (2010) TSIS-CORSIM
37. Omrani R, Kattan L (2013) Simultaneous calibration of microscopic traffic simulation model and estimation of origin/destination (od) flows based on genetic algorithms in a high-performance computer. In: *16th international IEEE conference on intelligent transportation systems—(ITSC)*, pp 2316–2321. doi:10.1109/ITSC.2013.6728573
38. Wang Z, Koutsopoulos HN (2011) Calibration of urban rail simulation models: a methodology using SPSA algorithm. *IEEE*, pp 3699–3709
39. Ma X, Huang Z, Koutsopoulos H (2014) Integrated traffic and emission simulation: a model calibration approach using aggregate information. *Environ Model Assess* 19(4):271–282
40. Cipriani E, Florian M, Mahut M, Nigro M (2011) A gradient approximation approach for adjusting temporal origin destination matrices. *Transp Res Part C* 19(2):270–282
41. Ben-Akiva M, Gao S, Wei Z, Wen Y (2012) A dynamic traffic assignment model for highly congested urban networks. *Transp Res Part C Emerg Technol* 24:62–82
42. McTrans (2012) Traffic software integrated system—corridor simulation. <http://mctrans.ce.ufl.edu/>
43. Hourdakakis J, Michalopoulos PG, Kottommannil J (2003) Practical procedure for calibrating microscopic traffic simulation models. *Transp Res Rec J Transp Res Board* 1852(1):130–139

# Optimizing Leisure Travel: Is BigData Ready to Improve the Joint Leisure Activities Efficiency?

K. Gkiotsalitis and A. Stathopoulos

**Abstract** Over the past years we are witnessing an upsurge on the volume of travelers' generated data. The upsurge of user-generated data from Smart Cards, Smart phones, personal navigators and social media has drawn the attention of the scientific community and new methods for utilizing such data in the areas of citizen-sensing, mobility understanding and travelers' behavioral analysis have been developed and tested. Stepping ahead from the central problem of leveraging user-generated data for improving the scheduling of transport services, this survey paper tries to investigate the importance of big-data on improving the organizational efficiency of physical meetings among multiple travelers in urban environments. First, this work examines the state-of-the-art on capturing travelers' patterns based on their data traces and the expected gains from leveraging user-generated data for optimizing leisure travel. Then, the problem of optimizing joint leisure travel is formulated and presented in an algorithmic form concluding to the suggestion of new research directions for future work.

## 1 Introduction

Today's metropolis with complex transport networks and numerous places for leisure activities pose great challenges to individuals who are willing to organize and participate in joint activities. In this study, we consider as joint leisure activities all activities conducted out-of-work involving the participation of two or more travelers.

Transport for London [1] posed that 29.2% of all daily trips are related to leisure activities, while 28% were conducted for shopping and personal business and 10.7% for other activities including escort. Similar results were observed on the New York Regional Travel survey [2]. Given the surveys' insights, it is evident that almost

---

K. Gkiotsalitis · A. Stathopoulos (✉)  
National Technical University of Athens, Iroon Polytechniou 5, 15773 Zografou, Greece  
e-mail: kgki1987@central.ntua.gr

A. Stathopoulos  
e-mail: a.stath@transport.ntua.gr



70% of all conducted trips (*typical weekday trips* =  $2.51 \times$  number of inhabitants in the city of London) can strengthen agents' interpersonal relations via shifting the general, non-recurrent trips to joint leisure trips. The aforementioned action is expected to promote the interpersonal relations among agents via increasing the number of physical meetings and improving the planning efficiency of out-of-work activities.

Fixed trips with recurrent characteristics (i.e., trips to work/school) can be easily recorded enabling the central transport authority or the individual agent to act on easing congestion or reducing traveling times via altering the transport/working schedules or shifting the departure times respectively. In contrary, leisure trips have a non-recurrent nature and that complicates the implementation of policy measures for congestion relief beforehand.

In the case of recurrent trips, travelers observe the repeated congestion patterns since they confront them on a daily basis while traveling over similar areas and adapt their schedules in order to reduce their waiting times. For instance, travelers are well-informed regarding the traffic conditions for trips to/from work due to their prior experience on traversing the same path on a daily basis, while they are less aware of the feasible set of trip-selection options when planning their out-of-work or other non-recurrent trips. Consequently, the lack of information yields three main inefficiencies:

- **Fluctuation of Travel Demand:** Out-of-work trips cannot be easily predicted from the central operator since travelers' actions cannot be forecasted and vary heavily from day to day
- **Interpersonal Activities Loss:** Not aware of the daily schedules of other individuals, one examined agent is either not able to schedule a joint leisure activity or schedules an inefficient one with high opportunity cost and limited participants
- **Trip Selection Inefficiency:** Agents enumerate a number of possible trips and select a most-preferred option via simple permutation or perceived utility-maximization without holding perfect information during the decision-making process

To that point, it should be stated that the individual-level planning of trips in metropolitan areas cannot be perceived as fully inefficient since it is based on a perceived utility-maximization approach; however, the lack of perfect information on the decision-making phase affects the efficiency of trip selection while attempting to maximize the utility function. Failure to construct a utility function which corresponds to the real-world conditions leads to the maximization of a non well-defined problem.

In this paper, it is assumed that the utility function is perceived correctly if the individual is well-informed during the trip-selection via holding information over three separate dimensions (refer to Fig. 1):

- The current traffic on the road network and delays on public transport services
- The exact location of all places of interest for leisure activities in the examined metropolis (i.e., location of bars, restaurants, cinemas)

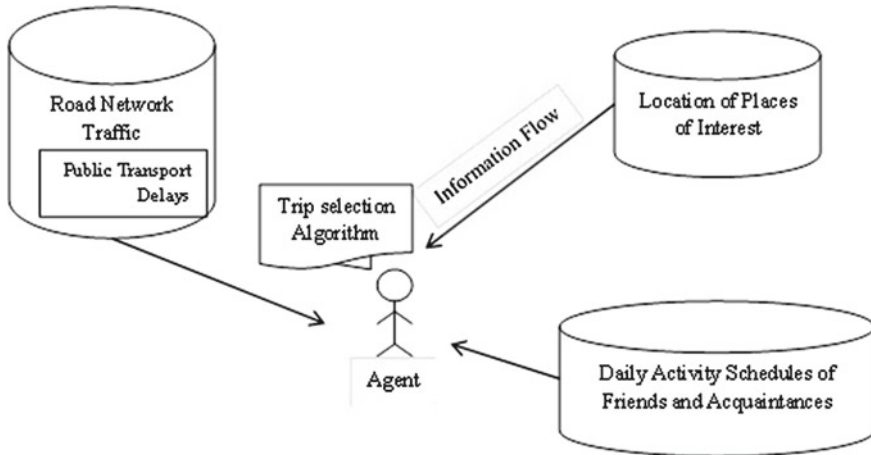


Fig. 1 The three dimensions of information flow for trip selection at the individual level

- The daily schedules and the preferences of all friends and acquaintances with whom a joint leisure activity can be organized (the degrees of freedom might differ depending on the social network of the examined agent)

Several attempts have been made to define special laws to model and explain the movement of people (refer to [3]). However, the lack of information at the trip-selection phase hinders the maximization of utility. At this stage, the utilization of new information streams can be seen as a valuable resource for improving the awareness of agents over the three dimensions of the decision-making process. In an example of improving a service via raising the level information dissemination, early research in a survey with bus riders demonstrated various positive effects such as increased ridership and traveler satisfaction attributed to enhanced information availability pointing out that easy access to relevant travel information is a decisive factor for the success and adoption of public transport systems (refer to [4]).

Given the above, this study examines the state-of-the-art in the area of non-recurrent trips which can be turned to opportunities for leisure joint activities with the use of insights from user-generated data. Attention is given to searching for studies on utilizing near real-time user-generated data (i.e., data from smartphones, smartcards, PDAs) for tackling transportation problems. The aim is to formulate the problem of optimizing leisure travel considering the provision of user-generated data, present the direction of the state-of-the-art, understand why user-generated data has not been used for increasing the volume and the efficiency of joint leisure activities and propose actions to move towards this direction.

In Sect. 2, the utilization of user-generated Cellular Data (CD) in transportation problems is examined. In Sect. 3, we are investigating works utilizing Social Media (SM) data and in Sect. 4 works in the area of Smart Card (SC) data. In Sect. 5, the use of Geo-location data via personal navigators and smartphones is examined. In

Sect. 6, a problem formulation for the joint leisure travel optimization with the use of user-generated data is proposed. Finally, the use of data from personal navigators is examined and a detailed catalog with future work directions is presented.

## 2 Utilizing Cellular Data in Transportation Problems

Cellular data is the form of user-generated data which have been studied the most for predicting individuals' mobility patterns even if the mobile tracking via cell towers is not as accurate as the satellite-based positioning. Regardless the posed challenges, cellular data have been utilized to improve the understanding on human mobility and develop individual-level models for capturing the mobility and activity habits of individuals.

The most common individual-level models for predicting the mobility of individuals-which are not based solely on spatio-temporal travel pattern recognition-are the activity-based models (refer to [5–8]). Those models are the basis for forecasting individuals' daily trip schedules from cellular data and perceive each trip as a means to participate at pre-scheduled activities.

In the literature, Musolesi and Mascolo [9] utilized Cellular data logs for correlating the mobility patterns of an individual with the mobility patterns of his friends and acquaintances. The underpinning theory of the correlation process includes the assumption that users' travel patterns do not depend on time and space, but also on the travel patterns of other individuals inside their social network. The findings of the research showed that the mobility patterns of one examined agent can be predicted more accurately when the mobility patterns of his/her social network are considered as explanatory variables. De Domenico et al. [10] worked also on the same direction using data from the Nokia Mobile Data Challenge dataset. The work of Musolesi and Mascolo [9] can provide some evidence on the theoretical concepts developed by Carrasco et al. [11], Arentze and Timmermans [12] and Chen et al. [13] on predicting agents' mobility based on their social networks. Those theoretical concepts place the traveler to the center of decision-making (ego-centric approaches) and offer a new framework for microsimulation, while harvesting large-scale user-generated data is expected to facilitate their implementation.

In addition, Carrasco et al. [14, 15], González et al. [3], Zhang et al. [16], Pan et al. [17] and White and Wells [18] utilized cellular data for predicting the mobility patterns of individuals in urban scenarios over time and space. Those studies, including studies of White and Wells [18] and Djuknic and Richton [19], attempted to exploit the emergence of cell tower positioning and the market penetration of mobile phones by developing methods for estimating the OD matrices in study areas. In the same way, Sohn and Kim [20] used cellular communication system and cell phone tower to transfer information and estimate OD matrices. To give a practical example, in the work of Calabrese et al. [14], an algorithm for estimating a population's travel demand in terms of ODs from aggregating the trips of individual mobile phone users in the Boston Metropolitan area was developed. During the validation, it was shown

also that the OD flows correlated well with the US Census estimates.

The limits of predictability in human dynamics by analyzing mobility patterns of mobile phone users were also analyzed and evaluated by Song et al. [21]. More recently, Dong et al. [22] and Wu et al. [23] proposed a methodology for using mobile phone data to analyze the mechanism of trip generation, trip attraction and the OD information with a pilot study at Beijing via using the K means clustering algorithm to divide the traffic zones. In addition, Ohashi et al. [24] worked on a method for separating trips (capturing the starting and ending points of a trip) on the basis of GPS data collected from smartphones by considering that even when the subject stays into a place, the collected GPS coordinates are not always exactly the same according to the surrounding environment assuming 81 % percision and recall rate of 62 %. Apart from detecting departure and arrival times, methods for classifying automatically modes of transportation on the basis of smartphone GPS data were also proposed by Ohashi et al. [25].

Finally, in another set of studies from [26, 27], Bluetooth devices were distributed to people to collect mobility data and study the characteristics of co-location patterns among them.

To summarize, works on utilizing cellular data in transportation have been focused on different areas:

- Estimating the OD matrix in a study area
- Exploring the mobility patterns of one individual based on the mobility patterns of his/her social network
- Extracting the current mode of transportation
- Extracting the starting and ending time of a trip

However, there is no work in our knowledge in the area of activity-participation analysis which can facilitate the development of new applications for suggesting common activities to users with social ties based on their willingness to participate simultaneously in similar activities in close proximity locations.

### **3 User-Generated Data from Social Media and Its Applications in Transportation**

The research on data from social networks on understanding users' mobility is in its early stages. The first studies focused on the power of micro-blogging on offering near real-time insights on crisis events when all other means of communication have failed. Routinely, the importance and the volume of the crisis event is captured through the magnitude of micro-blogging messages and their content information. A study from [28] explored crisis informatics using Twitter data after the Oklahoma Moore tornado demonstrating the potential of social media data on extracting relevant information during natural disasters.

In a similar fashion, social media data from social networks like Facebook, Twitter, and the image sharing service, Flickr, have already been used in research works

describing crisis or natural disasters such as Virginia Tech shooting ([29]), Southern California wildfires ([30]), major Earthquakes in China ([31, 32]), Red River floods and Oklahoma grassfires ([33]).

In another set of works, [34] utilized the Internet as resource to capture the crowd levels during planned special events. In general, local events are not tracked from transport authorities since manual, labor-intensive tracking is needed. Pereira et al. [34] utilized the Internet as a resource for contextual information about special events and developed a model that predicts public transport arrivals in event areas. The results were demonstrated with a case study from the city-state of Singapore using public transport tap-in/tap-out data coupled with local event information obtained from the Internet performing primitive data fusion.

In another work, Gkiotsalitis and Alexandrou [35] focused on developing and testing analytic techniques for fusing user-generated data from Social Media and smart-cards in order to capture the mobility patterns in urban areas. Automatic models for retrieving users' mobility patterns from historic, user-generated data logs, comparing user' profiles based on the similarity of their observed mobility patterns and categorizing users in clusters were developed. During the testing phase, user-generated data from London Smart Card and Social Media users collected between November 2012 and February 2014 were utilized to cluster users based on their mobility-activity pattern similarities. Results showed that it is possible to integrate data logs from multiple sources to capture the main mobility-activity patterns observed in an area. However the topic of joint participation in non-recurrent activities has not been addressed until now.

Social media have also been used for capturing the activity types performed by users at different locations via advanced spatio-temporal analysis and educated rules (refer to [36]). In the same work, techniques for estimating individuals' daily schedules and the sequence of activities were developed. Alesiani et al. [37] focused also on the same topic introducing a probabilistic model for modeling individuals' daily schedules based on input data from several sources (i.e., Social Media, Cellular Data).

Summarizing, social media data which is individualistic in nature has been utilized for:

- Capturing the volume and the effects of crisis events
- Estimating individuals' mobility patterns and correlating them with with patterns observed with the use of other datasets
- Retrieving activity types of users
- Capturing the arrival times and the expected demand at local events

## 4 User-Generated Data from Smart Cards

With the deployment of automatic fare collection systems, large-scale data becomes available for real-world transport usage ([38]). As more and more sensors have been integrated into public transport infrastructures, large-scale transport data is produced

at high rates ([39]). Nonetheless, studies of estimating individual travel patterns with smartcard data are sparse in public transport research compared to studies on cellular data and social media.

In the past, research has mainly focused on aggregate demand forecast ([40]). Based on a gravity model, Smith et al. [41] showed that some of the variation in mobility flows is influenced by distance and population of local residents via analyzing smartcard data, while Ceapa et al. [42] analyzed time series of automated fare collection data to identify events of overcrowding at public transport stations. Morency et al. [43] and Jang [44] also measured the transit use variability with smart-card data.

The potential of smart card data for travel behavior analysis in Britain was studied by Bagchi and White [45] where the pensioner concessionary pass in Southport, Merseyside, and the commercially operated scheme in Bradford were examined. There was stated that the nature of smart card data puts an emphasis on concept definition and rules-based processing; but limitations, such as the trip lengths which are not recorder to the system, were also recognized. The latter implicates also the efforts on performing individual-level analysis and predicting individuals daily travel schedules.

Foell et al. [46] utilized travel card data from a large population of bus riders from Lisbon, Portugal. The main intention of the work was to predict the future bus stops accessed by individual drivers and it was demonstrated that accurate predictions can be delivered by combining knowledge from personal ride histories and the mobility patterns of other riders. In another work, Ivanchev et al. [47] utilized smart card data from a bus line in Singapore for developing a modeling platform for testing bus transportation.

Finally, as discussed before, in the work of Gkiotsalitis and Alexandrou [35] a more individual-based approach was considered for clustering users based on the similarities of their mobility patterns as they were retrieved from pattern recognition on their historic smart card data logs. For the case study, data from 200 Oyster card users in London were utilized.

It is evident that smartcard data offers less qualitative information compared to social media or cellular data generating problems for predicting the daily schedules and the social networks of individuals. Nevertheless, it has great potential on the first scale of information retrieval: “Capture in real-time the traffic on road networks and the deals on public transport”.

## **5 Use of Geo-Location Data via Personal Navigators and Smartphones in Transportation**

More classic methods on dynamic OD estimation using automatic vehicle identification data can be found in the work of Zhou and Mahmassani [48] and Baek et al. [49]. Schuessler and Axhausen [50], Zheng et al. [51], and Brunauer et al. [52] proposed

methods for distinguishing pedestrians, bicycles, cars, buses, and trains on the basis of GPS data only. Stenneth et al. [53] introduced an idea of using GIS information to enhance the accuracy of classification. They utilized information about the real-time location of buses and locations of rail lines and bus stops where they reported that they could improve the classification accuracy by 17%.

Nitsche et al. [54] and Feng and Timmermans [55] proposed methods that use acceleration data together with GPS data following the work of Wu et al. [56] on estimating individuals' activity patterns. Wu et al. [56] attempted to estimate the activity patterns of smartphone users. They developed a method for classifying "indoor", "outdoor static", "outdoor walking", and "in-vehicle" status. Similarly, Hato [57] developed a special device, called a behavioral context addressable logger (BCALs), for collecting various kinds of data such as GPS coordinates, acceleration, atmospheric pressure, angular velocity, UV index, direction, and loudness. BCALs can distinguish situations in which smartphone users are classified as "walking", "up/down-staircase", "bicycling", and "in-store".

There are also studies on trip-separation methods (mainly by capturing the starting and ending time of a trip) by Li et al. [58], Bohte et al. [59], Chen et al. [60] and Li et al. [61]. Among these studies, only Witayangkurn et al. [62] reported an evaluation of a trip-separation method. The basic idea forming the basis of their method is to find the so-called "stay points". They regard consecutive GPS coordinates as stay points if they satisfy the following two conditions: (i) they fall within a circle with diameter of 196 m; and (ii) the time difference between the first and last stay points is more than 14 min. The key idea behind this stay-point detection is that it eliminates outliers, which can cause mis-detection. This trip-separation method achieved precision of 92.4% and recall rate of 90.5%.

Given the above, one can conclude that geo-location data from smartphones or personal navigators have been mainly utilized for:

- Estimating OD matrices
- Capturing the type of utilized transportation
- Activity-pattern estimation
- Separation of trips

## 6 Optimizing Joint Leisure Travel with BigData

Continuous updated, user-generated data can be utilized to capture less frequent trips and improve the understanding of individuals' mobility behavior. Collected data from Smartphones, Social Media, personal navigators and Smartcards has an individualistic nature since it is generated from distinct users. In a generalized example, it is assumed that the generated data footprint from an individual at each time instance returns information about the timestamp of data publishing, the utilized transport mode, the geo-location and the user ID.

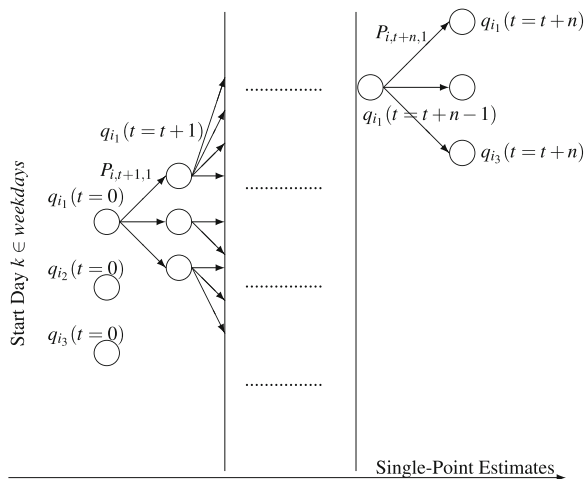
$$\delta_{i,t} = \begin{cases} t \\ \zeta \\ L \\ i \end{cases}$$

where  $\delta_{i,t}$  is agent's  $i$  generated data footprint,  $t$  the timestamp,  $\zeta$  the transport mode, where  $\zeta \in Z$  and  $L$  the geo-location where  $L \in \Lambda$  and  $\Lambda$  is the set of geo-locations defined by a pair of coordinates.

Following the above notation, individuals' data generation can offer mobility insights regarding his/her daily mobility patterns via utilizing un-supervised pattern recognition models. Those models can be trained on datasets containing historical data from one individual's data footprints accumulated over a significant time period (i.e., more than 6 months). The outcome of the pattern recognition phase can be summarized in a probability matrix with spatio-temporal characteristics,  $P_{i,t}[L, \zeta]$ , which returns the probability of individual,  $i$ , to be at location,  $L$ , and use transport mode,  $\zeta$ , at time  $t$ . Since individuals mobility patterns can vary significantly on weekends, for each individual,  $i$ , two matrices can be assigned—one capturing the travel patterns of the user during the week and one during the weekend. Further discretization is allowed and can be decided in a case by case basis if certain individuals have significantly different mobility patterns over some days of the week.

Each matrix  $P_{i,t}[L, \zeta]$  has  $[T \times \lambda \times Z]$  elements, where  $T$  is the sum of time instances over a day and  $Z$  the set of available transport modes including on foot travel. Having calculated one individual's probability matrix at day type  $k$ , the daily mobility plan of the individual can be estimated with deterministic modeling via using single point estimates. For each single-point estimate the matrix  $P_{i,t}$  is utilized and the output is a sequence of states,  $q_{i,\rho}(t)$ , where  $\rho$  the list of feasible states at time  $t$  and day type  $k$  as they are derived from the analysis of individual's data footprints  $\delta_{i,t}$  (refer to Fig. 2).

**Fig. 2** Estimating the daily evolution of states over a weekday with the use of probability matrix  $P_{i,t}$





Apart from the probabilistic matrix, historical, user-generated data can offer insights on the social network of individuals. For instance the list of friends, acquaintances and common preferences can be retrieved from user-generated data analysis (refer for instance to the released dataset from [63] using data from Facebook.com). Golder et al. [64] showed also that users only message to a small number of friends on Facebook (close friends) while they have a large number of declared friends (acquaintances) and Huberman et al. [65] showed that most of the links declared in Twitter are meaningless from an interaction point of view but hidden social networks can be revealed when tracing the spread of ideas. In the same direction, different forms of user-generated data can be utilized to identify the social network of one individual (ego-centric approach) and attach a weight representing the strength of bonds among individuals:

$$W_i = \begin{cases} w_{i,j} \\ \dots \\ \dots \\ w_{i,N} \end{cases}$$

where  $\{j, \dots, N\}$  is the set of individuals having social ties with user  $i$  and  $W_i \geq 0$  the weight symbolizing the connection strength among them.

On another note, the preferred undertaken activity of one individual  $i$  at one re-visited location  $L \in A$  can be estimated after analyzing historical user-generated data. In the work of Gkiotsalitis and Stathopoulos [66] empirical rules for allocating one activity  $A_m \in A$  at one re-visited location  $L$  were defined by categorizing all activities in a discrete set of four (Fixed; Quasi-Fixed; Flexible; Home-related). Hence, each re-visited location  $L$  by one individual  $i$  is associated to one and only one activity  $A_m$ :

$$i[L] = A_m \in A \quad (1)$$

For allocating activities to locations, one can utilize spatio-temporal analysis on historical data. Such approach had been used in the case of user-generated data from social media ([36]) and cellular data ([67]).

Moving further towards that direction, user-generated data can also provide information on how far one individual can travel to participate at a leisure activity at different day times and day types. For instance, one individual might not be willing to travel more than 500m at working hours during the week for participating in leisure activities. In an attempt to model one individual's choice of traveling a certain distance for participating in a leisure activity, a utility function can be defined. After applying a time discretization scheme  $t = \{1, \dots, T\}$ , the choice options can be indexed by  $j = \{1, \dots, J\}$  where  $F_j$  is the traveled distance between two consecutive activities. The distance between two consecutive activities can be either calculated with the Haversine formula (Great-circle distance) or via the map-based shortest path distance with the use of a shortest path algorithm (refer to [68] for such algorithms).

$$j = \begin{cases} 1 : F_j \leq 250 \text{ m} \\ 2 : 250 \text{ m} < F_j \leq 500 \text{ m} \\ 3 : 500 \text{ m} < F_j \leq 750 \text{ m} \\ 4 : 750 \text{ m} < F_j \leq 1 \text{ km} \\ 5 : \dots \\ 6 : \dots \end{cases}$$

For each day type,  $k$ , an index of satisfaction for participating in different activity types with respect to their distance from the previous location can be defined in the form of a linear utility function:

$$V_{tj}(k) = \alpha_j(k) + \beta_j(k)A_t(k) \tag{2}$$

where  $A_t(k)$  varies across different times of the day and represents the activity type (i.e., home, fixed, quasi-fixed or flexible) in the form of a categorical variable. In addition,  $\alpha$  is a scalar utility term representing individual’s preference for alternative  $j$ .

The random utility of traveling distance  $F_j$ , for an individual can be described by a random utility model:

$$U_{tj}(k) = V_{tj}(k) + \varepsilon_{tj}(k) \tag{3}$$

where  $\varepsilon_{tj}(k)$  is the unobserved component of the utility function and can be treated as a random variable since it includes the impact of all the unobserved variables which influence the utility of selecting a specific alternative.

With the assumptions that errors follow a Gumbel distribution, are independent and identically distributed, the probability of selecting an alternative  $\lambda = F_j$  at a certain point in time,  $\rho_{t\lambda}(k)$ , can be expressed via a multinomial logit model:

$$\begin{aligned} \rho_{t\lambda}(k) &= \rho \left( V_{t\lambda} + \varepsilon_{t\lambda}(k) \geq \max_{j \in \{1, \dots, J\}} V_{jt}(k) + \varepsilon_{tj}(k) \right) \\ &= \frac{e^{V_{t\lambda}(k)}}{\sum_{j=1}^J e^{V_{tj}(k)}} \end{aligned} \tag{4}$$

The parameters  $\alpha_j(k)$ ,  $\beta_j(k)$  can be estimated for each individual as the values that maximize the log-likelihood function:

$$\max_{\alpha_j(k), \beta_j(k)} \ell(\alpha_j(k), \beta_j(k)) \tag{5}$$

resulting to a non-linear optimization problem for which the optimization algorithm BHHH proposed by Berndt et al. [69] can be applied. The coefficient values are

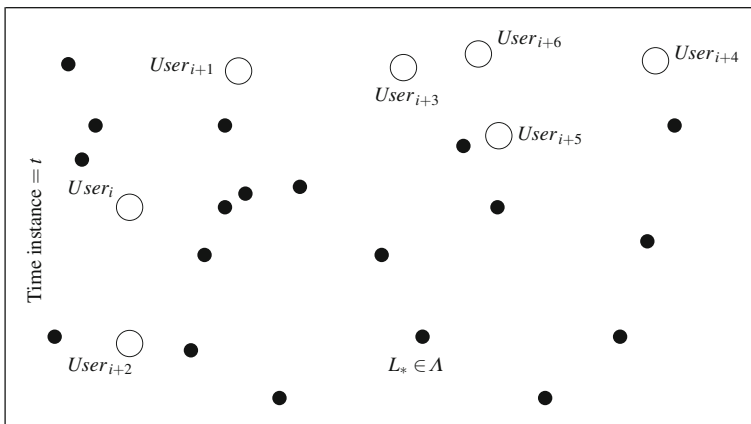
updated in an iterative approach beginning with a starting set of values and iterations continue until convergence.

To summarize, pattern recognition models can be applied to generate some value from user-generated data:

- Estimate individuals’ daily schedule over different day types via single-point estimates of their state evolution over time
- Capture the gravity of personal relationships and assign weights to friends and acquaintances of each individual
- Replicate the decision-making process of each individual and return their willingness to travel certain distances at different times of the day to participate in leisure activities

To optimize joint leisure travel, a time and place for a joint leisure activity which maximizes the gain of all attendees should be defined. For such undertaking, the perceived utility of all individuals participating at one activity  $L \in \Lambda$  at each point in time  $t$  over a day should be estimated for selecting the spatio-temporal set  $L_*, t_*$  which maximizes the perceived utility among all attendees. The computational cost of it is  $\lambda \times T \times N^2$  where  $T$  is the discretized time scheme and  $N$  the number of individuals (refer to Fig. 4). Although the effect of time discretization to the overall time complexity is linear, a discretization every 30 min to one hour is proposed to avoid significant computational cost increases. Therefore, the time should be discretized and at each step the utility of attending one location of leisure activity can be computed (refer to Fig. 3).

In the problem of estimating the location of a leisure activity and the time of day that maximizes the utility of performing a joint leisure activity, the relationship weights among the attendees can be perceived as positive factors while the required



**Fig. 3** Selecting place of interest  $L_*$  among a set of locations  $L \in \lambda$  at time instance  $t$  which maximizes the perceived utility among all attendees

travel distance from each individual's current location to the meeting place can be perceived as negative.

Let us assume that the probability of individual  $i$  to travel a certain distance  $\tau_i \in F_j$  at time instance  $t$  is  $\rho_{tj}(k)$  as it is already derived from his/her utility-maximization model. Then, a threshold value  $\mathcal{Y}$  can be introduced and if  $\rho_{tj}(k) < \mathcal{Y}$  for one place of interest  $L \in \Lambda$  for which the distance from the previous individual's location is within  $F_j$ , then location  $L$  is perceived as non-feasible place for transition. Hence, for each individual  $i$  the distance between his/her current location,  $c$ , and a place of interest  $L \in \Lambda$  is calculated. Then, if distance  $\tau(c, L) \in F_j$  and  $\rho_{tj}(k) < \mathcal{Y}$  or  $L$  is not a leisure activity location for individual  $i$ , the place of interest  $L$  is assigned to the list of in-feasible transitions for time  $t$ :  $\phi_{i,t} = \phi_{i,t} + \{L\}$ . Then, a location  $L_* \in \Lambda$  and time  $t_* \in T$  is the optimal joint leisure activity set if:

$$\{L_*, t_*\} = \operatorname{argmax}(\alpha \sum_{i=1}^N \sum_{m=1}^N \frac{1}{2} w_{i,m}(L, t) - \beta \sum_{i=1}^N \tau_i(L, t)) \quad (6)$$

where:

$$w_{i,m}(L, t) = \begin{cases} w_{i,m} \geq 0: \text{weight of connection strength between users } i, m \\ 0 \text{ if } L \in \phi_{i,t} \text{ or } L \in \phi_{m,t} \end{cases}$$

$$\tau_i(L, t) = \begin{cases} \tau_i \geq 0: \text{the traveled distance between the current location and } L \\ 0 \text{ if } L \in \phi_{i,t} \end{cases}$$

and  $\alpha, \beta > 0$  objective function coefficients. It is evident that  $\alpha$  is more significant than  $\beta$  if one considers the activity participation of attendees with strong social ties as the main objective, while  $\beta$  is more significant if the scope is to reduce the covered travel distance. Algorithm 1 summarizes the optimization procedure.

The computational cost of Algorithm 1 for joint leisure travel optimization was tested on a 2556 MHz processor machine with 1024 Megabytes RAM. During the testing, the number of locations was,  $\Lambda = 200$  locations, and the time was discretized into ninety-six periods of 30-min. duration,  $T = 96$ . The main variable is the number of friends and acquaintances for which Algorithm 1 computes the location and time for a leisure activity and the computational cost is plotted in Fig. 4. Figure 4 provides an indication of the number of individuals which can be served within a reasonable time frame. Finally, it should be mentioned that Algorithm 1 runs centrally to avoid unnecessary re-computations (in general, the approach follows a central architecture where user-generated data is stored centrally and the travel patterns, list of friends and acquaintances and willingness to travel certain distances to participate at leisure activities are estimated after processing the stored data; therefore, enabling the implementation of Algorithm 1 at a central level.

```

for each user  $i \in N$  do
  for each time instance  $t \in T$  do
    Estimate location  $L_{i,t}$  where individual  $i$  is expected to be located at time  $t$  using the
    single-point estimate based on matrix  $P_{i,t}$ ;
    for each location  $L \in \Lambda$  do
      if transition from  $L_{i,t}$  to  $L \in F_j$  and  $\rho_{t+1,j}(k) < \gamma$  then
         $\phi_{i,t+1} = \phi_{i,t+1} + \{L\}$ ;
      end
    end
  end
end
Set the cost of optimal solution  $Q = -\infty$ ;
Set optimal meeting location  $L_* = \emptyset$  and optimal timing as  $t_* = \emptyset$ ; for each time instant
 $t \in T$  do
  for each location  $L \in \Lambda$  do
    Set  $E = 0$ ;
    for each user  $i \in N$  do
      for each user  $m \in N$  do
        if location  $L \notin \phi_{i,t}$  and  $L \notin \phi_{m,t}$  then
           $E = E + (\frac{1}{2}\alpha w_{i,m}(L, t) - \beta \tau_i(L, t))$ ;
        end
      end
    end
    if  $E > Q$  then
       $Q = E$ ;
       $L_* = L$ ;
       $t_* = t$ ;
    end
  end
end

```

**Algorithm 1:** Calculate the optimal cost for participation in a joint leisure activity

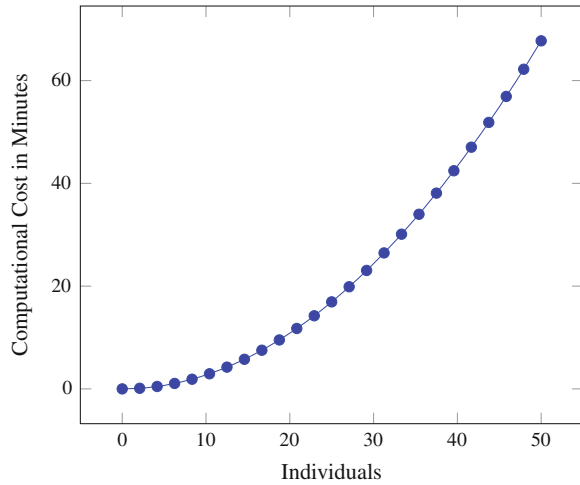
## 7 Discussion and Conclusions

This survey study attempted to investigate how different forms of user-generated data (cellular, social media, smart card and personal navigator data) have been utilized until now and examine if the data sources and the developed techniques have some potential on increasing the efficiency of joint leisure activities in today's metropolis.

In a first attempt to summarize the results, Table 1 provides aggregated information on the usage of user-generated data from different sources according to the state-of-the-art studies.

In the introduction section of the survey paper, three information dimensions were considered for assuming that an individual is perfectly informed for making an optimal decision on selecting a leisure joint activity. In Table 2, we show which kind of information is expected to be retrieved from different sources of user-generated data. From Tables 1 and 2 one can observe that although the full information for

**Fig. 4** Computational cost considering  $\Lambda = 200$  and  $T = 96$  for different numbers of individuals (tested on a 2556MHz processor machine with 1024 Megabytes RAM).



**Table 1** Aggregated information on the usage of user-generated data from different sources according to the state-of-the-art studies

	Cellular	Social media	Smart card	GPS positioning
Estimating OD matrices	★	×	★	★
Extracting the utilized mode of transportation	★	×	★	×
Capturing the trip separation	×	×	★	★
Real-time traffic estimation	★	×	×	★
Estimating the daily schedule of agent's social network	×	★	×	×
Crisis events analysis	×	★	×	×
Capturing individuals' mobility patterns	★	★	×	×
Retrieving the performed activities by users	★	★	×	×
Forecasting the expected demand at local events	×	★	×	×
Separating trips	★	×	×	★
Activity-pattern estimation	★	★	★	★

forming an objective function is obtainable, research work has not been focused on that direction.

Due to the above, the importance of developing new models for tapping the potential of user-generated data for improving the efficiency of joint leisure activity planning is highlighted. New models and techniques are recommended to focus on the following:

**Table 2** Potential of user-generated data on providing information for joint leisure activity planning

	Cellular	Social media	Smart card	GPS positioning
Real-time traffic and public transport schedules	×	★	★	★
Location of places of interest	×	★	×	★
Daily schedules of agent's social network	×	★	×	×

- Data processing tools
- Algorithmic tools for data aggregation and fusion
- Processing tools that can calculate the maximum of the utility function and return an optimal joint leisure activity to the traveler

Considering those issues, we tried to formulate the problem of optimizing joint leisure travel by taking into consideration the special characteristics of user-generated data. The proposed formulation is flexible and can handle inputs even after data fusion since it requires a minimum information set (UserID; Timestamp; Geo-location; transport mode). The proposed algorithm is also designed to ensure scalability by enabling the computation of leisure travel optimization for up to 30 individuals in less than 20 minutes considering a 15-minute time discretization and up to 200 locations to choose from.

Proceeding towards this direction, around 70% of the total number of trips in metropolitan areas can be planned more efficiently and the interpersonal activities can be heavily increased in numbers yielding remarkable gains for both the individual traveler and the central transport authorities.

## References

1. Transport for London (2014) Travel in London, London travel demand survey. <http://www.tfl.gov.uk/cdn/static/cms/documents/london-travel-demand-survey.pdf>. Accessed 30 Sept 2014
2. New York Regional Travel Survey (2014). <http://www.nymtc.org/project/surveys/survey.html>. Accessed 30 Sept 2014
3. González MC, Hidalgo CA, Barabási A-L (2008) Understanding individual human mobility patterns. *Nature* 453(7196):779–782. ISSN 0028–0836
4. Ferris B, Watkins K, Borning A (2010) Onebusaway: results from providing real-time arrival information for public transit. In: Proceedings of the 28th international conference on human factors in computing systems (CHI '10), June 2010
5. Axhausen KW, Gärling T (1992) Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. *Transp Rev* 12(4):323–341. ISSN 0144–1647
6. Arentze TA, Timmermans HJP (2004) A learning-based transportation oriented simulation system. *Transp Res Part B: Methodol* 38(7):613–633
7. Pendyala R, Kitamura R, Kikuchi A, Yamamoto T, Fujii S (2005) Florida activity mobility simulator: overview and preliminary validation results. *Transp Res Rec J Transp Res Board* 1921(-1):123–130

8. Lin D-Y, Eluru N, Waller S, Bhat C (2008) Integration of activity-based modeling and dynamic traffic assignment. *Transp Res Rec J Transp Res Board* 2076(-1):52–61
9. Musolesi M, Mascolo C (2007) Designing mobility models based on social network theory. *SIGMOBILE Mob Comput Commun Rev* 11(3):5970. ISSN 1559–1662
10. De Domenico M, Lima A, Musolesi M (2013) Interdependence and predictability of human mobility and social interactions. *Pervasive Mobile Comput* 9(6):798–807
11. Carrasco JA, Hogan B, Wellman B, Miller EJ (2008) Collecting social network data to study social activity-travel behavior: an egocentric approach. *Environ Plann B Plann Des* 35(6):961–980
12. Arentze T, Timmermans H (2008) Social networks, social interactions, and activity-travel behavior: a framework for microsimulation. *Environ Plann B Plann Des* 35(6):1012
13. Chen Y, Frei A, Mahmassani HS (2014) From personal attitudes to public opinion: information diffusion in social networks toward sustainable transportation. In: *Transportation research board 93rd annual meeting*, number 14–3566
14. Calabrese F, Di Lorenzo G, Liu L, Ratti C (2011a) Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Comput* 10(4):36–44. ISSN 1536–1268
15. Calabrese F, Colonna M, Lovisolo P, Parata D, Ratti C (2011b) Real-time urban monitoring using cell phones: a case study in rome. *IEEE Trans Intell Transp Syst* 12(1):141–151. ISSN 1524–9050
16. Zhang Y, Qin X, Dong S, Ran B (2010) Daily o-d matrix estimation using cellular probe data. In: *Transportation research board: 89th annual meeting*
17. Pan C, Lu J, Di S, Ran B (2006) Cellular-based data-extracting method for trip distribution. *Transp Res Rec J Transp Res B* 1945(-1):33–39
18. White J, Wells I (2002) Extracting origin destination information from mobile phone data. In: *Road transport information and control, 2002. Eleventh international conference on (Conf. Publ. No. 486)*, pp 30–34
19. Djuknic GM, Richton RE (2001) Geolocation and assisted gps. *Computer* 34(2):123–125
20. Sohn K, Kim D (2008) Dynamic origin-destination flow estimation using cellular communication system. *IEEE Trans Veh Technol* 57(5):2703–2713
21. Song C, Qu Z, Blumm N, Barabasi AL (2010) Limits of predictability in human mobility. *Science* 327(5968):1018–1021
22. Dong H, Ding X, Wu M, Shi Y, Qin Y, Chu L (2014) Urban traffic commuting analysis based on mobile phone data. In: *Proceedings of the 17th international IEEE conference on intelligent transportation systems*, pp 605–610
23. Wu M, Dong H, Ding X, Shan Q, Chu L, Li-min J, Qin Y (2014) Traffic semantic analysis based on mobile phone base station data. In: *Proceedings of the 17th international IEEE conference on intelligent transportation systems*, pp 611–616
24. Ohasi H, Akiyama T, Yamato M, Sato A (2014) Trip-separation method using sensor data continuously collected by smartphone. In: *Proceedings of the 17th international IEEE conference on intelligent transportation systems*, pp 2966–2972
25. Ohashi H, Akiyama T, Yamamoto M, Sato A (2013) Modality classification method based on the model of vibration generation while vehicles are running. In: *Proceedings of the sixth ACM SIGSPATIAL international workshop on computational transportation science*, ACM, 2013, p 37
26. Hui P, Chaintreau A, Scott J, Gass R, Crowcroft J, Diot C (2005) Pocket switched networks and human mobility in conference environments. In: *Proceedings of the 2005 ACM SIGCOMM workshop on delay-tolerant networking, WDTN '05*, New York, NY, USA, ACM, p 244251
27. Chaintreau A, Hui P, Crowcroft J, Diot C, Gass R, Scott J (2005) Pocket switched networks: real-world mobility and its consequences for opportunistic forwarding. Technical report, Technical Report UCAM-CL-TR-617, University of Cambridge, Computer Laboratory
28. Ukkusuri SV, Zhan X, Sadri AM, Ye Q (2014) Exploring crisis informatics using social media data: a study on 2013 oklahoma tornado. In: *Transportation research board 93rd annual meeting*, number 14–2099



29. Vieweg S, Palen L, Liu SB, Hughes AL, Sutton J (2008) Collective intelligence in disaster: an examination of the phenomenon in the aftermath of the 2007 virginia tech shootings. In: Proceedings of the information systems for crisis response and management conference (ISCRAM)
30. Hughes AL, Palen L, Sutton J, Liu SB, Vieweg S (2008) Sight-seeing in disaster: an examination of on-line social convergence. In: Proceedings of the information systems for crisis response and management conference (ISCRAM)
31. Qu Y, Wu PF, Wang X (2009) Online community response to major disaster: a study of tianya forum in the 2008 sichuan earthquake. In: System sciences, 2009. HICSS'09. 42nd Hawaii international conference on IEEE, pp 1–11
32. Qu Y, Huang C, Zhang P, Zhang J (2011) Microblogging after a major disaster in china: a case study of the 2010 yushu earthquake. In: Proceedings of the ACM 2011 conference on computer supported cooperative work, ACM, pp 25–34
33. Vieweg S, Hughes AL, Starbird K, Palen L (2010) Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In: Proceedings of the SIGCHI conference on human factors in computing systems, ACM, pp 1079–1088
34. Pereira FC, Rodrigues F, Ben-Akiva M (2014) Using data from the web to predict public transport arrivals under special events scenarios. *J Intell Transp Syst*, 1–16
35. Gkiotsalitis K, Alexandrou A (2014) Mobility demand prediction in urban scenarios through multi-source, user-generated data. In: 18th Pan-American conference of traffic and transportation engineering and logistics, PANAM 2014, June 2014
36. Gkiotsalitis K, Alesiani F, Baldessari R (2014) Educated rules for the prediction of human mobility patterns based on sparse social media and mobile phone data. In: Transportation research board: 93rd annual meeting, pp 14–745
37. Alesiani F, Gkiotsalitis K, Baldessari R (2014) A probabilistic activity model for predicting the mobility patterns of homogeneous social groups based on social network data. In: Transportation research board: 93rd annual meeting, pp 14–1013
38. Pelletier MP, Trpanier M, Morency C (2011) Smart card data use in public transit: a literature review. *Transp Res Part C Emerg Technol* 19(4):557–568
39. Wilson NHM, Zhao J, Rahbee A (2009) The potential impact of automated data collection systems on urban public transport planning. *Sched-Based Model Transp Netw* 46:1–25
40. Chatterjee A, Venigalla MM (2011) Travel demand forecasting for urban transportation planning. In: Handbook of transportation engineering, volume i: systems and operations
41. Smith C, Quercia D, Capra L (2012) Anti-gravity underground? In: Proceedings of the second workshop on pervasive urban applications. PURBA '12
42. Ceapa I, Smith C, Capra L (2012) Avoiding the crowds: understanding tube station congestion patterns from trip data. In: Proceedings of the ACM SIGKDD international workshop on urban computing, ACM, pp 134–141
43. Morency C, Trepanier M, Agard B (2007) Measuring transit use variability with smart-card data. *Transp Policy* 14(3):193–203
44. Jang W (2010) Travel time and transfer analysis using transit smart card data. *Transp Res Rec J Transp Res Board* 2144(1):142–149
45. Bagchi M, White PR (2005) The potential of public transport smart card data. *Transp Policy* 12(5):464–474
46. Foell S, Phithakkintunukoon S, Kurtuem G, Bento C (2014) Catch me if you can: predicting mobility patterns of public transport users. In: Proceedings of the 17th international IEEE conference on intelligent transportation systems, pp 1983–1990
47. Ivanchev J, Aydt H, Knoll A (2014) Stochastic bus traffic modelling and validation using smart card fare collection data. In: Proceedings of the 17th international IEEE conference on intelligent transportation systems, pp 1983–1990
48. Zhou X, Mahmassani HS (2006) Dynamic origin-destination demand estimation using automatic vehicle identification data. *IEEE Trans Intell Transp Syst* 7(1):105–114
49. Baek S, Lim Y, Rhee S, Choi K (2010) Method for estimating population OD matrix based on probe vehicles. *KSCE J Civil Eng* 14(2):231–235

50. Schuessler N, Axhausen KW (2009) Processing raw data from global positioning systems without additional information. *Transp Res Rec J Transp Res Board* 2105(1):28–36
51. Zheng Y, Chen Y, Li Q, Xie X, Ma W-Y (2010) Understanding transportation modes based on gps data for web applications. *ACM Trans Web (TWEB)* 4(1):1
52. Brunauer R, Hufnagl K, Rehr K, Wagner A (2013) Motion pattern analysis enabling accurate travel mode detection from gps data only. In: *Proceedings of the 16th international IEEE conference on intelligent transportation systems*
53. Stenneth L, Wolfson O, Yu PS, Xu B (2011) Transportation mode detection using mobile phones and gis information. In: *Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems*, ACM, pp 54–63
54. Nitsche P, Widhalm P, Breuss S, Brändle N, Maurer P (2014) Supporting large-scale travel surveys with smartphones-a practical approach. *Transp Res Part C Emerg Technol* 43( 2):212–221
55. Feng T, Timmermans HJP (2013) Transportation mode recognition using gps and accelerometer data. *Transp Res Part C Emerg Technol* 37:118–130
56. Wu J, Jiang C, Houston D, Baker D, Delfino R (2011) Automated time activity classification based on global positioning system (gps) tracking data. *Environ Health* 10:101
57. Hato E (2010) Development of behavioral context addressable loggers in the shell for travel-activity analysis. *Transp Res Part C Emerg Technol* 18(1):55–67
58. Li Q, Zheng Y, Xie X, Chen Y, Liu W, Ma W-Y (2008) Mining user similarity based on location history. In: *Proceedings of the 16th ACM SIGSPATIAL international conference on advances in geographic information systems*, ACM, p 34
59. Bohte W, Maat K, Quak W (2008) A method for deriving trip destinations and modes for gps-based travel surveys. *Res Urb S* 1(1):127–143
60. Chen C, Gong H, Lawson C, Bialostozky E (2010) Evaluating the feasibility of a passive travel survey collection in a complex urban environment: lessons learned from the new york city case study. *Transp Res Part A Policy Pract* 44(10):830–840
61. Li M, Dai J, Sahu S, Naphade M (2011) Trip analyzer through smartphone apps. In: *Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems*, ACM, pp 537–540
62. Witayangkurn A, Horanont T, Ono N, Sekimoto Y, Shibasaki R (2013) Trip reconstruction and transportation mode extraction on low data rate gps data from mobile phone. In: *Proceedings of the international conference on computers in urban planning and urban management (CUPUM 2013)*
63. Lewis K, Kaufman J, Gonzalez M, Wimmer A, Christakis N (2008) Tastes, ties, and time: a new social network dataset using facebook. *com. Soc Netw* 30(4):330–342
64. Golder SA, Wilkinson DM, Huberman BA (2007) Rhythms of social interaction: messaging within a massive online network. In: *Communities and technologies*, Springer, pp 41–66
65. Huberman BA, Romero DM, Wu F (2008) Social networks that matter: Twitter under the microscope. [arXiv:0812.1045](https://arxiv.org/abs/0812.1045) (arXiv preprint)
66. Gkiotsalitis K, Stathopoulos A (2015) A utility-maximization model for retrieving users' willingness to travel for participating in activities from big-data. *Transp Res Part C Emerg Technol*. Available at <http://www.sciencedirect.com/science/article/pii/S0968090X14003568>
67. Zhu Y, Zhang Y, Shang W, Zhou J, Ying C (2009) Trajectory enabled service support platform for mobile users' behavior pattern mining. In: *Mobile and ubiquitous systems: networking and services, MobiQuitous, 2009. MobiQuitous' 09. 6th annual international, IEEE*, pp 1–10
68. Gkiotsalitis K, Stathopoulos A (2014) A mobile application for real-time multimodal routing under a set of users preferences. *J Intell Transp Syst (ahead-of-print)*:1–18
69. Berndt EK, Hall BH, Hall RE, Hausman J (1974) Estimation and inference in nonlinear structural models. *Ann Econom Soci Meas NBER* 3(4):103–116

# Air Traffic Flow Management Data Mining and Analysis for In-flight Cost Optimization

Leonardo L.B.V. Cruciol, Li Weigang, John-Paul Clarke and Leihong Li

**Abstract** As the air traffic volume has increased significantly over the world, the great mass of traffic management data, named as Big Data, have also accumulated day by day. This factor presents more opportunities and also challenges as well in the study and development of Air Traffic Management (ATM). Usually, Decision Support Systems (DSS) are developed to improve the efficiency of ATM. The main problem for these systems is the data analysis to acquisition sufficient knowledge for the decision. This paper introduces the application of the methods of Data Mining to get the knowledge from air traffic Big Data in management processes. The proposed approach uses a Bayesian network for the data analysis to reduce the costs of flight delay. The process makes possible to adjust the flight plan such as the schedule of arrival at or departure from an airport and also checks the airspace control measurements considering weather conditions. An experimental study is conducted based on the flight scenarios between Los Angeles International Airport (LAX) and Miami International Airport (MIA).

## 1 Introduction

Air Traffic Management (ATM) is a complex process involving many attributes with on-line operation. Moreover, it is a chain with various factors that impacts the environment. A wrong or not previously evaluated decision in an interval could

---

L.L.B.V. Cruciol · L. Li · J.-P. Clarke  
Georgia Institute of Technology, Atlanta, Georgia, USA  
e-mail: leocruciol@gmail.com

L. Li  
e-mail: leihong.li@gatech.edu

J.-P. Clarke  
e-mail: johnpaul@gatech.edu

L. Weigang (✉)  
TransLab, University of Brasília, Brasília, DF, Brazil  
e-mail: weigang@unb.br

generate unexpected or unknown results in future instants. Hence, ATM is time contingent. Air traffic controllers do not have enough time to discover, analyze and evaluate potential impacts of previous decisions. Using well developed decision support system (DSS), the suggested actions to air traffic controllers can improve the air traffic flow management, safety, and also reducing the operational costs, etc.

To illustrate the possible chained impacts could be cited the overloaded maneuvering area, remote boarding and landing when this procedure is possible at the airport, retention of flights at the origin airport to wait for the flight crew or while the delayed flights are properly accommodated within the available air traffic flow. As the need to hold or forward aircraft in flight or wait on the runway, the operational cost with fuel and crew is affected and causes circular waiting en route near the airport until get authorization to land, and others.

The development of knowledge management has influenced many areas. However, there are two opportunities to scientific community: how lead with an amount of data so big in real-time and achieve useful results; and with Big Data available how improve the real-time decision support systems using historical information.

In the last decade, there has been a large increase in the number of databases, especially the unstructured data. To discover useful knowledge from these data is the new task for the government organizations and enterprises. This great mass of data, called Big Data, is presented in ATM environment too, which are from air traffic control process, whether information and airlines. In ATM systems, the study of Big Data is with the focus on the following two aspects: (1) ATM creates a huge amount of digital data such as radar data, restrict measurements applied by air traffic controllers, communication between pilots and controllers, flight plans, etc. [1]; (2) ATM needs to use information from various data sources such as meteorological data, GPS guidance, historical monitor images, etc.

Some approaches can be integrated to solve ATM problems such as to reduce operational in-flight costs for the airlines and passengers, improve airspace management and control with safety and cheaper air traffic fluency and reduce impact of decisions in airspace scenarios. Nowadays, there are conditions, data and knowledge to be used as input for intelligent systems to support decision process in air traffic management. The increasing amount of historical data provides both opportunities and challenges to improve the decision support systems.

The DSS comes as a great tool in the whole ATM environment, which can support in the automation processes with quick and easy information to controllers by impact evaluations, prediction analysis, improve the control on chained processes, and others. An important point of success in this domain is the air traffic controller confidence about each suggestion made by the system.

Considering the proposed suggestions are based on historical information, it will improve the acceptance by specialists day-by-day and also the speed of knowledge acquiring by new controllers. Hence, the knowledge acquired by the controller is transferred to the knowledge base. So, DSS will learn, adapt and suggest more appropriate decisions based on historical actions applied. The learning process of the system can be accomplished either on daily tasks as with the previously acquired knowledge.

This paper presents an approach in which Big Data structures are used to compile an appropriated knowledge for DSS in a real-time manner. The presented approach is developed in two steps. First, Bayesian network is used to conduct data mining in Big Data; second, a prediction rule structure is constructed in a real-time environment. The two-step approach involves the proposal of adjustment in a flight plan such as schedule of arrival/departure airport and checking in airspace controls considering schedule and/or weather conditions. The approach is demonstrated with air traffic between Los Angeles International Airport (LAX) and Miami International Airport (MIA).

The paper is organized in the following structure. Section 2, briefly reviews relevant research and concepts of Data Mining, Big Data, Bayesian network and ATFM. Section 3 proposes a Data Mining model for ATFM. Section 4 presents the case study and results. Section 5 concludes the paper with summaries and the direction of future study.

## 2 Related Concepts

This section briefly describe the related concepts of Data Mining, Big data, Bayesian network and also Air Traffic Flow Management.

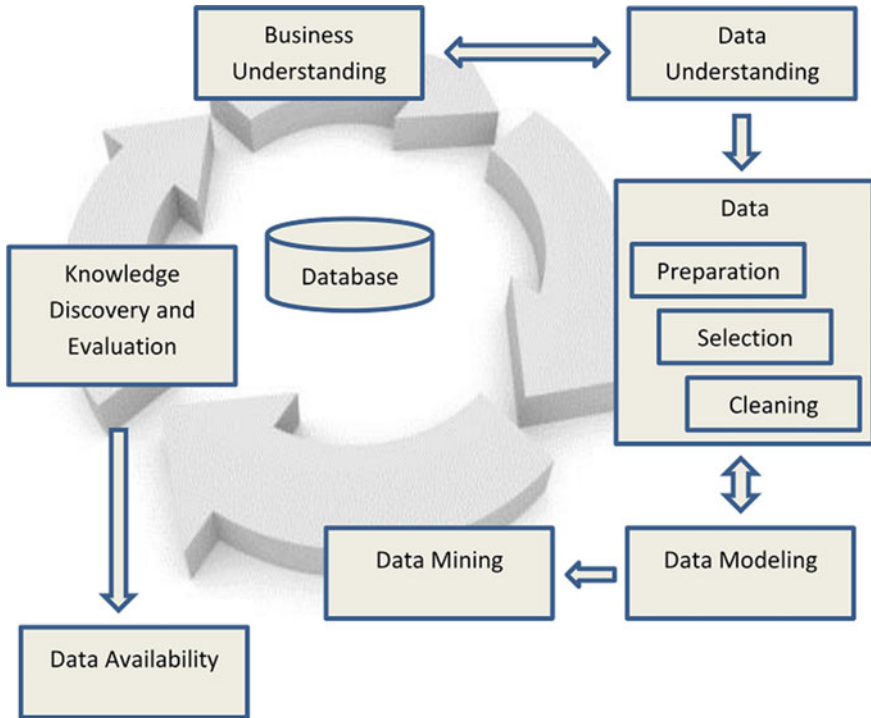
### 2.1 Data Mining

Data mining is a process that aims to discover useful patterns and correlations through historical data [2–4]. This technique makes possible to discover relationships between business attributes and understand its process to take better actions based on real and specific knowledge for each situation.

The steps of data mining can be summarized such as business and data understanding; data preparation, selection, cleaning and modeling; knowledge discovery and evaluation; and data availability for use of specialists and/or decision support systems.

The Fig. 1 presents a basic Data Mining flow. It is possible to verify the process interactivity, which it is continuously improved, i.e., each phase or whole process is repeated according to how satisfactory it was the results or looking for business improvements.

Through Data Mining process is essential that DM experts work together with business specialists to achieve a better understanding about the business particularities. These specialists will interpret the achieved results, support the data correlation process, and others. The DM process can be explained in the following six steps.



**Fig. 1** Data mining flow

1. Business Understanding: This process will discover which goals might be achieved. The initial analysis of available data can determine which strategy will be used to select interest variables, evaluate information subgroups that are needed to develop the relationship among the data, and others.
2. Data Understanding: This step will analyze the data structures which will be processed and its computational requirements to be handled. Considering Big Data structures, it is important to perform tests in a reduced case study to evaluate and demonstrate that achieved results are relevant. Therefore, it will be possible a better data and business understanding about how all data are related and reduces the effort and time to develop the DM model.
3. Data Preparation, Selection and Cleaning: Considering Big Data structures, this step is the longest and hardest to complete due some reasons such as many different data sources, notations, values and meanings. The cleaning process will deal with missing, errors, outliers and integrity of preloaded data. This data processing must consider the business goals and its relationship among data.
4. Data Modeling: Considering Big Data structures, this step will create some Data Marts, which it will make possible to handle easier the data. These new structures organize the data as a Data Mining goal, which it can exists two or more goals in

the same DM process. Each Data Mart has all necessary related information to achieve the results of DM process.

5. **Data Mining:** This step is responsible to discover useful information in databases by data mining process and generate knowledge for decision process [4–8]. It is possible to be more specific about the tasks and its results such as description, classification, prediction, group and/or link. DM is a general concept which it can use many strategies and approaches to execute chosen task [9, 10]. There is a high computational demand to process all information in this phase.
6. **Knowledge Discovery, Evaluation and Data Availability:** After the Data Mining step, the information discovered will need to be analyzed by a business specialist to judge and understand the achieved results. This evaluation will determine if the process will need to be repeated from some specific step or the whole DM process again. In the evaluation task, it is possible to use statistics methods to prove and explain some discovered information which it will base the decision process with more confidence.

## 2.2 *Big Data*

The amount of available data is so big in many companies and authorities that a special term Big Data is referred to those data. Basically, Big Data is historical and useful data. As the time accumulated, the scale of data is so big and relates to many ones in the society. Data Mining can be used as a powerful technique to analyze and learn with Big Data and make structures that could be used as input for decision support system.

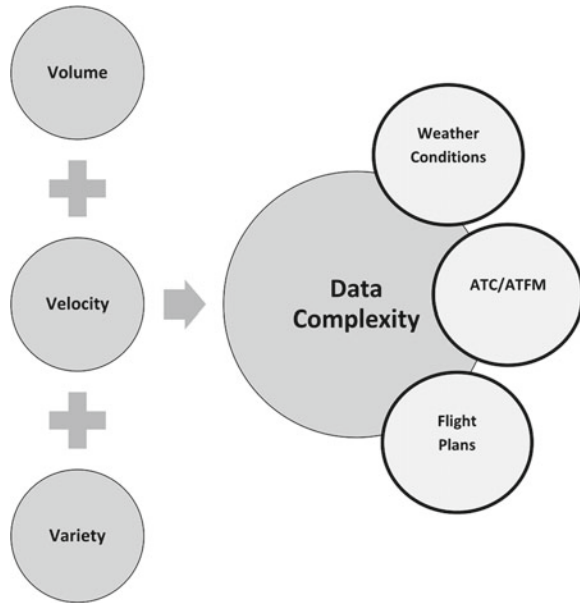
The major concern of using Big Data in real-time situations is how quick to achieve acceptable results. As it is necessary analyze so big and not-structured data or from many data sources, the DSS does not get to read these data and make available for the specialist in a real-time and critical environment, if it not be analyzed in a previous moment.

There is not a formal and unique Big Data definition for while. It can be described as a formal manner for knowledge discovery in so big data structures. Another way to explain this definition it is a manner used by companies to define strategies and tools to structure, handle, analyze and present the achieved results, expected or not, which it was discovered from big data structures. The complexity of analyzing big data is based on three factors: volume, velocity and variety, in order to base business specialists in their decision process.

- **Volume:** The size of data.
- **Velocity:** The speed of change in historical data.
- **Variety:** The number of data sources and how hard is to understand and merge.

These three factors make possible a better understanding about data fast increasing, variety about how these data are created, storage and made available for use and

**Fig. 2** Big data dimensions related with ATM



the impact of velocity on data that will be analyzed [11, 12]. Figure 2 presents the relationship of Big Data dimensions related with ATM.

To improve the data processing results, one manner is making the knowledge discovery process before the necessary time in two steps. First, it used a technique called Bayesian network that it will mining all available data and discover useful patterns and correlations. Second, it is created simple and fast structures to be used by decision support system in real-time. By this process, it will be created a prediction database, which contains rules identified by the first step ready to be read by DSS.

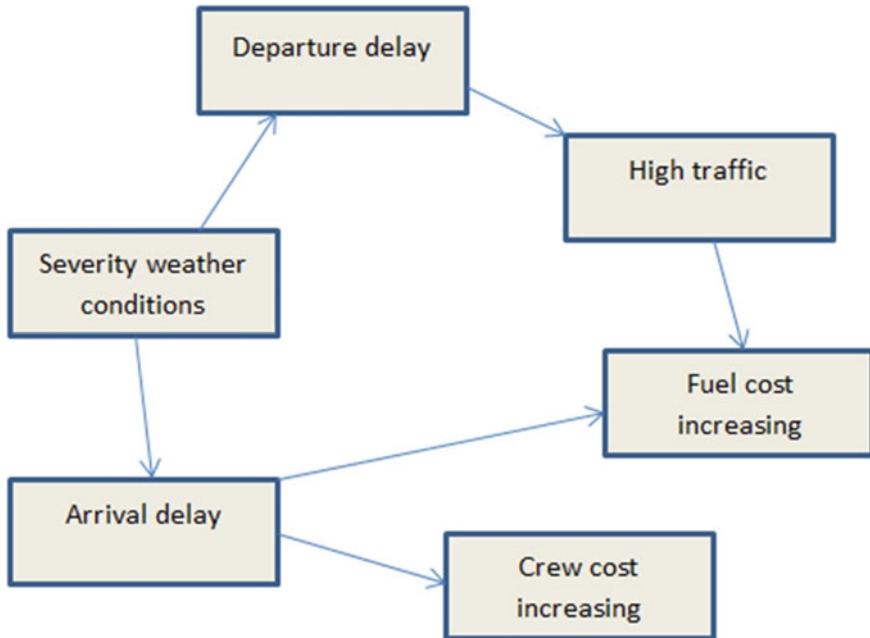
The use of historical data is an important step to improve and achieve a next step in decision support system. It is common using Data Mining techniques to acquire knowledge, however these data could be useful as input for other kind of systems, as it is proposed in this paper [13–15].

In the Air Traffic Management domain exists many opportunities to improve the decision support systems. Considering the critical real-time environment, the DSS suggestions might be clear and self-explained for air traffic controllers. Thus, the historical actions can improve the confidence of suggestions, once it is based on better similar historical actions.

### 2.3 Bayesian Network

Bayesian network is a structure which represent the correlations between attributes, in a specific domain, by using conditional probabilities [16–19]. Through these correlations are possible identify and understand how the domain is based on





**Fig. 3** Example of bayesian network

probability model and use this knowledge to take actions in similar situations. Figure 3 presents a basic example of how a Bayesian network could be constructed using ATM domain.

As it is detected probable association among variables and related uncertainty, Bayesian network arises as an important tool to identify and infer useful correlations which can be definitive, it usually happens due some aspects, or temporal, it was happening due some unusual environment.

To construct the network, it is necessary to perform a priori probability attribution for each correlation or use a learning algorithm. A Bayesian network is composed by following aspects [20]:

- Set of variables defined on a directed acyclic graph.
- The variables states are finite and mutually exclusive.
- For each variable  $X$ , with ascending  $Y_1, \dots, Y_n$ , There is a conditional probability associated in  $P(X - Y_1, \dots, Y_n)$ .

The Bayes' Theorem can be applied as a way to calculate the posterior probability distribution based on the product proportion of priori distribution and the similarity function [21].

The priori distribution is an ad-hoc probability associated based on usual events in the environment, so using the theorem is possible to get a normalized probability which will represent better the probability for data analyzed.

$$\Pr(A|B) = \frac{\Pr(B|A) \Pr(A)}{\Pr(B|A) \Pr(A) + \Pr(B|\neg A) \Pr(\neg A)} \quad (1)$$

where:

$\Pr(A|B)$  it is the posterior probability distribution

$\Pr(A)$  it is the priori probability distribution

$\Pr(B|A)$  it is the conditional probability

## 2.4 Air Traffic Flow Management

ATM focuses on providing means to manage air traffic, taking into consideration factors such as security, planning, justice, finance and meteorology [22, 23]. By ATM the airspace can be monitored, controlled and the aircraft flow can be managed in an integrated manner. The ATM environment can be divided into three sectors:

- Air Space Management: ASM focuses on increasing the capacity of aircraft in the airspace, with the purpose of provide sufficient services for demand within the available structure.
- Air Traffic Control: ATC focuses on controlling the aircraft flight, providing mandatory information to preserve the safety.
- Air Traffic Flow Management: ATFM focuses on providing information to maintain the air traffic flow with safety and reduced impact on future scenarios.

ATFM is a complex procedure to avoid exceeding air traffic capacity and focuses on the supply of information to maintain the traffic flow with safety and less impact on scenarios that are necessary to take unexpected actions. The ATFM environment can be organized into three phases:

- Strategic Level: Considering tactical planning of flights and covering the period of forty-eight hours until the time before the flight.
- Operational Level: Focusing on strategic decision making and covering the period from forty-eight to two hours before the flight.
- Tactical Level: Considering tactical decision making and covering the period from 2h before the flight until the aircraft arrives at its destination.

ATFM is responsible to assure aircraft traveling in a safe, quick, and economic way. It is responsible to avoid overloading facility capacity, optimize airspace usage, and provide information to responsible authority.

ATFM can guarantee that flights are conducted in a safe, quick, orderly and economic way. It is possible to avoid overloading in the air traffic capacity, optimize airspace and provide information to responsible authority [24–30].

Some activities from ATFM can be automate, partially or not, or improve using DSS. So, air traffic controller can monitor and analyze all aspects involved in the environment, such as meteorological aspects, evaluation of restrictive measures before to take some action, and verify alternatives for air traffic flows.

### 3 ATFM Data Mining Model

We proposed to use Big Data structures to discover an appropriated knowledge that is applied in real-time DSS for ATFM. The proposed method has two steps. First, Bayesian network is used to conduct data mining in Big Data; second, a prediction rule structure is constructed for real-time application environment. Figure 4 presents architecture that integrates Big Data analysis in decision support system for ATFM.

These rules are used to provide real-time knowledge for DSS in future times in order to reduce operating costs and increase safety with better-informed decisions. The proposed approach is performed in two stages: Preliminary Analysis and Data Analysis in Real Time.

It was developed a mechanism to discover patterns from historical information, considering big volume of data partial available. It aims to identify patterns on flights based on schedule, weather conditions, airports, and others, to use this information to conduct predictions.

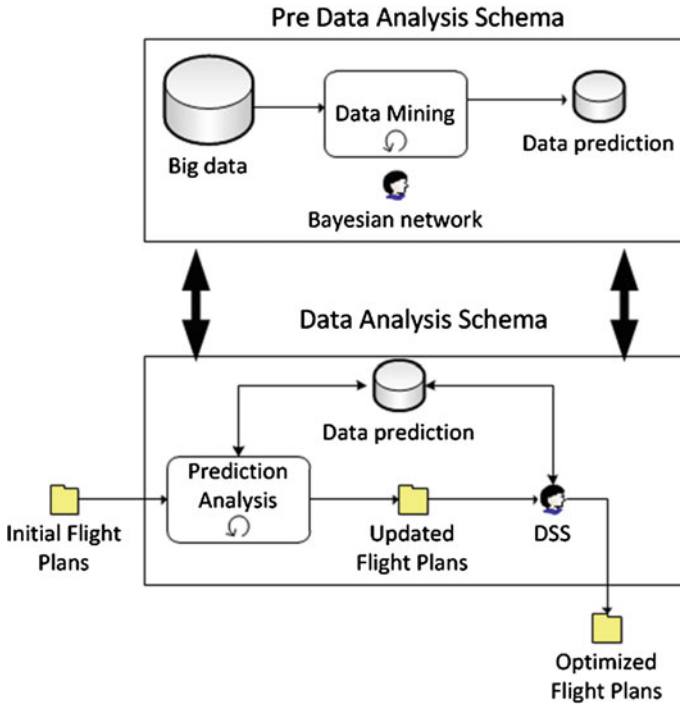


Fig. 4 ATFM data mining model overview

In the pre data analysis schema it is possible to develop a model which previous knowledge is analyzed offline with more time to create as many rules predictions as possible from data. During this first stage, Big Data structures are organized and analyzed to be handling by data mining process which uses Bayesian network to create this data prediction database. This step will be responsible for cleaning, organizing and structuring data and execute the processes of knowledge discovery to create forecasting rules, this will be used for Data Mining (DM) with the technique of Bayesian network. Data Analysis will use the knowledge generated in the previous step through the identified prediction rules. These structured rules will assist the decision making process of air traffic controller to be used, such as Multiagent Systems (MAS), Reinforcement Learning (RL) and Markov Decision (MDP) Processes.

Thus, it will be stored prediction rules based on historical information. The major objective is to discovery patterns that could be used to improve suggestions from DSS to airspace controllers. The model will combine the prediction rules and flight plans in a DSS simulated environment and suggest integrated actions to better decisions, i.e., considering the smallest impact on future scenarios. The smallest impact will be based on safety and reduce operating costs by improving the knowledge acquisition process in Big Data structures by own adaptation of decision support systems in real-time environments and improving on air traffic flow management.

At the end of this process, it had been created a data prediction database. Second phase will receive a group of flight plans to be analyzed in real-time environment. The prediction analysis process will verify the initial flight plans and compare with data prediction database, which it had been stored prediction rules based on acquired knowledge from Data Mining process.

At this moment, it generated updated flight plans that it could be used as input for DSS, which it can search more rules and create optimized flight plan as output from process.

The second phase aims to discovery similar situations and its variables looking for possible correlations between current situations and probability of achieve the results again. Considering this correlation, it is possible to select some actions as suggestions for airspace controllers, which will store all decisions taken to improve suggestions and learn with specialists.

## 4 Case Study

An experimental study is demonstrated with air traffic between Los Angeles International Airport (LAX) and Miami International Airport (MIA). This experiment is to identify patterns and correlations between departing and arriving flights in those two airports, and to update flight plans so that delay costs associated with extra crew hours and fuel burned.

More than 600,000 flights between 50 US airports were processed for this research, it was generated 25 tables with 25 GB of data. 719 flights from available data were analyzed. The major objective of this experiment is to create a schema that works and to achieve great results from a piece of big data, due the high cost of processing in this kind of structures. In this case study a big data structure was studied, however tests with a really big amount of data will be studied in future works.

The software were chosen with WEKA as it is one of the most popular applications in Data Mining area and the UnBBayes to generate the network. The task was association with Apriori algorithm, minimum support equals 90 % and minimum confidence equals 80 %. It was identified 42 attributes and chosen 7 to be used in first study: temperature, original departure, estimated departure, published departure, original arrival, estimated arrival and published arrival.

## 4.1 Results

The first study achieved promising results by identifying 6 rules from database structure available. These rules will compose data prediction database, which will be responsible to provide fast knowledge for DSS in real-time.

1. Original arrival between 5 and 7:30 pm and published departure delayed between 3 and 9 min, the published arrival delayed between 3 and 5 min in 64 % of cases.
2. Temperature is lower 55 F in arrival airport and published departure is delayed more than 4 min, the published arrival time increase about 20 %.
3. Estimated arrival is between 7 and 11 am and published departure was delayed until 7 min, the estimated arrival will be same as estimated arrival in 59 % of cases.
4. Temperature is lower 40 F in departure airport, the published departure increase more than 5 min in 27 % of cases.
5. Original departure between 6 and 8:15 pm, the flight period increases about 4 min in 70 % of cases.
6. Published arrival is delayed until 6 min from original arrival and temperature is higher 62 F, the aircraft will arrive in original time.

When initial flight plans are inside prediction analysis process, it will be evaluated if some flight plans match with some rule. In positive case, it will be adjusted by creating updated flight plans. This will be used in decision support system as suggestions for airspace controller verify and compare with original plan and take needed actions based on previous knowledge.

Considering these relationships the Bayesian network was developed. Figure 5 presents the influence correlation between each attribute. It is possible to verify that all attributes are much related, and this is a point that confirms a great chosen of attributes but this could limit more important and different patterns to be discovered.

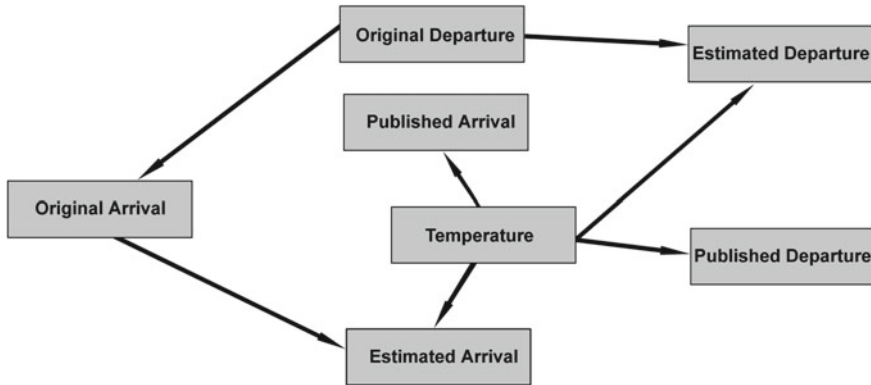


Fig. 5 Bayesian network

### 5 Conclusions

As the air traffic controller daily tasks are complex, there is necessary to provide decision support systems that could assists and provide suggestions and knowledge data to support their decisions. Nowadays, there are big amount of historical data which can be used to improve the decision process. Thus, it is possible to learn with this history, identify useful patterns, and make forecasts based on statistical events.

This proposal for ATFM domain is an experiment research to model these complex attributes and variables, which aims to create a fast process of data clean and load; data mining process that could identify attributes correlated as the influence of temperature, wind speed and forward a delayed flight in order to reduce operational costs; create a robust schema of predictions rules to support DSS operation; Based on these steps, the statistic makes better suggestions and adjustment in flight plans initially defined.

From the results achieved initially, it promises how to evaluate the complex process and to get the solutions for this kind of applications. It was identified 6 rules in the first study that it will be used to reduce in-flight costs considering costs of fuel and crew. When it is identified that some weather conditions repeats with great probability, the air traffic controller could take actions to make previous adjustment before aircraft take off, which will reduce many related risks.

The next steps for this research include more attributes for Bayesian network to identify new useful patterns and correlations, improve tests about minimum confidence and support to catch more possible patterns, increase amount of airports and flights, include more attributes and reports from METAR and TAF, and others. The DSS to support this kind of task will be developed based on presented approach, and used Reinforcement Learning and Multiagent System to model this approach. Moreover it will be created functions related with effectively crew and fuel cost to

verify the financial impact of delays. Also, include the knowledge of the aircraft manufacturers, e.g., the speed at which the aircraft must fly at a certain altitude to have a great fuel consumption.

**Acknowledgments** This work has been partially supported by the Brazilian National Council for Scientific and Technological Development - CNPq by the processes of No. 304903/2013-2 and No. 232494/2013-4. This paper is dedicated to the memory of Professor M.G. Karlaftis for his friendship and professional exemplar to the community.

## References

1. Pozzi S, Valbonesi C, Beato V, Volpini R, Giustizieri FM, Lieutaud F, Licu A (2011) Safety monitoring in the age of big data. In: Ninth USA/Europe air traffic management research and development seminar (ATM2011)
2. Chung HM, Gray P (1999) Special section: data mining. *J Manage Inf Syst* 16(1):11–17
3. Agrawal R, Shafer JC (1996) Parallel mining of association rules. *IEEE Eng Med Biol Mag Trans Knowl Data Eng* 8:962–969
4. Fayyad U, Piatetsky-Shapiro G, Smith P, Uthurusamy R (1996) Advances in knowledge discovery and data mining. In: Association for the advancement of artificial intelligence conference (AAAI). MIT Press
5. Berry MJA, Linoff G (1997) *Data mining techniques*. Wiley, New York (1997)
6. Groth R (1998) *Data mining*. Prentice Hall, Saddle River
7. Goebel M, Gruenwald L (1999) A survey of data mining and knowledge discovery software tools. Association for computing machinery's special interest group on knowledge discovery and data mining (SIGKDD) explorations
8. Hand D, Mannila H, Smyth P (2001) *Principles of data mining*. MIT Press, Cambridge
9. Schaffer C (1994) A conservation law for generalization performance. In: The 1994 international conference on machine learning. Morgan Kaufmann
10. Kibler D, Langley P (1988) Machine learning as an experimental science. In: Proceedings of the third European working session on learning. Glasgow Pittman, vol 1, pp 81–92
11. Laney D (2014) 3D data management: controlling data volume, velocity, and variety. Meta Group (2001) Available via Gartner Group. <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>. Accessed 10 Jul 2014
12. Laney D (2014) The importance of 'big data': a definition. (2012) Available via Gartner Group. <https://www.gartner.com/doc/2057415/importance-big-data-definition>. Cited. Accessed 03 Jul 2014
13. Pozzi S, Valbonesi C, Beato V, Volpini R, Giustizieri FM, Lieutaud F, Licu A (2011) Safety monitoring in the age of big data: from description to intervention. In: Ninth USA/Europe air traffic management research and development seminar (ATM2011)
14. Lavalley S, Hopkins MS, Lesser E, Shockley R, Kruschwitz N (2010) Big data, analytics and the path from insights to value. *MIT Sloan Manage Rev*
15. Pozzi S, Lotti G, Matrella G, Save L (2008) Turning information into knowledge: the case of automatic safety data gathering. EUROCONTROL annual safety R&D seminar
16. Jordan MI (2007) *Learning in graphical models*. SAE technical paper, MIT Press
17. Pearl J (1987) Evidential reasoning using stochastic simulation of causal models. *Artif Int* 32(2):245–258
18. Ye X, Kamath G, Osadciw LA (2009) Using bayesian inference for sensor management of air traffic control systems. In: Computational intelligence in multi-criteria decision-making (MCDM), pp 23–29

19. Han S, DeLaurentis D (2011) Air traffic demand forecast at a commercial airport using bayesian networks. In: 11th AIAA aviation technology, integration and operations (ATIO) conference, Virginia Beach, VA
20. Jensen FV (2001) Bayesian networks and decision graphs. Springer, Berlin
21. Alba E, Mendoza M (2007) Bayesian forecasting methods for short time series. *Int J Appl Forecast* 8:41–44
22. Agogino A, Tumer K (2009) Learning indirect actions in complex domains: action suggestions for air traffic control. *Adv Complex Syst* 12(4–5):493–512 (World Scientific Company)
23. Agogino A, Tumer K (2008) Regulating air traffic flow with coupled agents. *Advances in complex systems*. In: Proceedings of 7th international conference on autonomous agents and multiagent systems
24. DECEA—Air Traffic Control Department of the Brazilian Air Force: Regras do ar e serviços de tráfego aéreo: ICA 100–12 (2012). Available via DECEA. <http://publicacoes.decea.gov.br/?i=publicacao&id=2558>. Accessed 19 Jun 2014
25. Piatetsky-shapiro G, Brachman R, Khabaza T, Kloesgen W, Simoudis E (1996) An overview of issues in developing industrial data mining and knowledge discovery applications. In: Proceedings of knowledge discovery in databases 96. AAAI Press, Menlo
26. Cheng T, Cui D, Cheng P (2003) Data mining for air traffic flow forecasting: a hybrid model of neural network and statistical analysis. In: Proceedings 2003 IEEE intelligent transportation systems, vol 1, pp 211–215
27. Weigang L, Dib MVP, Cardoso DA (2004) Grid service agents for real time traffic synchronization. In: Proceedings of the 2004 IEEE/WIC/ACM international conference on web intelligence, pp 619–623
28. Kulkarni D (2007) Integrated use of data mining and statistical analysis methods to analyze air traffic delays. SAE technical paper
29. Crespo AMF, Weigang L, Barros A (2012) Reinforcement learning agents to tactical air traffic flow management. *Int J Aviat Manage* 1(3):145–161
30. Zanin M, Perez D, Kolovos D, Paige R, Chatterjee K, Horst A, Rumpe B (2011) On demand data analysis and filtering for inaccurate flight trajectories. In: Proceedings of the SESAR innovation days, EUROCONTROL



# Simulation Optimization of Car-Following Models Using Flexible Techniques

Vasileia Papathanasopoulou and Constantinos Antoniou

**Abstract** Car-following behavior is a key component of microscopic traffic simulation. Numerous models based on traffic flow theory have been developed for decades in order to represent the longitudinal interactions between vehicles as realistically as possible. Nowadays, there is a shift from conventional models to data-driven approaches. Data-driven methods are more flexible and allow the incorporation of additional information to the estimation of car-following models. On the other hand, conventional car-following models are founded on traffic flow theory, thus providing better insight into traffic behavior. The integration of data-driven methods in applications of intelligent transportation systems is an attractive perspective. Towards this direction, in this research an existing data-driven approach is further validated using another training dataset. Then, the methodology is enriched and an improved methodological framework is suggested for the optimization of car-following models. Machine learning techniques, such as classification, locally weighted regression (loess) and clustering, are innovatively integrated. In this chapter, validation of the proposed methods is demonstrated on data from two sources: (i) data collected from a sequence of instrumented vehicles in Naples, Italy, and (ii) data from the NGSIM project. In addition, a conventional car-following model, the Gipps' model, is used as reference in order to monitor and evaluate the effectiveness of the proposed method. Based on the encouraging results, it is suggested that machine learning methods should be further investigated as they could ensure reliability and improvement in data driven estimation of car-following models.

---

V. Papathanasopoulou (✉) · C. Antoniou  
Laboratory of Transportation Engineering, National Technical University of Athens,  
15780 Zografou, Greece  
e-mail: vasileia.papathanasopoulou@gmail.com

C. Antoniou  
e-mail: antoniou@central.ntua.gr

## 1 Introduction

Simulation models play an important role in traffic engineering and recently in the development of Intelligent Transportation Systems [45]. They are divided into microscopic, mesoscopic and macroscopic models according to the modeling detail. Microscopic models describe in high level of detail interactions between individual vehicles, including interactions between vehicles and roads as well [11]. They consist of lane changing, gap-acceptance, overtaking, speed adaptation, ramp merging and car-following models [57]. On the other hand, macroscopic models represent traffic states in a lower level of detail using aggregated variables (traffic flow, density, speed) and theories of fluid dynamics [15]. Mesoscopic models provide an intermediate level of detail using speed-density relationships and queuing models [15].

Appropriate models are chosen according to the requirements of each application. This research is directed at microscopic traffic simulation, which gives the opportunity of detailed analysis required in the development of Intelligent Transportation Systems. Focusing on optimization of car-following models and the key elements of microscopic simulation [7, 16, 45] an alternative methodological framework is suggested. Car-following models generally represent driving behavior influenced by the preceding vehicle moving in the same lane so as a crash to be avoided. According to [57], they are grouped into categories such as Gazis-Herman-Rothery models [33], safe distance models [34, 44], psycho-physical models [32, 87], and fuzzy logic models [2, 43].

Over the years, many researches have been demonstrated aiming at the optimization of car-following models. Recently, it has been clarified that driving behavior varies in different traffic conditions, such as free-flowing, approaching, emergency braking, and stop-and-go [1, 45, 75, 85, 88]. Therefore, there has been a shift from single state models [61, 67] to more flexible models. The lack of models capable of capturing various traffic states and correspondingly various driving behaviors has led to the development of multi-regime approaches [48]. Nowadays, the generalization of these multi-regime approaches is a challenge issue.

Restrictions, related to the number of regimes and their complexity, have been the motivation for this research on estimation of car-following models. An alternative methodology based on data-driven approaches is proposed; actually an existing methodology has been modified to address these problems. Data-driven methods have been already used in applications in the field of transportation (e.g. [5, 26, 81–84]). These methods are more flexible than conventional models and allow the incorporation of additional information. The development of data-driven methods has also been benefited from technological advancements such as differential GPS and real time kinematic, which allow the collection and the availability of high quality traffic data [3, 66].

In this chapter, an existing methodology based on a machine learning method is further validated and enriched for optimization of car-following models. The historical background of car-following models and the development of data-driven approaches is first presented. The existing methodology is applied to a number of

available NGSIM data sets and the different nature of data is discussed regarding the impact on the efficiency of the method. In addition, the existing methodology is further extended and improved for the development of more reliable car-following models. The revised innovative methodology integrates data-driven methods such as loess method, clustering and classification, and is validated to Naples data.

## 2 Historical Background

A historical review of car-following models has been performed by Brackstone and McDonald [16]. Reuschel [67] and Pipes [61] introduced the idea of car-following models. Representative microscopic traffic models between the 1950s and the 1970s have been developed by Bifulco et al. [14], Chandler et al. [17], Colombaroni and Fusco [23], Kometani and Sasaki [44] and Zhang et al. [90], Herman et al. [40], [27, 35, 55, 79]. Most of them are defined by an acceleration function, which includes the difference of position  $x_{i+1} - x_i$  and the difference of speed  $v_{i+1} - v_i$  between a vehicle  $i$  and its lead vehicle  $i + 1$ : the difference of position  $x_{i+1} - x_i$  and the difference of speed  $v_{i+1} - v_i$ . Other models have been developed including only one variable such as the difference of speed [17, 90] or the difference of position [35]. Gazis et al. [33] proposed a General Motors model (GM) with doubtful efficiency both in low and high-speed networks [48]. Several extensions to the GM framework followed [13]. Leutzbach [47] and Wiedemann [86] introduced psycho-physical models in order to address restrictions of GM models. Wiedemann and Reiter [87] suggested that there are longitudinal interactions in four traffic states: free flowing, approaching, car-following and emergency situation.

After 1990, [76] identified a different tendency in car-following models due to technological advancements. New microscopic methods are considered as multi-agent and are defined by a system of differential equations, each of which captures a different state. Treiber et al. [78] clarified that reaction time and time steps should have various values in the simulation process. Gipps' model [34] is a safety distance model described by two speed equations correspondingly to free flowing and car following state [71]. In this research this model is used as a reference for the framework developed in this research. Rakha and Wang [65] tried to modify Gipps' model. A detailed analysis of the model evolution is presented by [18]. Bando et al. [8] and Bando et al. [9] developed a nonlinear model, the Optimal Velocity model, to deal with stop-and-go traffic states. Further research was performed later [24, 39, 42, 46, 58, 69, 91].

According to Subramanian, [45, 73], drivers' reaction time is differentiated under acceleration or deceleration conditions. Ahmed [1] suggested an acceleration model both for free-flowing and car-following situations. Newell [56] clarified that the trajectory of a vehicle depends on a time and a minimum distance of spacing. Treiber et al. [77] proposed the Intelligent Driver Model, which determines driver's acceleration in relation with the gap, the speed and the speed difference between a pair of vehicles moving in sequence. Aw et al. [6] proposed a new general model. Zhang and Kim [89] developed a multi-regime car-following model, which is determined by a gap-distance function and the traffic state. Hamdar and Mahmassani [36]

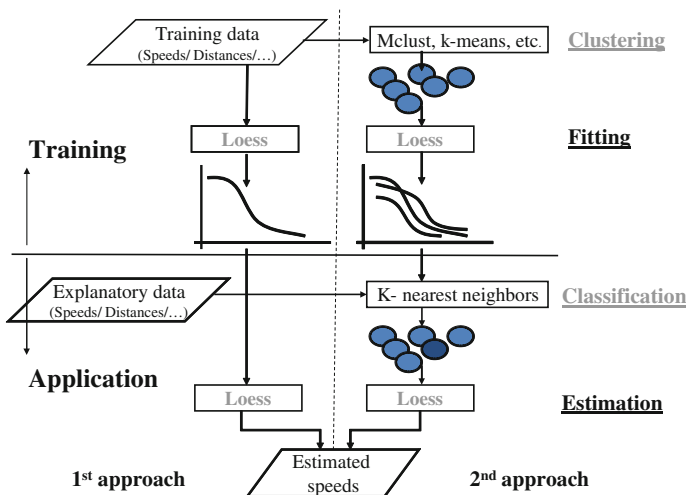
demonstrated calibration and validation of existing car-following models using NGSIM data. Tordeux et al. [76] proposed the impact of the vehicle type on driving behavior. Moreover, the assumption of the GM model that a driver will accelerate if the speed of the preceding vehicle is higher is re-examined.

More and more parameters and traffic states should be integrated in simulation process. This need has led to the development of multi regime models and by extension to data-driven approaches. A multimodal regression to speed-flow data has been performed by Einbeck and Tutz [28]. Sun and Zhou[73] used cluster analysis in order to determine the regime boundaries for traditional speed–density models. [4] suggested a data-driven approach as an alternative to the classic speed–density models. Zhang et al. [90] have demonstrated the use of machine learning methods to support the development of data-driven intelligent transportation system. Data-driven approaches have already been used in a fully adaptive cruise control system by [14] or in car-following modeling with artificial neural networks by Colombaroni and Fusco [23]. Finally [59] have performed a data-driven approach based on loess method for speed estimation using Naples data. This research is further extended in this chapter.

### 3 Methodology

#### 3.1 Methodological Framework

Two data-driven approaches are presented, outlined in Fig. 1. Regarding the first one approach is an existing method based on locally weighted regression (loess), which has been already proposed and analyzed in an earlier research [59]. The second data-



**Fig. 1** Overall methodology framework for data-driven estimation of car-following models with machine learning approaches

driven method is an extension and improvement of the earlier method and comprises a combination of computational methods, such as locally weighted regression, model-based clustering and classification.

Both methodological approaches include two parts: training and application. In the training step the estimation of car-following models is achieved using a training dataset with triples  $\langle v_i, v_{i-1}, d_{i,i-1} \rangle$  (leader and follower speed and their distance) per each time instant. The problem to be addressed is the speed estimation of the third vehicle when there are available the speeds of the preceding and the following vehicle and its distance from them. In the application process, when new observations arise, the appropriate calibrated models are retrieved from the knowledge base and are applied to provide speed predictions  $v_i$  for the following vehicle and the next time instant. The proposed methods rely on non-parametric approaches and do not include any fixed functional form. They might be considered as generalization of the multi-regime approaches [4, 5].

As concerns as the second methodological approach, it includes a clustering step to identify portions of the available data that correspond to traffic states with similar characteristics. Then, a locally weighted regression is applied to each cluster separately and representative models are formed for each group fitting to the data (fitting). The application step follows, when new measurements arise. New data are classified to the appropriate classes based on their characteristics. The flexible model that has been estimated for that class is then retrieved from the knowledge base and applied to the new data for the estimation of the speeds of the following vehicle.

The performance of the each approach is evaluated using the root-mean square error (RMSN) of speeds. This assesses the overall error of each method estimating the difference between the observed ( $Y^{obs}$ ) and simulated values ( $Y^{sim}$ ),  $N$  is the number of observations [41, 60]:

$$RMSN = \frac{\sqrt{N \cdot \sum_{n=1}^N (Y_n^{obs} - Y_n^{sim})^2}}{\sum_{n=1}^N Y_n^{obs}} \quad (1)$$

### 3.2 Methodological Components

*Locally weighted regression* could be considered as a generalization of the k-nearest neighbor method [53]. It was firstly introduced by Cleveland [19] and the following analysis is based on [20]. Locally weighted regression  $y_i = g(x_i) + \varepsilon_i$ , where  $i = 1, \dots, n$  index of observations,  $g$  is the regression function and  $\varepsilon_i$  are residual errors, provides an estimate  $g(x)$  of each regression surface at any value  $x$  in the  $d$ -dimensional space of the independent variables. Correlations between observations of the response variable  $y_i$  and the vector with the observations  $d$ -tuples  $x_i$  of  $d$  predictor variables are identified. Local regression provides an estimation of function  $g(x)$  near  $x = x_0$  according to its value in a particular parametric class. This estimation could be achieved by adapting a regression surface to the data points within

a neighborhood of the point  $x_0$ , which is bounded by a smoothing parameter: span  $\alpha$ . The span determines the percentage of data that are considered for each local fit and hence the smoothness of the estimated surface is influenced [22]. Each local regression uses either a first or a second degree polynomial that is specified by the value of the “degree” parameter of the method.

The data are weighted according to their distance from the center of neighborhood  $x$ , therefore a distance and a weight function are required. As a distance function  $p$ , Euclidean distance could be used for a single independent variable; otherwise, for the multiple regression case, any variable should be evaluated on a scale before applying a standard distance function [21].

A weight function defines the size of influence on fit for each data point taking for granted that nearby points have higher influence than the most distant. Therefore the weight function calculates the distances between each point and the estimation point and higher values in a scale from 0 to 1 are set for the nearest observations. A weight function should meet the requirements determined by Chandler et al. [17] and the most common one is the tri-cube function:

$$W(u) = \begin{cases} (1 - u^3)^3, & 0 \leq u \leq 1 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

The weight of each observation  $(y_i, x_i)$  is defined as following:

$$w_i(x) = W[p(x, x_i)/d(x)] = (1 - (\frac{x_i - x}{d(x)})^3)^3 \quad (3)$$

where  $d(x)$  is the distance of the most distant predictor value within the area of influence.

In the loess method, weighted least squares are used so as linear or quadratic functions of the independent variables could be fitted at the centers of neighborhoods [17]. The objective function that should be minimized is:

$$\sum_{i=1}^n w_i \cdot \varepsilon_i^2 \quad (4)$$

Fraley and Raftery [30, 31] suggest a model based *clustering* which combines hierarchical clustering, expectation-maximization algorithm (EM algorithm) for mixture models and Bayesian information Criterion (BIC) for selection of models and number of classes [70]. Hierarchical clustering, used for model-based hierarchical agglomeration, is initialized by default with each observation of the data in a cluster by itself and finished when all observations have been merged into a cluster. A classification maximum likelihood approach is required to determine which two groups are merged at each stage [10, 29, 52]. EM algorithm is included in the R Mclust package and is applied for maximum likelihood clustering with parameterized Gaussian

mixture models [25, 52]. The EM algorithm is implemented in two steps: E-step which calculates a matrix  $z_{ik}$ , which corresponds to the likelihood of an observation  $i$  to be merged into a cluster  $k$  given the current parameter estimates, and M-step, which calculates maximum likelihood parameter estimates given  $z$ . Each cluster is represented by a Gaussian model  $\phi_{\kappa}(x|\mu_{\kappa}, \Sigma_{\kappa})$ , where  $x$  are the data,  $k$  an integer indicating a cluster centered at means  $\mu_{\kappa}$  and covariances  $\Sigma_{\kappa}$ . Then the maximum likelihood values for the Gaussian mixture model is given by Eq. (5) [30], where  $\tau_{\kappa}$  are the mixing proportions.

$$\prod_{i=1}^n \sum_{k=1}^G \tau_{\kappa} \phi_{\kappa}(x_i/\mu_{\kappa}, \Sigma_{\kappa}) \quad (5)$$

Banfield and Raftery [10] suggested a clustering strategy based on a maximization algorithm and Bayes factors. This strategy was upgraded by Fraley [29], Fraley and Raftery [30, 31] and could be carried out with the following steps:

- A maximum number of clusters and a subset of covariance structures are considered
- A hierarchical agglomeration that maximizes the classification likelihood for each model is performed and the appropriate classifications are illustrated up to  $M$  groups.
- The EM algorithm is applied for each model and each number of clusters  $2, \dots, M$ . The procedure is initialized from the classification result of hierarchical agglomeration.
- The Bayesian information Criterion BIC is calculated for the one-cluster case for each model and for the mixture model with the optimal parameters from EM for  $2, \dots, M$  clusters. Each combination corresponds to a unique probability model.
- The model with the highest BIC is selected and the best classification is recovered. Although in such a way the optimal number of classes is determined, a lower number of classes could be chosen, aiming at the development of more parsimonious models.

Another clustering algorithm is k-means. As its name suggests, the k-means algorithm Hartigan [37], Hartigan and Wong [38] and MacQueen [49] minimizes the distance between each point and the center of its cluster for  $k$  given clusters. This is achieved by assigning each point to the nearest mean and re-estimating or moving the mean to the center of its cluster. It is regarded as a maximum likelihood clustering. The objective function to be minimized is:

$$\min_{\mu_1, \dots, \mu_k} \sum_{h=1}^k \sum_{x \in X_h} \|x - \mu_h\|^2 \quad (6)$$

where  $\mu_i$  is the mean of cluster  $i$

A hypothesis  $h_1 = \langle \mu_1, \dots, \mu_k \rangle$  with the means of the  $k$  different normal distributions is requested. A random hypothesis is assumed for the initialization of the procedure. Each instance could be written as  $\langle x_i, z_{i1}, z_{i2}, \dots, z_{ik} \rangle$  where  $x_i$  is

the observed variable and  $z_{ij}$  is equal to 1 if it was obtained by the  $j$ th normal distribution or 0 otherwise. A maximum-likelihood hypothesis is sought after iterative re-estimations of the expected values of  $z_{ij}$ . Then, a new maximum likelihood hypothesis  $h_2$  is calculated using the expected values in the previous step. Finally, the new hypothesis replaces the earlier one and iterations are going on until the algorithm converges to a value for the hypothesis.

One of the most common methods of *classification* is k-nearest neighbors [53]. According to this method, all observations correspond to points in n-dimensional space. Future data points are registered in the class of nearest neighbors of the already grouped data. Especially, the point of the nearest neighbor classification is the calculation of the correlation map:

$$f(z) = \arg \min_{y \in M} d(z, y) \quad (7)$$

In a pattern space  $P$ , where  $M \subseteq P$ ,  $z \in P$  and  $d()$  is a metric in  $P$ -dimensional space. The evaluation of Eq. (7) could be easily achieved on a computer following three steps: computation of an array with distances from  $z$  to each  $y \in M$ , finding the minimum distance after comparisons and exporting the final result  $y^* \in M$  [54].

The nearest neighbors could be defined according to the Euclidean distance [68], if a point  $x$  is described as  $\langle a_1(x), a_2(x), \dots, a_n(x) \rangle$  where  $a_r(x)$  corresponds to the value of the  $r$ th attribute of  $x$ . Attributes of  $x$  could include density, traffic flow, and time. The distance between two points is defined by Eq. 8 [53]. Thus the class of a new observation  $x_i$  is the same as the class of point  $x_j$ , which minimizes the distance  $\|x_i - x_j\|$ .

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^n [a_r(x_i) - a_r(x_j)]^2} \quad (8)$$

## 4 Experimental Set-Up

The data used in this survey are available from two sources: (i) an experiment carried out in Naples, Italy [63] and (ii) from the “Next Generation SIMulation (NGSIM) program” [80]. Naples Data are used for the validation of the second methodological approach, while NGSIM data for further validation of the first methodological approach.

### 4.1 Naples Data

A series of data-collection experiments were carried out on roads surrounding the city of Naples, in Italy [62]. All data were collected under real traffic conditions in



October 2002. Although traffic conditions and driving routes may be different in each dataset, the platoon consisted of four vehicles is unchanged regarding the vehicles, the drivers and the sequence. Datasets with index A, C correspond to one-lane urban road, while datasets with index B to a two-lane extraurban highway. However, all selected roads have one lane per direction in order to avoid effects on driving behavior by lane changing. GPS receivers located on the vehicles were recording the coordinates X, Y, Z of each vehicle per 0.1 s (in 10 Hz). Thus, the speed of each vehicle and the distances between each pair of vehicles could be calculated at each moment. The setup included five dual frequency GPS+GLONASS receivers (1 base station + 4 rovers) with expected accuracy in real time kinematic 10 mm + 1.0 ppm horizontally and 15 mm + 1.0 ppm vertically.

In this research, data used are readily available observations from the field. No corrections and no interpolation have been occurred. Therefore, only segments with consecutive measurements have been considered. Six data series were used, one for calibration and five for validation. A detailed description of the data could be found in [62], who kindly provided the data for this research.

## 4.2 NGSIM Data

The “Next Generation SIMulation (NGSIM)” program (<http://ngsim.fhwa.dot.gov>.) includes vehicle trajectories in real traffic conditions, which—along with other output of the project- have become available to the scientific community for research of microscopic driving behavior. As this data-set is rather different than the Naples data (different road type, vehicle fleet composition and driving population) it provides an opportunity to assess the transferability of the car-following models estimated on the Naples data.

The considered NGSIM data were collected on eastbound I-80 in the San Francisco Bay area in Emeryville on April 13, 2005 [80]. The study area extends approximately 500 m in length and consists of six freeway lanes. Seven modern digital cameras were mounted on the top of a 30-story-building adjacent to the freeway and were recording passing vehicles. The custom NG-VIDEO software application transformed video to vehicle trajectories data (also at 10 Hz). These data were recorded mainly in congested conditions. 45 min of data are available in a data set divided into three periods of 15 min and particularly in accordance with the register time, 4:00–4:15 pm, 5:00–5:15 pm, and 5:15–5:30 pm.

For each vehicle the available data which are taken into account are: vehicle ID, type of vehicle (only cars are taken into consideration), time (ms), global coordinate X (feet), global coordinate Y (feet), length of vehicle, vehicle velocity (feet/s), distance between the front side of a vehicle and the front side of the preceding vehicle, number of the preceding vehicle, number of the following vehicle, lane identification. (The data were converted to SI units prior to our application). Due to the large amount of available NGSIM data, 17 tetrads of vehicles moving consecutively were selected randomly for this analysis. The vehicles, which compose a tetrad, are considered only when they are moving in the same lane and in sequence one after the other.

This is easily recognizable from the lane identification and the number of preceding and following vehicles.

NGSIM data have been used in many studies for calibration or validation of existing models (e.g. [12]). In the years 2007–2008 more than 30 studies used the NGSIM data [63]. However, only few studies have raised the issue of their accuracy [36, 63, 74]. Although the way that the velocities and accelerations of vehicles were calculated and the errors were reduced is not known, studies suggest the existence of residual noise and errors in the data [12, 63]. In the context of this work the existence of noise in data is not addressed, presuming that if there are errors, they are included in both methods (model Gipps, proposed method) and therefore may not affect the comparison but the result of each method separately. Also, this implies that the presented approach can work directly with collected data, without requiring copious data-cleaning efforts.

### 5 Validation Results

The first methodological approach has been already demonstrated using Naples data by [59]. The authors have presented a sensitivity analysis both of Gipps’ model and Loess method and their calibration process as well. For Gipps’ model the following two combinations of parameters have been chosen as optimal:  $\tau = 0.4\text{ s}$ ,  $V_n = 14\text{ m/s}$ ,  $\alpha_n = 0.8\text{ m/s}^2$ ,  $s_{n-1} = 5.6\text{ m}$ ,  $b_n = -5.2\text{ m/s}^2$  and  $\hat{b} = -3.0\text{ m/s}^2$  or  $\tau = 1.0\text{ s}$ ,  $V_n = 16\text{ m/s}$ ,  $\alpha_n = 1.6\text{ m/s}^2$ ,  $s_{n-1} = 5.6\text{ m}$ ,  $b_n = -5.2\text{ m/s}^2$  and  $\hat{b} = -3.0\text{ m/s}^2$ . For Loess method degree = 1 and span = 0.75 have been specified.

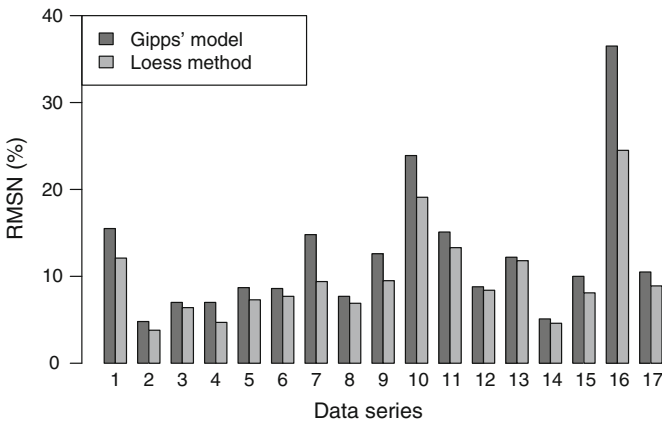
Both methods have been calibrated using the most representative data series B1695 and for the speed estimation the same factors have been used (speed  $v_2(t)$  and  $v_3(t)$  of vehicles 2 and 3 and distance  $D_{23}(t)$  between vehicles 2 and 3). The results encourage the application of the data-driven approaches and are summarized in Table 1. Loess method outperforms Gipps’ model for all the available data series.

**Table 1** Results for speed estimation for all Naples data sets using the first methodological approach

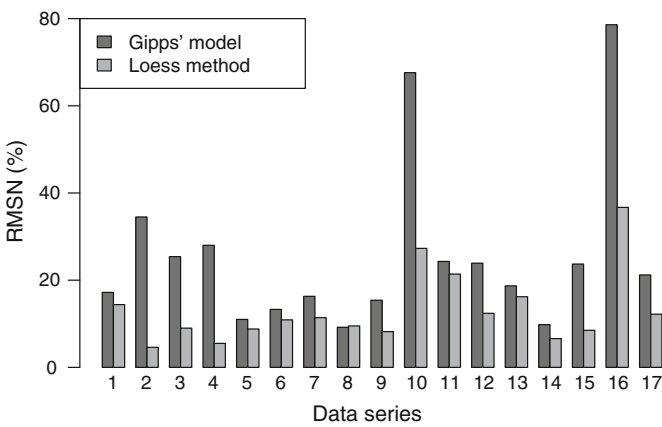
Data series	Reaction time $\tau = 0.4\text{ s}$			Reaction time $\tau = 1.0\text{ s}$		
	RMSN (%)		Improvement of estimation (%)	RMSN (%)		Improvement of estimation (%)
	Gipps’ model	Loess method		Gipps’ model	Loess method	
B1695	2.7	1.6	40.7	4.9	3.0	38.8
C621	6.6	4.3	34.8	14.4	6.7	53.5
A358	2.7	2.1	22.2	12.7	3.7	70.9
A172	4.6	3.4	26.1	16.0	6.3	60.6
C168	2.3	1.8	21.7	4.9	3.1	36.7
C171	7.2	6.2	13.9	31.6	6.7	78.8

The calibrated models are now validated to another data set from the US (NGSIM data) and it is demonstrated how the different nature of data affect the accuracy of speed estimation. NGSIM data and data from Naples are of different nature, as the former refer to freeway in congestion conditions and the latter to roads with one lane per direction. Moreover, as [51] suggested, differences between car-following headways and times-to-collision are identified between different sites. In this section, the transferability of the models estimated in Naples to the NGSIM data set is tested. Two models are presented: (i) Gipps', (ii) a loess model with the same data as those used by Gipps' model.

The results are presented in Figs. 2 and 3 for time reaction 0.4 s and 0.1 s accordingly and several observations can be drawn. As expected, the RMSN values are



**Fig. 2** Comparison of RMSN by applying Gipps' model and loess method for NGSIM data for reaction time  $\tau = 0.4$  s



**Fig. 3** Comparison of RMSN by applying Gipps' model and loess method for NGSIM data for reaction time  $\tau = 1.0$  s

higher than in the Naples data, as model calibration and validation/application was performed on dissimilar data. The proposed loess method seems to provide better results than Gipps' model. The machine learning approach seems to be more robust, while the effectiveness of the conventional car-following models may depend significantly on the chosen parameter values. Using additional data would be easy with the proposed data-driven model and improves the performance even further; on the other hand, reformulating Gipps' model to consider additional parameters would be a tedious exercise.

The degree by which the proposed approach outperforms the reference model varies across data series. In order to develop some insight into this, an exploration of the speed profiles of the various vehicles was performed.

Figure 4 presents the speed profile for the considered vehicle in the longest sequence of the Naples data-set (B1695) used for calibration, while Figs. 5 and 6 present similar speed profiles for data series that showed satisfactory performance (Fig. 5) and less satisfactory performance (Fig. 6). Data series with lower performance have high frequency of low speeds (0–2 m/s), reflecting congested conditions, while data series with higher speeds naturally provided better fits. This could be addressed by using clustered models, in which individual sub-models are estimated on suitable

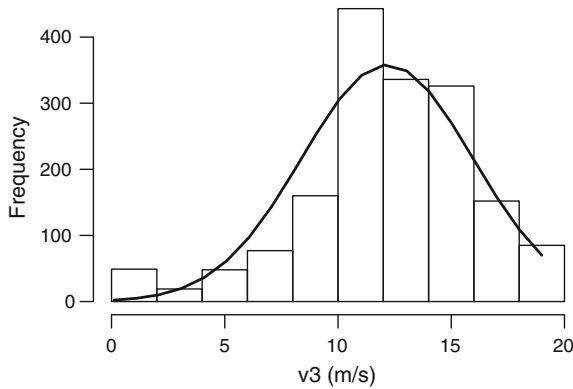


Fig. 4 Histogram of speeds for data series B1695

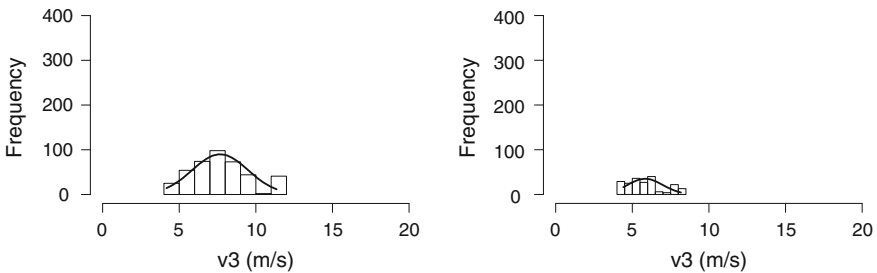
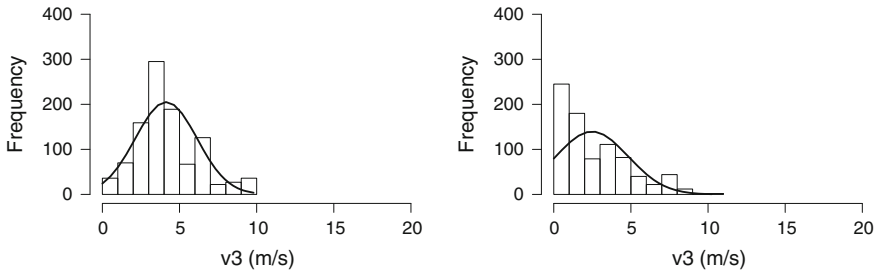


Fig. 5 Histogram of speeds for data series 2, 14 for which satisfactory speed estimation is achieved



**Fig. 6** Histogram of speeds for data series 10, 16 for which unsatisfactory speed estimation is produced

data series with similar characteristics. An approach to accomplish this is presented in the next section.

## 6 Application of Clustered Model

### 6.1 Model-Based Clustering

The limitation of dealing with heterogeneous data can be addressed by the second methodological approach, presented in Fig. 1, which comprises methods such as clustering, loess and classification and allows the adaptation of more flexible and case-specific car-following models. In this section, we use data from Naples to verify that this approach could indeed provide better results than the first methodological approach.

First, a model-based clustering is applied to the longest data series (B1695). Traffic states with different characteristics are recognized and data are divided into groups. The factors which are taken into account for the clustering are the speeds of the second and third vehicle ( $v_2$  and  $v_3$ ) and their distance  $D_{23}$ , since they are considered as the most relevant for driving behavior according to the preceding analysis. In the clustering algorithm various combinations of models were examined and the optimal number of classes was researched. The BIC index [70] was calculated and the number of classes, which minimizes the index was selected. The classification results for different combinations of models and different number of classes (components) are illustrated in Fig. 7. Although the lowest value of BIC index corresponds to 9 classes, the fit on the data is similar for classes between 7 and 9 classes. In addition, even for 4 classes there is not a great loss in relation to the optimal number of classes; therefore the performance of fewer classes could be tested aiming at parsimonious models.

Figure 8 presents the results of clustering for different number of classes. As expected, fewer classes result in simpler clustering, in which case the characteristics of each class are more distinct and easily recognizable. In contrast, the traffic

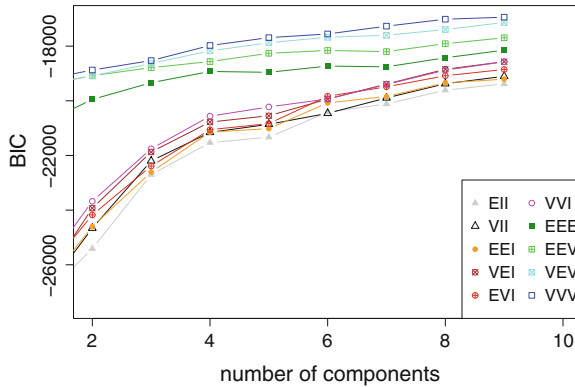


Fig. 7 Choice of optimal number of classes

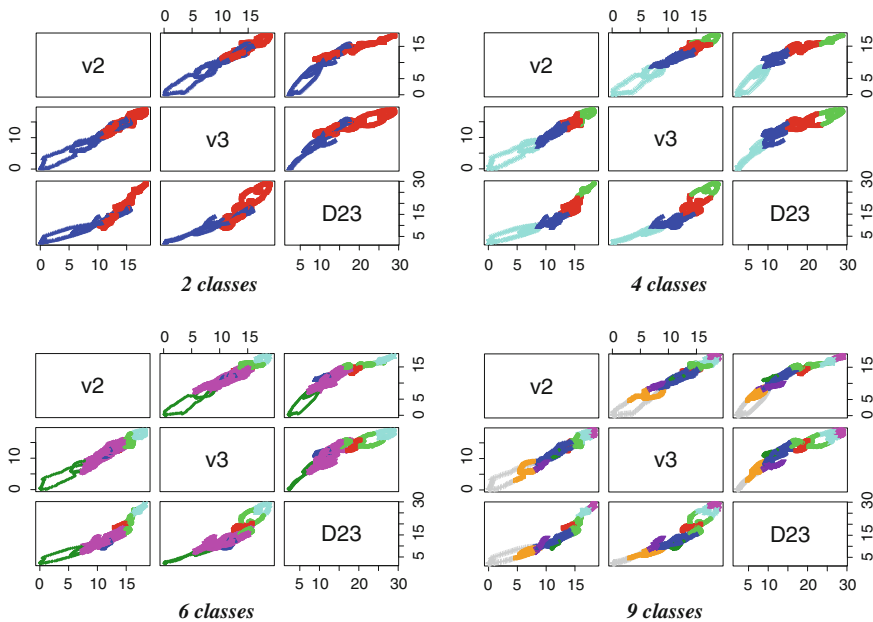


Fig. 8 Clustering results for different number of classes

characteristics of each class appear subtler when a greater number of classes (eg. 6 or 9) is used.

Specific loess models are then calibrated for each traffic state, resulting in a number of models. The other available datasets are then classified into the existing classes created by the B1695 data set. The classification is implemented using the k-nearest neighbor method. Then, the appropriate flexible model is retrieved and applied to the new data for speed estimation.

**Table 2** Results for speed estimation for all Naples data sets using the second methodological approach and mclust package ( $\tau = 0.4$  s)

Data series	RSMN (%)									
	Gipps' model	Loess method	Clustered method (Number of classes)							
			2	3	4	5	6	7	8	9
B1695	2.70	1.59	1.55	1.53	1.48	1.48	1.46	1.42	1.40	1.37
C621	6.60	4.34	4.37	4.41	<b>3.99</b>	4.59	4.54	4.10	3.99	5.36
A358	2.70	2.08	<u>2.10</u>	2.31	3.10	2.34	2.25	2.33	2.33	3.61
A172	4.60	3.40	3.48	3.06	<b>2.44</b>	3.14	2.85	3.30	3.13	9.39
C168	2.30	1.78	<u>1.87</u>	2.04	1.95	2.02	2.05	2.09	2.08	2.08
C171	7.20	6.23	<u>6.31</u>	6.60	7.35	6.50	6.47	8.22	8.19	8.73

The expected result would be that the estimation error would be reduced over the previous case, as a higher number of classes would lead to a more precise estimate, because the models are applied to more homogeneous sub-data-sets. On the other hand, in this case less data per group are available and the calibrated models may be too “narrow” to have a good fit to other data, possibly indicating over-fitting. Furthermore, a larger number of classes would lead to difficulties in identifying the distinct underlying behaviors.

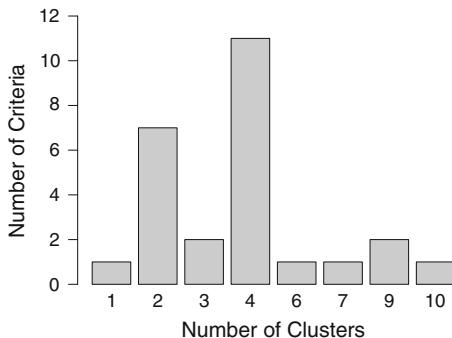
The results are summarized in Table 2, indicating that the expected result is not achieved for all datasets. The B1695 data series, which was used for model calibration, provides the best correspondence between the traffic states. For three of the other datasets the best performance was obtained by a single class (and a model with two classes provided very similar results), while for the remaining two the best performance was achieved with four classes. Overall, the clustered approach appears to outperform the simpler loess approach in some cases, and perform similarly in the remaining cases.

## 6.2 K-Means Algorithm

The second methodological approach was revised using the k-means algorithm for the clustering step. The optimal number of clusters was determined using the NbClust package [50] in R software [64]. Twenty six indices were taken into account for determining the optimal number of clusters. The results are presented in Fig. 9 for dataset B1695. According to the majority of indices estimated by NbClust algorithm the optimal number of clusters are four.

The results are summarized in Table 3, indicating probably a slightly clearer clustering using the k-means algorithm than the model-based one. The behavior of the B1695 data series is similar with the model based clustering. For three of the

**Fig. 9** Optimal number of clusters chosen by 26 criteria for dataset B1695 using NbClust package



**Table 3** Results for speed estimation for all Naples data sets using the second methodological approach and k-means algorithm ( $\tau = 0.4s$ )

Data series	RSMN (%)									
	Gipps' model	Loess method	Clustered method (Number of classes)							
			2	3	4	5	6	7	8	9
B1695	2.70	1.59	1.55	1.51	1.45	1.46	1.43	1.39	1.46	<b>1.31</b>
C621	6.60	4.34	4.75	<b>3.99</b>	<b>4.16</b>	5.24	5.01	5.41	5.44	5.47
A358	2.70	2.08	2.91	<u>2.11</u>	2.60	3.17	2.84	2.45	3.42	3.42
A172	4.60	3.40	3.28	3.56	<b>3.12</b>	3.68	5.00	5.09	6.03	6.03
C168	2.30	1.78	1.84	<b>1.78</b>	<b>1.78</b>	1.79	2.10	1.79	1.81	1.86
C171	7.20	6.23	6.56	7.10	6.69	6.00	<b>5.94</b>	8.08	8.08	7.61

other datasets a better performance was achieved using the second methodological approach but for different number of clusters (3–6 clusters). However, the number of four classes seems to be the most appropriate overall. For the remaining two data series there was almost the same result for the simple loess method and the clustered model with three classes.

Both methods indicate that four clusters are indeed the most effective, though generating gain in some cases and loss in other cases. There is need for further investigation of the best clustering method.

## 7 Discussion and Conclusion

Data driven approaches could be a promising tool for optimization of car-following models, as it may lead to more robust and reliable representation of driving behavior. In this research, an existing methodology for estimation of car-following models has been validated to some NGSIM datasets. This simpler approach outperforms



the reference (Gipps') model for all available datasets. The extended methodology, more elaborate approach, combines clustering, loess and classification, and further improves the performance of the simpler approach in some cases (while providing essentially the same performance as the simpler approach in the remaining cases).

Additional testing on richer data should be performed to determine the factors that determine its performance, as well as develop guidelines for the selection of one or the other approach and the best way of clustering. The proposed methodological framework is more flexible, less time-consuming and allows the incorporation of additional parameters that may influence driving behavior (such as drivers' age, road infrastructure etc.). Resorting cumbersome reformulations of a fixed model form could be impractical. However, conventional models such as Gipps' model may provide better insight into driving behavior, as they are relied on traffic flow theory. The integration of data-driven methods in traffic micro simulation could be very helpful, though additional research should be conducted.

**Acknowledgments** The authors would like to thank Prof. Vincenzo Punzo from the University of Napoli–Federico II for kindly providing the data collected from Napoli and the FHWA for making the NGSIM data-sets freely available. This research has been supported by the Action: ARISTEIA-II (Action's Beneficiary: General Secretariat for Research and Technology), co-financed by the European Union (European Social Fund – ESF) and Greek national funds project.

## References

1. Ahmed KI (1999) Modeling drivers' acceleration and lane changing behavior. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, Mass
2. Al-Shihabi T, Mourant RR (2003) Toward more realistic driving behavior models for autonomous vehicles in driving simulators. In: 82nd annual meeting of the transportation research board, Washington, DC
3. Antoniou C, Balakrishna R, Koutsopoulos HN (2011) A synthesis of emerging data collection technologies and their impact on traffic management applications. *Eur Trans Res Rev* 3(3):139–148. doi:10.1007/s12544-011-0058-1
4. Antoniou C, Koutsopoulos HN (2006) Estimation of traffic dynamics models with machine learning methods. *Transp Res Rec: J Transp Res Board* 1965:103–111 (Washington, DC)
5. Antoniou C, Koutsopoulos HN, Yannis G (2013) Dynamic data-driven local traffic state estimation and prediction. *Transp Res C: Emerg Technol* 34:89–107
6. Aw A, Klar A, Rascle M, Materne T (2002) Derivation of continuum traffic flow models from microscopic follow-the-leader models. *SIAM J Appl Math* 63(1):259–278
7. Aycin MF, Benekohal RF (1999) Comparison of car-following models for simulation. *Transp Res Rec: J Transp Res Board* 1678(1):116–127
8. Bando M, Hasebe K, Nakayama A, Shibata A, Sugiyama Y (1995) Dynamical model of traffic congestion and numerical simulation. *Phys Rev E* 51(2):1035–1042
9. Bando M, Hasebe K, Nakanishi K, Nakayama A (1998) Analysis of optimal velocity model with explicit delay. *Phys Rev E* 58(5):5429–5435
10. Banfield JD, Raftery AE (1993) Model-based gaussian and non gaussian clustering. *Biometrics* 49:803–821
11. Bellemans T, De Schutter B, De Moor B (2002) Models for traffic control. *J A* 43(3–4)13–22
12. Bevrani K, Chung E (2011) Car following model improvement for traffic safety metrics reproduction. In: Proceedings of the Australasian transport research forum 2011. PATREC, Adelaide Hilton Hotel, Adelaide, SA, pp 1–14

13. Bierley RL (1963) Investigation of an inter vehicle spacing display. *Highw Res Rec* 25:58–75
14. Bifulco GN, Pariota L, Simonelli F, Di Pace R (2013) Development and testing of a fully adaptive cruise control system. *Transp Res C* 29(2013):156–170
15. Boxill SA, Yu L (2000) An evaluation of traffic simulation models for supporting ITS development. Center for Transportation Training and Research, Texas Southern University
16. Brackstone M, McDonald M (1999) Car-following: a historical review. *Transp Res F* 2(4):181–196
17. Chandler RE, Herman R, Montroll EW (1958) Traffic dynamics: studies in car following. *Oper Res* 6(2):165–184
18. Ciuffo B, Punzo V, Montanino M (2012) 30 years of the gipps' car-following model: applications, developments and new features. *TRB 2012 Ann Meet*, Paper number: 12–3350
19. Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc* 74(1978):829–836
20. Cleveland WS, Devlin SJ (1988) Locally weighted regression: an approach to regression analysis by local fitting. *J Am Stat Assoc* 83(1988):596–610
21. Cleveland WS, Devlin SJ, Grosse E (1988) Regression by local fitting: methods, properties and computational algorithms. *J Econometrics* 37(1988):87–114
22. Cohen RA (1999) An Introduction to PROC LOESS for local regression. In: *Proceedings of the 24th SAS users group international conference*, Paper 273
23. Colombaroni C, Fusco G (2013) Artificial neural network models for car following: experimental analysis and calibration issues. *J Int Transp Syst* 18(1) (2014)
24. Davis LC (2003) Modifications of the optimal velocity traffic model to include delay due to driver reaction time. *Phys A: Stat Mech Appl* 319:557–567
25. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the E-M algorithm (with discussion). *J R Stat Soc Ser B* 39:1–38
26. Dunne S, Ghosh B (2012) Regime-based short-term multivariate traffic condition forecasting algorithm. *J Transp Eng* 138(4):455–466
27. Edie LC (1961) Car-following and steady-state theory for non-congested traffic. *Oper Res* 9(1):66–76. doi:[10.2307/167431](https://doi.org/10.2307/167431)
28. Einbeck J, Tutz G (2004) Modelling beyond regression functions: an application of multimodal regression to speed-flow data. *SFB Discussion Paper* 395
29. Fraley C (1998) Algorithms for model-based gaussian hierarchical clustering. *SIAM J Sci Comput* 20:270–281
30. Fraley C, Raftery AE (2002) Model-based clustering. Discriminant analysis and density estimation. *J Am Stat Assoc* 97(458):611–631
31. Fraley C, Raftery AE (2003) Enhanced software for model-based clustering, density estimation, and discriminant analysis: MCLUST. *J Class* 20(263–286):2003
32. Fritzsche HT (1994) A model for traffic simulation. *Traffic Eng Control* 5:317–321
33. Gazis DC, Herman R, Rothery RW (1961) Nonlinear follow-the-leader models of traffic flow. *Oper Res* 9(4):545–567. <http://www.jstor.org/stable/167126>
34. Gipps PG (1981) A behavioral car-following model for computer simulation. *Transp Res B* 15:105–111
35. Greenberg H (1959) An analysis of traffic flow. *Oper Res* 7:79–85
36. Hamdar SH, Mahmassani HS (2008) Driver car-following behavior: from discrete event process to continuous set of episodes. In: *Proceedings of the 87th annual meeting of the transportation research board* (CD, Paper No. 08-3134), January, Washington, DC
37. Hartigan JA (1975) *Clustering algorithms*. Wiley, New York
38. Hartigan JA, Wong MA (1979) A K-means clustering algorithm. *Appl Stat* 28:100–108
39. Helbing D, Tilch B (1998) Generalized force model of traffic dynamics. *Phys Rev E* 58(1):133–138
40. Herman R, Montroll EW, Potts RB, Rothery RW (1959) Traffic dynamics: analysis of stability in car following. *Oper Res* 7(1):86–106
41. Huang E, Antoniou C, Wen Y, Ben-Akiva M, Lopes J, Bento J (2009) Real-time multi-sensor multi-source network data fusion using dynamic traffic assignment models. In: *12th international IEEE conference on intelligent transportation systems, ITSC'09, 2009*. IEEE, pp 1–6

42. Jiang R, Wu Q, Zhu Z (2001) Full velocity difference model for a car-following theory. *Phys Rev E* 64(1):017101
43. Kikuchi C, Chakroborty P (1992) Car following model based on a fuzzy inference system. *Transp Res Rec* 1365:82–91
44. Kometani E, Sasaki T (1958) On the stability of traffic flow. Report no. 1. *J Oper Res Jpn* 2(1):11–26
45. Koutsopoulos NH, Farah H (2012) Latent class model for car following behavior. *Transp Res B* 46(2012):563–578
46. Lenz H, Wagner CK, Sollacher R (1999) Multi-anticipative car-following model. *Eur Phys J B* 7(2):331–335
47. Leutzbach W (1988) *Introduction theory traffic flow*. Springer, Berlin
48. Liu R, Li X (2013) Stability analysis of a multi-phase car-following model. *Phys A: Stat Mech Appl* 392(11):2660–2671
49. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Le Cam LM, Neuman J (eds) *Proceedings 5th Berkeley symposium on mathematical statistics and probability*, vol 1. University of California Press, Berkeley, pp 281–297
50. Malika C, Nadia G, Veronique B, Azam N (2014) NbClust package for determining the best number of clusters, R package version 2.0.2. <http://CRAN.R-project.org/package=NbClust>
51. Marsden GR, McDonald M, Brackstone M (2003) A comparative assessment of driving behaviours at three sites. *Eur J Transp Res* 3(1):5–20. ISSN 1567–7141
52. McLachlan GJ, Krishnan T (1997) *The EM algorithm and extensions*. Wiley, New York
53. Mitchell T (1997) *Machine learning*, McGraw Hill, New York
54. Muezzinoglu MK, Zurada JM (2005) A recurrent RBF network model for nearest neighbor classification, IJCNN '05. In: *Proceedings of the 2005 IEEE international joint conference on neural networks* 1:343–348
55. Newell GF (1961) Nonlinear effects in the dynamics of car following. *Oper Res* 9:209–229
56. Newell GF (2002) A simplified car-following theory: a lower order model. *Transp Res B: Methodol* 36(3):195–205
57. Olstam JJ, Tapani A (2004) *Comparison of Car-following models*. Swedish National Road and Transport Research Institute, VTI meddelande 960A
58. Orosz G, Krauskopf B, Wilson RE (2005) Bifurcations and multiple traffic jams in a car-following model with reaction-time delay. *Phys D: Nonlinear Phenom* 211(3):277–293
59. Papatathanasopoulou V, Antoniou C (2015) Towards data-driven car-following models. *Transp Res C: Emer Technol*
60. Pindyck RS, Rubinfeld DL (1997) *Econometric models and economic forecasts*, 4th edn. Irwin McGraw-Hill, Boston
61. Pipes LA (1953) An operational analysis of traffic dynamics. *J Appl Phys* 24(3):274–281
62. Punzo V, Formisano DJ, Torrieri V (2005) A non-stationary kalman filter for the estimation of accurate multiple car-following data. In: *Proceedings of the 84th annual meeting TRB*, Washington, D.C
63. Punzo V, Borzacchiello MT, Ciuffo B (2011) On the assessment of vehicle trajectory data accuracy and application to the next generation simulation (NGSIM) program data. *Transp Res C: Emer Technol* 19(6):1243–1262
64. R Development Core Team (2014) *R: a language and environment for statistical computing*. R foundation for statistical computing, Vienna, Austria. [www.R-project.org](http://www.R-project.org). Accessed 26 Sept 2014
65. Rakha H, Wang W (2009) Procedure for calibrating Gipps car-following model. *Transp Res Rec* 2124:113–124
66. Ranjithkar P, Suzuki H, Nakatsuji T (2005) Microscopic traffic data with real-time kinematic global positioning system. In: *Proceedings of annual meeting of infrastructure planning and management*, Japan Society of Civil Engineer, Miyazaki, Preprint C.D., Dec 2005
67. Reuschel R (1950) Fahrzeugbewegungen in der Kolonne. *Osterreichisches Ing Archiv* 4:193–215

68. Roughan M, Sen S, Spatscheck O, Duffield N (2004) Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification. In: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement. ACM, pp 135–148
69. Sawada S (2002) Generalized optimal velocity model for traffic flow. *Int J Mod Phys C* 13(01):1–12
70. Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6(2):461–464
71. Spyropoulou I (2007) Gipps car-following model—an in-depth analysis. *Transportmetrica* 3(3):231–245
72. Subramanian H (1996) Estimation of car-following models (Doctoral dissertation, Massachusetts Institute of Technology)
73. Sun L, Zhou J (2005) Development of multiregime speed-density relationships by cluster analysis. *Transp Res Rec: J Trans Res Board* 1934(1):64–71
74. Thiemann C, Treiber M, Kesting A (2008) Estimating acceleration and lane-changing dynamics from next generation simulation trajectory data. *Transp Res Record* 90–101
75. Toledo T (2003) Integrated driving behaviour modelling. Ph.D. thesis, Massachusetts Institute of Technology
76. Tordeux A, Lassarre S, Roussignol M (2010) An adaptive time gap car-following model. *Transp Res B* 44(8–9):1115–1131
77. Treiber M, Hennecke A, Helbing D (2000) Congested traffic states in empirical observations and microscopic simulations. *Phys Rev E* 62(2):1805
78. Treiber M, Kesting A, Helbing D (2006) Delays, inaccuracies and anticipation in microscopic traffic models. *Phys A* 360(1):71–88
79. Underwood RT (1961) Speed volume and density relationships: quality and theory of traffic flow. Bureau of highway traffic, Yale University, New Haven, pp 141–188
80. US Department of Transportation (2012) NGSIM—Next generation simulation. <http://www.ngsim.fhwa.dot.gov>
81. van Lint JWC (2005) Accurate freeway travel time prediction with state-space neural networks under missing data. *Transp Res C: Emer Technol* 13:347–369
82. van Lint JWC (2008) Online learning solutions for freeway travel time prediction. *IEEE Trans Intell Transp Syst* 9(1):38–47
83. Vlahogianni EI, Karlaftis MG, Golias JC (2005) Optimized and meta-optimized neural networks for short-term traffic flow prediction: a genetic approach. *Transp Res C* 13(3):211–234
84. Vlahogianni EI, Karlaftis MG, Golias JC (2008) Temporal evolution of short-term urban traffic flow: a nonlinear dynamics approach. *Comput Aided Civ Infrastruct Eng* 23:536–548
85. Wang L, Rong J, Liu X (2005) The classification of car-following behavior in urban expressway based on fuzzy clustering analysis. In: Proceedings of the 84th annual meeting of the transportation research board, Washington, DC
86. Wiedemann R (1974) Simulation des Straenverkehrsflusses. *Schriftenreihe des Instituts fuer Verkehrswesen, Universitaet Karlsruhe Heft 8*
87. Wiedemann R, Reiter U (1992) Microscopic traffic simulation: the simulation system MISION, background and actual state. CEC Project ICARUS (V1052), Final Report, vol 2. CEC, Brussels (Appendix A)
88. Yang Q, Koutsopoulos HN (1996) A microscopic traffic simulator for evaluation of dynamic traffic management systems. *Transp Res C* 4(3):113–129
89. Zhang HM, Kim T (2005) A car-following theory for multiphase vehicular traffic flow. *Transp Res B* 39:385–399
90. Zhang J, Wang FY, Wang K, Lin WH, Xu X, Chen C (2011) Data-driven intelligent transportation systems: a survey. *IEEE Trans Int Transp Syst* 12(4):1624–1639
91. Zhao X, Gao Z (2005) A new car-following model: full velocity and acceleration difference model. *Eur Phys J B-Condens Matter Complex Syst* 47(1):145–150

# Computational Intelligence and Optimization for Transportation Big Data: Challenges and Opportunities

Eleni I. Vlahogianni

**Abstract** With the overwhelming amount of transportation data being gathered worldwide, Intelligent Transportation Systems (ITS) are faced with several modeling challenges. New modeling paradigms based on Computational Intelligence (CI) that take advantage of the advent of big datasets have been systematically proposed in literature. Transportation optimization problems form a research field that has systematically benefited from CI. Nevertheless, when it comes to big data applications, research is still at an early stage. This work attempts to review the unique opportunities provided by ITS and big data and discuss the emerging approaches for transportation modeling. The literature dedicated to big data transportation applications related to CI and optimization is reviewed. Finally, the challenges and emerging opportunities for researchers working with such approaches are also acknowledged and discussed.

## 1 Introduction

With a vast number of diverse Intelligent Transportation Systems (ITS) operating Worldwide, web-based, mobile, and sensor generated data arrive at and overwhelming scale. This availability allows for new science paradigms to be introduced and novel insights to be gained. Traditionally, turning data into knowledge relies on classical statistical analysis and interpretation; this fundamentally requires analysts to become intimately familiar with the data and serve as an interface between the data and the users. With the recent availability of very large data sets (big data), this form of manual probing becomes slow, expensive, and frequently unfeasible. Methodologically, new approaches are needed to efficiently deal with some of the challenging issues related to big data; some of them are data size, high dimensionality, overfitting,

---

E.I. Vlahogianni (✉)

Department of Transportation Planning and Engineering, School of Civil Engineering,  
National Technical University of Athens, Athens, Greece  
e-mail: elenivl@central.ntua.gr

assessing statistical significance, rapidly changing, missing and noisy data, complex relationships between fields, user interaction and prior knowledge, and system integration.

Big Data is growing exponentially due to the growth of both existing and new data sources (e.g. geospatial, social media comments, mobile). To build a smarter planet, we need smarter computing—computing that is tuned to operate, managed through the cloud and, importantly, designed for big data. Novel modeling paradigms will have to: i. Capture and manage high volume multi-source data encompassing text, images, sounds, generated impulses etc. ii. Understand patterns unfolding in time across a complex transportation system (spatial unfolding) and produce critical information and alerts.

In this context, Computational Intelligence (CI) offers an excellent alternative to traditional hypothesis-driven (i.e. deductive) statistical data analyses and attempts to extract meaningful patterns in big data. In Transportation, there has been increased interest among both researchers and practitioners in exploring the feasibility of CI algorithms in transportation problems, especially related to optimization. The advantage of CI data analysis applications over other alternatives lies in their flexibility, their ability to discover unknown mechanisms and covariations elusive to statistical approaches, their accuracy, and their ability to handle dynamically changing big data. Still, the development of efficient CI applications in Transportation is complex, rarely taught in transportation programs in Academia, while model development and validation are frequently done ad hoc and do not follow universally accepted procedures.

In this paper, the unique opportunities created by the data obtained from modern ITS are discussed and some of the emerging approaches for handling big data are reviewed. The literature dedicated to big data transportation applications related to CI and optimization is reviewed. Finally, the challenges and emerging opportunities for researchers working with such approaches are also acknowledged and discussed.

## **2 The “New” Transportation Landscape**

Urbanization, smart cities and disruptive technologies may be considered as the three pillars transforming the transportation arena. Urban areas are, nowadays, considered as the dominant type of settlement for humanity. In this context, optimizing transportation and mobility play an imperative role in sustainable urban development. Second, cities are becoming smarter, in terms of their infrastructure, with the aim to maximize resources and actively support sustainable growth and high quality of life, through participatory action and engagement, while preserving natural resources [18].

To be able to fully benefit of the above, a transportation system should be instrumented, interconnected and intelligent. In this context, there is an increasing interest in finding novel technologies to support the transportation arena. Some of the most prominent are mobile communications, cloud technologies, energy storage,

autonomous vehicles and the Internet of Things (IoT). The latter is a novel concept straightforwardly applicable to transportation applications; IoT consists of a variety of devices or objects—such as Radio-Frequency Identification (RFID) tags, sensors, actuators, mobile phones, and so on—which, through unique addressing schemes, are able to interact with each other and cooperate with their neighbors to reach common goals [4, 106]. By continuously collecting, analyzing and redistributing transportation information, IoT networks can offer valuable, real time information to both travelers and operators, and, thus, support and improve the operations of ITS, traffic and public transportation systems.

### **3 Big Data and Transportation**

#### ***3.1 A Definition***

Most widely available definitions of “big data” converge to the following: any collection of data is big or may become big, when it becomes difficult or impossible to model its complexity using traditional data processing tools. This definition leave much room for arguments and misconceptions about what data can be considered as big and how big are the available data.

A more scrutinized look at big data introduces the concept of three V’s: big data are quantities amounts (Volume), of any type (Variety), that are collected at unprecedented speed and must be dealt with in a timely manner (Velocity) [71]. The V’s can be extended to include acyclic or irregular temporal data (Variability), the uncertainty stemming from the difficulty in controlling the quality and accuracy of the data (Veracity).

#### ***3.2 Sources and Applications of Big Data in Transportation***

The big data phenomenon is not new in Transportation and Traffic Engineering. The leading edge of transportation data has for long been streaming data coming from a variety of sensors (loop detectors, video cameras, weather stations etc.). What has changed over the years is the cost of new monitoring systems (more economic ways of producing streaming data, such as the passive data produced by personal GPS), the data granularity (very detailed information collected in real time) and the availability of new sources of unstructured or semi-structured data, such as logs, clickstreams, and social media data (tweets, Facebook posts etc.). A detailed classification of Big Data sources may be found in Hashem et al. [50].

The intrusion of big data and analytics to the transportation research and industry is significant. Large companies including Google, IBM, SAS, INRIX etc. systematically fund research and applications on how to leverage big data of all forms

(structured and unstructured) to improve transportation services and customer satisfaction, manage transportation infrastructure, as well as predict or estimate traffic conditions. The gains from using big data in transportation are numerous for road users, authorities and private sector. Road users can make informed decisions to save time and reduce their personal trip cost based on continuously available traffic information from various sources of the road network with extended spatio-temporal coverage. Road authorities may take advantage of big data to understand travel patterns to identify policy interventions, control traffic and manage demand and congestion, or even change the users' behavior. Finally, private sector may gain significant competitive advantage by identifying prevailing trends or increase productivity by improving their route planning and logistics.

A field that has profited the most from the advent of big data is travel demand estimation; various approaches to derive OD matrix and mobility patterns have been based on mobile phone and personal GPS data [16, 42, 61, 74, 76, 88]. Papinski et al. [89] and Bierlaire et al. [13] developed a route choice behavior based on personal GPS traces, whereas Hood et al. [53] used GPS traces to develop a bicycle route choice model. Liu et al. [76, 79] studied land uses based by analyzing GPS-enabled taxi data in Shanghai. Cai et al. [17] analyzed the manner travel patterns may influence the electric vehicle charging infrastructure development using trajectory data from taxis in Beijing. Chen and Chen [24] utilized taxi GPS traces for nigh bus routes planning.

Regarding traffic, mobile phone counts have been systematically used for extracting traffic information in the form of volume, speed and density in both urban and suburban road networks [3, 9, 10, 51]. Castro et al. [19] used taxi GPS traces to estimate the traffic flow conditions in urban areas. Guido et al. [47] attempted to infer speeds using GPS smartphone traffic probes.

Location based services and social media are the new hype for collecting transportation related data. Cheng et al. [27] and Cheng et al. [28] addressed issues of urban mobility by analyzing twitter and social networking data. Collins et al. [33] proposed a sentiment analysis approach to measure transit rider satisfaction by quantifying twitter feeds. Hasan and Ukkusuri [49] demonstrated the use of a large-scale geo-location data set to analyze and understand individual activity patterns. Recently, Yang et al. [122] analyzed Foursquare data to derive OD information for non-commuting trips.

A new field of research that emerged from gathering individual data collection—either through smartphones or instrumented vehicles—is the extraction of driver's profiles during driving [5, 83, 84, 95, 100, 104, 110, 115, 119]. The scope of such profiles is to improve the efficiency during driving and mitigate risky behaviors that may lead to near misses or crashes. Driving big data has also been systematically used to develop advanced insurance systems based on the time and manner a user drives (pay as you drive, pay how you drive) [6, 85, 86].



## 4 Transportation Big Data Analytics

Analyses based on data, regardless of being big or not, have been recognized as a valuable tool for transportation operations. The stake when using big data is to be able to transform data into knowledge. Transforming data into knowledge involved a set of processes that are described in Fig. 1.

Each step towards the ultimate goal involves a set of tasks. For example data capturing and management involves indexing, searching, querying and visualization. The analysis stage may target to detect anomalies, reveal patterns and complex relationships. The prediction step entails complex and flexible data driven models that may consistently and accurately provide information on the future conditions, whereas mechanisms to create and disseminate information are the final step.

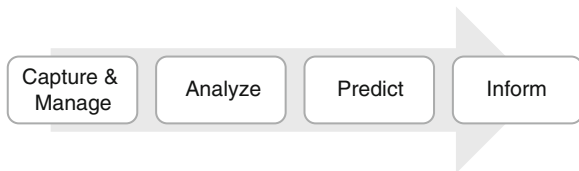
From a modeling standpoint, the problem faced with big data are numerous; first, these datasets are frequently of high dimensionality, meaning that they are difficult to visualize and understand. Moreover, having an extended dataset may not always mean having a representative dataset or a dataset with “perfect” information. The latter signifies that there is a need for a powerful preprocessing stage to assure that the models developed may be estimated and generalize real world conditions. Finally, assessing the statistical fit in big multi-dimensional datasets is not an easy task. Even when using data driven models, the surplus of data may lead to overfitting and models with reduced generalization power.

### 4.1 From Statistics to Computationally Intelligent Models

Usually, the statistical tools implemented entail several structural constraints and are unable to work on quirky and messy data with little or no structure. The lack of diversified statistical tools for big data analyses lead statisticians to see big data as a burdensome rather than a source of valuable information. A typical example is the time series of road traffic characteristics; typical autoregressive statistical models suppress or ignore nonlinearity and irregularities, whereas literature has systematically underlined the usefulness of these irregularities to understand the transitional nature of traffic flow [64, 107, 110, 111, 113, 114].

Evidently, with the advent of multi-source data collection systems, transportation datasets will not become perfect. Treating big data brings forward the focus on size,

**Fig. 1** Processes in big data analytics



the ability to model messiness and multi-dimensionality in datasets, as well as the importance of correlations along with causation; we do not have to always understand the underlying mechanisms of the data to make them work to our benefit. To this end, new flexible and powerful modeling paradigms are imperative that are robust to imperfections and hypothesis free. The need to develop new analysis paradigms for the rapidly growing datasets has been underlined since the late 90s' [39]. This road contains either new forms of statistical thinking or data mining and computational intelligent models. Computational intelligence (CI) is the new hype in transportation modeling. CI includes neural networks, fuzzy logic, swarm intelligence, evolutionary algorithms, expert systems, agent based modeling etc. These models are applicable to many data mining problems, from warehousing to prediction and decision making, and may be proven more efficient due to their non-parametric hypothesis free nature.

Contrary to common thinking, some CI tools may bare significant similarities to classical statistical models, an issue frequently disregarded by connectionists that are more interested in producing accurate results rather than judging on the quality of their models and the properties of the error [15]. With the use of statistical inference, researchers may construct CI models equivalent to many popular statistical models [66]. For example, a single Perceptron is a linear regression model [93], while a Multilayer Perceptron with one hidden unit and a logistic function at the output layer is equivalent to a logit model [107].

The importance of CI to transportation is significant; CI may be used to develop scalable, manageable, adaptable and affordable transportation systems using common sense reasoning, perception and learning, as well as autonomy. One of the many advantages of CI, which is among the main differences with statistical thinking, is the ability of the latter to treat many “non-algorithmizable” problems (natural language processing, visual perception, character recognition etc.). Their ability to augment or replace human skills reflects to gains in computations, accelerates processing and increases productivity. These features may lead to providing results with improved accuracy and quality in a timely manner.

## **5 Computational Intelligent Optimization for Big Data Problems**

In the entire process of mining knowledge from data, several modeling stages may be formulated as optimization problems. Optimization targets the “optimum” solution(s) for a given problem within allowable time. The issue is that each problem may have several local optimal solutions. The difficulty in converging relates to the problem's dimension and the number of objectives (large-scale multi-objective optimization). Evidently, large-scale optimization processes are affected by the curse of dimensionality in numerous ways [29]; the larger the dimensions of the phenomenon, the larger the solution space will be. The larger the dimension of a problem, the greater the risk of some problem characteristics to be altered with the scale.

Moreover, most traditional methods can only be applied to continuous and differentiable functions. Nevertheless, these conditions do not hold for most real world. The above complexities may be treated by problem decomposition strategies, surrogate-based fitness evaluations, data transformations etc. [58]. Another issue that may increase the complexity of the optimization problems is the spatio-temporal evolution of the datasets. In non-stationary environments and transportation problems (e.g. traffic flow) the dynamics may impose different optimal solutions in relation to time and space. This means that an optimization strategy should be able to treat dynamic problems and continuously converge to a solution.

CI approaches have both the structural flexibility and learning capability to deal with complex, time varying multi-objective problems [128]. CI applications to transportation include nature-inspired algorithms (evolutionary algorithms, particle swarm optimization etc.) and non-linear mapping and knowledge embedding approaches (neural networks, fuzzy algorithms etc.). CI have been found to perform well in non-stationary and highly nonlinear problems due to their robustness (impose little or no requirements on the objective function) and flexibility to handle highly non-linear mappings [54]. Moreover, self-adaptation and parallel operation are among the most important characteristics that enable CI to improve their performance and decompose complex tasks into simpler ones. Nevertheless, literature systematically underlines the need to cautiously apply CI to transportation problems as their proper development is frequently tedious and involves significant parametrization [66].

## ***5.1 Computational Intelligent Optimization in Transportation Problems***

Numerous efforts dedicated to CI optimization approaches to transportation applications can be traced in literature. Table 1 is a non-exhaustive list of the most recent research attempts related to CI and optimization. These applications are categorized by the transportation problem they aim to solve, the CI algorithms implemented, as well as the type of data used to evaluate the proposed approach. Special attention is given to whether the listed applications involve the full big data perspective (5 Vs).

Genetic algorithms may be considered the first and leading CI techniques in transportation optimization problems systematically applied to network design problems [67], vehicle routing and allocation problems [2, 44, 65, 78], signalization optimization [21, 22, 91, 99, 101] and highway alignment optimization [55, 63], pricing [68] and so on.

Significant interest from transportation modelers has been placed on Swarm Intelligence (SI). SI is an innovative branch of meta-heuristics derived from imitating the behavioral pattern of natural insects. Teodorović [102] reviews the literature on swarm intelligence and transportation and traffic engineering applications, whereas Zhang et al. [127] conduct a thorough review on the swarm intelligence applications to transportation logistics.

**Table 1** Classification of literature on computational intelligent application to transportation optimization problems

Authors	Date	Problem	CI method	Data
Bai et al. [7]	2014	Transportation asset management	NSGA II	Numerical example
Chen et al. [25, 26]	2014	Trip planning	Heuristic Algorithm	Location-based social network, taxi GPS digital footprints <sup>a</sup>
Chira et al. [31]	2014	Vehicle routing	Evolutionary algorithms, ant colony	Real world case study
Danalet et al. [36]	2014	Pedestrian routing	Bayesian networks	Wi-fi data <sup>a</sup>
Doolan and Muntean [37]	2014	Vehicle routing	Ant-colony optimization	Simulation
Fagnant and Kockelman [38]	2014	Share autonomous vehicles	Agent-based model	Simulation
Forcael et al. [40]	2014	Tsunami evacuation routes	Ant colony	Real world case study
Galland et al. [43]	2014	Car pooling	Agent-based model	Simulation
Kallioras et al. [59]	2014	Emergency inspection scheduling	Harmony search	Real world case study
Kammoun et al. [60]	2014	Traffic routing	Ant-hierarchical fuzzy model	Simulation
Lin and Ku [75]	2014	Stopping patterns for passenger rail transportation	Genetic algorithm	Real world case study
Liu et al. [78]	2014	Emergency medical service allocation	Genetic algorithms	Real world case study
Pahlavani and Delavar [87]	2014	Route planning	Weed colonization	Simulation
Stolfi and Alba [98]	2014	Traffic routing	Evolutionary algorithm	Simulation
Terzi and Serin [103]	2014	Maintenance works on pavements	Ant colony	Numerical example
Yang et al. [122, 123]	2014	Highway alignment optimization	Genetic algorithm	Real world case study
Yin et al. [124]	2014	Hurricane evacuation	Agent-based model	Simulation
Zhang et al. [125]	2014	Transit network design	Agent-based model	Simulation
Zhou et al. [128]	2014	Mobile traffic sensor routing	Ant colony, PSO	Simulation
Arango et al. [2]	2013	Berth allocation	Genetic algorithms	Simulation

(continued)

**Table 1** (continued)

Authors	Date	Problem	CI method	Data
Chevrier et al. [30]	2013	Railway scheduling	Evolutionary algorithm	Real world case study
Cong et al. [34]	2013	Traffic routing	Ant colony algorithm	Simulation
Goksal et al. [46]	2013	Vehicle routing	PSO algorithm	Numerical example
Jia et al. [56]	2013	Transportation-distribution planning	NSGA II algorithm	Numerical example
Kontou et al. [69]	2013	Transit depot allocation	Genetic algorithm	Real world case study
Lagaros et al. [70]	2013	Fund allocation	PSO algorithm	Real world case study
Levin and Kanza [73]	2013	Vehicle routing	Heuristic algorithm	Location-based network <sup>a</sup>
Liu et al. [77]	2013	Freeway corridor diversion control	Genetic algorithms	Real world case study
Shafahi and Bagherian [94]	2013	Highway alignment optimization	PSO algorithm	Numerical example
Ceylan and Ceylan [20]	2012	Signalization optimization	Harmony search algorithm	Simulation
D'Acerno et al. [35]	2012	Signalization optimization	ACO-based algorithm	Simulation
Kang et al. [61, 62]	2012	Highway alignment optimization	Genetic algorithm	Real world case study
Putha et al. [91]	2012	Traffic signal optimization	Ant colony, GA	Numerical example
Balseiro et al. [8]	2011	Vehicle routing	Ant colony	Numerical example
Geroliminis et al. [44]	2011	Transit mobile repair units allocation	Genetic algorithm	Real world case study
Mesbah et al. [82]	2011	Transit priority	Genetic algorithm	Numerical example
Deshpande et al. [122]	2010	Scheduling pavement rehabilitation	Multi-objective genetic algorithm	Numerical example
García-Nieto et al.	2010	Traffic light scheduling	PSO algorithm	Simulation
Kepaptsoglou et al. [68]	2010	Pricing policy optimization	Genetic algorithm	Real world case study
Meng and Khoo [81]	2010	Ramp metering	NSGA-II	Real world case study

(continued)

**Table 1** (continued)

Authors	Date	Problem	CI method	Data
Pishvae et al. [90]	2010	Logistics network design	Memetic algorithm	Numerical example
Shimamoto et al. [96]	2010	Transit network design	NSGA-II	Ticket-based travel data <sup>a</sup>
Kang et al. [63]	2009	Highway alignment optimization	Genetic algorithm	Real world case study
Karlaftis et al. [65]	2009	Vehicle routing	Genetic algorithm	Real world case study
Kepaptsoglou and Karlaftis [67]	2009	Transit network design	Genetic algorithm	Real world case study
Lau et al. [72]	2009	Vehicle routing	Genetic algorithm, fuzzy algorithm	Simulation

<sup>a</sup>Big data applications

Another domain of CI that has attracted significant attention in transportation and traffic engineering is agent based modeling. Agent and multi-agent systems have been applied to many traffic and transportation fields including dynamic routing and congestion management. Chen et al. [24] and Bazzan and Klüge [12] reviewed the literature related to agent-based traffic modelling and simulation, and agent-based traffic control and management. However, as stated in Bazzan [11], the “agentification” of transportation problems may hinder several challenging issues (e.g. the number of agents is high, the extent and magnitude of collective behavioral patterns is immense and probably unpredictable etc.) that should be carefully examined and taken into consideration.

A significant portion of literature refers to the optimization of learning processes involved in transportation models. Learning from extensive transportation and traffic datasets involve multi-source data distributed in many different locations and involve too many data points and extensive spatial coverage. Learning strategies inside traffic and transportation prediction models, as well as dimensionality reduction approaches and imputation problems have been systematically addressed using computationally intelligent techniques [23, 52, 80, 105, 107, 108, 110, 112, 118].

The analysis of literature indicates that there are very few big data applications to transportation optimization problems that are treated with CI methods. Shimamoto et al. [96] introduce a NSGA II algorithm to solve the transit assignment problem using ticket-based travel data. Levin and Kanza [73] implemented heuristic algorithms for the vehicle outing problem using location based data. Danalet et al. [36] leveraged campus wi-fi data to solve the pedestrian routing problem, whereas Chen et al. [26] used GPS traces and location based data for trip planning. The limited number of studies on transportation optimization using big data does not signify limited interest on the specific subject, but reflects two distinct challenges: first, large-scale optimization problems involving a significant number of modeling parameters are difficult to be estimated in a global search context; even CI that are more robust that

classical approaches, may fail or become extremely time consuming, especially in a multi-objective framework [128]. Second, transportation optimization problems are complex and involve a tedious procedure for evaluating the quality of solutions when dealing with global population based search algorithms.

## 6 Opportunities and Challenges

### 6.1 *The Changing Nature of Transportation Problems*

Conceptually, the methodological change that big data brings to transportation is the need to automatically process and analyze data. This has significant effects on the knowledge that may be or needs to be extracted from the available data. Several solutions to problems in transportation science that were founded on static univariate data may not be applicable to dynamically changing multivariate datasets leading to the need to reexamine several phenomena or even change the way we think of transportation problems.

Three promising research fields that will most likely benefit from the data deluge area are:

- User experience mining for improving transportation services,
- Naturalistic driving experiments for monitoring driver's behavior, constructing driver's profile and identifying risk in driving, and
- Autonomous driving for congestion mitigation and safety.

The deluge of big data may not signify that some scientific questions are to be better modeled, but, a more detailed modeling approach to various phenomena may be accomplished [97]; some examples are OD surveys home interviews, census surveys, and so on. The ability to monitor the transportation and traffic related characteristics of individual road users will significantly affect the manner transportation research problems are articulated. Nevertheless, to turn data into knowledge some old dilemmas and challenges extend to big data science. These refer to model selection, real time operation, the quality and availability of the data, the quality of optimization solutions, the inference mechanisms, as well as ethical and social issues.

### 6.2 *Big Data Analytics Versus Models*

The changing nature of transportation problems often drives the need to test and evaluate new modeling paradigms robust to big data and imperfections. CI and data mining has taken a large part of the related transportation literature frequently leaving less ground to classical statistics and models. This may hinder the danger to consider that models, either statistics or borrowed by laws of physics, traditionally used to treat

transportation problems are now obsolete. The truth is rather in the middle and relates to the type and extent of information needed. Evidently, a deeper understanding of the transportation problems will dictate the use of models that may translate data into causal relationships. Towards this direction, literature has emphasized the need to develop synergies with statistics to enhance the explanatory power of many CI applications [66]. Statistics may enhance the inference mechanisms of CI approaches and assure the reliability of the models developed and their generalization power.

### ***6.3 From Batch to Real-Time Computations***

The challenging task in big data analysis is not only to produce knowledge, but to produce it in a timely manner. The time to produce results relates to the size and the complexity of the datasets. Processes that may take long, but can claim increase accuracy and reliability are of limited use, if they are provided with delay. Batch model building with either data mining or statistical approaches has been the dominant approach to transportation problems. Modeling has been traditionally based on historical data, that where leveraged using different modeling paradigms to extract knowledge. In this framework, by the time new data arrive, these were batch processed to produce the output. This approach seems to be conceptually at arms with the computational needs of modern ITS systems that require timely and accurate information to disseminate to centers and users in a highly dynamic transportation environment. Data driven ITS and individual driven ITS systems are founded on real time computations, developing real-time new models that may not only respond in real-time, but learn to change their behavior in real-time (retraining strategies for CI short-term forecasting models) [108, 117]. In such conditions, optimization challenges are numerous and involve optimizing models to include new phenomena and forget past—probably incorrect or trivial—knowledge.

### ***6.4 Data Quality, Availability, Representativeness and Relevance***

Data unceasingly coming from multiple sources, at a variety of forms and in high resolutions are inhomogeneous and may contain noise and erroneous values. Noise and errors mask the significant information hindering in the data. The usual approach is to filter and apply data reduction techniques to eliminate the effect of noise and errors [48]. Data cleaning is a long standing problem but with significance in cases of big data. Data cleaning may include several tasks, such as irregularities (anomalies) detection, incomplete data filling, duplicates removal, conflicting values detection and so on. Nevertheless, these tasks are not so easy to be accomplished in the big data framework [121]; first, because many data cleaning strategies are not suitable for



big data, and second, because in the big data framework, many error types (incomplete data, missing data, erroneous data duplicate data etc.) coexist, while existing techniques are focused on treating a specific error type at a time.

Furthermore, there is a thin line between information and extreme data. Noise and extreme values may contain useful information for the phenomenon under investigation. The use of advanced techniques to automatically preprocess the data and transform them to a more “analyzable” form may lead to datasets that have significantly distorted information about real world conditions [109].

Having large datasets may not always mean having a representative sample to study a phenomenon. Quality is linked to the sample size that needs to be accounted for. The collected data may account for a small part of the phenomenon both spatially and demographically. A typical example is data gathered from tweets and Facebook posts; those that do not possess a profile in social media will not be captured and included in big datasets.

The big data frequently dictate the modeling approach to follow. Nevertheless caution should be given to the modeling strategy; the belief that analyses suited for small datasets may be done with the same or better accuracy to larger datasets is misleading. There are models that have traditionally work well for small datasets, but could become unfeasible with more massive data, whereas in some modeling cases with clear underlying dynamics, simple models, such as linear regression with distinct causal implications could approximate with comparable accuracy and effectiveness the given data. Hand [48] defines the unintelligent data analysis as the one that results to over-specified models or over-idealized problems and underlines that intelligent analysis is dependent of a “good” strategy that defines the steps, decisions and actions taken to analyze a given dataset.

## ***6.5 Inference from Data: Correlations and Causation***

In the era of “big data” several researchers may claim that correlations will be enough to provide information and a deeper look to causations that may help researcher to acquire a thorough understanding of the different phenomena may not be necessary. This misconception deriving from data enthusiasts is tricky and contradicts the true intentions of data analytics. With data analytics we aim to extract information for making better and more informed decisions. Such decision based solely on correlations and deprived from causalities may be far from being accurate and intelligent.

Even if CI approaches are to be implemented, interpretation remain a focal point in transportation engineering. CI using big data can easily reveal correlations; the larger the datasets the greater number of correlations between different variables may be revealed. This does not, however, imply that causations may be achieved [116]. Moreover, several correlations may be also coincidental (spurious) [1]. The lack of straightforward inference mechanisms in CI approaches may lead to misinterpretations and erroneous results. This is a major shortcoming of applying CI methods to transportation data and should be taken into consideration. Big data are complex and

causation can be distorted by various factors such as latent variables, indirect influences imposed by various systems acting simultaneously, multi-collinearity, missing values and so on.

## ***6.6 Quality of Optimization Solutions and Uncertainties***

Evaluating optimization solutions is a time consuming and costly task. The more complex the optimization problems the less efficient the global population based approaches become. To reduce the time and effort needed to provide optimization solutions of high quality, surrogate modeling often qualifies as a viable solution. Surrogate modeling is a macro-modeling technique that aims to minimize the time and computational load to develop simulations to replicate input-output relationships [41]. The aim is to produce a faster and simpler approximation of a simulator to make optimization, design space exploration, etc. feasible.

Another critical issue to consider is the robustness of the produced solutions over time. Most transportation phenomena has significant spatio-temporal dependencies that may influence the quality and consistency of the produced solutions. As such, robustness over time is a critical characteristics of the optimization strategies. This may be tackled by selecting the optimization approach that produces results that are the least affected by the varying conditions (changes in variables etc.). The use of dynamic optimization strategies that are computationally intensive seem to be out of the context of real-time ITS applications. Evidently, achieving a tradeoff between the best solution and the optimum solution over time—that will change only when a solution will provide results that are no longer acceptable—is a viable approach [57, 128].

## ***6.7 Ethics, Privacy, Inequalities***

The big data deluge in transportation comes with significant ethical and institutional challenges. As in all disciplines, big data, especially those coming from participatory sensing, have serious ethical and privacy issues that are frequently addressed but rarely understood. Nowadays, a legislative framework that will dictate the ethical boundaries of using personal data streams is missing.

Moreover, until recently, data was a key advantage of a scientific work because several phenomena, especially those dealing with behavioral aspects, were difficult to be monitored. Nowadays, having data still provides a competitive advantage, but for different reasons. Although the technological means to achieve a detailed monitoring of complex phenomena exist, they are not accessible to everyone. The digital divides created by those who possess technology and data are significant for achieving innovation [14]. Moreover, inequalities will progressively extend to research institutions and Academia between those that may fund big data systems

and those that do not possess the economic means to penetrate the market of big data and use them to their benefit. Significant competitive advantage will have those companies and organizations that may not only possess big data, but also can analyze them.

## 7 The Road Ahead

In the near future, every moving object (both humans and machines) is planned to have a unique identity and operate in a smart social and environmental setting. In this framework, advanced skills in data analytics and optimization will be required to solve complex problems and materialize advanced transportation ideas. The road ahead contains CI, but they have to be applied with caution. Some drivers for success will be: i. develop real-time modeling efforts and efficient solutions to complex phenomena and settings, ii. the development of synergies and the use of intuition to enhance explanatory power, iii. the development of test beds and test data to battle inequalities and evaluate ongoing development, iv. the integration of nature inspired algorithms to enable the full abilities of CI, v. cloud and parallel computing for increasing computational power and reducing the cost of transportation services, and vi. the development of new educational paradigms so as transportation researchers and practitioners can cope with the demanding algorithms for treating big data.

The rapid growing of transportation data impose delivering computations and results that reflect the dynamically evolving transportation phenomena in real-time. In this framework researchers should focus on responsive new methods and model building techniques. Moreover, the spatio-temporal complexity seen in most transportation datasets impose the decomposition of a problem to many simpler ones; this decomposition should extend to model building. Ensembles of models rather than a single approach should be evaluated to deliver reliable and accurate models and predictions. As for optimization, literature review underlined that although CI global optimization techniques may well cope with the complexities seen in transportation datasets, they have been rarely used in big transportation data due to the high computational cost they entail. It is of great importance to use big data to develop more flexible and computationally less costly CI meta-optimization techniques—for example surrogates—or improve the manner to formulate optimization problems.

The rise of CI techniques to handle big data does not make statistics obsolete. Several researchers have systematically underlined that the statistical thinking is the means to justify the inferential leap from data to knowledge. Possible synergies between these two different schools of thought will increase the explanatory power of CI models and their transparency [66]. Statistics may be useful for enhancing the clarity about the modeling goals, assessing for the reliability of the model developed, accounting for sources of uncertainty in the underlying data mechanism and models [45]. In the model development and evaluation stages, statistics can provide the theoretical means for testing for optimality and suitability of the learning algorithms. Moreover, statistics may be used to extract causalities, if necessary, an issue largely

disregarded in the CI literature. In this spirit, intuition has a great role to play in the understanding of the huge streams of data. The CI approaches should be tied to human intuition so as results to be reflect reality and not a myopic look at the different phenomena.

Research using big data in transportation should be supported by publically available testbeds and test data. Test beds of varying size and complexity are a critical tool for reducing inequalities, supporting innovation, but also evaluating ongoing research and may serve as a proof-of-concept tool [115]. To this end, open data is considered today as the greatest enabler of research in intelligent transportation systems. A typical example of the direction towards freely available data is the European Open Government Data Initiative (EU OGD). This initiative targets to create a transparent environment without discrimination and exclusivity constraints where both data and software can be freely stored to improve practices and implemented policies across EU member countries. The concept of open big data multiplies the sources of creativity and collective innovation, as new applications and algorithms are produced by both established providers (e.g. Google, IBM, SAS etc.) and public authorities, but also by individual initiatives from programmers (e.g. applications on smart phones).

Another critical issue that will dictate the future of CI in transportation is the ability to fully benefit from artificial intelligence (AI), a key technology to improve the efficiency, safety, and environmental-compatibility of transportation systems [92]. Until now, CI and AI applications have been limited to specific modules of ITS applications, especially for data analysis and prediction disregarding their powerful capabilities for data managing and decision making [32]. Extended usage of CI and AI techniques is needed to fully benefit from their unique capabilities. Towards this direction, concepts such as cloud (computation, software, data access, and storage services) that do not require end-user knowledge of the physical location and configuration of the system that delivers the services, and parallel computing (clusters of computers), can enable the implementation of complex network level ITS [50, 120, 126].

In the instrumented future, transportation engineers and researchers are challenged to be capable of applying both transportation science and interdisciplinary data analyses for the realization and evaluation of their advanced ideas. Evidently, the advent of the new “big data” area in transportation dictates the need to develop new educational paradigms to produce qualified transportation researchers and practitioners able cope with the demanding algorithms for treating big data. The aim is not to replace other disciplines but to be able to produce engineers that may understand and efficient use the full potential of big datasets and the accompanying modeling tools.

**Acknowledgments** This work is part of research co-financed by the European Union (European Social Fund—ESF) and the Hellenic National Funds, through the Operational Program “Education and Lifelong Learning” of the National Strategic Reference Framework (NSRF)—Research Funding Program “Aristeia I”. This paper is dedicated to the memory of my mentor and friend, Professor Matthew G. Karlaftis.

## References

1. Aldrich J (1995) Correlations genuine and spurious in Pearson and Yule. *Stat Sci* 10:364–376
2. Arango C, Cortés P, Onieva L, Escudero A (2013) Simulation-optimization models for the dynamic berth allocation problem. *Comput Aided Civil Infrastruct Eng* 28(10):769–779
3. Astarita V, Bertini L, d’Elia S, Guido G (2006) Motorway traffic parameter estimation from mobile phone counts. *Eur J Oper Res* 175:1435–1446
4. Atzori L, Iera A, Morabito F (2010) The internet of things: a survey. *Comput Netw* 54(15):2787–2805
5. Aupetit S, Riff J, Buttelli O, Espié S (2013) Naturalistic study of rider’s behaviour in initial training in France: evidence of limitations in the educational content. *Accid Anal Prev* 58:206–217
6. Ayuso M, Guillén M, Pérez-Marín AM (2014) Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance. *Accid Anal Prev* 73:125–131
7. Bai Q, Ahmed A, Li Z, Labi S (2014) A hybrid pareto frontier generation method for trade-off analysis in transportation asset management. *Comput Aided Civil Infrastruct Eng*
8. Balseiro SR, Loiseau I, Ramonet J (2011) An ant colony algorithm hybridized with insertion heuristics for the time dependent vehicle routing problem with time windows. *Comput Oper Res* 38(6):954–966
9. Bar-Gera H (2007) Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: a case study from Israel. *Transp Res Part C* 15:380–391
10. Basyoni Y, Talaat H (2014) A bi-level traffic data extraction procedure via cellular phone network for inter-city travel. *J Intell Transp Syst Technol Plann Oper* (forthcoming)
11. Bazzan AL (2009) Opportunities for multiagent systems and multiagent reinforcement learning in traffic control. *Auton Agent Multi-Agent Syst* 18(3):342–375
12. Bazzan AL, Klügl F (2014) A review on agent-based technology for traffic and transportation. *Knowl Eng Rev* 29(03):375–403
13. Bierlaire M, Chen J, Newman J (2013) A probabilistic map matching method for smartphone GPS data. *Transp Res Part C: Emerg Technol* 26:78–98
14. Boyd D, Crawford K (2012) Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Inf Commun Soc* 15(5):662–679
15. Breiman L (2001) Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat Sci* 16(3):199–231
16. Caceres N, Wideberg JP, Benitez FG (2007) Deriving origin-destination data from a mobile phone network. *Intell Transp Syst* 1(1):15–26
17. Cai H, Jia X, Chiu AS, Hu X, Xu M (2014) Siting public electric vehicle charging stations in Beijing using big-data informed travel patterns of the taxi fleet. *Transp Res Part D: Transp Environ* 33:39–46
18. Caragliu A, Del Bo C, Nijkamp P (2009) Smart cities in Europe. *Serie Research Memoranda 0048* (VU University Amsterdam, Faculty of Economics, Business Administration and Econometrics)
19. Castro PS, Zhang D, Li S (2012) Urban traffic modelling and prediction using large scale taxi GPS traces. In: *Pervasive computing*. Springer, Berlin, Heidelberg, pp 57–72
20. Ceylan H, Ceylan H (2012) A hybrid harmony search and TRANSYT hill climbing algorithm for signalized stochastic equilibrium transportation networks. *Transp Res Part C: Emerg Technol* 25:152–167
21. Ceylan H (2006) Developing combined genetic algorithm hill-climbing optimization method for area traffic control. *J Transp Eng ASCE* 132(8):663–671
22. Ceylan H, Bell MGH (2004) Traffic signal timing optimization based on genetic algorithm approach, including drivers’ routing. *Transp Res Part B* 38(4):329–342
23. Chan KY, Dillon T, Chang E, Singh J (2013) Prediction of short-term traffic variables using intelligent swarm-based neural networks. *IEEE Trans Control Syst Technol* 21(1):263–274
24. Chen B, Cheng HH (2010) A review of the applications of agent technology in traffic and transportation systems. *IEEE Trans Intell Transp Syst* 11(2):485–497

25. Chen C, Zhang D, Guo B, Ma X, Pan G, Wu Z (forthcoming) TripPlanner: personalized trip planning leveraging heterogeneous crowdsourced digital footprints. *IEEE Trans Intell Transp Syst*
26. Chen C, Zhang D, Li N, Zhou ZH (2014) B-Planner: planning bidirectional night bus routes using large-scale taxi GPS traces. *IEEE Trans Intell Transp Syst* 15(4):1451–1465
27. Cheng Z, Caverlee J, Lee K (2010) You are where you tweet: a content-based approach to geo-locating twitter users. In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, Toronto, ON, Canada
28. Cheng Z, Caverlee J, Lee K, Sui D (2011) Exploring millions of footprints in location sharing services. In: *Proceedings of the fifth international conference on weblogs and social media*. AAAI, Barcelona, Spain
29. Cheng S, Shi Y, Qin Q, Bai R (2013) Swarm intelligence in big data analytics. In: *Intelligent data engineering and automated learning-IDEAL 2013*. Springer, Berlin, Heidelberg, pp 417–426
30. Chevrier R, Pellegrini P, Rodriguez J (2013) Energy saving in railway timetabling: a bi-objective evolutionary approach for computing alternative running times. *Transp Res Part C: Emerg Technol* 37:20–41
31. Chira C, Sedano J, Villar JR, Cámara M, Corchado E (2014) Urban bicycles renting systems: modelling and optimization using nature-inspired search methods. *Neurocomputing* 135:98–106
32. Chowdhury M, Sadek AW (2012) Advantages and limitations of artificial intelligence. *Artif Intell Appl Crit Transp Issues* 6
33. Collins C, Hasan S, Ukkusuri SV (2013) A novel transit rider satisfaction metric. *J Public Transp* 16(2):21–45
34. Cong Z, De Schutter B, Babuška R (2013) Ant colony routing algorithm for freeway networks. *Transp Res Part C: Emerg Technol* 37:1–19
35. D’Acerno L, Gallo M, Montella B (2012) An ant colony optimisation algorithm for solving the asymmetric traffic assignment problem. *Eur J Oper Res* 217(2):459–469
36. Danalet A, Farooq B, Bierlaire M (2014) A Bayesian approach to detect pedestrian destination-sequences from WiFi signatures. *Transp Res Part C: Emerg Technol* 44:146–170
37. Doolan R, Muntean GM (2014) Time-ants: an innovative temporal and spatial ant-based vehicular routing mechanism. In: *Intelligent vehicles symposium proceedings, 2014 IEEE*. IEEE, pp 951–956
38. Fagnant DJ, Kockelman KM (2014) The travel and environmental implications of shared autonomous vehicles, using agent-based model scenarios. *Transp Res Part C: Emerg Technol* 40:1–13
39. Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. *AI Mag* 17(3):37
40. Forcael E, González V, Orozco F, Vargas S, Pantoja A, Moscoso P (2014) Ant colony optimization model for tsunamis evacuation routes. *Comput Aided Civil Infrastruct Eng* 29(10):723–737
41. Forrester A, Sobester A, Keane A (2008) *Engineering design via surrogate modelling: a practical guide*. Wiley, Chichester
42. Friedrich M, Immisch K, Jehlicka P, Otterstatter T, Schlaich J (2010) Generating origin-destination matrices from mobile phone trajectories. *Transp Res Rec* 2196:93–101
43. Galland S, Knapen L, Yasar AUH, Gaud N, Janssens D, Lamotte O, Wets G (2014) Multi-agent simulation of individual mobility behavior in carpooling. *Transp Res Part C: Emerg Technol*
44. Geroliminis N, Kepaptsoglou K, Karlaftis MG (2011) A hybrid hypercube-genetic algorithm approach for deploying many emergency response mobile units in an urban network. *Eur J Oper Res* 210(2):287–300
45. Glymour C, Madigan D, Pregibon D, Smyth P (1997) Statistical themes and lessons for data mining. *Data Min Knowl Discov* 1(1):11–28

46. Goksal FP, Karaoglan I, Altiparmak F (2013) A hybrid discrete particle swarm optimization for vehicle routing problem with simultaneous pickup and delivery. *Comput Ind Eng* 65(1):39–53
47. Guido G, Gallelli V, Saccomanno F, Vitale A, Rogano D, Festa D (2014) Treating uncertainty in the estimation of speed from smartphone traffic probes. *Transp Res Part C: Emerg Technol* 47:100–112
48. Hand DJ (2000) Data mining. New challenges for statisticians. *Soc Sci Comput Rev* 18(4):442–449
49. Hasan S, Ukkusuri SV (2014) Urban activity pattern classification using topic models from online geo-location data. *Transp Res Part C: Emerg Technol* 44:363–381
50. Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Khan SU (2015) The rise of “big data” on cloud computing: review and open research issues. *Inf Syst* 47:98–115
51. Herrera JC, Work DB, Herring R, Ban X, Jacobson Q, Bayen A (2010) Evaluation of traffic data obtained via GPS-enabled mobile phones: the mobile century field experiment. *Transp Res Part C* 18:568–583
52. Hong WC (2012) Application of seasonal SVR with chaotic immune algorithm in traffic flow forecasting. *Neural Comput Appl* 21(3):583–593
53. Hood J, Sall E, Charlton B (2011) A GPS-based bicycle route choice model for San Francisco, California. *Transp Lett* 3(1):63–75
54. Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. *Neural Netw* 2(5):359–366
55. Jha MK, Schonfeld P (2004) A highway alignment optimization model using geographic information systems. *Transp Res Part A: Policy Pract* 38(6):455–481
56. Jia L, Feng X, Zou G (2013) Solving multiobjective bilevel transportation-distribution planning problem by modified NSGA II. In: 2013 9th international conference on computational intelligence and security (CIS). IEEE, pp 303–307
57. Jin Y, Tang K, Yu X, Sendhoff B, Yao X (2013) A framework for finding robust optimal solutions over time. *Memetic Comput* 5(1):3–18
58. Jin Y, Sendhoff B (2009) A systems approach to evolutionary multiobjective structural optimization and beyond. *IEEE Comput Intell Mag* 4(3):62–76
59. Kallioras NA, Lagaros ND, Karlaftis MG (2013) An improved harmony search algorithm for emergency inspection scheduling. *Eng Optim (ahead-of-print)* 1–23
60. Kammoun HM, Kallel I, Casillas J, Abraham A, Alimi AM (2014) Adapt-Traf: an adaptive multiagent road traffic management system based on hybrid ant-hierarchical fuzzy model. *Transp Res Part C: Emerg Technol* 42:147–167
61. Kang C, Ma X, Tong D, Liu Y (2012) Intra-urban human mobility patterns: an urban morphology perspective. *Phys A: Stat Mech Appl* 391(4):1702–1717
62. Kang MW, Jha MK, Schonfeld P (2012) Applicability of highway alignment optimization models. *Transp Res Part C: Emerg Technol* 21(1):257–286
63. Kang MW, Schonfeld P, Yang N (2009) Prescreening and repairing in a genetic algorithm for highway alignment optimization. *Comput Aided Civil Infrastruct Eng* 24(2):109–119
64. Karlaftis MG, Vlahogianni EI (2009) Memory properties and fractional integration in transportation time-series. *Transp Res Part C: Emerg Technol* 17(4):444–453
65. Karlaftis MG, Kepaptsoglou K, Sambracos E (2009) Containership routing with time deadlines and simultaneous deliveries and pick-ups. *Transp Res Part E: Logist Transp Rev* 45(1):210–221
66. Karlaftis MG, Vlahogianni EI (2011) Statistics versus neural networks in transportation research: differences, similarities and some insights. *Transp Res Part C: Emerg Technol* 19(3):387–399
67. Kepaptsoglou K, Karlaftis MG (2009) The bus bridging problem in metro operations: conceptual framework, models and algorithms. *Public Transp* 1(4):275–297
68. Kepaptsoglou K, Karlaftis MG, Li Z (2010) Optimizing pricing policies in Park-and-Ride facilities: a model and decision support system with application. *J Transp Syst Eng Inf Technol* 10(5):53–65

69. Kontou E, Kepaptsoglou K, Charalampakis AE, Karlaftis MG (2014) The bus to depot allocation problem revisited: a genetic algorithm. *Public Transp* 6(3):237–255
70. Lagaros ND, Kepaptsoglou K, Karlaftis MG (2012) Fund allocation for civil infrastructure security upgrade. *J Manage Eng* 29(2):172–182
71. Laney D (2001) 3D data management: controlling data volume, velocity and variety. Gartner. Accessed 6 Feb
72. Lau HC, Chan TM, Tsui WT, Chan FT, Ho GT, Choy KL (2009) A fuzzy guided multi-objective evolutionary algorithm model for solving transportation problem. *Expert Syst Appl* 36(4):8255–8268
73. Levin R, Kanza Y (2014) TARS: traffic-aware route search. *GeoInformatica* 18(3):461–500
74. Liang X, Zheng X, Lv W, Zhu T, Xu K (2012) The scaling of human mobility by taxis is exponential. *Phys A: Stat Mech Appl* 391(5):2135–2144
75. Lin DY, Ku YH (2014) Using genetic algorithms to optimize stopping patterns for passenger rail transportation. *Comput Aided Civil Infrastruct Eng* 29(4):264–278
76. Liu Y, Kang C, Gao S, Xiao Y, Tian Y (2012) Understanding intra-urban trip patterns from taxi trajectory data. *J Geogr Syst* 14(4):463–483
77. Liu Y, Li P, Wehner K, Yu J (2013) A generalized integrated corridor diversion control model for freeway incident management. *Comput Aided Civil Infrastruct Eng* 28(8):604–620
78. Liu Y, Roshandeh AM, Li Z, Kepaptsoglou K, Patel H, Lu X (2014) Heuristic approach for optimizing emergency medical services in road safety within large urban networks. *J Transp Eng* 140(9):04014043
79. Liu Y, Wang F, Xiao Y, Gao S (2012) Urban land uses and traffic ‘source-sink areas’: evidence from GPS-enabled taxi data in Shanghai. *Landscape Urban Plann* 106(1):73–87
80. Lv Y, Duan Y, Kang W, Li Z, Wang FY (2014) Traffic flow prediction with big data: a deep learning approach. *IEEE Trans Intell Transp Syst* (forthcoming)
81. Meng Q, Khoo HL (2010) A Pareto-optimization approach for a fair ramp metering. *Transp Res Part C: Emerg Technol* 18(4):489–506
82. Mesbah M, Sarvi M, Currie G (2011) Optimization of transit priority in the transportation network using a genetic algorithm. *IEEE Trans Intell Transp Syst* 12(3):908–919
83. Musicant O, Bar-Gera H, Schechtman E (2010) Electronic records of undesirable driving events. *Transp Res Part F: Traffic Psychol Behav* 13(2):71–79
84. Musicant O, Bar-Gera H, Schechtman E (2014) Temporal perspective on individual driver behavior using electronic records of undesirable events. *Accid Anal Prev* 70:55–64
85. Paefgen J, Staake T, Fleisch E (2014) Multivariate exposure modeling of accident risk: insights from Pay-as-you-drive insurance data. *Transp Res Part A: Policy Pract* 61:27–40
86. Paefgen J, Staake T, Thiesse F (2013) Evaluation and aggregation of pay-as-you-drive insurance rate factors: a classification analysis approach. *Decis Support Syst* 56:192–201
87. Pahlavani P, Delavar MR (2014) Multi-criteria route planning based on a driver’s preferences in multi-criteria route selection. *Transp Res Part C: Emerg Technol* 40:14–35
88. Pan CX, Lu JG, Di S, Ran B (2006) Cellular-based data-extracting method for trip distribution. *Traffic Urban Data* 1945:33–39
89. Papinski D, Scott DM, Doherty ST (2009) Exploring the route choice decision-making process: a comparison of planned and observed routes obtained using person-based GPS. *Transp Res Part F: Traffic Psychol Behav* 12(4):347–358
90. Pishvaei MS, Farahani RZ, Dullaert W (2010) A memetic algorithm for bi-objective integrated forward/reverse logistics network design. *Comput Oper Res* 37(6):1100–1112
91. Putha R, Quadrioglio L, Zechman E (2012) Comparing ant colony optimization and genetic algorithm approaches for solving traffic signal coordination under oversaturation conditions. *Computer Aided Civil Infrastruct Eng* 27(1):14–28
92. Sadek AW (ed) (2007) Artificial intelligence in transportation: information for application. Transportation Research Board Circular (E-C113), TRB, National Research Council, Washington, DC. <http://onlinepubs.trb.org/onlinepubs/circulars/ec113.pdf>
93. Sarle WS (1994) Neural networks and statistical models. In: Proceedings of the nineteenth annual SAS users group international conference, 1–13 April



94. Shafahi Y, Bagherian M (2013) A customized particle swarm method to solve highway alignment optimization problem. *Computer Aided Civil Infrastruct Eng* 28(1):52–67
95. Shichrur R, Sarid A, Ratzon NZ (2014) Determining the sampling time frame for in-vehicle data recorder measurement in assessing drivers. *Transp Res Part C: Emerg Technol* 42:99–106
96. Shimamoto H, Murayama N, Fujiwara A, Zhang J (2010) Evaluation of an existing bus network using a transit network optimisation model: a case study of the Hiroshima city bus network. *Transportation* 37(5):801–823
97. Shuldiner AT, Shuldiner PW (2013) The measure of all things: reflections on changing conceptions of the individual in travel demand modeling. *Transportation* 40(6):1117–1131
98. Stolfi DH, Alba E (2014) Red swarm: reducing travel times in smart cities by using bio-inspired algorithms. *Appl Soft Comput* 24:181–195
99. Sun D, Benekohal RF, Waller ST (2006) Bi-level programming formulation and heuristic solution approach for dynamic traffic signal optimization. *Comput Aided Civil Infrastruct Eng* 21(5):321–333
100. Takeda K, Miyajima C, Suzuki T, Angkititrakul P, Kurumida K, Kuroyanagi Y, Komada Y (2012) Self-coaching system based on recorded driving data: learning from one's experiences. *IEEE Trans Intell Transp Syst* 13(4):1821–1831
101. Teklu F, Sumalee A, Watling D (2007) A genetic algorithm approach for optimizing traffic control signals considering routing. *Comput Aided Civil Infrastruct Eng* 22(1):31–43
102. Teodorović D (2008) Swarm intelligence systems for transportation engineering: principles and applications. *Transp Res Part C: Emerg Technol* 16(6):651–667
103. Terzi S, Serin S (2014) Planning maintenance works on pavements through ant colony optimization. *Neural Comput Appl* 25:1–11
104. Toledo T, Musicant O, Lotan T (2008) In-vehicle data recorders for monitoring and feedback on drivers' behavior. *Transp Res Part C: Emerg Technol* 16(3):320–331
105. Tselentis DI, Vlahogianni EI, Karlaftis MG (2014) Improving short-term traffic forecasts: to combine models or not to combine? *IET Intell Transp Syst* (forthcoming)
106. Vlacheas P, Giaffreda R, Stavroulaki V, Kelaidonis D, Foteinos V, Poullos G, Demestichas P, Somov A, Biswas AR, Moessner K (2013) Enabling smart cities through a cognitive management framework for the internet of things. *IEEE Commun Mag* 51(6):102–111
107. Vlahogianni EI, Karlaftis MG (2012) Comparing freeway lane speed patterns under fine and adverse weather conditions. *Nonlinear Dyn* 69(4):1949–1963
108. Vlahogianni EI (2009) Enhancing predictions in signalized arterials with information on short-term traffic flow dynamics. *J Intell Transp Syst: Technol Plann Oper* 13(2):73–84
109. Vlahogianni EI, Karlaftis MG (2011) Aggregating temporal and spatial data: implication for statistical characteristics and model choice. *Transp Lett: Int J Transp Res* 3(1):37–49
110. Vlahogianni EI, Karlaftis MG (2013) Testing and comparing neural network and statistical approaches for predicting transportation time series. *Transp Res Rec: J Transp Res Board* 2399(1):9–22
111. Vlahogianni EI, Golias JC, Karlaftis MG (2004) Short-term traffic forecasting: overview of objectives and methods. *Transp Rev* 24(5):533–557
112. Vlahogianni EI, Karlaftis MG, Golias JC (2005) Optimized and meta-optimized neural networks for short-term traffic flow prediction: a genetic approach. *Transp Res Part C: Emerg Technol* 13(3):211–234
113. Vlahogianni EI, Karlaftis MG, Golias JC (2006) Statistical methods for detecting nonlinearity and non-stationarity in univariate short-term time-series of traffic volume. *Transp Res Part C: Emerg Technol* 14(5):351–367
114. Vlahogianni EI, Karlaftis MG, Golias JC (2008) Temporal evolution of short-term urban traffic flow: a nonlinear dynamics approach. *Comput Aided Civil Infrastruct Eng* 23(7):536–548
115. Vlahogianni EI, Karlaftis MG, Golias JC (2014) Short-term traffic forecasting: where we are and where we're going. *Transp Res Part C: Emerg Technol* 43:3–19
116. Vlahogianni EI, Karlaftis MG, Orfanou FP (2012) Modeling the effects of weather and traffic on the risk of secondary incidents. *J Intell Transp Syst* 16(3):109–117

117. Vlahogianni EI, Karlaftis MG, Golias JC, Kourbelis ND (2006) Pattern-based short-term urban traffic predictor. In: Intelligent transportation systems conference, 2006, ITSC'06. IEEE, pp 389–393
118. Vlahogianni EI, Yannis G, Golias JC (2013) Critical power two wheeler riding patterns at the emergence of an incident. *Accid Anal Prev* 58:340–345
119. Vlahogianni EI, Yannis G, Golias JC (2014) Detecting powered-two-wheeler incidents from high resolution naturalistic data. *Transp Res Part F: Traffic Psychol Behav* 22:86–95
120. Wang FY (2010) Parallel control and management for intelligent transportation systems: concepts, architectures, and applications. *IEEE Trans Intell Transp Syst* 11(3):630–638
121. Wang H, Li M, Bu Y, Li J, Gao H, Zhang J (2014) Cleanix: a big data cleaning parfait. In: Proceedings of the 23rd ACM international conference on conference on information and knowledge management. ACM, pp 2024–2026
122. Yang F, Jin PJ, Cheng Y, Ran B (2014) Origin-destination estimation for non-commuting trips using location-based social networking data. *Int J Sustain Transp* (just-accepted)
123. Yang N, Kang MW, Schonfeld P, Jha MK (2014) Multi-objective highway alignment optimization incorporating preference information. *Transp Res Part C: Emerg Technol* 40:36–48
124. Yin W, Murray-Tuite P, Ukkusuri SV, Gladwin H (2014) An agent-based modeling system for travel demand simulation for hurricane evacuation. *Transp Res Part C: Emerg Technol* 42:44–59
125. Zhang G, Zhang H, Li L, Dai C (2014) Agent-based simulation and optimization of urban transit system. *IEEE Trans Intell Transp Syst* 15(2)
126. Zhang L, Wu C, Li Z, Guo C, Chen M, Lau F (2013) Moving big data to the cloud: an online cost-minimizing approach. *IEEE J Sel Areas Commun* 31(12):2710–2721
127. Zhang S, Lee CKM, Chan HK, Choy KL, Wu Z (2015) Swarm intelligence applied in green logistics: a literature review. *Eng Appl Artif Intell* 37:154–169
128. Zhou Z, Chawla N, Jin Y, Williams G (2014) Big data opportunities and challenges: discussions from data analytics perspectives. *IEEE Comput Intell Mag* 9(4):62–74

# Nature-Inspired Algorithms: Success and Challenges

Xin-She Yang

**Abstract** The simplicity and flexibility of nature-inspired algorithms have made them very popular in optimization and computational intelligence. Here, we will discuss the key features of nature-inspired metaheuristic algorithms by analyzing their diversity and adaptation, exploration and exploitation, attractions and diffusion mechanisms. We also highlight the success and challenges concerning swarm intelligence, parameter tuning and parameter control as well as some open problems.

**Keywords** Algorithm · Adaptation · Bat algorithm · Cuckoo search · Diversity · Firefly algorithm · Metaheuristic · Optimization

## 1 Introduction

Many applications concern hard optimization problems, which may require sophisticated optimization techniques to deal with. However, traditional algorithms usually cannot cope with such highly nonlinear and multimodal problems. Alternative approaches have to be found. In recent years, nature-inspired metaheuristic algorithms have gained huge popularity, and these algorithms include ant colony optimization, particle swarm optimization, cuckoo search, firefly algorithm, bat algorithm, bee algorithms and others [4, 14, 17, 28]. There are many reasons for such popularity. From the algorithm analysis point of view, these algorithms tend to be flexible, efficient and highly adaptable, and yet easy to implement. The high efficiency of these algorithms makes it possible to apply them to a wide range of problems in diverse applications.

The main purpose of this chapter is to highlight some key issues in adaptation and diversity in swarm intelligence. Therefore, the chapter is organized as follows. Section 2 outlines some widely used nature-inspired algorithms, followed by a brief discussion of the main mechanisms of generating new solutions in Sect. 3. Section 4

---

X.-S. Yang (✉)

School of Science and Technology, Middlesex University, London NW4 4BT, UK  
e-mail: x.yang@mdx.ac.uk

analyzes adaptation and diversity in swarm intelligence in detail. Section 5 discusses the parameter tuning and control, and finally some conclusions will be drawn briefly, with some discussions for open problems in Sect. 6.

## 2 Some Recent Algorithms Based on Swarm Intelligence

Before we proceed to carry out any analysis, let us briefly introduce some popular nature-inspired, swarm-intelligence-based algorithms for global optimization.

From a mathematical point of view, an algorithm  $A$  is an iterative process, which aims to generate a new and better solution  $\mathbf{x}^{t+1}$  to a given problem from the current solution  $\mathbf{x}^t$  at iteration or (pseudo)time  $t$ . In general, an algorithm can be written as

$$\mathbf{x}^{t+1} = A(\mathbf{x}^t, p), \quad (1)$$

where  $p$  is an algorithm-dependent parameter. A good example is the so-called quasi-Newton method with a step size parameter.

The above formula is for a trajectory-based, single agent system. For population-based algorithms with a swarm of  $n$  solutions  $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ , we can extend the above iterative formula to a more general form

$$\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix}^{t+1} = A\left((\mathbf{x}_1^t, \mathbf{x}_2^t, \dots, \mathbf{x}_n^t); (p_1, p_2, \dots, p_k); (\epsilon_1, \epsilon_2, \dots, \epsilon_m)\right) \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix}^t, \quad (2)$$

where  $p_1, \dots, p_k$  are  $k$  algorithm-dependent parameters and  $\epsilon_1, \dots, \epsilon_m$  are  $m$  random variables. An algorithm can be viewed as a dynamical system, Markov chains and iterative maps [28], and it can also be viewed as a self-organized system [1].

Most nature-inspired algorithms nowadays are swarm intelligence based. Their updating equations vary significantly. However, most algorithms have linear updating equations. For example, particle swarm optimization has two linear equations in terms of  $\mathbf{x}$ . On the other hand, some algorithms such as the firefly algorithm use nonlinear updating equations, which can lead to rich characteristics and potentially higher efficiency.

Linear systems are easier to analyze, while nonlinear systems can be more challenging to analyze. At the moment, it still lacks in-depth understanding how different systems work. In the rest of this section, we will introduce some nature-inspired algorithms.

## 2.1 PSO

Particle swarm optimization (PSO) is one of the first algorithms that are based on swarm intelligence. PSO was developed by Kennedy and Eberhart in 1995 [14], based on the swarm behaviour of fish or bird schooling in nature. Each particle updates its position  $\mathbf{x}_i$  and velocity  $\mathbf{v}_i$ , and their evolution is controlled by two learning parameter  $\alpha$  and  $\beta$  with typical values of  $\alpha \approx \beta \approx 2$  and two random vectors  $\epsilon_1$  and  $\epsilon_2$  that are uniformly distributed in  $[0,1]$ . Briefly speaking, the main equations of PSO are as follows:

$$\mathbf{v}_i^{t+1} = \mathbf{v}_i^t + \alpha\epsilon_1[\mathbf{g}^* - \mathbf{x}_i^t] + \beta\epsilon_2[\mathbf{x}_i^* - \mathbf{x}_i^t], \quad (3)$$

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \mathbf{v}_i^{t+1}. \quad (4)$$

There are more than two dozen variants of PSO. For example, Yang et al. developed the accelerated PSO [20], while Fister Jr. et al. used some reasoning techniques to improve the efficiency of PSO [11].

## 2.2 Firefly Algorithm

The firefly algorithm (FA) is simple, flexible and easy to implement. FA was developed by Yang in 2008 [17], which was based on the flashing patterns and behaviour of tropical fireflies.

One of the main advantages of the FA is that FA can naturally deal with nonlinear multimodal optimization problems. The movement of a firefly  $i$  is attracted to another more attractive (brighter) firefly  $j$  is determined by

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \beta_0 e^{-\gamma r_{ij}^2} (\mathbf{x}_j^t - \mathbf{x}_i^t) + \alpha \epsilon_i^t, \quad (5)$$

where the second term is due to the attraction, and  $\beta_0$  is the attractiveness at  $r = 0$ . The third term is randomization with  $\alpha$  being the randomization parameter, and  $\epsilon_i^t$  is a vector of random numbers drawn from a Gaussian distribution at time  $t$ . Other studies also use the randomization in terms of  $\epsilon_i^t$  that can easily be extended to other distributions such as Lévy flights.

A comprehensive review of the firefly algorithm and its variants has been carried out by Fister et al. [6–8]. One novel feature of FA is that attraction is used, and this is the first of its kind in any SI-based algorithms. Since local attraction is stronger than long-distance attraction, the population in FA can automatically subdivide into multiple subgroups, and each group can potentially swarm around a local mode. Among all the local modes, there is always a global best solution which is the true optimality of the problem. Thus, FA can deal with multimodal problems naturally and efficiently.

### 2.3 Cuckoo Search

The cuckoo search (CS) was developed in 2009 by Yang and Deb [23]. CS is based on the brood parasitism of some cuckoo species. In addition, this algorithm is enhanced by the so-called Lévy flights [15], rather than by simple isotropic random walks.

Recent studies show that CS is potentially far more efficient than PSO and genetic algorithms [24, 25]. Mathematically speaking, CS uses a balanced combination of a local random walk and the global explorative random walk, controlled by a switching parameter  $p_a$ . The local random walk can be written as

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \alpha s \otimes H(p_a - \epsilon) \otimes (\mathbf{x}_j^t - \mathbf{x}_k^t), \quad (6)$$

where  $\mathbf{x}_j^t$  and  $\mathbf{x}_k^t$  are two different solutions selected randomly by random permutation,  $H(u)$  is a Heaviside function,  $\epsilon$  is a random number drawn from a uniform distribution, and  $s$  is the step size. On the other hand, the global random walk is carried out by using Lévy flights:

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \alpha L(s, \lambda), \quad (7)$$

where

$$L(s, \lambda) = \frac{\lambda \Gamma(\lambda) \sin(\pi \lambda / 2)}{\pi} \frac{1}{s^{1+\lambda}}, \quad (s > 0). \quad (8)$$

Here  $\alpha > 0$  is the step size scaling factor, which should be related to the scales of the problem of interest.

If we look at the CS equations from a mathematical point of view, we can analyze their key features and characteristics, and thus highlight their advantages. CS has two distinct advantages over other algorithms such as GA and SA, and these advantages are: efficient random walks and balanced mixing. Since Lévy flights are usually far more efficient than any other random-walk-based randomization techniques, CS can be efficient in global search. In fact, recent studies show that CS can have guaranteed global convergence [28].

On the other hand, the similarity between eggs can produce better new solutions, which is essentially fitness-proportional generation with a good mixing ability. In other words, CS has a varying mutation rate realized by Lévy flights, and the fitness-proportional generation of new solutions based on the solution similarity provides a subtle form of crossover. In addition, simulations also show that CS can have an autozooming ability in the sense that new solutions can automatically zoom into the region where the promising global optimality is located.

Using the framework of Markov chains and probability, we can see that equation (7) is essentially simulated annealing in the framework of Markov chains. In Eq. (6), if  $p_a = 1$  and  $\alpha s \in [0, 1]$ , CS can degenerate into a variant of differential evolution. Furthermore, if we replace  $\mathbf{x}_j^t$  by the current best solution  $\mathbf{g}^*$ , then (6) can further degenerate into accelerated particle swarm optimization (APSO) [20]. This means that SA, DE and APSO are special cases of CS, and that is one of the reasons why

CS is so efficient. A brief literature review has been carried out by Yang and Deb [26] and Fister Jr. et al. [9].

## 2.4 Bat Algorithm

The bat algorithm (BA) is the first algorithm of its kind to use frequency tuning for the optimization purpose. BA was developed by Yang in 2010 [18], inspired by the echolocation behavior of microbats. Each bat is associated with a velocity  $\mathbf{v}_i^t$  and a location  $\mathbf{x}_i^t$ , at iteration  $t$ , in a  $d$ -dimensional search or solution space. Among all the bats, there exists a current best solution  $\mathbf{x}_*$ . Therefore, the updating equations for  $\mathbf{x}_i^t$  and velocities  $\mathbf{v}_i^t$  can be written as

$$f_i = f_{\min} + (f_{\max} - f_{\min})\beta, \quad (9)$$

$$\mathbf{v}_i^t = \mathbf{v}_i^{t-1} + (\mathbf{x}_i^{t-1} - \mathbf{x}_*)f_i, \quad (10)$$

$$\mathbf{x}_i^t = \mathbf{x}_i^{t-1} + \mathbf{v}_i^t, \quad (11)$$

where  $\beta \in [0, 1]$  is a random vector drawn from a uniform distribution.

The motion of bats are updated by the above equations, but when to update and which branch is updated first are controlled by the loudness and pulse emission rate of each bat. In the most simplest case, the loudness and pulse emission rates are regulated by the following equations:

$$A_i^{t+1} = \alpha A_i^t, \quad (12)$$

and

$$r_i^{t+1} = r_i^0 [1 - \exp(-\gamma t)], \quad (13)$$

where  $0 < \alpha < 1$  and  $\gamma > 0$  are constants. Loosely speaking, here  $\alpha$  is similar to the cooling factor of a cooling schedule in simulated annealing.

There have been a lot of interest in the study of BA in recent years, and BA has been extended to multiobjective optimization [19] and various variants. For example, Fister et al. have extended to a hybrid bat algorithm [10, 12]. The preliminary results suggested that they are very efficient [21].

Obviously, there are other nature-inspired algorithms such as the flower pollination algorithm [22]. However, as the main purpose of this chapter is to analyze adaptation and diversity in metaheuristic algorithms, we will now focus on the analysis and discussion of the forms of adaptation and diversity and their roles/representations in the actual algorithms.

### 3 Mechanisms for Generating New Solutions

There are many ways for generating new solutions. However, from the locality point of view, they can be divided into the following subcategories:

- Modification of selected solutions (from the existing population).
- Local modifications.
- Global modifications.
- Mixed (both local and global as well as selected).

One of the most widely used methods for generating new solutions is to select a subset of existing solutions from the evolving population. For example, if two solutions are randomly selected from the existing population, they can be combined to form two new solutions by crossover or recombination. This is one of fundamental mechanisms in genetic algorithms and many evolutionary algorithms. In the simplest case when one solution is selected, some modification on a part (or a few parts) of the solution can be carried out. This is the main mechanism for mutation in genetic algorithms. In fact, these two ways of generating new solutions have paved the ways for most modern evolutionary algorithms.

The above operations can be converted to mathematical equations. Mathematically speaking, crossover can be written as

$$\begin{pmatrix} \mathbf{x}_i^{t+1} \\ \mathbf{x}_j^{t+1} \end{pmatrix} = C(\mathbf{x}_i^t, \mathbf{x}_j^t, p_c), \quad (14)$$

where  $p_c$  is the crossover probability, though the exact form of  $C()$  depends on the actual crossover manipulations. Mutation can be written schematically as

$$\mathbf{x}_i^{t+1} = M(\mathbf{x}_i^t, p_m), \quad (15)$$

where  $p_m$  is the mutation rate. However, the form  $M()$  depends on the coding and the number of mutation sites.

On the other hand, the fitness-dependent reproduction of the offsprings may depend on the relative fitness of the parents in the population. In this case, the function form can be even more complex. For example,  $C()$  can depend on all the individuals in the population, which may lead to  $C(\mathbf{x}_1^t, \mathbf{x}_2^t, \dots, \mathbf{x}_n^t, p_c)$  where  $n$  is the population size.

From the mathematical point of view, local modifications are local random walks, which can take many different forms and are usually around an existing solution. For example, from an existing solution  $\mathbf{x}_i^t$ , new solutions can be generated locally by using

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t + s(\mathbf{x}_i, \alpha), \quad (16)$$



where  $s(x_i^t, \alpha)$  is a step size function that can depend on the current solution and a parameter  $\alpha$ . If  $s$  is small enough, the distance  $d = \|x_i^{t+1} - x_i^t\|$  is small, which means the new solutions are limited to a neighborhood of the existing solution  $x_i^t$ . As random walks are widely used for randomization and local search in metaheuristic algorithms [17, 18], a proper step size is very important. As different algorithms use different forms of randomization techniques, it is not possible to provide a general analysis for assessing randomness. In addition, the above form of equation can in general be written in a more compact form as

$$x_i^{t+1} = N(x_i^t, w, \alpha), \tag{17}$$

where  $N(\cdot)$  depends on the random variable  $w$  with a parameter  $\alpha$ .

Here, randomness increases the diversity of the solutions and thus enables an algorithm to have the ability to jump out of any local optimum. However, too much randomness may slow down the convergence of the algorithm and thus can waste a lot of computational efforts. Therefore, there is some tradeoff between deterministic and stochastic components, though it is difficult to gauge what is the right amount of randomness in an algorithm? In essence, this question is related to the optimal balance of exploration and exploitation, which still remains an open problem.

Global modifications can also take many forms. For example, the simplest form of global modification or global randomization is

$$x_i = L + (U - L)\varepsilon, \tag{18}$$

where  $\varepsilon$  is a random number drawn in  $[0, 1]$ . This equation gives new solutions between the lower bound  $L$  and the upper bound  $U$ . On the other hand, random walks can be both local and global simultaneously. For example, the method in the cuckoo search uses Lévy flights in terms of

$$x_i^{t+1}(\text{new solution}) = x_i^t(\text{old solution}) + \alpha L(s, \lambda), \tag{19}$$

which can generate both local and global solutions, controlled by  $\alpha$  and the intrinsic nature of Lévy flights that provides occasional long-jumps. However, this is just a simple case where the new solution only depends on one existing solution and the randomization term. In general, the solutions can be generated in parallel by random permutation, and thus we may have a more generic form

$$\begin{pmatrix} x_1^{t+1} \\ x_2^{t+1} \\ \vdots \\ x_n^{t+1} \end{pmatrix} = G(x_1^t, x_2^t, \dots, x_n^t, w, \beta), \tag{20}$$

where  $G(\cdot)$  can be very complex, which also depends on the random variable and parameter  $\beta$ . For example, the mutation operator in differential evolution takes the form

$$\mathbf{x}_k^{t+1} = \mathbf{x}_k + F(\mathbf{x}_i - \mathbf{x}_j), \quad (21)$$

where  $i$ ,  $j$  and  $k$  are random permutations among  $1, 2, \dots, n$ , and  $F$  is a parameter or constant.

It is worth pointing out that the difference between global or local modifications are subtle. When the step sizes are large enough, local modifications can become global. Furthermore, these mechanisms for generating new solutions do not always belong to a single mechanism, and they can be a mixture of two or more components. For example, Lévy flights in the cuckoo search can be considered as a mixture of both local and global modifications, while the bat algorithm uses a combination of simple global randomization in one branch and the local modification in another branch, with the additional control for switching between these two branches depending on the loudness and pulse emission rate.

In fact, all good algorithms use a combination of the above components, not just a simple component. However, how to combine different modification methods is a challenging problem and what is the most efficient combination is yet to be discovered (if it ever exists). Furthermore, such effective combinations may be problem dependent and should be adaptive as well.

## 4 Adaptation and Diversity in Swarm Intelligence

Adaptation and diversity in metaheuristic algorithms can take many forms, including the balance of exploration and exploitation, generations or moves of new solutions, the right amount of randomness, parameter adjustment and parameter control, and other subtle forms. We will discuss the role of adaptation and diversity in such cases.

### 4.1 Diversity and Adaptation

The effectiveness of swarm intelligence based algorithms can be attributed to two important characteristics: adaptation and diversity of nature-inspired optimization algorithms.

Adaptation in nature-inspired algorithms can take many forms. For example, the ways to balance exploration and exploitation are the key form of adaptation [2]. As diversity can be intrinsically linked with adaptation, it is better not to discuss these two features separately. If exploitation is strong, the search process will use problem-specific information (or landscape-specific information) obtained during the iterative process to guide the new search moves; this may lead to the focused search and thus reduce the diversity of the population, which may help to speed up the convergence

of the search procedure. However, if exploitation is too strong, it can result in the quick loss of diversity in the population and thus may lead to the premature convergence. However, if new search moves are not guided by local landscape information, it can typically increase the exploration capability and generate new solutions with higher diversity. However, too much diversity and exploration may result in meandered search paths, thus lead to the slow convergence. Therefore, adaptation of search moves so as to balance exploration and exploitation is crucial. Consequently, to maintain the balanced diversity in a population is also important.

On the other hand, adaptation can also be in terms of the representations of solutions of a problem. In genetic algorithms, representations of solutions are usually in binary or real-valued strings [2, 13], while in swarm-intelligence-based algorithms, representations mostly use real number solution vectors. For example, the population size used in an algorithm can be fixed or varying. Adaptation in this case may mean to vary the population size so as to maximize the overall performance. For a given algorithm, adaptation can also occur to adjust its algorithm-dependent parameters. As the performance of an algorithm can largely depend on its parameters, the choice of these parameter values can be very important.

Parameter values can be varied so as to adapt the landscape type of the problem and thus may lead to better search efficiency. Such parameter tuning is in essence parameter adaptation. Once a parameter is tuned, it can remain fixed. However, there is no particular reason why parameters should be fixed. In fact, adaptation in parameter can be extended to parameter control. That is to control the parameter values in such a way that their values vary during the iterations so that optimal performance of the algorithm can be achieved.

Similarly, diversity in metaheuristic algorithms can also take many forms. The simplest diversity is to allow the variations of solutions in the population by randomization. For example, solution diversity in genetic algorithms is mainly controlled by the mutation rate and crossover mechanisms, while in simulated annealing, diversity is achieved by random walks. In most swarm-intelligence-based algorithms, new solutions are generated according to a set of deterministic equations, which also include some random variables. Diversity is represented by the variations, often in terms of the population variance. Once the population variance is getting smaller (approaching zero), diversity also decreases, leading to converged solution sets. However, if diversity is reduced too quickly, premature convergence may occur. Therefore, a right amount of randomness and the right form of randomization can be crucial.

From a different perspective, we can also say that adaptation and diversity can also be related to the selection of solutions among the population and the replacement of the old population. If the selection is based on the fitness, parent solutions with a higher level of fitness will be more likely to pass onto the next generation. In the extreme case, only the best solutions can be selected, which is a kind of elitism. If the replacement of worst solutions by new (hopefully better) solutions, this will ensure that better solutions will remain in the population. The balance of what to replace and what to pass on can be tricky, which requires good adaptation so as to maintain good diversity in the population.

## ***4.2 Exploration and Exploitation***

Adaptation and diversity are just one side of the coin. In the context of nature-inspired metaheuristics, the characteristics of an algorithm can also be analyzed in terms of basic components: exploitation and exploration, which are also referred to as intensification and diversification [3, 17].

Roughly speaking, exploitation uses any information obtained from the problem of interest so as to help to generate new solutions that are better than existing solutions. However, this process is typically local, and information (such as gradients) is also local. Therefore, it is for local search. For example, hill-climbing is a method that uses derivative information to guide the search procedure. In fact, new steps always try to climb up the local gradient. The advantage of exploitation is that it usually leads to very high convergence rates, but its disadvantage is that it can get stuck in a local optimum because the final solution point largely depends on the starting point. On the other hand, exploration makes it possible to explore the search space more efficiently, and it can generate solutions with enough diversity and far from the current solutions. Therefore, the search is typically on a global scale. The advantage of exploration is that it is less likely to get stuck in a local mode, and the global optimality can be more accessible. However, its disadvantages are slow convergence and waste of lot computational efforts because many new solutions can be far from global optimality.

Therefore, a fine balance is required so that an algorithm can achieve the best performance. Too much exploitation and too little exploration means the system may converge more quickly, but the probability of finding the true global optimality may be low. On the other hand, too little exploitation and too much exploration can cause the search path meander with very slow convergence. The optimal balance should mean the right amount of exploration and exploitation, which may lead to the optimal performance of an algorithm. Therefore, a proper balance is crucially important.

In essence, the optimal balance is itself a higher-level optimization problem. However, how to achieve such a balance is still an open problem. In fact, no algorithm can claim to have achieved such an optimal balance in the current literature. In essence, the balance itself is a hyper-optimization problem, because it is the optimization of an optimization algorithm. In addition, such a balance may depend on many factors such as the working mechanism of an algorithm, its setting of parameters, tuning and control of these parameters and even the problem to be considered. Furthermore, such a balance may not universally exist [16], and it may vary from problem to problem, thus requiring an adaptive strategy.

## ***4.3 Attraction and Diffusion***

The novel idea of attraction via light intensity as an exploitation mechanism was first used by Yang in the firefly algorithm (FA) in 2007 and 2008. In FA, the attractiveness

(and light intensity) is intrinsically linked with the inverse-square law of light intensity variations and the absorption coefficient. As a result, there is a novel but nonlinear term of  $\beta_0 \exp[-\gamma r^2]$  where  $\beta_0$  is the attractiveness at the distance  $r = 0$ , and  $\gamma > 0$  is the absorption coefficient for light [17]. The main function of such attraction is to enable an algorithm to converge quickly because these multi-agent systems evolve, interact and attract, leading to some self-organized behaviour and attractors. As the swarming agents evolve, it is possible that their attractor states will move towards to the true global optimality.

The novel attraction mechanism in FA is the first of its kind in the literature of nature-inspired computation and computational intelligence. This also motivated and inspired others to design similar or other kinds of attraction mechanisms. Other algorithms that were developed later also used inverse-square laws, derived from nature. For example, the charged system search (CSS) used Coulomb's law, while the gravitational search algorithm (GSA) used Newton's law of gravitation.

Whatever the attraction mechanism may be, from the metaheuristic point of view, the fundamental principles are the same: that is, they allow the swarming agents to interact with one another and provide a forcing term to guide the convergence of the population. Attraction mainly provides the mechanisms for exploitation, but, with proper randomization, it is also possible to carry out some degree of exploration. However, the exploration is better analyzed in the framework of random walks and diffusive randomization. From the Markov chain point of view, random walks and diffusion are both Markov chains. In fact, Brownian diffusion such as the dispersion of an ink drop in water is a random walk. Lévy flights can be more effective than standard random walks. Therefore, different randomization techniques may lead to different efficiency in terms of diffusive moves. In fact, it is not clear what amount of randomness is needed for a given algorithm.

All these unresolved issues and problems discussed so far may motivate more research in this area and thus the relevant literature can be expected to expand in the near future.

## 5 Parameter Tuning and Parameter Control

Adaptation and diversity can also take the form of parameter tuning and parameter control. In fact, one of the most challenging issues when designing metaheuristic algorithms is probably to control exploration and exploitation properly in terms of controlling algorithm-dependent parameters, which is still an open question. It is possible to control attraction and diffusion in algorithms that use such features so that the performance of an algorithm can be influenced in the right way.

Ideally we should have some mathematical relationships that can explicitly show how parameters can affect the performance of an algorithm, but this is an unresolved problem. In fact, unless for very simple cases under very strict, (often) unrealistic assumptions, no theoretical results exist at all. Obviously, one of the key questions

is how to tune parameters to gain the best parameter values so that an algorithm can perform in the most effective way.

### 5.1 Parameter Tuning

As an algorithm is a set of interacting Markov chains, we can in general write an algorithm as

$$\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix}^{t+1} = A[\mathbf{x}_1, \dots, \mathbf{x}_n, p_1(t), \dots, p_k(t), \epsilon_1, \dots, \epsilon_m] \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{pmatrix}^t, \quad (22)$$

which generates a set of new solutions  $(\mathbf{x}_1, \dots, \mathbf{x}_n)^{t+1}$  from the current population of  $n$  solutions. In principle, the behaviour of an algorithm is largely determined by the eigenvalues of the matrix  $A$  that are in turn controlled by the parameters  $p_k(t)$  and the randomness vector  $\epsilon = (\epsilon_1, \dots, \epsilon_m)$ .

From the Markovian theory, we know that the first largest eigenvalue is typically 1, and therefore the convergence rate of an algorithm is mainly controlled by the second largest eigenvalue  $0 \leq \lambda_2 < 1$  of  $A$ . However, it is extremely difficult to find this eigenvalue in general. Therefore, the tuning of parameters becomes a very challenging task. In fact, parameter tuning, or tuning of parameters, is an important topic under active research [5, 27]. The aim of parameter tuning is to find the best parameter setting so that an algorithm can perform most efficiently for a wider range of problems. At the moment, parameter tuning is mainly carried out by detailed, extensive parametric studies, and there is no efficient method in general.

In essence, parameter tuning itself is an optimization problem which requires higher-level optimization methods to tackle. However, a recent study shows that a framework for self-tuning algorithms can be established with promising results [27]. For example, Yang et al. used the firefly algorithm to tune itself so that the firefly algorithm can achieve optimal performance for a given set of problems. This framework can be expected to be applicable to other algorithms and a range of applications.

### 5.2 Parameter Control

Related to parameter tuning, there is another issue of parameter control. Parameter values after parameter tuning are often fixed during iterations, while parameters should vary for parameter control.

The main idea of parameter control is to vary the parameters so that the algorithm of interest can provide the best convergence rate and thus may achieve the best performance. Again, parameter control is another tough optimization problem to be yet resolved. In the bat algorithm, some basic form of parameter control has been attempted and found to be very efficient [18]. By controlling the loudness and pulse emission rate, BA can automatically switch from explorative moves to local exploitation that focuses on the promising regions when the global optimality may be nearby. Similarly, the cooling schedule in simulated annealing can be considered as a form of basic parameter control.

Both parameter tuning and parameter control are crucial to the performance of all algorithms, and thus deserve more research attention.

## 6 Discussions and Open Problems

As we have seen from the above detailed analysis, proper adaptation and diversity are crucial to ensure the good performance of an algorithm. Adaptation can be carried out in different components (of an algorithm), such as the generation of the population, selection of solutions, elitism, replacement of solutions, adjustment of parameters and overall balance of exploration and exploitation. Diversity can also appear in many places such as the ways to generate new solutions, selection and replacement of existing solutions, explorative moves, randomization, and most importantly to maintain a good balance in exploration and exploitation.

Despite the success of nature-inspired algorithms, there are still some challenging, open problems that need to be addressed. These open problems include the balance of exploration and exploitation, selection mechanisms, right amount of randomization, and parameter tuning as well as parameter control.

As mentioned in the main text, one of the most challenging problems is how to balance exploration and exploitation in an algorithm so that it can deal with a vast range of problems efficiently. In reality, the amount of exploration and exploitation may depend on the type of problem, and therefore, some a priori knowledge of the problem to be solved can help to determine such a balance. However, it is not known how to incorporate such knowledge effectively. For example, gradient/derivative information obtained from the objective function can be very useful for exploitation, but if such exploitation is too strong, it can cause the system to be trapped in a local optimum, thus sacrificing the possibility of finding the true global optimality. In order to balance exploration and exploitation, a right amount of randomness is needed. However, no one knows what amount is the right amount. At one extreme, if there is no randomness, an algorithm becomes a deterministic algorithm, and thus loses the ability to explore. At the other extreme, if the search is dominated by a high level of randomness, the algorithm becomes a random search, and thus significantly reduces its ability to exploit the landscape information. In fact, it is not known how to control randomness properly so as to balance exploration and exploitation most effectively.

Another important issue is the selection mechanism and it is not known what selection is most effective. A proper selection pressure is crucial to maintain a healthy population. For example, when many solutions have similar fitness, numerically speaking, their fitness values may almost be the same, thus how to select certain solutions becomes tricky. Typical approaches include re-scaled fitness values, ranking of solutions, and adaptive elitism [2]. However, it is not clear if they can work for all algorithms and if there is other better ways to handle selection.

On the other hand, as the performance of almost any algorithm will depend on its parameter settings, how to tune these parameters to achieve the best performance is a higher level optimization problem. In fact, this is the optimization of an optimization algorithm. It is still an open question. Similarly, how to control the parameters by varying their values to achieve the best overall performance is also a key challenging issue.

From the landscape point of view, the problems that have been solved in the current literature usually have fixed landscape. That is, once the problem is defined, its landscape in the search space remain unchanged. However, for dynamic problems and problems with noise, the search landscape can change with time. In such cases, adaptation can be more sophisticated and challenging. It is not clear if most current methods can still work well in such time-dependent, noisy environments.

It is worth pointing out that whatever the algorithms may be, the role of adaptation and diversity may be subtle in affecting the performance of an algorithm. Therefore, in-depth understanding and theoretical results are needed. Possible research routes may require a combination of mathematical analysis, numerical simulations, empirical observations as well as other tools such as dynamical system theories, Markov theory, self-organization theory and probability. It may even require a paradigm shift in analyzing metaheuristic algorithms.

Obviously, there are other issues and open problems as well. The above discussion has just focused a few key issues. All these challenges can present golden opportunities for further research in analyzing adaptation and diversity in metaheuristic algorithms. It can be expected more theoretical results will appear in the future, and any theoretical results will provide tremendous insight into understanding metaheuristic algorithms. It is hoped that efficient tools can be developed to solve a wide range of large-scale problems in real-world applications. Future research directions should focus on such key issues and challenges.

## References

1. Ashby WR (1962) Principles of the self-organizing system, in: Principles of self-organization: transactions of the University of Illinois symposium Von Foerster H, Zopf Jr. GW (eds) Pergamon Press, London, pp 255–278
2. Booker L, Forrest S, Mitchell M, Riolo R (2005) Perspectives on adaptation in natural and artificial systems. Oxford University Press, Oxford
3. Blum C, Roli A (2003) Metaheuristics in combinatorial optimisation: overview and conceptual comparison. *ACM Comput Surv* 35:268–308



4. Dorigo M, Di Caro G, Gambardella LM (1999) Ant algorithms for discrete optimization. *Artif Life* 5(2):137–172
5. Eiben AE, Smit SK (2011) Parameter tuning for configuring and analyzing evolutionary algorithms. *Swarm Evolutionary Comput* 1(1):19–31
6. Fister I, Fister I Jr, Yang XS, Brest J (2013) A comprehensive review of firefly algorithms. *Swarm Evol Comput* 13(1):34–46
7. Fister I, Yang X-S, Brest J, Fister I Jr (2013) Modified firefly algorithm using quaternion representation. *Expert Syst Appl* 40(18):7220–7230
8. Fister I, Yang XS, Fister D, Fister Jr. I (2014) Firefly algorithm: a brief review of the expanding literature. In: *Cuckoo Search Firefly Algorithm: Theor Appl Stud Comput Intell* 516:347–360 (Springer, Heidelberg)
9. Fister Jr I, Yang XS, Fister D, Fister I (2014) Cuckoo search: a brief literature review. In: *Cuckoo Search Firefly Algorithm: Theor Appl Stud Comput Intell* 516:49–62 (Springer, Heidelberg)
10. Fister I Jr, Fister D, Yang XS (2013) A hybrid bat algorithm. *Elektrotehniski Vestn* 80(1–2):1–7
11. Fister Jr I, Yang XS, Ljubič K, Fister D, Brest J, Fister I (2014) Towards the novel reasoning among particles in PSO by the use of RDF and SPARQL. *Sci World J* 2014, article ID 121782. doi:[10.1155/2014/121782](https://doi.org/10.1155/2014/121782)
12. Fister Jr I, Fong S, Brest J, Fister I (2014) A novel hybrid self-adaptive bat algorithm, *Sci World J*, 2014, article ID 709738. doi:[10.1155/2014/709738](https://doi.org/10.1155/2014/709738)
13. Holland J (1975) *Adaptation in natural and artificial systems*. University of Michigan Press, Ann Arbor
14. Kennedy J, Eberhart RC (1995) Particle swarm optimization. In: *Proceedings of IEEE international conference on neural networks*, Piscataway, NJ, pp 1942–1948
15. Pavlyukevich I (2007) Lévy flights, non-local search and simulated annealing. *J. Comput Phys* 226(12):1830–1844
16. Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. *IEEE Trans Evol Comput* 1(1):67–82
17. Yang XS (2008) *Nature-Inspired metaheuristic algorithms*. Luniver Press, Bristol
18. Yang XS (2010) A new metaheuristic bat-inspired algorithm. In: *Nature inspired cooperative strategies for optimisation (NICSO 2010)*, vol. 284. Springer, Berlin, *Studies in Computational Intelligence*, pp 65–74
19. Yang XS (2011) Bat algorithm for multi-objective optimisation. *Int J Bio-Inspired Comput* 3(5):267–274
20. Yang XS, Deb S, Fong S (2011) Accelerated particle swarm optimization and support vector machine for business optimization and applications. *Netw Digital Technol* 2011, *Commun Comput Inf Sci* 136:53–66
21. Yang XS, Gandomi AH (2012) Bat algorithm: a novel approach for global engineering optimization. *Eng Comput* 29(5):1–18
22. Yang XS (2012) Flower pollination algorithm for global optimization. In: *Unconventional computation and natural computation*, Springer, Berlin, pp. 240–249
23. Yang XS, Deb S (2009) Cuckoo search via Lévy flights. In: *Proceedings of world congress on nature & biologically inspired computing (NaBIC 2009)*. IEEE Publications, USA
24. Yang XS, Deb S (2010) Engineering optimization by cuckoo search. *Int J Math Model Numer Optimisation* 1(4):330–343
25. Yang XS, Deb S (2013) Multiobjective cuckoo search for design optimization. *Comput Oper Res* 40(6):1616–1624
26. Yang XS, Deb S (2014) Cuckoo search: recent advances and applications. *Neural Comput Appl* 24(1):169–174
27. Yang XS, Deb S, Loomes M, Karamanoglu M (2013) A framework for self-tuning optimization algorithm. *Neural Comput Appl* 23(7–8):2051–2057
28. Yang XS (2014) *Nature-Inspired optimization algorithms*. Elsevier, London

# Comparative Study on Recent Metaheuristic Algorithms in Design Optimization of Cold-Formed Steel Structures

M.P. Saka, S. Carbas, I. Aydogdu, A. Akin and Z.W. Geem

**Abstract** Sustainable construction aims at reducing the environmental impact of buildings on human health and natural environment by efficiently using energy, resources and reducing waste and pollution. Building construction has the capacity to make a major contribution to a more sustainable future of our World because this industry is one of the largest contributors to global warming. The use of cold-formed steel framing in construction industry provides sustainable construction which requires less material to carry the same load compare to other materials and reduces amount of waste mimum design algorithms are developed for cold-formed steel frames made of thin-walled sections using the recent metaheuristic techniques. The algorithms considered are firefly, cuckoo search, artificial bee colony with levy flight, biogeography-based optimization and teaching-learning-based optimization algorithms. The design algorithms select the cold-formed thin-walled C-sections listed in AISI-LRFD (American Iron and Steel Institution, Load and Resistance Factor Design) in such a way that the design constraints specified by the code are satisfied and the weight of the steel frame is the minimum. A real size cold-formed steel building is optimized by using each of these algorithms and their performance in attaining the optimum designs is compared.

---

M.P. Saka

Department of Civil Engineering, University of Bahrain, Isa Town, Bahrain  
e-mail: mpsaka@uob.edu.bh

S. Carbas

Civil Engineering Department, Karamanoglu Mehmetbey University, Karaman, Turkey  
e-mail: scarbas@kmu.edu.tr

I. Aydogdu

Civil Engineering Department, Akdeniz University, Antalya, Turkey  
e-mail: aydogdu@akdeniz.edu.tr

A. Akin

Thomas & Betts Corporation, Meyer Steel Structures, Memphis, TN 38125, USA  
e-mail: alperakin@yahoo.com

Z.W. Geem (✉)

Department of Energy IT, Gachon University, Seongnam, South Korea  
e-mail: geem@gachon.ac.kr

© Springer International Publishing Switzerland 2015

N.D. Lagaros and M. Papadrakakis (eds.), *Engineering and Applied Sciences Optimization*, Computational Methods in Applied Sciences 38,  
DOI 10.1007/978-3-319-18320-6\_9

**Keywords** Structural optimization · Discrete optimization · Cold-formed thin-walled steel frames · Metaheuristic techniques · Swarm intelligence · Firefly algorithm · Cuckoo search algorithm · Artificial bee colony algorithm · Biogeography-based optimization algorithm · Teaching-learning-based optimization algorithm

## 1 Introduction

Structural optimization aims at producing buildings that can be built by using the least amount of materials. This aim is of prime importance today because of the reason that buildings and construction works have the largest single share in global resource use and pollution emission [1]. World's climate is visible changing and its ecosystem is currently leading towards irreversible damages due to global warming. Carbon dioxide is the primary greenhouse gas emitted through human activities which is blamed for the global warming. Although energy production and transportation are two of the major source of carbon dioxide emissions, the construction industry also play important role in this respect. The importance of sustainable construction becomes even more apparent when one considers the fact that urban population swells by around one million people every week.

Steel is one of the most sustainable materials in the world. Since the early 1990s, the steel industry has reduced its energy use to produce a ton of steel by approximately one third. More than 95 % of the water used in the steel making process is recycled and returned. Every piece of steel used in construction contains recycled content. Further, all steel can be recovered and recycled again and again into new high quality products. Steel structures require less material to carry the same load as concrete or masonry or wood structures. The use of cold-formed steel framing in the construction industry even provides further economy. Furthermore cold-formed steel framing construction reduces the amount of waste generated at a site. This is due to the fact that almost the entire building project is pre-engineered and prepared using modern and efficient technology as framing members and panels in workshops or factories which are then transported and assembled in the site [2]. Cold-formed steel framing refers specifically to members in light-frame building construction that are made entirely of sheet steel formed to various shapes at ambient temperatures. The most common shape for cold-formed steel framing members is a lipped channel section although "Z", "C", "tubular", "hat" and other shapes have been used. Figure 1 shows an example of such construction which is environmentally friendly and has high sound and heat insulation.

Cold-formed members are produced from very thin steel sheets where the thickness varies between 0.4 and 6.4 mm. This thickness is very small compare to the widths of walls of member that they buckle before the stresses reach to yield stress when they are subjected to axial load, shear, bending or bearing. Therefore one of their major design criteria is based on the local buckling of walls of these sections [3, 4]. Furthermore open sections whether hot-rolled or cold-formed in general has relatively small torsional rigidity compare to closed sections. Plane sections do not remain



**Fig. 1** Steel building of cold-formed thin-walled open sections

plane and warping distortion takes place when subjected to torsional moments. Large warping deformations cause normal stresses in the cross section in addition to shear stresses. *Vlasov's* theory provides simple way of calculating these stresses [5, 6]. This theory extends the simple bending stress formula to cover the normal stresses that come out due to warping by just adding a similar term to the same formula. This additional term necessitates computation of two new cross sectional properties that are called the sectorial coordinate and warping moment of inertia of the cross-section. Normal stresses develop in thin-walled open section due to warping can be larger than the bending stress depending on the magnitude of the torsional moment section subjected to [7, 8]. It is shown in the literature that warping has substantial effect in the optimum design of steel frames made of thin-walled open sections [9]. In [10] strength and stability problems of mono-symmetrical complex thin-walled open section are studied using *Vlasov's* theory. Local instability is described according to the theory of thin plates and shells. The analytical solution is compared with the one attained from the finite element model constructed using shell elements. It is stated that the analytical results differ from that of finite element model on the stress distribution. However, the differences between maximum stress values are not so large. Lateral buckling of thin-walled beam under uniformly distributed transverse load, small longitudinal force and two different moments located at its both ends is studied in [11].

Several studies are carried out on the optimum shape design of thin-walled open sections of different shapes in last decade [12, 13]. In [14] cross-sectional design optimization is carried out for cold-formed steel channel and lipped channel columns under axial compression passing through the centroid of the cross-section. The design problem is formulated according to the provisions of AISI (American Iron and Steel Institute) [15, 16]. Flexural, torsional and torsional-flexural buckling of columns and flat-width-to-thickness ratio of web, flange and lip are considered as constraints as they are described in [15]. Micro-genetic algorithm is used in obtaining the solution

of design optimization problem. Micro-genetic algorithm uses relatively smaller population size compare to genetic algorithm which results in less computational time. In [17] three optimization methods steepest descent, genetic algorithm and simulated annealing are applied to obtain the optimum shape of cold-formed thin-walled steel columns under AISI provisions. Interesting optimum shapes are obtained by the algorithms developed and performances of optimization methods are compared. This work is extended to cover different cross-sectional geometries and boundary conditions in [18]. The literature review carried out reveals the fact that deterministic as well as stochastic optimization techniques are used to determine the solution of the shape optimization problem of cold-formed thin-walled sections. Furthermore, it is also noticed that most of the research has considered cold-formed single beam with different boundary conditions subjected to axial force, bi-axial bending moment and torsional moment. There are not many works on steel frames made of cold-formed sections. In one of the recent study real-coded genetic algorithm is utilized to develop optimum design algorithm for cold-formed steel portal frames which minimizes its cost [19]. The design variables consist of continuous and discrete variables. The spacing between main frames and pitch of the frame are taken as continuous design variables while the section sizes are to be selected from cold-formed steel section list are treated as discrete design variables. Constraints are implemented from Australian Code of Practice for cold-formed steel.

In this study the optimum design algorithm is developed for cold-formed steel frames made of thin-walled open sections. The design constraints are implemented from AISI-LRFD (American Iron and Steel Institute, Load and Resistance Factor Design, American Institute of Steel Construction) [20, 21]. Design constraints include the displacement limitations, inter-story drift restrictions, effective slenderness ratio, strength requirements for beams and combined axial and bending strength requirements which includes the elastic torsional lateral buckling for beam-columns. Furthermore additional constraints are considered to satisfy practical design requirements. The design algorithm selects the cold-formed sections for the frame members from the cold-formed thin-walled C-sections listed in AISI [22] such that the design constraints are satisfied and the weight of the steel frame is the minimum. Five recent metaheuristic algorithms are employed to determine the optimum solution of the design problem formulated and their performance is compared.

## **2 Discrete Optimum Design of Cold-Formed Steel Frames to AISI-LRFD**

The selection of cold-formed thin-walled C-sections for the members of steel frame is required to be carried out in such a way that the frame with the selected C-sections satisfies the serviceability and strength requirements specified by the code of practice while the economy is observed in the overall or material cost of the frame. When the

constraints are implemented from AISI-LRFD [15] in the formulation of the design problem the following discrete programming problem is obtained.

Find a vector of integer values  $\mathbf{I}$  (Eq. 1) representing the sequence numbers of C-sections assigned to ng member groups

$$\mathbf{I}^T = [I_1, I_2, \dots, I_{ng}] \quad (1)$$

to minimize the weight (W) of the frame

$$\text{Minimize} \quad W = \sum_{k=1}^{ng} m_k \sum_{i=1}^{nk} L_i \quad (2a)$$

Subject to

- **Serviceability Constraints:**

$$\frac{\delta_{jl}}{L/Ratio} - 1.0 \leq 0 \quad j = 1, 2, \dots, nsm, l = 1, 2, \dots, nlc \quad (2b)$$

$$\frac{\Delta_{jl}^{top}}{H/Ratio} - 1.0 \leq 0, \quad j = 1, 2, \dots, nj_{top}, l = 1, 2, \dots, nlc \quad (2c)$$

$$\frac{\Delta_{jl}^{oh}}{h_{sx}/Ratio} - 1.0 \leq 0, \quad j = 1, 2, \dots, n_{st}, l = 1, 2, \dots, nlc \quad (2d)$$

where,  $\delta_{jl}$  is the maximum deflection of  $j$ th member under the  $l$ th load case,  $L$  is the length of member,  $nsm$  is the total number of members where deflections limitations are to be imposed,  $nlc$  is the number of load cases,  $H$  is the height of the frame,  $nj_{top}$  is the number of joints on the top story,  $\Delta_{jl}^{top}$  is the top story displacement of the  $j$ th joint under  $l$ th load case,  $n_{st}$  is the number of story,  $nlc$  is the number of load cases and  $\Delta_{jl}^{oh}$  is the story drift of the  $j$ th story under  $l$ th load case,  $h_{sx}$  is the story height and  $Ratio$  is limitation ratio for lateral displacements described in ASCE Ad Hoc Committee report [23]. According to this report, the accepted range of drift limits by first-order analysis is 1/750 to 1/250 times the building height  $H$  with a recommended value of  $H/400$ . The typical limits on the inter-story drift are 1/500 to 1/200 times the story height. 1/400 is used in this study.

- **Strength Constraints: Combined Tensile Axial Load and Bending**

It is stated in AISI-LRFD that when a cold-formed members are subject to concurrent bending and tensile axial load, the member shall satisfy the interaction equations given C5.1 of [15] which is repeated below.

$$\frac{M_{ux}}{\phi_b M_{nxt}} + \frac{M_{uy}}{\phi_b M_{nyt}} + \frac{T_u}{\phi_t T_n} \leq 1.0 \quad (2e)$$

$$\frac{M_{ux}}{\phi_b M_{nx}} + \frac{M_{uy}}{\phi_b M_{ny}} - \frac{T_u}{\phi_t T_n} \leq 1.0 \quad (2f)$$

where,

- $M_{ux}, M_{uy}$  the required flexural strengths [factored moments] with respect to centroidal axes.  
 $\phi_b$  for flexural strength [moment resistance] equals 0.90 or 0.95 [21].  
 $M_{nxt}, M_{nyt}$   $S_{ft}F_y$  (where,  $S_{ft}$  is the section modulus of full unreduced section relative to extreme tension fiber about appropriate axis and  $F_y$  is the design yield stress).  
 $T_u$  required tensile axial strength [factored tension].  
 $\phi_t$  0.95 [21].  
 $T_n$  nominal tensile axial strength [resistance].  
 $M_{nx}, M_{ny}$  nominal flexural strengths [moment resistances] about centroidal axes.

• **Strength Constraints: Combined Compressive Axial Load and Bending**

It is stated in AISI-LRFD that when a cold-formed members are subject to concurrent bending and compressive axial load, the member shall satisfy the interaction equations given in C5.2 of [15] which is repeated below.

For  $\frac{P_u}{\phi_c P_n} > 0.15$ ,

$$\frac{P_u}{\phi_c P_n} + \frac{C_{mx}M_{ux}}{\phi_b M_{nx}\alpha_x} + \frac{C_{my}M_{uy}}{\phi_b M_{ny}\alpha_y} \leq 1.0 \quad (2g)$$

$$\frac{P_u}{\phi_c P_{no}} + \frac{M_{ux}}{\phi_b M_{nx}} + \frac{M_{uy}}{\phi_b M_{ny}} \leq 1.0 \quad (2h)$$

For  $\frac{P_u}{\phi_c P_n} \leq 0.15$ ,

$$\frac{P_u}{\phi_c P_n} + \frac{M_{ux}}{\phi_b M_{nx}} + \frac{M_{uy}}{\phi_b M_{ny}} \leq 1.0 \quad (2i)$$

where,

- $P_u$  required compressive axial strength [factored compressive force].  
 $\phi_c$  0.85 [21].  
 $M_{ux}, M_{uy}$  the required flexural strengths [factored moments] with respect to centroidal axes of effective section.  
 $\phi_b$  for flexural strength [moment resistance] equals 0.90 or 0.95 [21]  
 $M_{nx}, M_{ny}$  the nominal flexural strengths [moment resistances] about centroidal axes.

and

$$\alpha_x = 1 - \frac{P_u}{P_{E_x}} > 0.0, \quad \alpha_y = 1 - \frac{P_u}{P_{E_y}} > 0.0 \quad (2j)$$

where,

$$P_{E_x} = \frac{\pi^2 E I_x}{(K_x L_x)^2}, \quad P_{E_y} = \frac{\pi^2 E I_y}{(K_y L_y)^2} \quad (2k)$$

where,

$I_x$	moment of inertia of full unreduced cross section about $x$ axis.
$K_x$	effective length factor for buckling about $x$ axis.
$L_x$	unbraced length for bending about $x$ axis.
$I_y$	moment of inertia of full unreduced cross section about $y$ axis.
$K_y$	effective length factor for buckling about $y$ axis.
$L_y$	unbraced length for bending about $y$ axis.
$P_{no}$	nominal axial strength [resistance] determined in accordance with Section C4 of AISI [22], with $F_n = F_y$ .
$C_{mx}, C_{my}$	coefficients taken as 0.85 or 1.0.

• **Allowable Slenderness Ratio Constraints:**

The maximum allowable slenderness ratio of cold-formed compression members has been limited to 200.

$$\frac{K_x * L_x}{r_x} \text{ or } \frac{K_y * L_y}{r_y} < 200 \quad (21)$$

where,

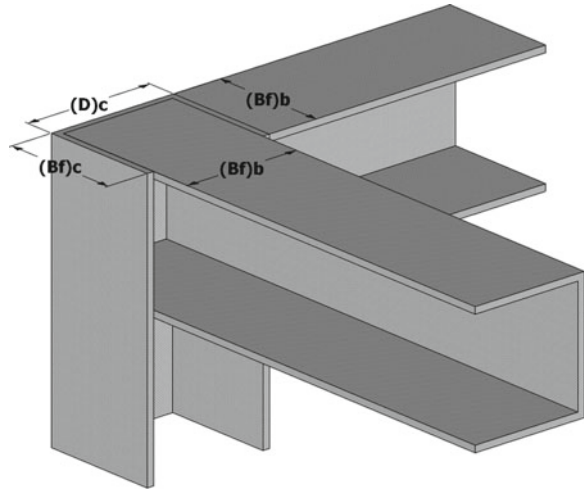
$K_x$	effective length factor for buckling about $x$ axis
$L_x$	unbraced length for bending about $x$ axis
$K_y$	effective length factor for buckling about $y$ axis
$L_y$	unbraced length for bending about $y$ axis
$r_x, r_y$	radius of gyration of cross section about $x$ and $y$ axes.

• **Geometric Constraints:**

Geometric constraints are required to make sure that C-section selected for the columns of two consecutive stories are either equal to each other or the one above storey is smaller than the one in the below storey. Similarly when a beam is connected to flange of a column, the flange width of the beam is less than or equal to the flange width of the column in the connection. Furthermore when a beam is connected to the web of a column, the flange width of the beam is less than or equal to  $(D - 2t_b)$  of the column web dimensions in the connections where  $D$  and  $t_b$  are the depth and the flange thickness of C-section as shown in Fig. 2.



**Fig. 2** Typical beam-column connection of C-section



$$\frac{D_i^a}{D_i^b} - 1 \leq 0 \quad \text{and} \quad \frac{m_i^a}{m_i^b} - 1 \leq 0, \quad i = 1, \dots, n_{ccj} \quad (2m)$$

$$\frac{B_i^{bi}}{D_i^{ci} - 2t_b^{ci}} - 1 \leq 0, \quad i = 1, \dots, n_{j1} \quad (2n)$$

$$\frac{B_f^{bi}}{B_f^{ci}} - 1 \leq 0, \quad i = 1, \dots, n_{j2} \quad (2o)$$

where  $n_{ccj}$  is the number of column-to-column geometric constraints defined in the problem,  $m_i^a$  is the unit weight of  $C$ -section selected for above story,  $m_i^b$  is the unit weight of  $C$ -section selected for below story,  $D_i^a$  is the depth of  $C$ -section selected for above story,  $D_i^b$  is the depth of  $C$ -section selected for below story,  $n_{j1}$  is the number of joints where beams are connected to the web of a column,  $n_{j2}$  is the number of joints where beams connected to the flange of a column,  $D_i^{ci}$  is the depth of  $C$ -section selected for the column at joint  $i$ ,  $t_b^{ci}$  is the flange thickness of  $C$ -section selected for the column at joint  $i$ ,  $B_f^{ci}$  is the flange width of  $C$ -section selected for the column at joint  $i$  and  $B_f^{bi}$  is the flange width of  $C$ -section selected for the beam at joint  $i$ .

Computation of nominal axial tensile strength  $T_n$ , nominal axial compressive strength  $P_n$ , nominal flexural strengths about centroidal axis  $M_{nx}$  and  $M_{ny}$  are given in [15] which requires consideration of elastic flexural buckling stress, elastic flexural-torsional buckling stress and distortional buckling strength. Each of these is calculated through use of certain expression given in the design code. Repetition of these expressions is not possible due to lack of space in the article. Hence reader is referred to references [3, 4, 15]. The design problem described through Eqs. 2a–2o

turns out to be discrete programming problem. The solution of the design program necessitates selection of cold-formed C-sections from the available list such that the design constraints (2b)–(2o) which are implemented from the design code are satisfied and the objective function given in Eq. 2a has the minimum value.

### 3 Metaheuristic Algorithms

Obtaining the solution of optimization problems with discrete variables is much harder than solving the optimization problems with continuous variables. Although mathematical programming techniques such as integer programming, branch and bound method and dynamic programming are available for attaining the solution of discrete programming problems, the literature survey related with these techniques reveals the fact they present numerical adversities in finding the solution of large and complex design optimization problems designer face in practice [24, 25]. On the other hand stochastic search methods that are known as metaheuristic techniques are quite efficient in determining the solution of discrete programming problems [26–31]. The fundamental properties of metaheuristic algorithms are that they imitate certain strategies taken from nature, social culture, biology or laws of physics which are used to direct the search process. Their goal is to efficiently explore the search space using these governing mechanisms in order to find near optimal solutions if not global optimum. They also utilize some strategies to avoid getting trapped in confined areas of search space. Furthermore they do not even require an explicit relationship between the objective function and the constraints. They are not problem specific and proven to be very efficient and robust in obtaining the solution of practical engineering design optimization problems with both continuous and discrete design variables [32–34]. In this study the solution of the design optimization problem described in the previous section is obtained by using five recent metaheuristic algorithms and their performance is compared. These are firefly algorithm, cuckoo search algorithm, artificial bee colony algorithm, biogeography-based optimization algorithm and teaching-learning-based optimization algorithms which are developed after 2005. Brief description of each algorithm is given in the following.

#### 3.1 Firefly Algorithm

Firefly algorithm is originated by Yang [35–37] and it is based on the idealized behaviour of flashing characteristics of fireflies. These insects communicate, search for pray and find mates using bioluminescence with varying flaying patterns. The firefly algorithm is based on three rules. These are:

1. All fireflies are unisex so they attract one another.
2. Attractiveness is propositional to firefly brightness. For any couple of flashing fireflies, the less bright one moves towards the brighter one. Attractiveness is proportional to the brightness and they both decrease as their distance increases. If there is no brighter one than a particular firefly, it will move randomly.
3. The brightness of a firefly is affected or determined by the landscape of the objective function.

**Attractiveness:** In the firefly algorithm attractiveness of a firefly is assumed to be determined by its brightness which is related with the objective function. The brightness  $i$  of a firefly at a particular location  $x$  can be chosen as  $I(x) \propto f(x)$  where  $f(x)$  is the objective function. However, the attractiveness  $\beta$  is relative; it should be judged by the other fireflies. Thus, it will vary with the distance  $r_{ij}$  between firefly  $i$  and firefly  $j$ . In addition, light intensity decreases with the distance from its source, and light is also absorbed in the media. In the firefly algorithm the attractiveness function is taken to be proportional to the light intensity by adjacent fireflies and it is defined as;

$$\beta(r) = \beta_0 e^{-\gamma r^m}, \quad (m \geq 1) \quad (3)$$

where  $\beta_0$  is the attractiveness at  $r = 0$ .

**Distance:** The distance between any two fireflies  $i$  and  $j$  at  $x_i$  and  $x_j$  is calculated as

$$r_{ij} = \|x_i - x_j\| = \sqrt{\sum_{k=1}^d (x_{i,k} - x_{j,k})^2} \quad (4)$$

where  $x_{i,k}$  is the  $k$ th component of the spatial coordinate  $x_i$  of the  $i$ th firefly.

**Movement:** The movement of a firefly  $i$  which is attracted to another brighter firefly  $j$  is determined by

$$x_i = x_i + \beta_0 e^{-\gamma r_{ij}^2} (x_j - x_i) + \alpha \left( rand - \frac{1}{2} \right) \quad (5)$$

where the second term is due to the attraction while the third term is randomization with  $\alpha$  being the randomization parameter. “rand” is a random number generator uniformly distributed in  $[0, 1]$ .

The values of parameters in the above equations are generally taken as  $\beta_0 = 1$  and  $\alpha \in [0, 1]$ . Randomization term can be extended to a normal distribution  $N(0, 1)$  or other distributions.  $\gamma$  characterizes the variation of the attractiveness, and its value determines the speed of convergence and performance of the firefly algorithm. In most applications its value is taken between 0 and 100. The pseudo code of the algorithm is given in [35–37] which is repeated in Fig. 3.

The firefly algorithm is applied to determine engineering as well as structural size, shape and topology design optimization problems [37–39]. In [37] firefly algorithm

**Firefly Algorithm**


---

```

Objective function  $f(x)$ ,  $\{x\} = \{x_1, \dots, x_d\}^T$ 
Generate initial population of fireflies  $x_i, (i = 1, \dots, n)$ 
Light intensity  $I_i$  at  $x_i$  is determined by  $f(x_i)$ 
Define light absorption coefficient  $\gamma$ 
  while (until the termination criteria is satisfied)
    for  $i = 1 : n$  all  $n$  fireflies
      for  $j = 1 : i$  all  $n$  fireflies
        if  $(I_j > I_i)$ 
          Move firefly  $i$  towards  $j$  in  $d$ -dimension
        end if
        Attractiveness varies with distance  $r$  via  $\exp[-\gamma r^2]$ 
        Evaluate new solutions and update light intensity
      end for  $j$ 
    end for  $i$ 
    Rank the fireflies and find the current best
  end while
Postprocess results and visualization

```

---

**Fig. 3** Pseudo code of firefly algorithm

is used to determine optimum solution of six engineering design problems that are taken from the literature and its performance is compared with other metaheuristic algorithms such as particle swarm optimizer, differential evolution, genetic algorithm, simulated annealing, harmony search method and others. It is stated that the results attained from the optimum solutions of these design examples firefly algorithm is more efficient than particle swarm optimizer, genetic algorithm, simulated annealing and harmony search method.

### 3.2 Cuckoo Search Algorithm

Cuckoo search algorithm is originated by Yang and Deb [40] which simulates reproduction strategy of cuckoo birds. Some species of cuckoo birds lay their eggs in the nests of other birds so that when the eggs are hatched their chicks are fed by the other birds. Sometimes they even remove existing eggs of host nest in order to give more probability of hatching of their own eggs. Some species of cuckoo birds are even specialized to mimic the pattern and color of the eggs of host birds so that host bird could not recognize their eggs which give more possibility of hatching. In spite of all these efforts to conceal their eggs from the attention of host birds, there is still a possibility that host bird may discover alien eggs. In such cases the host bird either throws these alien eggs away or simply abandons its nest and builds a new one somewhere else. In cuckoo search algorithm cuckoo egg represents a potential solution to the design problem which has a fitness value. The algorithm uses three idealized rules as given in [40]. These are: (a) each cuckoo lays one egg at a time and dumps it in a randomly selected nest. (b) the best nest with high quality eggs will be carried over to the next generation. (c) the number of available host nests is

**Cuckoo Search Algorithm****Begin;***Initialize a population of n host nests  $x_i, i = 1, 2, \dots, n$ ;***while** (*until the termination criterion is satisfied*);*Get a cuckoo randomly, (let it be  $x_i$ )**and generate a new solution by Levy flights;**Evaluate its fitness (let it be  $F_i$ );**Choose a nest among n nests randomly, (let it be  $x_j$ );***if** ( $F_i > F_j$ )*replace  $x_j$  by the new solution  $x_i$ ;***end***Abandon a fraction ( $P_a$ ) of worse nests and**built new ones at new locations via levy flights;**Keep the best nests (or solutions);**Rank the solutions and find the current best;***end while***Post process results;***end procedure;****Fig. 4** Pseudo code for cuckoo search algorithm

fixed and a host bird can discover an alien egg with a probability of  $p_a \in [0, 1]$ . In this case the host bird can either throw the egg away or abandon the nest to build a completely new one in somewhere else. The pseudo code of the cuckoo search algorithm is given in Fig. 4.

Cuckoo search algorithm initially requires selection of a population of n eggs each of which represents a potential solution to the design problem under consideration. This means that it is necessary to generate n solution vector of  $\mathbf{x} = \{x_1, \dots, x_{ng}\}^T$  in a design problem with  $ng$  variables. For each potential solution vector the value of objective function  $f(\mathbf{x})$  is also calculated. The algorithm then generates a new solution  $\mathbf{x}_i^{v+1} = \mathbf{x}_i^v + \beta\lambda$  for cuckoo  $i$  where  $\mathbf{x}_i^{v+1}$  and  $\mathbf{x}_i^v$  are the previous and new solution vectors.  $\beta > 1$  is the step size which is selected according to the design problem under consideration.  $\lambda$  is the length of step size which is determined according to random walk with Levy flights. A random walk is a stochastic process in which particles or waves travel along random trajectories consists of taking successive random steps. The search path of a foraging animal can be modeled as random walk. A Levy flight is a random walk in which the steps are defined in terms of the step-lengths which have a certain probability distribution, with the directions of the steps being isotropic and random. Hence Levy flights necessitate selection of a random direction and generation of steps under chosen Levy distribution.

Mantegna [41] algorithm is one of the fast and accurate algorithms which generate a stochastic variable whose probability density is close to Levy stable distribution characterized by arbitrary chosen control parameter  $\alpha$  ( $0.3 \leq \alpha \leq 1.99$ ). Using the Mantegna algorithm, the step size  $\lambda$  is calculated as

$$\lambda = \frac{x}{|y|^{1/\alpha}} \quad (6)$$

where  $x$  and  $y$  are two normal stochastic variables with standard deviation  $\sigma_x$  and  $\sigma_y$  which are given as

$$\sigma_x(\alpha) = \left[ \frac{\Gamma(1 + \alpha) \sin(\pi\alpha/2)}{\Gamma((1 + \alpha)/2) \alpha 2^{(\alpha-1)/2}} \right]^{1/\alpha} \quad \text{and} \quad \sigma_y(\alpha) = 1 \quad \text{for } \alpha = 1.5 \quad (7)$$

in which the capital Greek letter  $\Gamma$  represents the Gamma function ( $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ ) that is the extension of the factorial function with its argument shifted down by 1 to real and complex numbers. If  $z = k$  is a positive integer  $\Gamma(k) = (k-1)!$ .

Cuckoo search algorithm is applied to structural optimization problems as well as optimum design of steel frames in [42–44]. It is shown in these applications that cuckoo search algorithm performs better than particle swarm optimizer, big bang-big crunch algorithm and imperialist competitive algorithm. It finds lighter optimum designs.

### 3.3 Artificial Bee Colony Algorithm with Levy Flight

The artificial bee colony algorithm is suggested by Karaboga et al. [45–50]. It mimics the foraging behaviour of a honey bee colony. In a honey bee colony, there are three types of bees which carry out different tasks. The first group of bees are the *employed bees* that locate food source, evaluate its amount of nectar and keep the location of better sources in their memory. These bees when fly back to hive they share this information to other bees in the dancing area by dancing. The dancing time represents the amount of nectar in the food source. The second group are the *onlooker bees* who observe the dance and may decide to fly to the food source if they find it is worthwhile to visit the food source. Therefore food sources that are rich in the amount of nectar attract more onlooker bees. The third group are *scout bees* that explore new food sources in the vicinity of the hive randomly. The employed bee whose food source has been abandoned by the bees becomes a scout bee. Overall, scout bees carry out the exploration, employed and onlooker bees perform the task of exploitation. Each food source is considered as a possible solution for the optimization problem and the nectar amount of a food source represents the quality of the solution which is identified by its fitness value.

The artificial bee colony algorithm consists of four stages. These stages are initialization phase, employed bees phase, onlooker bees' phase and scout bees phase. These stages are summarized below for the optimization problem of  $Min. z = f(\mathbf{x})$  where  $\mathbf{x}$  is vector of  $n$  design variables.

1. **Initialization phase:** Initialize all the vectors of the population of food sources,  $\mathbf{x}_p$ ,  $p = 1, \dots, np$  by using Eq. 8 where  $np$  is the population size (total number of artificial bees). Each food source is a solution vector consisting of  $n$  variables ( $x_{pi}$ ,  $i = 1, \dots, n$ ) is a potential solution to the optimization problem.

$$x_{pi} = x_{\ell i} + rand(0, 1)(x_{ui} - x_{\ell i}) \quad (8)$$

where  $x_{\ell i}$  and  $x_{ui}$  are upper and lower bound on  $x_i$ .  $rand(0, 1)$  is a random number between 0 and 1.

3. **Employed bees phase:** Employed bees search new food sources by using Eq. 9.

$$v_{pi} = x_{pi} + \varphi_{pi}(x_{pi} - x_{ki}) \quad (9)$$

where  $k \neq i$  is a randomly selected food source,  $\varphi_{pi}$  is a random number in range  $[-1, 1]$ . After producing the new food source (solution vector) its fitness is calculated. If its fitness is better than  $x_{pi}$  the new food source replaces the previous one. The fitness value of the food sources is calculated according to Eq. 10.

$$fitness(x_p) = \frac{1}{1 + f(x_p)} \quad (10)$$

where  $f(x_p)$  is the objective function value of food source  $x_p$ .

4. **Onlooker bees' phase:** Unemployed bees consist of two groups. These are onlooker bees and scouts. Employed bees share their food source information with onlooker bees. Onlooker bees choose their food source depending on the probability value  $P_\ell$  which is calculated using the fitness values of each food source in the population as shown in Eq. 11.

$$P_\ell = \frac{fitness(x_p)}{\sum_{p=1}^{np} fitness(x_p)} \quad (11)$$

After a food source  $x_{pi}$  for an onlooker bee is probabilistically chosen, a neighbourhood source is determined by using Eq. 8 and its fitness value is computed using Eq. 10.

5. **Scout bees phase:** The unemployed bees who choose their food sources randomly called scouts. Employed bees whose solutions cannot be improved after predetermined number of trials (PNT) become scouts and their solutions are abandoned. These scouts start to search for new solutions.

The pseudo code of the artificial bee colony algorithm is given in Fig. 5.

Artificial bee colony algorithm is widely used to obtain the solutions of structural optimization problems [51–57]. It is concluded in these studies that artificial bee colony algorithm is robust and efficient technique that performs better than some other metaheuristic algorithms such as genetic algorithm, ant colony algorithm, particle swarm optimizer, big bang-big crunch and imperialist competitive algorithms.

**Artificial Bee colony Algorithm**


---

*Initialize the population of solutions  $x_{ij}$  and evaluate the population*

*while* (until termination criteria is satisfied)

- *Produce new solutions  $v_{ij}$  in the neighbourhood of  $x_{ij}$  for the employed bees using (9)*
- *Apply the greedy selection process between  $x_i$  and  $v_i$*
- *Calculate the probability values  $P_i$  for the solutions  $x_i$  using (11)*
- *Normalize  $P_i$  values into  $[0,1]$*
- *Produce the new solutions  $v_i$  for the onlookers from solutions  $x_i$  selected depending on  $P_i$  and evaluate their fitness*
- *Apply the greedy selection process for the onlookers between  $x_i$  and  $v_i$*
- *Determine the abandoned solution, if exists, and replace it with a new randomly produced solution  $x_i$  for the scout using (8)*
- *Memorize the best food source position (solution) achieved so far*

*end while*

---

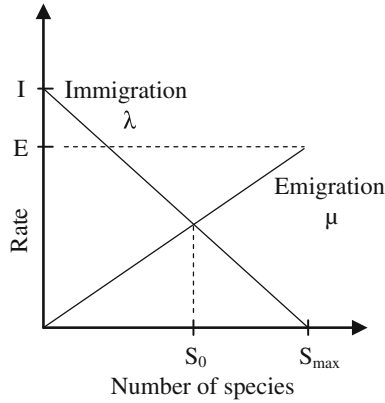
**Fig. 5** Pseudo code of the artificial bee colony algorithm

### 3.4 Biogeography-Based Optimization Algorithm

Biogeography-based optimization algorithm is developed by Simon [58] which is based on the theory of island biogeography. Mathematical model of biogeography describes the migration and extinction of species between islands. An island is any area of suitable habitat which is isolated from the other habitats. Islands that are friendly to life are said to have high habitat suitability index (HSI). Features that correlate with HSI include such factors as rainfall, diversity of vegetation, diversity of topographic features, land area, and temperature. The variables that characterize habitability are called suitability index variables (SIV). SIVs can be considered the independent variables of the habitat, and HSI can be considered the dependent variable. Naturally habitats with a high HSI tend to have a large number of species while those with a low HSI have a small number of species. Habitats with a high HSI have many species that emigrate to nearby habitats, simply by virtue of the large number of species that they host. Habitats with a high HSI have a low species immigration rate because they are already nearly saturated with species. Therefore, high HSI habitats are more static in their species distribution than low HSI habitats. This fact is used in biogeography based optimization for carrying out migration. Relationship between species count, immigration rate, and emigration rate is shown in Fig. 6 [58], where  $I$  refers to the maximum immigration rate,  $E$  is the maximum emigration rate,  $S_0$  is the equilibrium number of species and  $S_{max}$  is the maximum species count.

The decision to modify each solution is taken based on the immigration rate of the solution.  $\lambda_k$  is the immigration probability of independent variable  $x_k$ . If an independent variable is to be replaced, then the emigrating candidate solution is chosen with a probability that is proportional to the emigration probability  $\mu_k$  which is usually performed using roulette wheel selection.





**Fig. 6** Species model of a single habitat where  $\lambda$  is immigration rate and  $\mu$  is emigration rate

$$P(x_j) = \frac{\mu_j}{\sum_{i=1}^N \mu_i} \quad \text{for } j = 1, \dots, N \quad (12)$$

where  $N$  is the number of candidate solutions in the population.

Mutation is also another factor which is used to increase the species richness of islands. This increases the diversity among the population. Each candidate solution is associated with a mutation probability defined by

$$m(s) = m_{\max} \left( \frac{1 - P_s}{P_{\max}} \right) \quad (13)$$

$m_{\max}$  is a user defined parameter.  $P_s$  is the species count of the habitat,  $P_{\max}$  is the maximum species count. Mutation is carried out on the mutation probability of each habitat. The steps of the biogeography based optimization algorithm can be listed as follows [59].

1. Set up initial population; define the migration and mutation probabilities.
2. Calculate the immigration and emigration rates for each candidate solution in the population
3. Select the island to be modified based on the immigration rate.
4. Using roulette wheel selection on the emigration rate, select the island from which the SIV is to be immigrated.
5. Randomly select an SIV from the island to be emigrated.
6. Perform mutation based on the mutation probability of each island.
7. Calculate the fitness of each individual island
8. If the fitness criterion is satisfied go to step 2.

The pseudo code of biogeography-based optimization algorithm is given in Fig.7 [60].

**Biogeography-Based Optimization Algorithm**


---

```

For each solution  $y_k, k \in \{1, \dots, N\}$ , define emigration probability  $\mu_k \propto \text{fitness of } y_k, \mu_k \in [0,1]$ 
For each solution  $y_k$  define immigration probability  $\lambda_k = 1 - \mu_k$ 
 $z \leftarrow y$ 
For each solution  $z_k$ 
  For each solution feature  $s$ 
    Use  $\lambda_k$  to probabilistically decide whether to immigrate to  $z_k$ 
    If immigrating then
      Use  $\{\mu_j\}$  to probabilistically select the emigrating solution  $y_j$ 
       $z_k(s) \leftarrow y_j(s)$ 
    end if
  next solution feature
  Probabilistically mutate  $z_k$ 
next solution
 $y \leftarrow z$ 

```

---

**Fig. 7** Pseudo code for one generation of biogeography-based optimization algorithm

### 3.5 Teaching-Learning-Based Optimization Algorithm

Teaching-learning-based optimization algorithm is also population based process which mimics the influence of a teacher on learners [61]. The population represents class of learners. Different design variables in an optimum design problem are considered as different subjects offered to the learners. Learners' achievement is analogous to the fitness value of the objective function. In the entire population the best solution is considered as the teacher. The algorithm consists of two phases; teacher phase and learner phase. In the teacher phase class learns from a teacher and in the learner phase learning takes place through the interaction among the learners.

In the teacher phase the learning process of learners through a teacher is replicated. A good teacher puts an effort to bring the level of learners higher in terms of knowledge. However, in reality it is not only the effort of a teacher which can raise the level of knowledge of learners. The capability of learners also plays an important role in this process. Hence it is a random process. Supposing there are "m" number of subjects (design variables) offered to "n" number of learners (population size,  $k = 1, 2, \dots, n$ ). At any sequential teaching-learning cycle  $i$ , let  $T_i$  be the teacher and  $M_i$  be the mean of learners' achievements.  $T_i$  will try to move mean  $M_i$  to a higher level. After the teaching of  $T_i$  there will be a new mean, say  $M_{new}$ . The solution is updated according to the difference between the existing and the new mean as:

$$\text{Difference\_Mean} = r_i (M_{new} - T_F M_i) \quad (14)$$

where  $T_F$  is a teaching factor that decides the value of mean to be changed,  $r_i$  is a random number in the range of  $[0, 1]$ . The value of  $T_F$  can be either 1 or 2. It is not

a parameter in the algorithm which is computed randomly as  $T_F = \text{round} [1 + \text{rand}(0,1) \{2-1\}]$ . The difference calculated in Eq. 14 modifies the existing solution as

$$x_{new,i} = x_{old,i} + \text{Difference\_Mean} \quad (15)$$

In learners' phase the learning process of learners through interaction among themselves is imitated. A learner interacts randomly with other learners with the help of group discussions, presentations, and formal communications. It should be noticed that a learner can learn more unless the other learner has more knowledge than her or him. In this phase randomly two learners say  $x_i$  and  $x_j$  are selected where  $i \neq j$ . Learner modification is then expressed as follows:

$$x_{new,i} = x_{old,i} + r_i (x_i - x_j) \quad \text{if } f(x_i) < f(x_j) \quad (16)$$

$$x_{new,i} = x_{old,i} + r_i (x_j - x_i) \quad \text{if } f(x_i) > f(x_j) \quad (17)$$

$x_{new,i}$  is accepted if it gives a better function value. This process is repeated for the learners in the population. The pseudo code of the algorithm is given in Fig. 8.

Teaching-learning-based optimization algorithm is used to develop structural optimization algorithms in [62–64]. It is shown in these studies that teaching-learning-based optimization algorithm is robust and efficient algorithm that produced better optimum solutions that those metaheuristic algorithms considered for comparison.

#### Teaching-Learning-Based Optimization Algorithm

---

```

Initialize the population size and number of generations.
Generate a random population. Calculate the values of objective function for each learner.
While (number of generation is not reached)
    Calculate the mean of each design variable;  $x_{mean}$ 
    Identify the best solution as teacher [ $x_{teacher} \Rightarrow x$  with  $f(x)_{\min}$ ]
    for  $i = 1 \rightarrow n$ 
        Calculate teaching factor  $T_{F,i} = \text{round} [1 + \text{rand}(0,1)\{2-1\}]$ 
        Modify solutions based on teacher  $x_{new,i} = x_i + \text{rand}(0,1)[x_{teacher} - T_{F,i} x_{mean}]$ 
        Calculate the objective function value  $f(x_{new,i})$  for  $x_{new,i}$ 
        If  $f(x_{new,i}) < f(x_i)$  then replace  $x_i = x_{new,i}$ 
        Select a learner randomly, say  $x_j$  such that  $j \neq i$ 
        If  $f(x_i) < f(x_j)$  then
             $x_{new,i} = x_{old,i} + r_i (x_i - x_j)$ 
        Else
             $x_{new,i} = x_{old,i} + r_i (x_j - x_i)$ 
        End if
        If  $f(x_{new,i}) < f(x_i)$  then replace  $x_i = x_{new,i}$ 
    End for
End while

```

---

**Fig. 8** Pseudo code for teaching-learning-based optimization algorithm

## 4 Constraint Handling

Metaheuristic algorithms are developed to obtain the solution of unconstrained optimization problems. However, almost all of the structural design problems are constrained optimization problems. It is apparent that it becomes necessary to transform the constrained optimum design problem into unconstrained one if one intends to use metaheuristic algorithms for obtaining its solution. One way to achieve this is to utilize a penalty function. In this study the following function is used in this transformation.

$$W_p = W (1 + C)^\varepsilon \quad (18)$$

where  $W$  is the value of objective function of optimum design problem given in 2a.  $W_p$  is the penalized weight of structure,  $C$  is the value of total constraint violations which is calculated by summing the violation of each individual constraint.  $\varepsilon$  is penalty coefficient which is taken as 2.0 in this work.

$$C = \sum_{i=1}^{nc} c_i \quad (19)$$

$$c_i = \begin{cases} 0 & \text{if } g_j \leq 0 \\ g_j & \text{if } g_j > 0 \end{cases} \quad j = 1, \dots, nc \quad (20)$$

where  $g_j$  is the  $j$ th constraint function and  $nc$  is the total number of constraints in the optimum design problem. Constraint functions for the steel frame made of cold-formed sections are given through in Eqs. 2b–2o. It should be reminded that all the constraints are required to be normalized similar to constraint given in Eq. 2n before they are used in the metaheuristic algorithms.

## 5 Optimum Design Algorithms with Discrete Variables

Five optimum design algorithms are coded each of which is based on the metaheuristic algorithms summarized above. The solution of the discrete optimum design problem given in Eqs. 2a–2o is obtained using these algorithms. In all the optimum design techniques the sequence number of the steel C-sections in the standard list is treated as design variable. For this purpose complete set of 85 C-sections starting from 4CS2x059 to 12CS4x105 as given in AISI [22] is considered as a design pool from which the optimum design algorithms select C-sections for frame members. Once a sequence number is selected, then the sectional designation and properties of that section becomes available from the section table for the algorithm. The metaheuristic algorithms mentioned in Sect. 3 assume continuous design variables. However the design problem considered requires discrete design variables. This necessity is

resolved by rounding the numbers obtained through each algorithm. For example Eq. 8 of artificial bee colony algorithm is written as

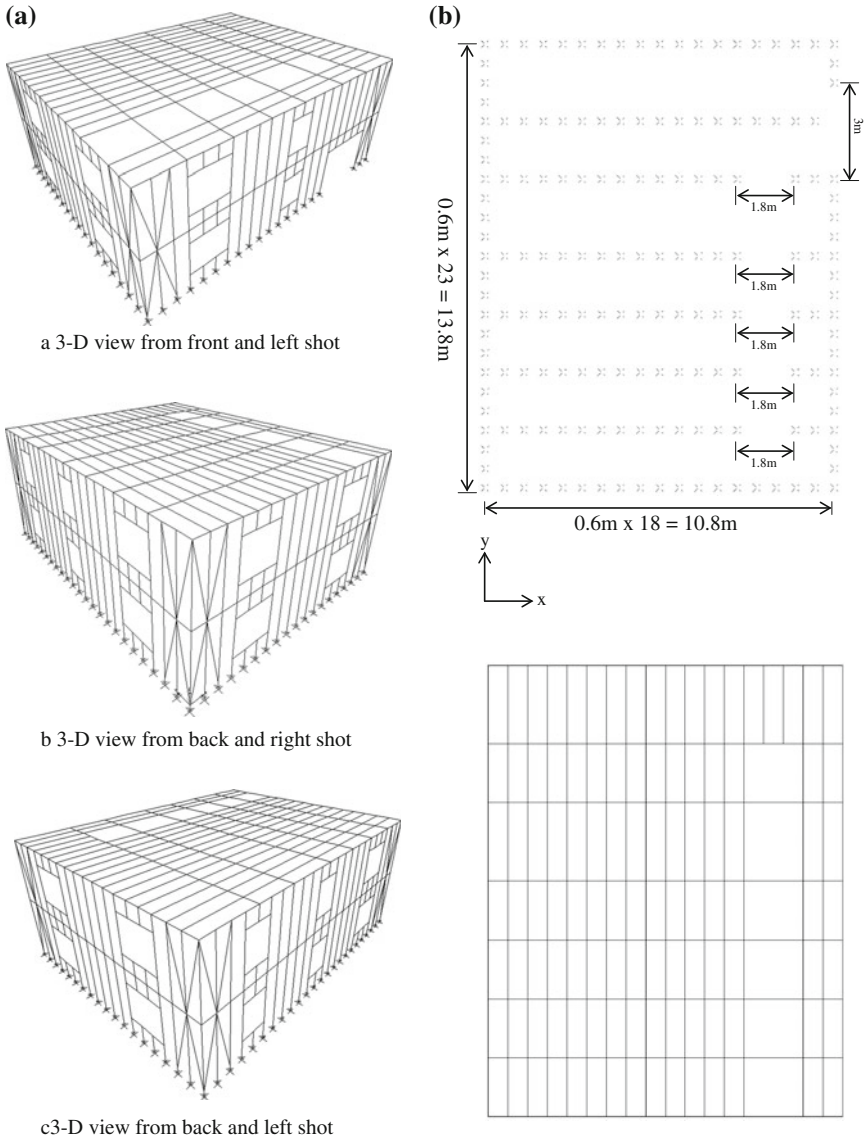
$$I_{pi} = I_{\min} + INT[rand(0, 1)(I_{\max} - I_{\min})], \quad i = 1, \dots, ng, \quad p = 1, \dots, np \quad (21)$$

where  $I_{pi}$  is the integer value for  $x_{pi}$ , the term  $rand(0, 1)$  represents a random number between 0 and 1,  $I_{\min}$  is equal to 1 and  $I_{\max}$  is the total number of values in the discrete set for C-section respectively which is equal to 85.  $ng$  is the total number of design variables and  $np$  is the number of bees in the colony which is equal to  $(neb+nob)$  where  $(neb)$  is the number of employed bees and  $(nob)$  is the number of onlooker bees. The similar adjustments are carried out in other metaheuristic algorithms wherever discrete value are needed for a design variable

The analysis of steel frames is achieved by using matrix displacement method. Noticing the fact that steel frames made of cold-formed thin-walled sections are quite slender structures, large deformations compare to their initial dimensions may take place under external loads. In structures with large displacements, although the material behaves linear elastic, the response of the structure becomes nonlinear [65]. Under certain types of loading, namely, even when small deformations are presumed, nonlinear behavior can be predicted. Changes in stiffness and loads occur as the structure deforms. In such structures, it is necessary to take into account the effect of axial forces to member stiffness. This is achieved by carrying out P- $\delta$  analysis in the application of the stiffness method. In each design cycle when the cross sectional properties of members is changed, steel frame is analyzed by constructing the nonlinear stiffness matrix where the interaction between bending moments and axial forces is considered through the use of stability functions. The details of the derivation of the nonlinear stiffness matrix and consideration of geometric nonlinearity in the analysis of steel frames made of thin-walled sections are given in [66].

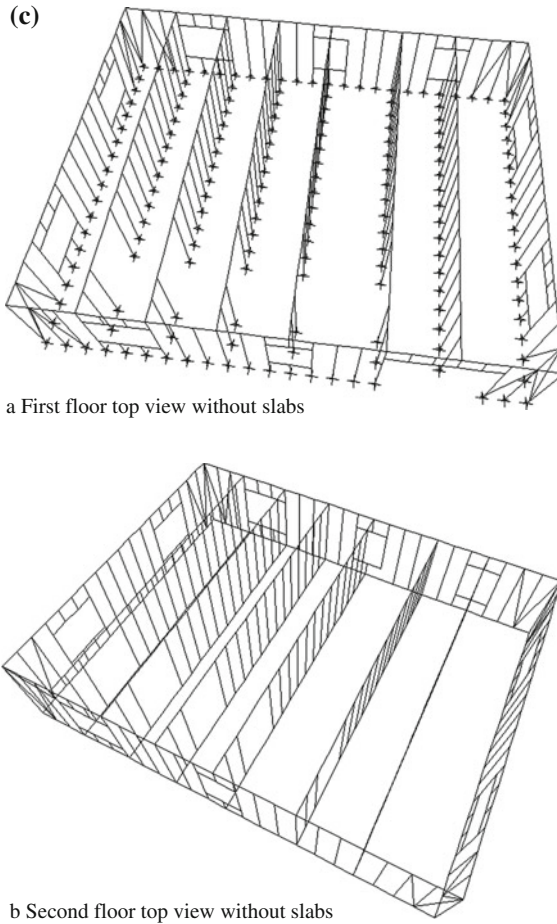
## 6 Design Example

Two-storey, 1211-member lightweight cold-formed steel space frame shown in Fig. 9 is selected to study the performance evaluation of five different metaheuristic algorithms. 3-D, plan and floor views of the frame are shown in the same figure respectively. The spacing between columns is decided to be 0.6 m span and each floor has 2.8 m height. The total height of the building is 5.6 m. The space frame consists of 708 joints (including supports) and 1211 members that are grouped into 14 independent member groups which are treated as design variables. The member grouping of the frame is illustrated in Table 1. The frame is subjected to gravity and lateral loads, which are computed as per given in ASCE 7-05 [67]. The loading consists of a design dead load of 2.89 kN/m<sup>2</sup>, a design live load of 2.39 kN/m<sup>2</sup>, a ground snow load of 0.755 kN/m<sup>2</sup>. Unfactored wind load values are taken as 0.6 kN/m<sup>2</sup>. The load and combination factors are applied according to code specifications of LRFD-AISC [21] as; Load Case 1: 1.2D+1.6L+0.5S, Load Case 2: 1.2D+0.5L+1.6S and Load



**Fig. 9** 1211-member three dimensional lightweight cold-formed steel frame, **a** 3-D views from different shots, **b** Plan views, **c** First and second floors top views without slabs

Case 3:  $1.2D+1.6WX+1.0L+0.5S$  where D represents dead load, L is live load, S is snow load and WX is the wind load applied on X global direction respectively. The top story drift in both X and Y directions are restricted to 14 mm and inter-story drift limitation is specified to 7 mm. The complete single C-section with lips list



**Fig. 9** (continued)

given in AISI Design Manual 2007 [22] which consists of 85 section designations is considered as a design pool for design variables.

The light weight cold-formed steel frame is designed by using five different optimum design algorithm each of which is based on one of the metaheuristic algorithms summarized in Sect. 3. Each metaheuristic algorithm has certain parameters to be initially decided by users. The values adopted for these parameters are given in Table 2 related to each metaheuristic algorithm. Maximum number of iterations is taken as 20,000 for all the algorithms to provide equal opportunity for these techniques. The optimum solutions are obtained after the number of iterations that is much smaller than 20,000.

The optimum designs determined by these five different optimization algorithms are listed in Table 3. It is interesting to notice that all the algorithms have almost found

**Table 1** The member grouping of 1211-member lightweight cold-formed steel frame

Storey	Beams outer short	Beams inner short	Beams inner gates	Beams windows	Beams outer gate
1	1	2	3	4	5
2	1	2	3	4	–
Storey	Columns connected short beams	Columns connected long beams	Columns near inner gates	Columns windows	Braces
1	6	7	8	9	14
2	10	11	12	13	14

**Table 2** Algorithm parameter values used in the design example

Firefly Algorithm (FFA)	Number of fireflies = 50, $\alpha = 0.5, \gamma = 1, \beta_{\min} = 0.2, \beta = 1.0$
Cuckoo Search Algorithm (CSA)	Number of nests = 40, $p_a = 0.90$
Artificial Bee Colony (ABC)	Total number of bees = 50, Maximum cycle number = 400 Limiting value for number of cycles to abandon food source = 250
Biogeography-Based Optimization (BBO)	Population size = 20, Maximum number of generation = 400, Elitism parameter = 2, Mutation probability = 0.01
Teaching-learning-based Optimization (TLBO)	Number of students = 50, Maximum number of generations = 200

optimum designs that are very close to each others. Among all, the Biogeography-Based Optimization (BBO) has attained the best global optimum design with the minimum weight of 53.584 kN (5464.05 kg). The second best solution is determined by Teaching-Learning-Based Optimization (TLBO) where the optimum weight of the frame is 53.677 kN (5473.53 kg) which is only 0.17 % heavier than the optimum design attained by BBO. In fact the difference between the lightest and the heaviest optimum designs is only 1.2 %. This indicates the fact that all these recent metaheuristic algorithms namely firefly algorithm, cuckoo search algorithm, artificial bee colony algorithm, biogeography-based optimization algorithm and teaching-learning-based optimization algorithm are robust and efficient metaheuristic algorithms that are can be used in confidence in solving structural design optimization problems. Inspection of the constraint values given in Table 3 clearly shows that the strength constraints are dominant in the design optimization problem. Almost in all the algorithms the maximum strength ratio is very close to 1.0 while displacement and inter-story drift constraints are much less than their upper bounds.

The convergence history of each algorithm is shown in Fig. 10. It is apparent from this figure that BBO and TLBO have much better convergence rate than firefly (FFA) and artificial bee colony (ABC) algorithms. Although it exhibits rapid conver-



**Table 3** Optimum design results of 1211-member lightweight steel frame

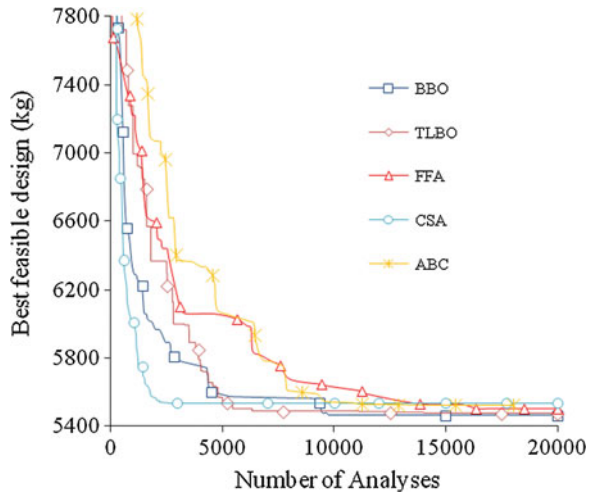
Group No	Group type	Sections selected by BBO algorithm	Sections selected by TLBO algorithm	Sections selected by FFA algorithm	Sections selected by ABC algorithm	Sections selected by CSO algorithm
1	1st and 2nd floors outer short beams	4CS2x085	4CS2x070	4CS2x105	4CS2.5x059	4CS2x070
2	1st and 2nd floors inner short beams	4CS2x059	4CS2x059	4CS2x059	4CS2.5x059	4CS2x059
3	1st and 2nd floors inner gates' beams	8CS2x059	6CS2.5x059	7CS2.5x059	6CS2.5x059	7CS2.5x059
4	1st and 2nd floors windows' beams	4CS2x059	4CS2.5x059	4CS2x059	4CS2x059	4CS2.5x059
5	1st floor outer gate beams	4CS2x059	12CS4x105	4CS2.5x065	4CS2x059	4CS2x059
6	1st floor columns connected short beams	4CS2x059	4CS2x059	4CS2x059	4CS2.5x059	4CS2x059
7	1st floor columns connected long beams	4CS4x059	4CS2x059	4CS2x059	4CS2.5x059	4CS2x059
8	1st floor columns near inner gates	4CS2x059	4CS2x059	4CS2x059	4CS2.5x059	4CS2x059
9	1st floor windows' columns	4CS2x059	4CS4x059	4CS2.5x070	4CS2.5x059	4CS4x059

(continued)

**Table 3** (continued)

Group No	Group type	Sections selected by BBO algorithm	Sections selected by TLBO algorithm	Sections selected by FFA algorithm	Sections selected by ABC algorithm	Sections selected by CSO algorithm
10	2nd floor columns connected short beams	4CS2x059	4CS2.5x059	4CS2x059	4CS2x059	4CS2x059
11	2nd floor columns connected long beams	4CS4x059	4CS2x059	4CS2x059	4CS4x059	4CS2x059
12	2nd floor columns near inner gates	12CS3.5x085	12CS4x105	10CS3.5x070	4CS2.5x059	12CS4x105
13	2nd floor windows' columns	4CS2x059	4CS2x059	4CS2.5x070	4CS2x059	4CS2.5x059
14	1st and 2nd floors braces	4CS2x059	4CS2x059	4CS2x059	4CS2x059	4CS2x059
	Minimum weight (kN (kg))	53.584 (5464.05)	53.677 (5473.53)	53.962 (5502.61)	54.161 (5522.91)	54.228 (5529.74)
	Maximum top storey drift (mm)	9.158	7.885	8.774	7.284	7.568
	Maximum inter-storey drift (mm)	1.759	2.433	2.022	2.761	1.882
	Maximum deflection (mm)	0.199	0.247	0.198	0.336	0.239
	Maximum strength ratio	0.938	0.998	0.937	0.999	0.938
	Maximum number of iterations	20,000	20,000	20,000	20,000	20,000

**Fig. 10** Search histories of 1211-member lightweight cold-formed steel frame



gence performance to reach optimum solution, the worst design is yielded by CSA producing optimum frame weight as 54.228 kN (5529.74 kg). However considering the fact that the difference between the lightest and heaviest optimum designs is only 1.2%, it can be concluded that the performance of the all metaheuristic algorithms considered in this study is efficient in this particular design optimization problem.

## 7 Conclusions

The use of cold-formed thin-walled steel framing in construction industry provides sustainable construction requiring less material to carry the same load. The concept of sustainable building has become quite important due to the rapid increase of human population. The optimum design algorithm developed for cold-formed light weight steel buildings reduces the required amount of material even further level helping the sustainability of the construction. The design procedure selects the optimum cold-formed C-section designations from the section list such that design constraints described in AISI-LRFD are satisfied and the light weight steel frame has the minimum weight. In view of the results obtained it can be concluded that the metaheuristic algorithms considered in this study that are firefly algorithm, cuckoo search algorithm, artificial bee colony optimization algorithm, biogeography-based optimization and teaching-learning-based optimization algorithm all yield an efficient and robust design optimization technique that can successfully be employed in optimum design of light weight cold-formed steel frames. The difference between the heaviest and the lightest optimum designs attained by these algorithms is only 1.2% which is not significant. The metaheuristic algorithms selected do not require initial selection of too many parameters. Except the firefly algorithm, the rest of

the metaheuristic techniques considered needs selection of two parameters, namely population size and maximum number of generations which is the minimum number of parameters that would be required in such procedures. The total number of structural analysis required to reach the optimum design is high similar to most of metaheuristic algorithms. This number may be reduced by carrying out some enhancements in these algorithms such as adding levy flights for random walk. It was not possible to perform comparison of the optimum designs attained in this study with other designs due to the fact that there is no other publication in literature that considers the same design code provisions.

## References

1. <http://www.isover.com/Our-Commitment-to-sustainability/Toward-sustainable-buildings/>. Accessed Oct 2014
2. <http://greenmaltese.com/2012/08/cold-formed-steel/>. Accessed Oct 2014
3. Ghersi A, Landolfo R, Mazzolani FM (2005) Design of metallic cold-formed thin-walled members. Spon Press, Great Britain
4. Yu W-W, LaBoube RA (2010) Cold-formed steel design, 4th edn. Wiley, New York
5. Vlasov VZ (1961) Thin-walled elastic beams. National Science Foundation, Washington
6. Zbirohowski-Koscia K (1967) Thin-walled beams. Crosby-Lockwood Ltd, London
7. Trahair NS, Bild S (1990) Elastic biaxial bending and torsion of thin-walled members. *Thin-Walled Struct* 9:269–307
8. Trahair NS (2003) Lateral buckling strengths of steel angle sections beams. *J Struct Eng ASCE* 129(6):784–791
9. Aydogdu I, Saka MP (2012) Ant colony optimization of irregular steel frames including elemental warping effect. *Adv Eng Softw* 44:150–169
10. Magnucki K, Szyz W, Stasiewicz P (2004) Stress state and elastic buckling of a thin-walled beam with mono-symmetrical open cross-section. *Thin-Walled Struct* 42(1):25–38
11. Magnucka-Blandzi E (2009) Critical state of a thin-walled beam under combined load. *Appl Math Modell* 33:3093–3098
12. Magnucki K, Monczak T (2000) Optimum shape of open cross section of thin-walled beam. *Eng Optim* 32:335–351
13. Al-Mosawi S, Saka MP (2000) Optimum shape design of cold-formed thin walled steel sections. *Adv Eng Softw* 31:851–862
14. Lee J, Kim S-M, Park H-S (2006) Optimum design of cold-formed steel columns by using micro genetic algorithms. *Thin-Walled Struct* 44:952–960
15. AISI (American Iron and Steel Institute) (2001) North American specification for the design of cold-formed steel structural members
16. AISI (American Iron and Steel Institute) (2002) Cold-formed steel design manual
17. Leng J, Guest JK, Schafer BW (2011) Shape optimization of cold-formed steel columns. *Thin-Walled Struct* 49:1492–1503
18. Moharrami M, Louhghalam A, Tootkaboni M (2014) Optimal folding of cold-formed steel cross sections under compression. *Thin-Walled Struct* 76:145–156
19. Phan DT, Lim JBP, Sha W, Siew CYM, Tanyimboh TT, Issa H, Mohammed FA (2013) Design optimization of cold-formed steel portal frames taking into account the effect of building topology. *Eng Optim* 45(4):415–433
20. AISI (American Iron and Steel Institute) S100–07 (2007) North American specification for the design of cold-formed steel structural members
21. AISC (American Institute of Steel Construction) (1991) LRFD, Volume I, Structural members, Specifications and code, Manual of steel construction

22. AISI (American Iron and Steel Institute) D100-08 (2008) Excerpts-gross section property tables, cold-formed steel design manual. Part I: Dimensions and properties
23. Ad Hoc Committee on Serviceability (1986) Structural serviceability: a critical appraisal and research needs. *J Struct Eng ASCE* 112(12):2646–2664
24. Saka MP (2003) Optimum design of skeletal structures: a review, Chapter 10. In: Topping BHV (ed) *Progress in civil and structural engineering computing*. Saxe-Coburg Publications, Stirlingshire, pp 237–284
25. Saka MP (2007) Optimum design of steel frames using stochastic search techniques based on natural phenomena: a review, Chapter 6. In: Topping BHV (ed) *Civil engineering computations: tools and techniques*. Saxe-Coburgh Publications, Stirlingshire, pp 105–147
26. Yang X-S (2008) *Nature-inspired metaheuristic algorithms*. Luniver Press, Bristol
27. Yang X-S (2010) *Engineering optimization: an introduction with metaheuristic applications*. Wiley, New York
28. Lamberti L, Pappalettere C (2011) Metaheuristic design optimization of skeletal structures: a review. *Comput Technol Rev* 4:1–32
29. Saka MP (2012) Recent developments in metaheuristic algorithms: a review. *Comput Technol Rev* 5:31–78
30. Saka MP, Geem ZW (2013) Mathematical and metaheuristic applications in design optimization of steel frame structures: an extensive review. *Math Probl Eng*
31. Saka MP, Dogan E, Aydogdu I (2013) Review and analysis of swarm-intelligence based algorithms, Chapter 2. In: Yang X-S, Cui Z, Xiao R, Gandomi AH, Karamanoglu M (eds) *Swarm intelligence and bio-inspired computation, theory and applications*. Elsevier, Amsterdam, pp 25–47
32. Hasançebi O, Çarbaş S, Doğan E, Erdal F, Saka MP (2009) Performance evaluation of metaheuristic search techniques in the optimum design of real size pin jointed structures. *Comput Struct* 87(5–6):284–302
33. Hasançebi O, Çarbaş S, Doğan E, Erdal F, Saka MP (2010) Comparison of non-deterministic search techniques in the optimum design of real size steel frames. *Comput Struct* 88(17–18):1033–1048
34. Lagaros ND (2014) A general purpose real-world structural design optimization computing platform. *Struct Multidiscip Optim* 49(6):1047–1066
35. Yang X-S (2009) Firefly algorithms for multimodal optimization, Chapter. In: Watanabe O, Zeugmann T (eds) *Stochastic algorithms: foundations and applications*. SAGA 2009, Lecture Notes in Computer Science 5792. Springer, Berlin, pp 169–178
36. Yang X-S (2010) Firefly algorithm, Lévy flights and global optimization, Chapter. In: Bramer M et al (eds) *Research and development in intelligent systems XXVI*. Springer, London, pp 209–218
37. Gandomi AH, Yang X-S, Alavi AH (2011) Mixed variable structural optimization using firefly algorithm. *Comput Struct* 89:2325–2336
38. Sayadi MK, Remazanian R, Ghaffari-Nasab N (2010) A discrete firefly metaheuristic with local search for make-span minimization in permutation flow shop scheduling problems. *Int J Ind Eng Comput* 1:1–10
39. Miguel LFF, Lopez RH, Miguel LFF (2013) Multimodal size, shape, and topology optimization of truss structures using the firefly algorithm. *Adv Eng Softw* 56:23–37
40. Yang X-S, Deb S (2010) Engineering optimization by cuckoo search. *Int J Math Model Numer Optim* 1(4):330–343
41. Mantegna RN (1994) Fast, accurate algorithm for numerical simulation of Levy stable stochastic processes. *Phys Rev* 49(5):4677–4683
42. Gandomi AH, Yang X-S, Alavi AH (2011) Cuckoo search algorithm: a metaheuristic approach to solve structural optimization problems. *Eng Comput* 29:17–35
43. Kaveh A, Bakhspoori T (2011) Optimum design of steel frames using cuckoo search algorithm with Levy flights. *Struct Des Tall Spec Build* 22(13):1023–1036
44. Saka MP, Dogan E (2012) Design optimization of moment resisting steel frames using a cuckoo search algorithm. In: Topping BHV (ed) *Proceedings of the eleventh conference on computational structures technology*, Paper 71. Dubrovnik, Croatia, 4–7 Sept

45. Karaboga D (2005) An idea based on honey bee swarm for numerical optimization. Technical report-TR06. Erciyes University, Engineering Faculty, Computer Engineering Department
46. Karaboga D, Basturk B (2007) A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. *J Global Optim* 39(3):459–471
47. Karaboga D, Akay B (2009) A comparative study of artificial bee colony algorithm. *Appl Math Comput* 214:108–132
48. Karaboga D, Basturk B (2007) Artificial bee colony (ABC) optimization algorithm for solving constrained optimization problems. *Adv Soft Comput: Found Fuzzy Logic Soft Comput* 4529:789–798
49. Karaboga D, Basturk B (2008) On the performance of artificial bee colony (ABC) algorithm. *Appl Soft Comput* 8(1):687–697
50. Akay B, Karaboga D (2012) Artificial bee colony algorithm for large scale problems and engineering design optimization. *J Intell Manuf* 23:1001–1014
51. Hadidi A, Azad SK, Azad SK (2010) Structural optimization using artificial bee colony algorithm. In: Proceedings of 2nd international conference on engineering optimization, 6–9 Sept
52. Sonmez M (2011) Artificial bee colony algorithm for optimization of truss structures. *Appl Soft Comput* 11(2):2406–2418
53. Sonmez M (2011) Discrete optimum design of truss structures using artificial bee colony algorithm. *Struct Multidiscip Optim* 43(1):85–97
54. Talathari S, Nouri M, Tadbiri F (2012) Optimization of skeletal structures using artificial bee colony algorithm. *Int J Optim Civil Eng* 2(4):557–571
55. Garg H (2014) Solving structural engineering design optimization problems using an artificial bee colony algorithm. *J Ind Manage Optim* 10(3):777–794
56. Yahya M, Saka MP (2014) Optimum construction site layout planning using multiobjective artificial bee colony algorithm with levy flights. *Autom Constr* 38(3):14–29
57. Aydogdu I, Akin A, Saka MP (2014) Design optimization of real size steel space frames using artificial bee colony algorithm with Levy flight distribution. *Steel Compos Struct* (under review)
58. Simon D (2008) Biogeography-based optimization. *IEEE Trans Evolut Comput* 12(6):702–713
59. Ammu PK, Sivakumar KC, Rejimoan R (2013) Biogeography-based optimization-a survey. *Int J Electron Comput Sci Eng* 2:154–160
60. Simon D, Rarick R, Ergezer M, Du D (2011) Analytical and numerical comparisons of biogeography-based optimization and genetic algorithms. *Inf Sci* 181:1224–1248
61. Rao RV, Savsani VJ, Vakhari DP (2011) Teaching-learning-based optimization: a novel method for constrained mechanical design optimization problems. *Comput Aided Des* 43:303–315
62. Togan V (2012) Design of planar steel frames using teaching-learning-based optimization. *Eng Struct* 34:225–232
63. Togan V (2013) Design of pin jointed structures using teaching-learning-based optimization. *Struct Eng Mech* 47(2):209–225
64. Degertekin O, Hayalioglu S (2013) Sizing truss structures using teaching-learning-based optimization. *Comput Struct* 119:177–188
65. Majid KI (1972) *Nonlinear structures*. Butterworth, London
66. Carbas S (2013) Optimum design of low rise steel frames made of cold-formed thin-walled steel sections. Ph.D. dissertation, Engineering Sciences Department, Middle East Technical University, Ankara, Turkey
67. ASCE 7–05 (2005) Minimum design loads for buildings and other structures, American Society of Civil Engineers

# Adaptive Switching of Variable-Fidelity Models in Population-Based Optimization

Ali Mehmani, Souma Chowdhury, Weiyang Tong and Achille Messac

**Abstract** This article presents a novel model management technique to be implemented in population-based heuristic optimization. This technique adaptively selects different computational models (both physics-based models and surrogate models) to be used during optimization, with the overall objective to result in optimal designs with high fidelity function estimates at a reasonable computational expense. For example, in optimizing an aircraft wing to obtain maximum lift-to-drag ratio, one can use low fidelity models such as given by the vortex lattice method, or a high fidelity finite volume model, or a surrogate model that substitutes the high-fidelity model. The information from these models with different levels of fidelity is integrated into the heuristic optimization process using the new adaptive model switching (AMS) technique. The model switching technique replaces the current model with the next higher fidelity model, when a stochastic switching criterion is met at a given iteration during the optimization process. The switching criterion is based on whether the uncertainty associated with the current model output dominates the latest improvement of the relative fitness function, where both the model output uncertainty and the function improvement (across the population) are expressed as probability distributions. For practical implementation, a measure of critical probability is used to regulate the degree of error that will be allowed, i.e., the fraction of instances where the improvement will be allowed to be lower than the model error, without having to change the model. In the absence of this critical probability, model man-

---

A. Mehmani · S. Chowdhury (✉)

Department of Mechanical and Aerospace Engineering, Syracuse University,  
Syracuse, NY 13244, USA  
e-mail: amehmani@syr.edu

S. Chowdhury

e-mail: chowdhury@bagley.msstate.edu

W. Tong · A. Messac

Department of Aerospace Engineering, Mississippi State University,  
Mississippi State, MS 39762, USA  
e-mail: wtong@syr.edu

A. Messac

e-mail: messac@ae.msstate.edu

© Springer International Publishing Switzerland 2015

N.D. Lagaros and M. Papadrakakis (eds.), *Engineering and Applied Sciences Optimization*, Computational Methods in Applied Sciences 38,  
DOI 10.1007/978-3-319-18320-6\_10

agement might become too conservative, leading to premature model-switching and thus higher computing expense. The proposed AMS-based optimization is applied to two design problems through Particle Swarm Optimization, which are: (i) Airfoil design, and (ii) Cantilever composite beam design. The application case studies of AMS illustrated: (i) the computational advantage of this method over purely high fidelity model-based optimization, and (ii) the accuracy advantage of this method over purely low fidelity model-based optimization.

## 1 Introduction

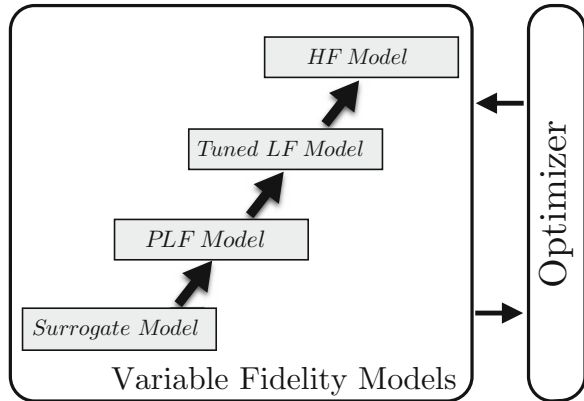
Population-based heuristic optimization algorithms, such as evolutionary algorithms and swarm optimization algorithms have been applied to diverse areas of science and engineering over the past few decades. They have been proven to be very effective in solving complex design optimization problems, especially those involving highly nonlinear functions. However, considering the computational cost of the high fidelity simulation models typically used to represent system behavior (e.g., CFD, FEA models), the large number of function evaluations often demanded by heuristic algorithms limit their applicability to practical complex system design (e.g., wing design of a high speed civil transport aircraft [1]). One approach to address this issue is *variable fidelity optimization*. In this approach, model management strategies adaptively integrate models of different fidelity and cost into the optimization process.

### 1.1 Variable Fidelity Models

*Variable fidelity models* refer to models with different levels of fidelity, where the computational cost of the model is generally related to the accuracy of the model estimation. In addition to low, medium, and high fidelity physics-based models, *surrogate models* (or mathematical approximation models) can also be used as candidates within a set of variable fidelity models. Surrogate models are purely mathematical models (i.e., not derived from the system physics) that are used to provide a tractable and inexpensive approximation of the actual system behavior. They are commonly used as an alternative to expensive computational simulations (e.g., CFD [2]) or to the lack of a physical model in the case of experiment-derived data (e.g., creation and testing of new metallic alloys [3]). Further description of the state of the art in surrogate modeling can be found in the following literature [4–6]. Major surrogate modeling methods include Polynomial Response Surfaces [7], Kriging [8, 9], Moving Least Square [10, 11], Radial Basis Functions (RBF) [12], Support Vector Regression(SVR) [13], Neural Networks [14] and hybrid surrogate models [15]. These methods have been applied to a wide range of disciplines, from aerospace design and automotive design to chemistry and material science [6, 16, 17].



**Fig. 1** Variable fidelity models



Besides direct implementation of a surrogate model as a black-box function (directly substituting a high fidelity model or data), low fidelity physics-based models can also be combined with a surrogate model to achieve a hybrid model of greater accuracy than its individual components (as illustrated in Fig. 1). Low fidelity physics-based models (e.g., the vortex lattice computational fluid dynamics method) are generally less complex than a high fidelity model and often provide a less faithful representation of the system behavior [18]. These models can be obtained by simplifying either the analysis model (e.g., using coarse finite element mesh) or the original physical formulation (e.g., using simplified boundary conditions or geometry). To their advantage, low fidelity physics-based models often inherit the major features of true models, while being significantly less expensive. Hence, these models could provide a reliable foundation for the construction of high-quality hybrid approximation models. These hybrid models, also called *tuned low fidelity models*, are expected to reflect the most prominent physical features of the system behavior, while preserving computational efficiency. Two well-known approaches for constructing tuned low fidelity (TLF) models are *multiplicative* and *additive* approaches, as given in Eqs. 1 and 2, respectively [19].

$$\text{Multiplicative approach: } y_{TLF} = A \times y_{LF} \quad (1)$$

$$\text{Additive approach: } y_{TLF} = B + y_{LF} \quad (2)$$

In both of these approaches, the tuning functions ( $A$  and  $B$ ) are trained using the associated values of the high and low fidelity models for a given DoE, as shown below:

$$\begin{aligned}
A(X) &= \frac{y_{HF}(X)}{y_{LF}(X)} \\
B(X) &= y_{HF}(X) - y_{LF}(X) \\
\text{where } X &= \{X_1, X_2, X_3, \dots, X_{N_S}\} \\
N_S &: \text{Number of sample points}
\end{aligned} \tag{3}$$

and where  $y_{HF}(\cdot)$  and  $y_{LF}(\cdot)$  respectively represent the functional responses of the low and the high fidelity models (where in the multiplicative scenario,  $y_{LF}$  is only allowed to take non-zero values). In surrogate-based tuned low fidelity models, the tuning (or correction) of a low fidelity model is performed using a surrogate model constructed through a DoE of the high fidelity model [20–22].

## 1.2 Model Management in Optimization

The major pitfall in using low fidelity models in optimization is that they can often mislead the search process, leading to suboptimal or infeasible solutions. To address this issue and provide optimum designs with high fidelity system evaluations, model management strategies can be applied. Different model management approaches have been reported in the literature, for integrating low fidelity models within optimization algorithms. One class of model management strategies are developed based on the *Trust-Region* methods [23–27]. The basic idea of the Trust-region is to solve an optimization problem,  $\text{Min}_{x \in \mathbb{R}^P} f(x)$ , using the high fidelity model ( $f(x)$ ). In solving this optimization problem using a gradient-based algorithm, the  $k$ th iteration is computed as  $x^{k+1} = x^k + \lambda \Delta x$ , where  $\lambda$  is the step length and  $\Delta x$  is the decent direction. As  $\Delta x$  is fixed, the problem reduces to a one-dimensional optimization problem:  $\text{Min}_{\lambda} f(x^k + \lambda \Delta x)$ . To improve the computational efficiency of the problem, the low fidelity model,  $\hat{f}(x)$ , can be used in the latter optimization problem. Assuming the low fidelity model is only valid in the vicinity of  $x^k$  (e.g.,  $x^k + \gamma$ ), the optimization search for  $\lambda$  is changed to the following constrained optimization problem:

$$\text{Min}_{\lambda} f(x + \lambda \Delta x), \quad \text{subject to: } \|\lambda \Delta x\| < \gamma \tag{4}$$

where  $\gamma$  is the *trust-region radius*. In the *Trust-Region* based model management methods developed by Alexandrov et al. [28] and by Toropove and Alvarez in 1998 [29], the parameter  $\gamma$  is adaptively increased (or decreased) depending on how well the low fidelity model,  $\hat{f}(x)$ , predicts the improvement in the high fidelity model. This criterion is estimated by computing the ratio of the actual to the predicted improvement in the objective function, as given by

$$\frac{f(x^k) - f(x^k + \lambda^k \Delta x^k)}{\hat{f}(x^k) - \hat{f}(x^k + \lambda^k \Delta x^k)} \tag{5}$$

The *Trust-Region* method seeks the agreement of the function and its gradient values estimated by the low fidelity model with those estimated by the high fidelity model. However, these techniques may not be directly applicable in problems where gradients are expensive to evaluate, or where zero-order algorithms are being used for optimization.

In another class of model management strategies, developed for non-physics-based low fidelity models (e.g., surrogate model and tuned-low fidelity model) the accuracy of the surrogate model (or metamodels) is improved during the optimization process by adding infill points, where additional evaluations of the high fidelity model is then performed. Infill points are generally added in (i) the region where the optimum is located (local exploitation); and/or (ii) the entire design space to improve the global accuracy of the surrogate (global exploration) [20, 30, 31]. Trosset and Torczon in 1997 [32] proposed an approach where the balance between exploitation and exploration was considered using the aggregate *merit function*,  $\hat{f}(x) - \rho d_{min}(x)$ , where,  $d_{min}(x) = \underset{x}{\text{Min}} \|x - x^i\|$ ,  $\rho > 0$ . It is important to note that, this technique is independent of the type of surrogate modeling technique being considered. Over the last two decades, different statistical model management strategies have been developed [33–36]. Among them, Jones et al. in 1998 [35] developed a well-known model management strategy that is based on an *Expected Improvement (EI)* criterion, and is called *Efficient Global Optimization (EGO)*. This powerful approach is however generally limited to surrogate models based on Gaussian processes. Assuming  $f_{min}$  is the objective function value of the optimum in the training data, the expected improvement in an infill point  $x$  is given by  $\mathbb{E}(I(x)) = \mathbb{E}(\max(f_{min} - F(x), 0))$ . In this case,  $F(x)$  is a Gaussian distribution,  $F(x) \sim \mathcal{N}(\hat{f}(x), \sigma^2(x))$ , where the posterior mean,  $\hat{f}(x)$ , is used as a surrogate model, and the posterior variance  $\sigma^2(x)$  gives an estimate of the uncertainty involved in the surrogate prediction. The expected improvement can be estimated by

$$\mathbb{E}(I(x)) = (f_{min} - \hat{f}(x))\Phi\left(\frac{f_{min} - \hat{f}(x)}{\sigma(x)}\right) + \sigma(x)\phi\left(\frac{f_{min} - \hat{f}(x)}{\sigma(x)}\right) \quad (6)$$

where  $\Phi(\cdot)$  and  $\phi(\cdot)$  denote the standard normal density and distribution functions, respectively [34]. Subsequently, an infill point can be found by maximizing the expected improvement,  $X^{infill} = \arg \max_x (\mathbb{E}(I(x)))$ .

The model management strategies used in heuristic optimization algorithms can be broadly classified into two different approaches which are (i) *individual-based evolution control*, and (ii) *generation-based evolution control* [37]. In the individual-based approach, selected individuals (*controlled individuals*) within a generation are evaluated using a high fidelity model. In the generation-based approach, the whole population at a certain generation (*controlled generation*) is evaluated using the high fidelity model. Graning et al. [38] explored different individual-based evolution frameworks such as (i) the Best Strategy [39], where the best individuals at each generation are selected as controlled individuals, (ii) the Pre-Selection method [40], where the offspring of the best individuals are selected as controlled individuals,

and (iii) the Clustering Technique [41], where the *k-means* clustering technique is used to find the “controlled individual cluster” based on the distance from the best individual.

In this section, a survey of existing model management strategies for integrating models with different levels of fidelity into an optimization process was provided. Several of the existing strategies are found to be defined for specific types of low fidelity model, e.g., EGO works primarily for Gaussian process-based surrogate models. On the other hand, existing techniques generally consider the combination of only two models of different fidelities (e.g., Trust-region methods, and individual- and generation-based techniques). This article seeks to address some of the above-stated crucial gaps in the variable-fidelity optimization paradigm. Specifically, the development of a model management strategy that can be coherently applied to different types of low fidelity models (i.e., physics-based and non-physics-based low fidelity models), and allows adaptive switching between more than two models is being pursued in this article.

### ***1.3 A New Approach to Global Model Switching***

The primary objective of this article is to investigate a new adaptive model management strategy that significantly reduces the computational cost of optimization while converging to the optimum with high fidelity model evaluation; in its current form, this method is designed to work with population-based optimization algorithms (e.g., GAs, PSOs). Additionally, this method assumes that models of different levels of fidelity are available to the user. Specifically, a new stochastic model switching metric, called Adaptive Model Switching (AMS), is formulated in this article. The AMS technique is implemented through a powerful version of the Particle Swarm Optimization (PSO) algorithm that involves explicit diversity preservation, called *Mixed-Discrete PSO* [42]. The effectiveness of this implementation is investigated by application to two engineering design optimization problems.

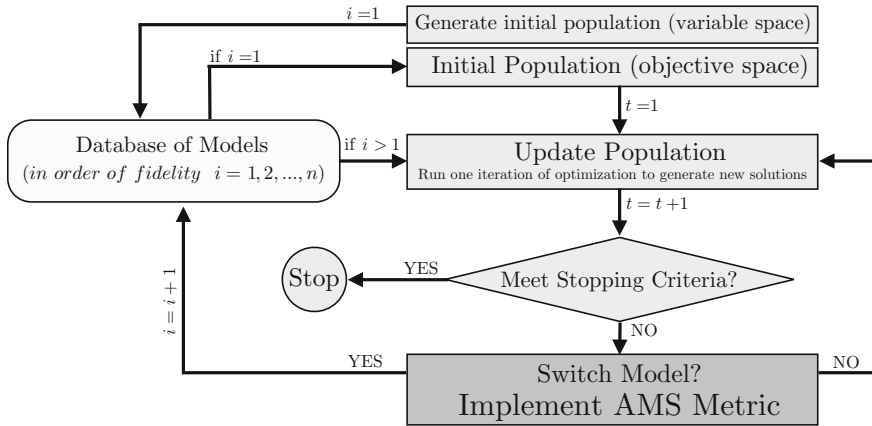
The remainder of the article is organized as follows: Sect. 2 presents the concept and the formulation of the new Adaptive Model Switching (AMS) metrics. Description of the model error quantification methods used in this article, including Predictive Error Estimation of Model Fidelity (PEMF), is provided in Sect. 2.3. Section 3 describes the practical problems to which AMS is applied; the numerical settings and case study results are illustrated and discussed in Sect. 3. Section 4 provides the concluding remarks.

## 2 Variable Fidelity Optimization with Adaptive Model Switching (AMS)

### 2.1 Major Steps in Optimization with AMS

In optimization based on variable fidelity models, the important question is when and where to integrate the models with different levels of fidelity. In this article, the “when to integrate” question is particularly addressed. Increasing fidelity too early in the design process can be computationally expensive while wasting resources to explore undesirable regions of the design domain. On the other hand, switching to a higher fidelity model too late might mislead the search process early on to suboptimal regions of the design domain (especially in multimodal problems), i.e., leading to scenarios where the global optimum is outside of the region spanned by the population of candidate solutions in later iterations. In this section, a novel model management strategy called, *Adaptive Model Switching (AMS)* metric is developed to avoid both these undesirable scenarios. AMS can be perceived as a decision-making tool for the timing of *model-switching* or *model integration*. The implementation of the proposed AMS in population-based algorithm involves the following five major steps:

- Step 1** Assuming the available models are non-dominated w.r.t. each other in terms of fidelity and computational expense, the models are first ranked from the lowest fidelity to the high fidelity, based on the error associated with each model- $M_i$  for  $i = 1, \dots, n$ . where model  $M_1$  has the lowest fidelity and model  $M_n$  has the highest fidelity. Assuming the distribution of model error is known for each model, the ranking is performed using the modal values of the error distributions.
- Step 2** The initial population is then generated at  $t = 1$ , using  $M_1$ .
- Step 3** At every iteration ( $t$ ) of the heuristic optimization algorithm, the current model,  $M_i$ , is used to update the function values of the population, and then set  $t = t + 1$ . In this article, Particle Swarm Optimization is the chosen heuristic optimization algorithm.
- Step 4** The following stopping criteria is checked after every iteration. The optimization algorithm stops when the relative changes in the fitness function value is less than a predefined function tolerance,  $\delta_F$ . To avoid termination before reaching the high fidelity model ( $M_n$ ), the function tolerance must be specified to be less than the modal error of the last but one model ( $M_{n-1}$ ).  
**IF** the termination criteria is satisfied, the current optimum (the best global solution in the case of PSO) is identified as the final optimum and the optimization process is terminated.  
**ELSE, Go To Step 5**
- Step 5** The switching metric (AMS metric) is evaluated in this step.  
**IF** the AMS metric is satisfied, a switching event occurs, and the algorithm



**Fig. 2** Adaptive model switching in population-based optimization

switches from model  $M_i$  to  $M_{i+1}$ .

### Go To Step 3

A flowchart of the algorithm for optimization with AMS is shown in Fig. 2. In practice, the AMS technique (Step 5) need not be applied at every iteration; the user can specify it to be applied after every  $K$  iteration (where  $K$  is a small positive integer). In the flowchart, AMS is shown to be applied at every iteration, for the sake of simplicity.

In the following subsection, the novel components of the AMS method (Fig. 2) are described. Subsequently, an overview of the Mixed-Discrete PSO algorithm, which is used for implementing and testing the AMS method, is provided.

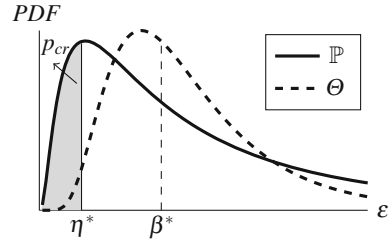
## 2.2 The Adaptive Model Switching (AMS) Metric

In this article, it is assumed that the uncertainty associated with each model ( $M_i$ ;  $i = 1, \dots, n$ ) is known or can be evaluated in the form of an error distribution,  $\mathbb{P}_i$ . Under this assumption, the fitness function values evaluated using the  $i$ th model can be related to the corresponding high fidelity estimation as

$$y_{HF}^i = \hat{y}_{LF}^i + \varepsilon^i \quad (7)$$

In Eq. 7,  $\hat{y}_{LF}^i$  and  $\varepsilon^i$  respectively represent the response of the  $i$ th low fidelity model and the stochastic error associated with it; and  $y_{HF}^i$  is the corresponding high fidelity model response. The relative improvement in the fitness function value ( $\Delta f$ ) can be considered to follow an unknown distribution,  $\Theta$ , over the population of solutions. Here,  $\Delta f$  in the  $t$ th iteration ( $t \geq 2$ ) can be expressed as

**Fig. 3** The illustration of the AMS metric



$$\Delta f_k^t = \begin{cases} \left| \frac{f_k^t - f_k^{t-1}}{f_k^t} \right| & \text{if } f_k^t \neq 0 \\ |f_k^t - f_k^{t-1}| & \text{if } f_k^t = 0 \end{cases} \quad (8)$$

where  $k = 1, 2, 3, \dots, N_{pop}$

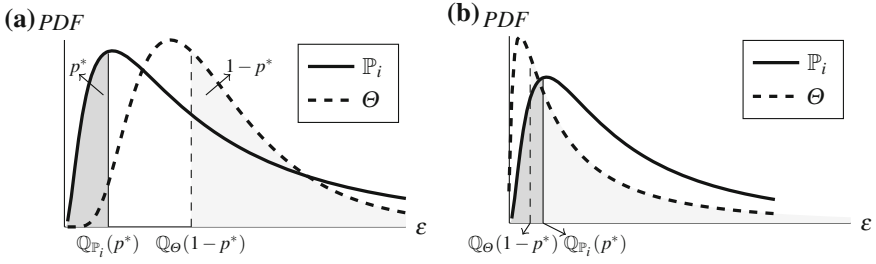
The model switching criteria is then defined based on “whether the uncertainty associated with a model response is higher than the observed improvement in the relative fitness function of the population”. Due to the practical unavailability of reliable local measures of model error (i.e.,  $\varepsilon$  as a function of  $x$ ), the model switching criteria is designed using the stochastic global measures of model error and the distribution of solution improvement. Based on prior experience or practical design requirements, the designer is likely to be cognizant of what levels of global model error,  $\eta$ , is acceptable for a particular low fidelity model in an optimization process. Hence,  $\eta$  can be perceived as a user-preference. The *critical probability*,  $p_{cr}$  for that low fidelity model with an error distribution  $\mathbb{P}$  is then defined as the probability of the model error to be less than  $\eta$ . This definition can be expressed as

$$p_{cr} = Pr[\varepsilon \leq \eta] = \int_0^\eta \mathbb{P}(\varepsilon') d\varepsilon' \quad (9)$$

The critical probability ( $p_{cr}$ ) essentially indicates a critical bound in the error distribution  $\mathbb{P}$  ( $0 \leq \varepsilon \leq \eta$ ). If the predefined cut-off value ( $\beta$ ) of the  $\Theta$  distribution lies inside this region, the current low fidelity model is considered to be no more reliable for use in the optimization process. As illustrated in Fig. 3, assuming that  $\Theta$  and  $\mathbb{P}$  follow a log-normal distribution,  $p_{cr} = Pr[\varepsilon \leq \eta^*]$ ; and  $\beta^*$  is the pre-computed cut-off value in the  $\Theta$  distribution. The model with the  $\mathbb{P}$  error distribution can be used in the optimization process provided that  $\eta^* \leq \beta^*$ .

The Adaptive Model Switching (AMS) metric is formulated as a hypothesis testing that is defined by a comparison between

- (I) the distribution of the relative fitness function improvement ( $\Theta$ ) over the entire population, and
- (II) the distribution of the error associated with the  $i$ th model ( $\mathbb{P}_i$ ) over the entire design space.



**Fig. 4** The illustration of the AMS hypothesis test (comparing the model error distribution ( $\mathbb{P}_i$ ) and the distribution of fitness function improvement ( $\Theta$ )); **a** Rejection of the text; **don't change a model.** **b** Acceptance of the text; **change a model**

This statistical test for the  $i$ th model can be stated as

$$\begin{aligned}
 H_0: \mathbb{Q}_{\mathbb{P}_i}(p_{cr}) &\geq \mathbb{Q}_{\Theta}(1 - p_{cr}) \\
 H_1: \mathbb{Q}_{\mathbb{P}_i}(p_{cr}) &< \mathbb{Q}_{\Theta}(1 - p_{cr}) \\
 0 < p_{cr} &< 1
 \end{aligned} \tag{10}$$

where  $\mathbb{Q}$  represents a quantile function of a distribution; The  $p$ -quantile, for a given distribution function,  $\Psi$ , is defined as

$$\mathbb{Q}_{\Psi}(p) = \inf\{x \in \mathbb{R}: p \leq \Psi_{(c.d.f.)}(x)\} \tag{11}$$

In Eq. 10,  $p_{cr}$  or the critical probability is an *Indicator of Conservativeness (IoC)*. The IoC is based on user preferences, and regulates the trade-off between optimal solution reliability and computational cost in the AMS-based optimization process. Generally, the higher the IoC (closer to 1), the higher the solution reliability and the greater the computational cost; under these conditions, model switching events will occur early on in the optimization process.

For the sake of illustration, assume  $\Theta$  and  $\mathbb{P}_i$  follow a log-normal distribution, and  $p_{cr} = p^*$ . In this case, the null hypothesis will be rejected, and the optimization process will use the current model ( $M_i$ ) **if**  $\mathbb{Q}_{\Theta} > \mathbb{Q}_{\mathbb{P}_i}$ , as illustrated in Fig. 4a. Conversely, **if**  $\mathbb{Q}_{\Theta} < \mathbb{Q}_{\mathbb{P}_i}$ , the null hypothesis will be accepted, and the optimization process will switch to the next higher fidelity model ( $M_{i+1}$ ), as shown in Fig. 4b.

In this article, Kernel Density Estimation (KDE) is adopted to model the distribution of the relative improvement in the fitness function over consecutive  $kt$  iterations. Since the distribution of fitness function improvement over the population (for different problems) may not follow any particular probability model, and is also observed to be multimodal at times, KDE is a suitable choice in this context. KDE is a standard non-parametric approach to estimate the probability density function of random variables. Here, it is assumed that  $\Delta f = (\Delta f_1, \Delta f_2, \Delta f_3, \dots, \Delta f_{N_{pop}})$  is an independent and identically distributed sample drawn from a distribution with an



unknown density  $\Theta_{\Delta f}$ . The kernel density estimator can then be used to determine  $\Theta_{\Delta f}$ , as given by

$$\tilde{\Theta}_{\Delta f}(x; H) = N_{pop}^{-1} \sum_{i=1}^{N_{pop}} K_H(x - x_i) \quad (12)$$

Here, the kernel  $K(x)$  is a symmetric probability density function,  $H$  is the bandwidth matrix which is symmetric and positive-definite, and  $K_H(x) = |H|^{-1/2} K(H^{-1/2}x)$ . The choice of  $K$  is not as crucial as the choice of the  $H$  estimator for the accuracy of the KDE [43]. In this article, we consider  $K(x) = (2\pi)^{-d/2} \exp(-\frac{1}{2}x^T x)$ , the standard normal throughout. The Mean Integrated Squared Error (MISE) method is used as a criterion for selecting the bandwidth matrix,  $H$  [44], where

$$MISE(H) = \mathbb{E} \left( \int [\tilde{\Theta}_{\Delta f}(x; H) - \Theta_{\Delta f}(x)]^2 \right) \quad (13)$$

### 2.3 Quantifying Model Uncertainties

In this article, the uncertainties associated with surrogate models and surrogate-based tuned low fidelity models are determined using an advanced surrogate error estimation method, called *Predictive Estimation of Model Fidelity* or *PEMF* [45]. The PEMF method is derived from the hypothesis that “the accuracy of approximation models is related to the amount of data resources leveraged to train the model”. A brief description of the PEMF method is provided in the following sub-section (Sect. 2.3.1). In the case of physics-based low fidelity (PLF) models, the uncertainty in their output is quantified through an inverse assessment process, by comparing the physics-based low fidelity model responses with the high fidelity model responses. In this case, the relative absolute error ( $RAE_{PLF}$ ) of a PLF model is estimated as

$$RAE_{PLF_i} = \begin{cases} \left| \frac{HF_i - PLF_i}{HF_i} \right| & \text{if } HF_i \neq 0 \\ |HF_i - PLF_i| & \text{if } HF_i = 0 \end{cases} \quad (14)$$

where  $i = 1, 2, 3, \dots, N_s$  (Number of sample points)

A DoE of  $N_S$  high fidelity evaluations is used to perform the above-stated error quantification, and also to train a surrogate models and a tuned low fidelity models. The uncertainty of a low fidelity physics-based models is represented by a log-normal distribution,  $\ln \mathcal{N}(\mu_{PLF}, \sigma_{PLF})$ , where the  $p$ -quantile of this distribution is defined as

$$\mathbb{Q}_{\mathbb{P}_{PLF}}(p) = \mathbb{P}_{PLF}^{-1}(p|\mu_{PLF}, \sigma_{PLF}) = \exp(\mu_{PLF} + \Phi^{-1}(p) \sigma_{PLF})$$

where

$$\mu_{PLF} = \ln\left(\frac{m_{RAR_{PLF}}^2}{\sqrt{v_{RAR_{PLF}}^2 + m_{RAR_{PLF}}^2}}\right) \quad \sigma_{PLF} = \sqrt{\ln\left(1 + \frac{v_{RAR_{PLF}}}{m_{RAR_{PLF}}^2}\right)} \quad (15)$$

In Eq. 15,  $\Phi^{-1}(\cdot)$  is the inverse of the c.d.f of the standard normal distribution with zero mean and unit variance, and  $m_{RAR_{PLF}}$  and  $v_{RAR_{PLF}}$  are the mean and the variance of  $RAE_{i=1,2,3,\dots,N_s}$ , respectively.

### 2.3.1 Predictive Estimation of Model Fidelity (PEMF)

In concept, the PEMF method [45] can be perceived as a novel sequential implementation of  $k$ -fold cross-validation, with carefully constructed error measures that are significantly less sensitive to outliers and the DoE (compared to Mean or Root Mean Square error measures). The PEMF method predicts the error by capturing the variation of the surrogate model error with an increasing density of training points (without investing any additional test points).

In the PEMF method, for a set of  $N_s$  sample points, intermediate surrogates are constructed at each iteration,  $r$ , using  $S^r$  heuristic subsets of  $n^r$  training points (called intermediate training points), where  $n^r < N_s$ . These intermediate surrogates are then tested over the corresponding remaining  $N_s - n^r$  points (called intermediate test points). The median error is then estimated for each of the  $S^r$  intermediate surrogates at that iteration, and a parametric probability distribution is fitted to yield the modal value,  $E_{med}^{mo,r}$ , and the median value,  $E_{med}^{med,r}$ , of the model error at that stage. The smart use of the modal value of the median error significantly reduces the occurrence of oscillations in the variation of error with sample density, unlike mean or root mean squared error which are highly susceptible to outliers [46]. This approach gives PEMF an important advantage over conventional cross-validation-based error measures, as illustrated by Mehmani et al. [45–47]. It is important to note that all error quantifications are performed in terms of the relative absolute error ( $E_{RAE}$ ), which is given by:

$$E_{RAE}(X_i) = \begin{cases} \left| \frac{F(X_i) - \hat{F}(X_i)}{F(X_i)} \right| & \text{if } F(X_i) \neq 0 \\ |F(X_i) - \hat{F}(X_i)| & \text{if } F(X_i) = 0 \end{cases} \quad (16)$$

where  $F$  is the actual function value at  $X_i$ , given by high fidelity model, and  $\hat{F}$  is the function value estimated by the surrogate model.

In order to control the computational expense of PEMF, the lognormal distribution is used to represent the surrogate model error; this distribution has been previously observed (from numerical experiments) to be one of the most effective choice in

representing the surrogate model error distribution. The PDFs of the median error,  $p_{med}$ , can thus be expressed as

$$p_{med} = \frac{1}{E_{med}\sigma_{med}\sqrt{2\pi}} \exp\left(-\frac{(\ln(E_{med} - \mu_{med}))^2}{2\sigma_{med}^2}\right) \quad (17)$$

In the above equation,  $E_{med}$  represents the median of the relative errors estimated over a heuristic subset of training points at any given iteration in PEMF. The parameters,  $(\mu_{med}, \sigma_{med})$  are the generic parameters of the log-normal distribution. The modal and median values of the median error distribution at any iteration,  $r$ , can then be expressed as

$$\begin{aligned} E_{med}^{mo}|_r &= \exp(\mu_{med} - \sigma_{med}^2)|_r \\ E_{med}^{med}|_r &= \exp(\mu_{med})|_r \end{aligned} \quad (18)$$

Once the history of modal and median errors at different sample size ( $<N_s$ ) are estimated, the variation of the modal and median values of the errors with sample density are then modeled using the multiplicative ( $E = a_0n^{a_1}$ ) or the exponential ( $E = a_0e^{a_1n}$ ) regression functions (depending on the best least-square fit). These regression functions are then used to predict the modal and the median values of the error distribution in the final surrogate, where the final surrogate is trained using all the  $N_s$  sample points. The predicted modal and the median error values,  $\varepsilon_{mod}$  and  $\varepsilon_{med}$ , are then used to define the distribution of the error in the final surrogate model, or in other words the response uncertainty of the surrogate model. The location and scale parameters of the error distribution is then given by

$$\begin{aligned} \mu_\varepsilon &= \ln \varepsilon_{med} \\ \sigma_\varepsilon &= \sqrt{\ln\left(\frac{\varepsilon_{med}}{\varepsilon_{mod}}\right)} \end{aligned} \quad (19)$$

Subsequently, the  $p$ -quantile of the error distribution associated with the surrogate model is given by

$$\mathbb{Q}_{\mathbb{P}_{SM}}(p) = \mathbb{P}_{SM}^{-1}(p|\mu_\varepsilon, \sigma_\varepsilon) = \exp(\mu_\varepsilon + \Phi^{-1}(p) \sigma_\varepsilon) \quad (20)$$

## 2.4 Optimization Algorithm: Particle Swarm Optimization

In the proposed model management methodology, optimization is performed using an advanced implementation of the Particle Swarm Optimization (PSO). PSO was originally developed for solving continuous nonlinear optimization problems by Eberhart and Kennedy in 1995 [48]. Several advanced versions of this algorithm have been reported in the literature since its inception. In this article, one particular advanced

implementation of the PSO algorithm called Mixed-Discrete PSO (MDPSO), which was developed by Chowdhury et al. [42], is used. The advantages that the MDPSO algorithm provides over a conventional PSO algorithm include: (i) an ability to deal with both discrete and continuous design variables, and (ii) an explicit diversity preservation capability that mitigates the possibility of premature stagnation of particles. Further description of the MDPSO algorithm can be found in the paper by Chowdhury et al. [42].

### 3 Numerical Case Studies

#### 3.1 Aerodynamic Shape Optimization of 2D Airfoil

This section describes a 2D airfoil design problem where the ratio of the coefficients of lift and drag ( $C_L/C_D$ ) of the Wortmann FX60.126 2D airfoil [49] is to be maximized. The lift-to-drag ratio ( $C_L/C_D$ ) is expressed as a function of four design variables, which include the angle of incidence (ranging from 0 to 10) and the three normalized shape variables (each ranging from  $-0.01$  to  $0.01$ ). As illustrated in Fig. 5, the three shape variables define the distances (i) between the middle of the suction side and the horizontal axis ( $x_1$ ), (ii) between the middle of pressure side and the horizontal axis ( $x_2$ ), and (iii) between the trailing edge and the horizontal axis ( $x_3$ ). These three shape variables allow a modification of the un-deformed airfoil profile. With respect to the initial airfoil design, two cubic splines are added to the suction and the pressure sides. Each of these splines is characterized by 3 points, defined on the leading edge, the middle span, and the trailing edge. The chord length of the airfoil is equal to 1 m. The design constraints are the side constraints on the design variables which are listed in Table 1.

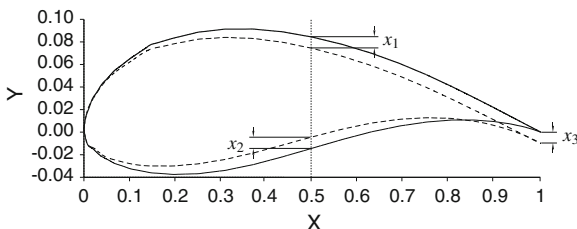


Fig. 5 Design variables governing the geometry of the airfoil

**Table 1** Design variables in airfoil optimization problem

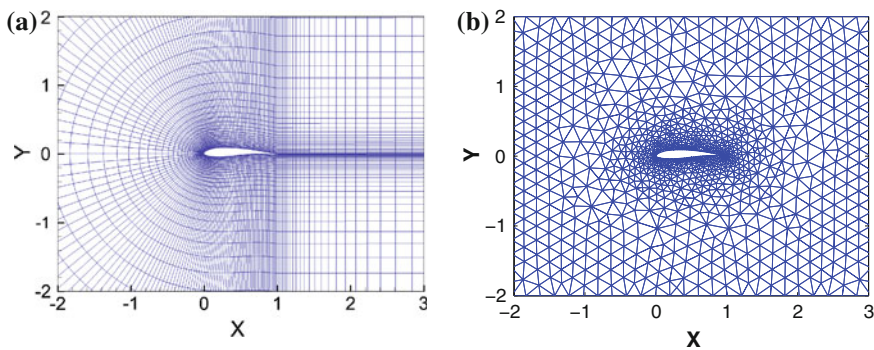
Description	Notation	Lower limit	Upper limit
Distance between the middle of suction side and horizontal axis	$x_1$	-0.01	0.01
Distance between the middle of pressure side and horizontal axis	$x_2$	-0.01	0.01
Distance between the trailing edge and horizontal axis	$x_3$	-0.01	0.01
Incidence angle	$x_4$	0°	10°

### 3.1.1 Aerodynamic Models with Different Level of Fidelity

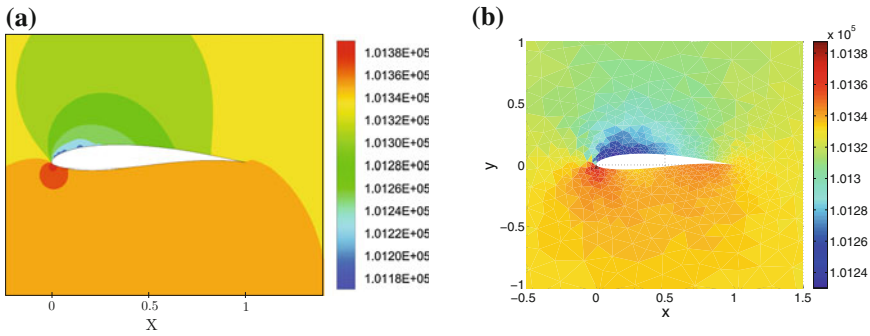
To develop a high fidelity aerodynamic model for determining  $C_L$  and  $C_D$  ( $M_{HF}^A$ ), the commercial Finite Volume Method package, FLUENT, is used. The Reynolds-averaged Navier-Stokes (RANS) formulation is used along with a Reynolds model to represent the turbulence. The CFD mesh is constructed using quadrangular cells [49], characterized by 9,838 quadrangular cells and 10,322 grid points (Fig. 6a).

The low fidelity physics-based model ( $M_{PLF}^A$ ) is constructed based on the assumptions that the fluid is steady, incompressible, and irrotational. In this model, the Navier-Stokes equations are solved using the Finite Element method. Triangular T3 elements are used for demonstration, as shown in Fig. 6b. The incoming velocity in the analysis is set to 25 m/s. The computational time of the High and Low fidelity physics-based models are approximately 300 and 30 s, respectively (i.e., an order of magnitude apart). The pressure field around the airfoil for the low and high fidelity aerodynamic models at a baseline design ( $x_1 = 0$ ,  $x_2 = 0$ ,  $x_3 = 0$ , and  $x_4 = 5^\circ$ ) are illustrated in Fig. 7.

The third model is a surrogate model ( $M_{SM}^A$ ) constructed using a DoE of high fidelity evaluation involving 30 sample points. The fourth model is a tuned low



**Fig. 6** Fine and coarse mesh for CFD of airfoil [49]; **a** High fidelity model mesh; **b** Low fidelity model mesh



**Fig. 7** Pressure field around the airfoil at a baseline design; **a** High fidelity model, **b** Low fidelity physics-based model

fidelity model ( $M_{TLF}^A$ ). In this article, the tuned low fidelity model is constructed using the *Multiplicative approach*, as given by

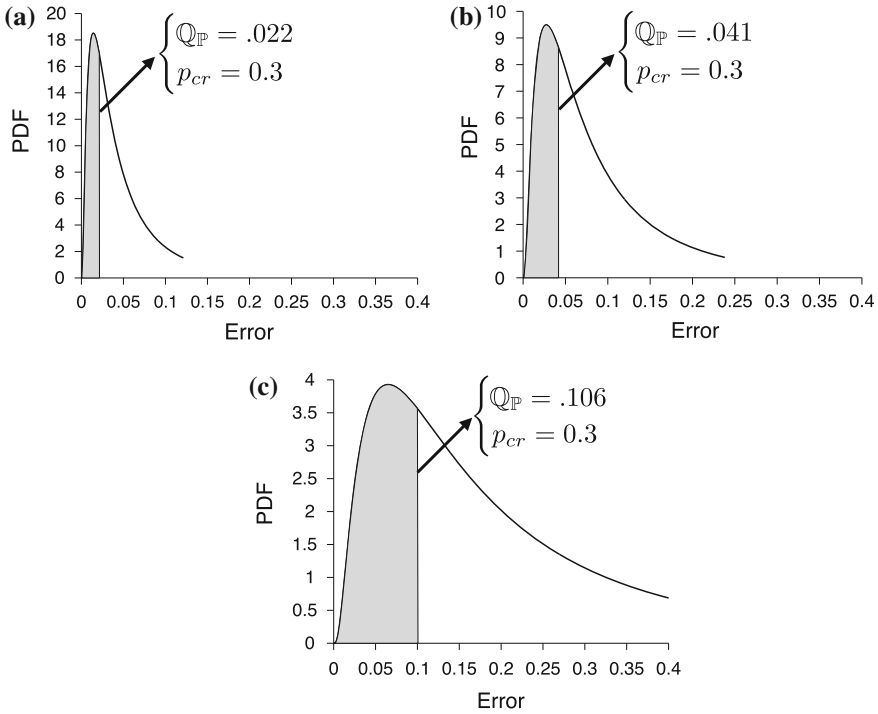
$$\tilde{F}(x, a) = f(x) \times C(x) \quad (21)$$

where  $\tilde{F}$  is a tuned low fidelity model;  $f(x)$  is a low fidelity model;  $C(x)$  is an explicit tuning surrogate constructed using the high fidelity samples, as shown below:

$$C(x) = \frac{\frac{C_L}{C_D}|_{HF}}{\frac{C_L}{C_D}|_{PLF}} \quad (22)$$

where  $C_L$  and  $C_D$  are respectively the lift and drag coefficients.

The surrogate model ( $M_{SM}^A$ ) and the surrogate component of the tuned low fidelity model ( $M_{TLF}^A$ ) are both constructed using Kriging with a Gaussian correlation function [8, 9]. Kriging is an interpolating method that is widely used for representing irregular data. Under the Kriging approach, the zero-order polynomial function is used as a regression model. In this article the Optimal Latin Hypercube is adopted to determine the locations of the sample points. The PEMF method is then applied to estimate the error in the surrogate models constructed using the high fidelity responses, and the tuned low fidelity model. To estimate the error in the physics-based low fidelity FEA model, the inverse assessment process defined in Sect. 2.3, is applied. Figure 8a–c illustrate the distributions of the error in the tuned low fidelity model, the surrogate model, and the physics-based low fidelity model. It is observed from Fig. 8 that the accuracy of the physics-based low fidelity model is less than that of the surrogate model. It is also readily evident that the computational cost of the physics-based low fidelity model is more than that of the surrogate model. Therefore, in this problem, the physics-based low fidelity model is dominated by the other three models and is hence not included as a model choice in the variable fidelity optimization.



**Fig. 8** Distribution of the model errors in evaluating the aerodynamic  $C_L/C_D$  ratio of the 2D airfoil: **a** Tuned LF model, **b** Surrogate model, **c** Physics-based LF model

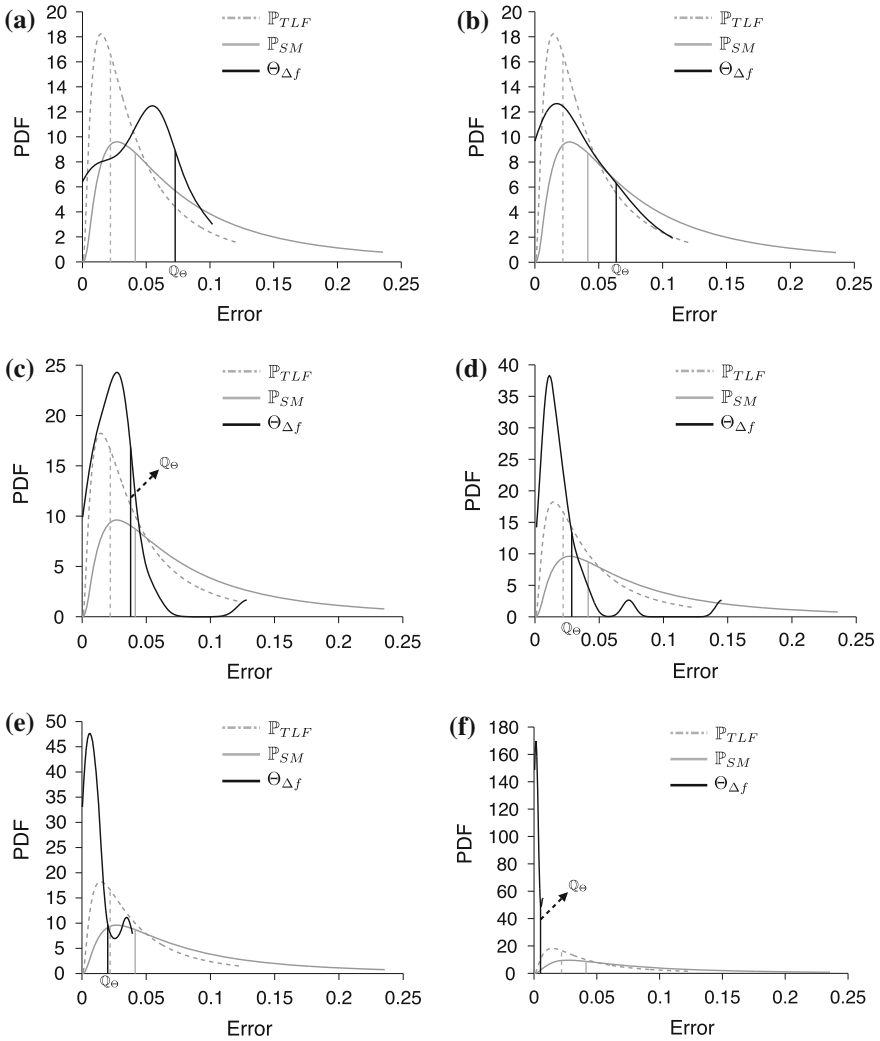
### 3.1.2 Airfoil Optimization Problem: Results and Discussion

In the airfoil optimization problem, the initial population of particles is generated using the fastest model, which is the surrogate model. The AMS technique adaptively switches the model type twice during optimization (over a total of 22 iterations), resulting in an optimum design with a high fidelity function estimate.

The model types, the error distribution parameters associated with each model, and the number of calls made to each model in this optimization are listed in Table 2.

**Table 2** Models with different levels of fidelity used in the airfoil optimization problem (the high fidelity model is assumed to be a true representation of the system behavior)

Model	Location parameter $\mu$	Scale parameter $\sigma$	$Q(p_{cr})$ $p_{cr} = 0.3$	No. of calls made $N_{pop} \times \# \text{ Iter.}$
Surrogate	-2.6793	0.9628	0.0414	$30 \times 13$
Tuned LF	-3.3197	0.9547	0.0219	$30 \times 6$
High fidelity	-	-	-	$30 \times 3$



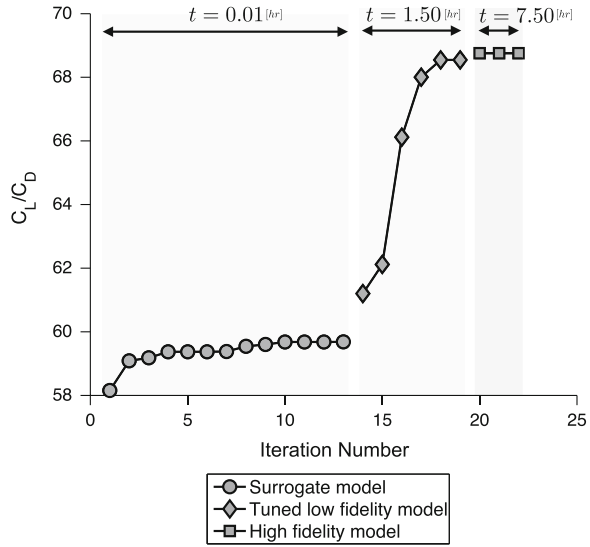
**Fig. 9** Distribution of the fitness function improvements in different iterations of the airfoil optimization with PSO-AMS (also showing the model error distributions); **a** 5th iteration, **b** 10th iteration, **c** 15th iteration, **d** 18th iteration, **e** 20th iteration, **f** 22th iteration

The total number of calls made to each model is equal to the product of the particle population and the number of iterations during which that particular model is used for system evaluation. In this problem, the AMS technique is applied at every iteration.

Figure 9a–f illustrate the distribution of the fitness function improvement at different iterations during the optimization process. In these figures,  $Q$  represents  $(1 - p_{cr})$ -quantile of the  $\Theta$  distribution. The error distributions of the surrogate model and the tuned low fidelity model, which are determined apriori, are also shown in these fig-



**Fig. 10** Optimization history of the airfoil design problem



ures. Through AMS, model switching from the surrogate model to the tuned low fidelity model and from the tuned low fidelity model to the high fidelity model occur at the 13th and the 19th iteration, respectively.

The convergence history of the airfoil optimization is illustrated in Fig. 10. This figure also indicates which model is active at each iteration. It is observed that, from the first iteration till the 13th iteration the surrogate model ( $M_{SM}^A$ ) is active, before switching to the tuned low fidelity model that remains active till the 19th iteration. Interestingly, most of the objective function improvement occurs under the tuned low fidelity model (more than 10% increase in the  $C_L/C_D$  ratio). The optimization uses the high fidelity model in the last 3 iterations before reaching convergence. In this case, the algorithm converges by satisfying the predefined function tolerance,  $\delta f = 10^{-5}$ .

Next, the performance of the AMS method is investigated and compared with the performances of running optimizations that solely rely on a low fidelity model or a high fidelity model. The results yielded by the PSO-AMS thus compared with the results yielded by separately running MDPSO solely using the surrogate model (PSO-SM), solely using the tuned low fidelity model (PSO-TLF), and solely using the high fidelity model (PSO-HF). The optimum results thus obtained, the computational cost, and the total number of function evaluations in each case are reported in Table 3. The final column of this table shows the high fidelity function estimate at the optimum design obtained under each optimization run (e.g.,  $y_{HF}^*(x_{SM}^*)$  and  $y_{HF}^*(x_{TLF}^*)$ ). It is observed that the PSO-AMS not only requires 185% less computing time compared to PSO-HF, it also provides the best optimum value that is 5% better than the next best value (where the 2nd best is obtained by PSO-TLF). It is also observed that, in the PSO-TLF approach, the optimum is located in the region where the TLF model

**Table 3** 2D Airfoil design: optimization results using single-fidelity and variable-fidelity optimization approaches

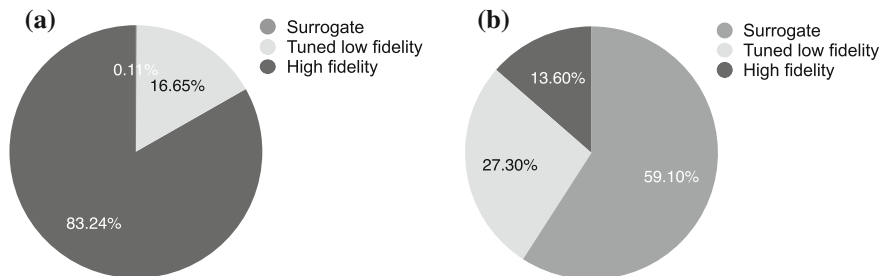
Approach	$x_1^*$	$x_2^*$	$x_3^*$	$x_4^*$	Optimum function $C_L/C_D (f^*)$	Model in last iteration	Computational time [hr] over function evaluation	HF response at optimum $(f_{HF}(x^*))$
PSO-SM	0.0003	0.0003	0.0003	4.8300	59.69	SM	0.275/990	59.43
PSO-TLF	-0.0028	-0.0058	-0.0014	2.7911	70.54	TLF	9.96/1380	65.20
PSO-HF	7.14E-5	-0.0021	-0.0018	4.8273	59.57	HF	25.7/360	59.57
PSO-AMS	-0.0020	-0.0004	-0.0009	2.8313	68.75	HF	9.01/660	68.75

*PSO-SM* optimization performed by MDPSO solely using the surrogate model

*PSO-TLF* optimization performed by MDPSO solely using the tuned low fidelity model

*PSO-HF* optimization performed by MDPSO solely using the high fidelity model

*PSO-AMS* optimization performed by MDPSO using AMS



**Fig. 11** Percentage of resources used by each model in the airfoil optimization problem performed through PSO-AMS: **a** Computing time resources; **b** Function evaluation resources

has more than 8% error. This optimum is in the vicinity of the high fidelity optimum yielded by the AMS method. The optimization performed solely using the tuned low fidelity model (PSO-TLF) also incurs a slightly higher computational time in comparison with that performed using the AMS method, which is attributed to the high number of function evaluations invested to satisfy the termination criterion in the former (1380 evaluations vs. 660 evaluations).

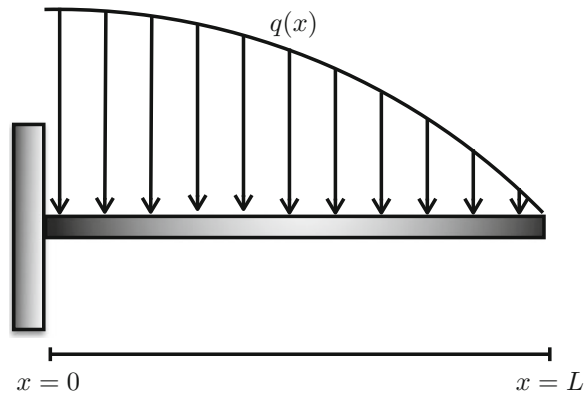
Figure 11 a, b illustrate the resources used in terms of computing time and function evaluations, by the three different models in the airfoil design optimization performed by PSO-AMS. These figures show that the overall computational cost is highly sensitive to the number of high fidelity model evaluations, which is expected. It is also observed that the surrogate model dominates the optimization process in terms of function calls, while the computational expense of this model is significantly lower than that of the tuned low fidelity and the high fidelity models. This observation supports the hypothesis that a probabilistic AMS technique can provide a significantly better balance between accuracy of the optimum and computational efficiency, compared to purely low fidelity or purely high fidelity optimizations.

### 3.2 Shape Optimization of a Cantilever Composite Beam

In the second optimization test problem, the maximum deflection of a cantilever composite beam (as shown in Fig. 12) is minimized. This beam is subjected to a parabolically-distributed load,  $q(x) = q_0(1 - \frac{x^2}{L^2})$  [22]. In this problem, the fiber direction Young's modulus,  $E_L$ , and the composite weight density,  $\rho$ , are given by

$$\begin{aligned}
 E_L &= E_f v_f + E_m(1 - v_f) \\
 \rho &= \rho_f v_f + \rho_m(1 - v_f) \\
 &\text{where} \\
 v_f + v_m &= 1
 \end{aligned} \tag{23}$$

**Fig. 12** Cantilever composite beam subjected to a parabolic distributed load



In Eq. 23,  $E_f$  and  $E_m$  are the elastic modulus for graphite and epoxy resin, respectively;  $\rho_f$  and  $\rho_m$  are the weight density of the graphite fiber and epoxy resin, respectively; and  $v_f$  and  $v_m$  respectively represent the fiber volume fraction and the matrix volume fraction in the continuous fiber composite material.

The design variables include (i) the second moment of area ( $x_1$ ), (ii) the depth of the beam ( $x_2$ ), and (iii) the fiber volume fraction ( $x_3$ ). The side constraints on the design variables and the values of the prescribed design parameters are listed in Table 4 and Table 5, respectively.

The beam optimization problem is defined as

**Table 4** Design variables for the beam design problem

Description	Notation	Lower limit	Upper limit
Second moment of area, $I$ [ $\text{mm}^4$ ]	$x_1$	$3.3E4$	$20.8E4$
Depth of the beam, $h$ [mm]	$x_2$	20	50
Fiber volume fraction, $v_f$	$x_3$	0.40	0.90

**Table 5** Prescribed design parameters for the beam design problem

Parameter	Value
Parabolic distributed load, $q_0$ [N/mm]	1
Length of the beam, $L$ [mm]	1000
Elastic modulus of graphite fiber, $E_f$ [N/mm <sup>2</sup> ]	$2.30E5$
Elastic modulus of epoxy resin, $E_m$ [N/mm <sup>2</sup> ]	$3.45E5$
Weight density of graphite fiber, $\rho_f$ [N/mm <sup>3</sup> ]	$1.72E - 5$
Weight density of epoxy resin, $\rho_m$ [N/mm <sup>3</sup> ]	$1.20E - 5$

$$\text{Minimize: } \frac{\delta_{max}}{\delta_0}, \quad [\delta_0] = 12.93 \quad (24)$$

subject to

$$W/W_0 \leq 1, \quad [W_0] = 2.9E4 \quad (25)$$

$$\sigma_{max}/\sigma_0 \leq 1, \quad [\sigma_0] = 200 \quad (26)$$

$$\frac{x_2^4}{1.2E6x_1} \leq 1 \quad (27)$$

$$x_i^{min} \leq x_i \leq x_i^{max}, \quad i = 1, 2, 3 \quad (28)$$

In this optimization formulation, the inequality constraints (Eqs. 25, 26, and 27) are related to the allowable weight, the maximum stress, and a geometric restriction on the beam design ( $depth \leq 10 \times width$ ). The weight and the maximum stress are given by

$$W = A\rho L = \frac{12I}{h^2} \times (12 + 5.2\nu_f)10^{-6} \times L = \frac{x_1}{x_2}(1440 + 624x_3) \quad (29)$$

$$\sigma_{max} = \frac{q_0 L^2 h}{8I} = \frac{1E6x_2}{8x_1} \quad (30)$$

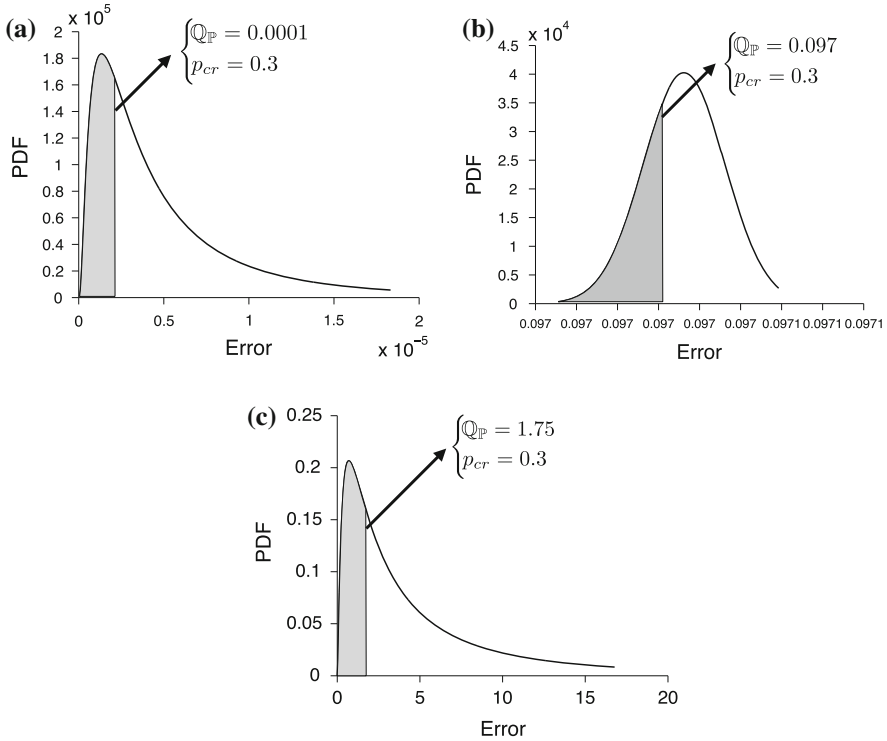
The models used to estimate the maximum deflection,  $d_{max}$ , are described next.

### 3.2.1 Structural Models with Different Levels of Fidelity

To develop the high fidelity physics-based structural model ( $M_{HF}^B$ ) and the low fidelity physics-based or PLF structural model ( $M_{PLF}^B$ ), the Finite Element Analysis package ANSYS is used. In ANSYS, the PLF Finite Element model is constructed using 2 beam elements, while the HF Finite Element model comprises 1000 beam elements. The third model ( $M_{SM}^B$ ) in this problem is a surrogate model constructed using Kriging with Gaussian correlation function. A set of 30 high fidelity function evaluations are used for this purpose. The fourth model ( $M_{TLF}^B$ ) is a tuned low fidelity model constructed using the *Multiplicative form* where

$$C(x) = \frac{\delta_{max}|_{HF}}{\delta_{max}|_{PLF}} \quad (31)$$

The distribution of the error in the tuned low fidelity model (TLF) and the surrogate model (SM) are estimated using PEMF (Sect. 2.3.1) and are illustrated in Fig. 13a, c, respectively. The distribution of the error in the Physics-based low fidelity model (PLF) is estimated using the inverse assessment process, by leveraging the same 30 high fidelity samples that were used to construct the TLF and SM; the PLF error distribution is shown in Fig. 13b.



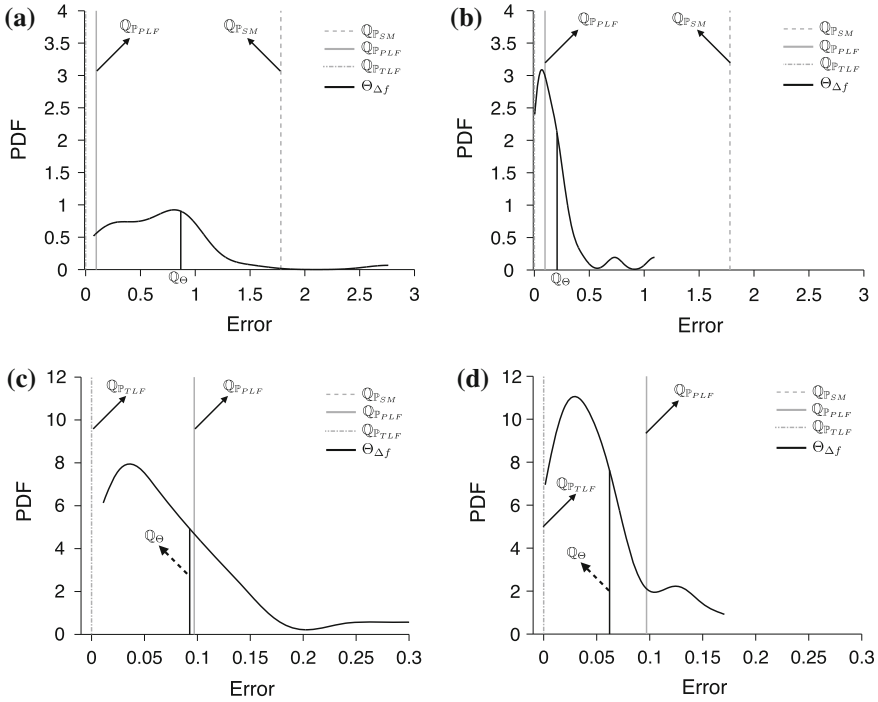
**Fig. 13** Distributions of the model errors for the cantilever beam design problem; **a** Tuned LF model, **b** Physics-based LF model, **c** Surrogate model

**Table 6** Models with different levels of fidelity used in the cantilever beam optimization problem (the high fidelity model is assumed to be a true representation of the system behavior)

Model	Location parameter $\mu$	Scale parameter $\sigma$	$Q(p_{cr})$ $p_{cr} = 0.3$	Number of calls made $N_{pop} \times$ <i>No. of Iter.</i>
Surrogate	1.22	1.20	1.75	$30 \times 3$
Physics-based LF	-2.30	0.001	0.097	$30 \times 6$
Tuned LF	-12.52	0.99	0.0001	$30 \times 7$
High fidelity	-	-	-	$30 \times 4$

### 3.2.2 Cantilever Beam Design: Results and Discussion

For the cantilever beam design problem, the four model types, the error distribution parameters and  $Q(p_{cr})$  associated with each model, and the number of calls made by AMS to each model are listed in Table 6. It can be seen from Fig. 13 and Table 6



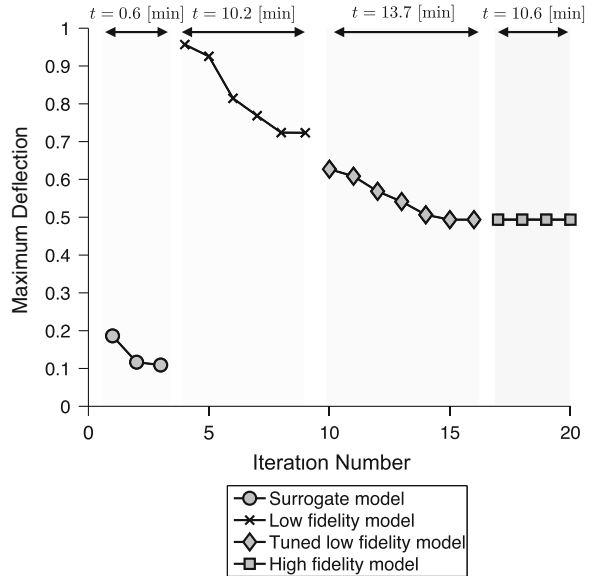
**Fig. 14** Distribution of the fitness function improvements in different iterations of the beam optimization with PSO-AMS (also showing the model error distributions); **a** 3rd iteration, **b** 8th iteration, **c** 10th iteration, **d** 14th iteration

that the tuned low fidelity model provides the highest degree of accuracy and the surrogate model is the least accurate among the three low fidelity models. Hence, the initial population of particles is generated using the surrogate model in this case.

Figure 14a–d illustrate the distribution of the relative fitness function improvements ( $Q_{\Theta}$ ) at different iterations during the optimization process. The  $(1 - p_{cr})$ -quantile of the  $Q_{\Theta}$  distribution, and the  $p_{cr}$ -quantile of the error distributions of the tuned low fidelity model, the surrogate model, and the physics-based low fidelity model are also shown in these figures.

The convergence history of the cantilever beam optimization performed by PSO-AMS is illustrated in Fig. 15. The AMS technique adaptively switches the model type three times (SM  $\rightarrow$  LF  $\rightarrow$  TLF  $\rightarrow$  HF) during the optimization process at the 3rd, the 9th, and the 16th iteration, therefore resulting in an optimum design with a high fidelity function estimate. There is a substantial discontinuity in the estimated function value at the first switching event (3rd iteration), which can be attributed to the significant uncertainty in the surrogate model—the  $Q(p_{cr})$  value of the surrogate model ( $M_{SM}^B$ ) is orders of magnitude higher than those of the other models ( $M_{PLF}^B$  and  $M_{TLF}^B$ ). To avoid the termination of PSO before reaching the high fidelity model

**Fig. 15** Optimization history of the cantilever beam optimization with PSO-AMS



( $M_{HF}^B$ ), the relative function tolerance is set to  $\delta = 10^{-5}$ , which is smaller than the modal error of the tuned low fidelity model.

In Table 7, the optimization results obtained by PSO-AMS is compared with the results yielded by running MDPSO solely using the surrogate model (PSO-SM), solely using the physics-based low fidelity model (PSO-PLF), solely using the tuned low fidelity model (PSO-TLF), and solely using the high fidelity model (PSO-HF). Interestingly, the PSO-AMS, PSO-TLF, and PSO-HF arrive at the same optimum design with  $f^* = 0.5435$ . It is seen from Table 7 that PSO-AMS reaches this optimum design at a 33% lower computational expense compared to PSO-TLF and a 119% lower computational expense compared to PSO-HF (both expense differences are estimated with respect to PSO-AMS expense). It is important to note from Table 7 that the performance of the surrogate model-based optimization (PSO-SM) is significantly worse than that of the others. The error in the surrogate model ( $M_{SM}^B$ ) at its optimum ( $X_{SM}^*$ ) is more than 99%, which is expected based on the predicted PEMF error of this model (Fig. 13c).

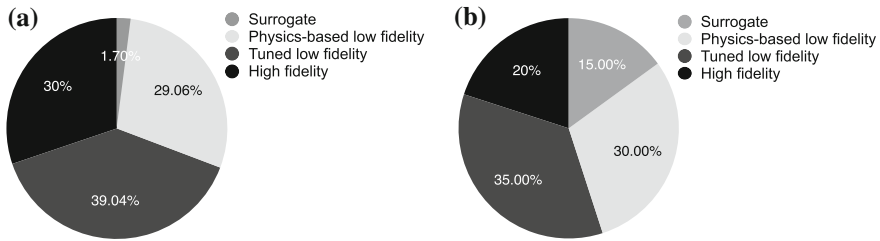
The resources used by the four different models, in terms of computing time and function calls, in the beam optimization performed by PSO-AMS are illustrated in Fig. 16a, b. It is observed that, unlike the airfoil problem, the surrogate model does not have a significant contribution in the beam optimization process in terms of function calls. Due to its high inaccuracy ( $Q(p_{cr}) = 1.75$ ), the fitness function improvement of the particles is quickly dominated by the error distribution of the surrogate model (in only 3 iterations). In this optimization process, the tuned low fidelity model ( $M_{TLF}^B$ ) makes the highest contribution in terms of computing time and function calls. This case study again shows that the uncertainty in the lower fidelity



**Table 7** Cantilever beam design: optimization results using single-fidelity and variable-fidelity optimization approaches

Approach	$x_1^* (e + 4)$	$x_2^*$	$x_3^*$	Optimum function $\delta_{max}/\delta_0$ ( $f^*$ )	Model in last iteration	Computational time [min] over function evaluation	HF response at optimum ( $f_{HF}(x^*)$ )
PSO-SM	2.82	43.28	0.71	0.0010	SM	5.7/990	0.8800
PSO-PLF	3.58	48.72	0.89	0.5011	PLF	18.4/330	0.5550
PSO-TLF	3.62	50.00	0.90	0.5435	TLF	46.8/720	0.5435
PSO-HF	3.62	50.00	0.90	0.5435	HF	76.85/870	0.5435
PSO-AMS	3.62	50.00	0.90	0.5435	HF	35.1/630	0.5435

*PSO-SM* optimization performed by MDPSO solely using the surrogate model  
*PSO-PLF* optimization performed by MDPSO solely using the physics-based low fidelity model  
*PSO-TLF* optimization performed by MDPSO solely using the tuned low fidelity model  
*PSO-HF* optimization performed by MDPSO solely using the high fidelity model  
*PSO-AMS* optimization performed by MDPSO using AMS



**Fig. 16** Percentage of resources used by each model in the cantilever beam optimization problem using PSO-AMS : **a** Computing time resources; **b** Function evaluation resources

models could exceed the relative function improvement across constitutive iterations way ahead of reaching convergence in practical optimization, and this behavior is also highly problem dependent. Such likely scenarios make this variable fidelity optimization technique (AMS) a unique and essential tool for designing complex systems, where fast low fidelity models are almost indispensable.

## 4 Conclusion

This article presented a novel model management technique that is implemented in population-based optimization algorithms to provide high fidelity optimum designs at a reasonable computational expense. The model pool is created with models that offer different (non-dominated) trade-offs between computational cost and fidelity. The optimization process is started using the model with the highest computational efficiency, which could be a physics-based low fidelity model or a surrogate model. A novel switching metric (called Adaptive Model Switching or AMS) is then used to determine when to switch to the next higher fidelity model during the optimization iterations. Assuming that the uncertainties associated with the lower fidelity models follow a probabilistic distribution (lognormal pdf is used here), the proposed model switching metric is defined as: “a probability estimate of whether the uncertainty associated with a model exceeds the improvement in the relative fitness function over the population of solutions”. The new adaptive model switching technique (AMS) is applied to: (i) 2D Airfoil design and (ii) Cantilever composite beam design. A powerful version of the Particle Swarm Optimization (mixed-discrete PSO) algorithm is used to implement and investigate the performance of AMS. The results indicate that AMS along with Mixed Discrete PSO improve the efficiency of the optimization process significantly when compared to optimization performed solely using high fidelity models, with up to 185 % reduction in computing time, while reaching the same or a better optimum. The value of the optimum with AMS is also better than that accomplished using only single low fidelity models for optimization. The current version of AMS is implemented primarily for optimization problems where multiple physics-based and/or surrogate models exist to represent the physical system behav-

ior. Future work will focus on problems where only a high fidelity physics-based model or experimental data is available, which can be used to construct different surrogates. A related notion is that of Surrogate-based design optimization, where surrogate models are improved through adaptive or sequential sampling during the optimization process. A more intuitive definition of the Indicator of Conservativeness (IoC) as a function of user's preferences regarding computational expense and robustness would further establish the wide potential of AMS for optimizing complex practical systems.

**Acknowledgments** Support from the National Science Foundation Awards CMMI-1100948 and CMMI-1437746 is gratefully acknowledged. Any opinions, findings, conclusions, or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the NSF.

## References

1. Hutchison MG, Unger ER, Mason WH, Grossman B, Haftka RT (1994) Variable-complexity aerodynamic optimization of a high-speed civil transport wing. *J Aircr* 31(1):110–116
2. Jeong S, Murayama M, Yamamoto K (2005) Efficient optimization design method using kriging model. *J Aircr* 42(2):413–420
3. Oktem H, Erzurumlu T, Kurtaran H (2005) Application of response surface methodology in the optimization of cutting conditions for surface roughness. *J Mater Proces Technol* 170(1):11–16
4. Simpson T, Booker A, Ghosh D, Giunta A, Koch P, Yang RJ (2004) Approximation methods in multidisciplinary analysis and optimization: a panel discussion. *Struct Multidiscip Optim* 27(5):302–313
5. Simpson T, Toropov V, Balabanov V, Viana F (2008) Design and analysis of computer experiments in multidisciplinary design optimization: a review of how far we have come or not. In: 12th AIAA/ISSMO multidisciplinary analysis and optimization conference, Victoria, Canada
6. Wang G, Shan S (2007) Review of metamodeling techniques in support of engineering design optimization. *J Mech Des* 129(4):370–381
7. Jin R, Chen W, Simpson TW (2000) Comparative studies of metamodeling techniques under multiple modeling criteria. *AIAA* (4801)
8. Forrester A, Keane A (2009) Recent advances in surrogate-based optimization. *Prog Aerosp Sci* 45(1–3):50–79
9. Simpson T, Korte J, Mauery T, Mistree F (2001) Kriging models for global approximation in simulation-based multidisciplinary design optimization. *AIAA J* 39(12):2233–2241
10. Choi K, Young B, Yang R (2001) Moving least square method for reliability-based design optimization. In: 4th world congress of structural and multidisciplinary optimization, Dalian, China, pp 4–8
11. Toropov VV, Schramm U, Sahai A, Jones RD, Zeguer T (2005) Design optimization and stochastic analysis based on the moving least squares method. In: 6th world congresses of structural and multidisciplinary optimization, Rio de Janeiro
12. Hardy RL (1971) Multiquadric equations of topography and other irregular surfaces. *J Geophys Res* 76:1905–1915
13. Clarke SM, Griebisch JH, Simpson TW (2005) Analysis of support vector regression for approximation of complex engineering analyses. *J Mech Des* 127(6): 1077–1087
14. Yegnanarayana B (2004) Artificial neural networks. PHI Learning Pvt, Ltd, New Delhi
15. Zhang J, Chowdhury S, Messac A (2012) An adaptive hybrid surrogate model. *Struct Multidiscip Optim* 46(2):223–238

16. Mehmani A, Chowdhury S, Messac A (2014) A novel approach to simultaneous selection of surrogate models, constitutive kernels, and hyper-parameter values. In: 55th AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials conference. National Harbor, MD, USA
17. Zhang J, Chowdhury S, Mehmani A, Messac A (2014) Characterizing uncertainty attributable to surrogate models. *J Mech Des* 136(3):031004
18. Barthelemy JF, Haftka R (1993) Approximation concepts for optimum structural design (in a review). *Struct Optim* 5(3):129–144
19. Haftka RT (1991) Combining global and local approximations. *AIAA J* 29(9):1523–1525
20. Keane A, Nair P (2005) Computational approaches for aerospace design: the pursuit of excellence. Wiley, Chichester
21. Zadeh PM, Mehmani A (2010) Multidisciplinary design optimization using variable fidelity modeling: application to a wing based on high fidelity models. In: Third international conference on multidisciplinary design optimization, Paris, France
22. Zadeh PM, Toropov VV, Wood AS (2009) Metamodel-based collaborative optimization framework. *Struct Multidiscip Optim* 38(2):103–115
23. Alexandrov NM, Lewis RM, Gumbert C, Green L, Newman P (1999) Optimization with variable-fidelity models applied to wing design. Technical report, ICASE, Institute for Computer Applications in Science and Engineering. NASA Langley Research Center, Hampton, Virginia
24. Booker AJ, Dennis JE, Frank PD, Serafini DB, Torczon V, Trosset MW (1999) A rigorous framework for optimization of expensive functions by surrogates. *Struct Optim* 17(1):1–13
25. Marduel X, Tribes C, Trepanier JY (2006) Variable-fidelity optimization: efficiency and robustness. *Optim Eng* 7(4):479–500
26. Robinson TD, Eldred MS, Willcox KE, Haimes R (2008) Surrogate-based optimization using multifidelity models with variable parameterization and corrected space mapping. *AIAA J* 46(11):2814–2822
27. Rodriguez JF, Perez VM, Padmanabhan D, Renaud JE (2001) Sequential approximate optimization using variable fidelity response surface approximations. *Struct Multidiscip Optim* 22(1):24–34
28. Alexandrov NM, Dennis JE, Lewis RM, Torczon V (1998) A trust-region framework for managing the use of approximation models in optimization. *Struct Optim* 15(1):16–23
29. Toropov VV, Alvarez LF (1998) Development of mars-multipoint approximation method based on the response surface fitting. *AIAA J* 98: 4769
30. Forrester A, Sobester A, Keane A (2008) Engineering design via surrogate modelling: a practical guide. Wiley, Chichester
31. Sugiyama M (2006) Active learning in approximately linear regression based on conditional expectation of generalization error. *J Mach Learn Res* 7:141–166
32. Trosset MW, Torczon V (1997) Numerical optimization using computer experiments. Technical report, DTIC Document
33. Bichon BJ, Eldred MS, Mahadevan S, McFarland JM (2013) Efficient global surrogate modeling for reliability-based design optimization. *J Mech Des* 135(1):011, 009
34. Duan Q, Sorooshian S, Gupta V (1992) Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resour Res* 28(4):1015–1031
35. Jones D, Schonlau M, Welch W (1998) Efficient global optimization of expensive black-box functions. *J Glob Optim* 13(4):455–492
36. Kleijnen JP, Beers WV, Nieuwenhuys IV (2012) Expected improvement in efficient global optimization through bootstrapped kriging. *J Glob Optim* 54(1):59–73
37. Jin Y, Olhofer M, Sendhoff B (2002) A framework for evolutionary optimization with approximate fitness functions. *IEEE Trans Evolut Comput* 6(5):481–494
38. Graning L, Jin Y, Sendhoff B (2007) Individual-based management of meta-models for evolutionary optimization with application to three-dimensional blade optimization. In: Evolutionary computation in dynamic and uncertain environments, pp 225–250

39. Jin Y (2005) A comprehensive survey of fitness approximation in evolutionary computation. *Soft Comput* 9(1):3–12
40. Ulmer H, Streichert F, Zell A (2004) Evolution strategies with controlled model assistance. In: *Evolutionary computation, 2004, IEEE congress on CEC2004, vol 2*, pp 1569–1576
41. Jin Y, Sendhoff B (2004) Reducing fitness evaluations using clustering techniques and neural network ensembles. In: *Genetic and evolutionary computation, GECCO 2004*, pp 688–699
42. Chowdhury S, Tong W, Messac A, Zhang J (2013) A mixed-discrete particle swarm optimization algorithm with explicit diversity-preservation. *Struct Multidiscip Optim* 47(3):367–388
43. Epanechnikov V (1969) Non-parametric estimation of a multivariate probability density. *Theory Probab Appl* 14:153–158
44. Duong T, Hazelton M (2003) Plug-in bandwidth matrices for bivariate kernel density estimation. *Nonparametric Stat* 15(1):17–30
45. Mehmani A, Chowdhury S, Messac A (2015) Predictive quantification of surrogate model fidelity based on modal variations with sample density. *Struct Multidiscip Optim* (Accepted)
46. Mehmani A, Chowdhury S, Zhang J, Tong W, Messac A (2013) Quantifying regional error in surrogates by modeling its relationship with sample density. In: *54th AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials conference, Boston, MA, USA*
47. Chowdhury S, Mehmani A, Messac A (2014) Concurrent surrogate model selection (cosmos) based on predictive estimation of model fidelity. In: *ASME 2014 international design engineering technical conferences (IDETC), Buffalo, NY*
48. Kennedy J, Eberhart RC (1995) Particle swarm optimization. In: *IEEE international conference on neural networks, vol 6*, pp 1942–1948
49. Coelho F, Breitkopf P, Knopf-Lenoir C (2008) Model reduction for multidisciplinary optimization: application to a 2d wing. *Struct Multidiscip Optim* 37(1):29–48

# Genetic Algorithms for the Construction of $2^2$ and $2^3$ -Level Response Surface Designs

Dimitris E. Simos

**Abstract** Response surface methodology is widely used for developing, improving and optimizing processes in various fields. In this paper, we present a general algorithmic method for constructing  $2^q$ -level design matrices in order to explore and optimize response surfaces where the predictor variables are each at  $2^q$  equally spaced levels, by utilizing a genetic algorithm. We emphasize on various properties that arise from the implementation of the genetic algorithm, such as symmetries in different objective functions used and the representation of the  $2^q$  levels of the design with a  $q$ -bit Gray Code. We executed the genetic algorithm for  $q = 2, 3$  and the produced four and eight-level designs achieve both properties of near-rotatability and estimation efficiency thus demonstrating the efficiency of the proposed heuristic.

**Keywords** Response surface designs · Genetic algorithms · Efficiency · Optimization

## 1 Introduction

Response surface methodology is used in experiments in which the main interests are to determine the relationship between the response and the settings of a group of experimental factors and to find the combination of the factor levels that gives the best expected response. Response surfaces can also provide information about the rate of change of the response variable and indicate the interactions between the treatment factors. This class of designed experiments has a wide range of applications in industrial and chemical engineering, agricultural experiments and biotechnological processes [1, 10, 12, 13, 18, 25].

In this paper we focus on the construction of  $2^q$ -level response surface designs by emphasizing on an algorithmic perspective of the problem. In such designs the design matrix columns are constituted of combinations of  $2^q$  distinct symbols and

---

D.E. Simos (✉)  
SBA Research, Favoritenstrasse 16, 1040 Vienna, Austria  
e-mail: dsimos@sba-research.org

correspond to the treatment factors, each at  $2^q$  equally spaced quantitative levels. Any combination of the levels of all factors under consideration is called a treatment combination. Let  $\mathbf{X} = [x_1, x_2, \dots, x_k]$  be the design matrix of the experiment in which, each row represents the  $n$  treatment combinations and each column gives the sequence of factor levels. For each factor, all level values are of equal interest and each experimental result should have equal influence. Thus we consider designs with the equal occurrence property, when for example we construct four-level designs we have that all columns consist of  $n/4$  elements equal to 1,  $n/4$  elements equal to  $-1$ ,  $n/4$  elements equal to  $1/3$ ,  $n/4$  elements equal to  $-1/3$ , if  $n$  is a multiple of four. The designs with the equal occurrence property are called balanced designs. Although  $2^q$ -level factors appear often in experimental problems, a minor work has been done in this specific area of response surface designs [10, 12, 15, 24].

The paper is organized as follows. In Sect. 2 the concepts and the measures of rotatability and efficiency of response surface designs are defined. A genetic algorithm approach for the construction of  $2^q$ -level response surface designs is presented in Sect. 3, while the obtained results are given in Sect. 4.

## 2 Model and Design Optimality Criteria

Suppose we want to test the effects of  $k$  predictor variables, coded to  $x_1, x_2, \dots, x_k$ , on a response variable  $y$  subject to random error. Generally the first attempt is to approximate the shape of the response surface by fitting a first-order model to the response,

$$y = \beta_0 + \sum_{j=1}^k \beta_j x_j + \varepsilon, \quad (1)$$

where  $\beta_0, \beta_j, j = 1, \dots, k$  are unknown parameters and  $\varepsilon$  is a random error term. When the first-order model appears inadequate to describe the true relationship between the response and the predictor variables due to the existence of surface curvature, it is upgraded to a second-order model

$$y = \beta_0 + \sum_{j=1}^k \beta_j x_j + \sum_{j=1}^k \beta_{jj} x_j^2 + \underbrace{\sum_{i=1}^k \sum_{j=1}^k \beta_{ij} x_i x_j}_{i < j} + \varepsilon, \quad (2)$$

where  $\beta_0, \beta_j, j = 1, \dots, k, \beta_{ij}, i = 1, \dots, k, j = 1, \dots, k$ , are unknown parameters and  $\varepsilon$  is a random error term.

Two of the most important characteristics that a response surface design should possess is rotatability and efficiency.

The concept of rotatability was introduced by Box and Hunter [3]. A  $k$ -dimensional design is called rotatable if the variance of the response estimated by the fitted polynomial at the point  $(x_1, \dots, x_k)$ ,  $\text{Var}[\hat{Y}(\mathbf{x})]$ , is a function only of  $\rho^2 = \sum_{i=1}^k x_i^2$ . Such a design insures that the estimated response has a constant variance at all points that are equidistant from the design center. One of the desirable features of rotatability is that the quality of the prediction, as measured by the magnitude of  $\text{Var}[\hat{Y}(\mathbf{x})]$ , is invariant to any rotation of the coordinate axes in the space of the input variables. In cases where exact rotatability is unattainable, it is important to measure how rotatable a design is. Khuri [17], Draper and Guttman [8] and Draper and Pukelsheim [9] proposed measures to test the near rotatability of a design. In this framework we use the rotatability measure  $Q^*$  provided by Draper and Pukelsheim [9] and given by the equation

$$Q^* = \frac{\|\bar{\mathbf{A}} - \mathbf{V}_0\|^2}{\|\mathbf{A} - \mathbf{V}_0\|^2} = \frac{\text{tr}(\bar{\mathbf{A}} - \mathbf{V}_0)^2}{\text{tr}(\mathbf{A} - \mathbf{V}_0)^2}, \quad (3)$$

where  $\bar{\mathbf{A}}$  is the rotatable component of the moment matrix  $\mathbf{A} = n^{-1}\mathbf{X}'\mathbf{X}$  and  $\mathbf{V}_0$  consists of a one in the (1, 1) position and zeros elsewhere. It is  $Q^* \leq 1$  and equality stands when the design is rotatable. For more details see [9].

Beyond testing the near rotatability of the designs in order to compare them, it is also needed to have an estimation efficiency measure for the same purpose. Box and Draper [4] discussed as a measure of design efficiency the choice of a design on the basis of maximizing the determinant of the information matrix. In this paper we adopt the following  $D$  criterion for determining the overall efficiency for estimating the set of the effects

$$|\mathbf{W}'\mathbf{W}|^{1/k}, \quad (4)$$

where  $\mathbf{W} = [x_0/\|x_0\|, x_1/\|x_1\|, \dots, x_k/\|x_k\|]$ ,  $x_0$  stands for the vector with all elements equal to 1, and  $x_i$  is the coefficient vector of the  $i$ th effect,  $i = 1, \dots, k$ . Since the columns of  $\mathbf{W}$  are standardized, the  $D$  criterion achieves its maximum value, which equals to 1, if and only if the  $x_i$  are orthogonal to each other. More details can be found in [26].

### 3 Optimization of Response Surface Designs by Means of Genetic Algorithms

Genetic algorithms form a powerful metaheuristic that mimicks processes from the Theory of Evolution to establish search algorithms by defining algorithmic analogues of biological concepts such as reproduction, crossover and mutation. Genetic Algorithms were introduced in 1970 by Holland [16] aiming to design an artificial system having properties similar to natural systems. In this paper, we assume some basic familiarity with Genetic Algorithm concepts. The concepts necessary



for a description of the Genetic Algorithm (GA) can be found in Goldberg [14], in Forrest's article [11] and in the Handbook of Genetic Algorithms edited by Davis [6].

GAs are attractive because of their robustness and flexibility in terms of a computer implementation and, mathematically, they do not require a differentiable objective function thereby reducing the chance of reporting local optima. Some earlier attempts utilizing a GA in the construction of response surface designs has been given by Drain et al. [7]. However, this approach, while promising, lacked of an efficient coding of the chromosomes i.e. the number of the experimental runs forming the design. In particular, the authors proposed utilizing and constructing the whole design; thus restricting the GA to evolve in finding optimal response surface designs in several cases. A successful reduction in terms of computational complexity of an efficient representation of the candidate design, has been proposed in [21–23] in a similar field of computational design theory with strong connection to statistical applications. In these applications, the authors integrated as a core ingredient of the GA the use of sequential juxtaposition of suitable generators, either forming circulant matrices [21] or block circulant matrices [22, 23].

### ***3.1 A Genetic Algorithm Framework for Response Surface Designs***

#### **Chromosomes Representation**

The respective generators considered in the case of response surface designs are the  $n/2^q$  column vectors which in the process form block circulant matrices of order  $k$ , when constructing an  $n \times k$  response surface design. This construction, is valid when  $n$  is a multiple of  $2^q$ . In particular, we form  $n/4$  and  $n/8$  column vectors when we consider four and eight-level response surface designs, respectively. However, in all previous constructions the generators, more precisely the genes forming a generator, consisted of binary variables since a two-level design was under development. In the case of response surface designs, the genes constitute of  $2^q$  possible values representing the  $2^q$ -levels of the designs.

#### **Chromosomes Encoding and Decoding**

A suitable encoding to binary variables was needed since the genetic operators behave better in binary arithmetic (Goldberg [14]). The answer to this vital question found in the field of Combinatorics and Computer Science in terms of representing a 2-bit Gray Code,  $GC_2 = \{00, 01, 11, 10\}$  when considering four-level designs; while in the case of eight-level designs we used a 3-bit Gray Code,  $GC_3 = \{000, 001, 011, 010, 110, 111, 101, 100\}$ . For more details, on Gray Codes we refer the interested reader to Carla [5]. More precisely, we mapped each level of a four-level design to a codeword of the 2-bit Gray Code, i.e.  $\{-1, -1/3, 1/3, 1\} \rightarrow \{00, 01, 11, 10\}$ , and each level of an eight-level design to a codeword of the 3-bit Gray Code i.e.  $\{-1, -1/3, -1/6, -1/9, 1/9, 1/6, 1/3, 1\} \rightarrow \{000, 001, 011, 010, 110, 111, 101, 100\}$ , thus transforming the problem on its binary equivalent which

allowed us to carry on with the next stages of utilizing a GA. It is made clear that we could repeat this procedure for  $2^q$ -level response surface designs by using a  $q$ -bit Gray Code. Details for constructing a  $q$ -bit Gray Code can be found in Knuth [19].

### Initial Population

consists of random chromosomes. We found it useful to generate these chromosomes by retrieving samples of binary sequences, after random permutations were applied to each of them.

### An Objective function for Response Surface Designs

The crucial choice of the objective function (OF) subject to be optimized arise naturally from the theoretical framework of rotatable and efficient designs. In particular, we have developed two versions of the algorithm each one depending on one of the two optimality design criteria, the rotatability measure  $Q^*$  and the  $D$  criterion. The genetic algorithm attempts in both cases to maximize the value of each criterion with respect to its upper bound which is equal to 1. Due to the theoretical background and statistical justifications when a value of  $Q^*$  was detected in the range of [0.95, 1.00] we considered we have found a global optimum solution, while in the case of  $D$  criterion we made some ramifications to accept a lower bound for the range of optimum solutions, i.e. [0.65, 1.00]. Thus we were able to detect both rotatable and efficient designs. In the following figure we give a comparison of the genetic algorithm performance in terms of contrasting the  $Q^*$  versus the  $D$  criterion by scaling on the evolving generations. From the figure we can conclude that we can use the  $Q^*$  and  $D$  criterion interchangeably as objective functions, since the fitness values for each case are similar.

We are now able to describe the three genetic operators of reproduction, crossover and mutation as specifically have been applied by the genetic algorithm we have used.

### Crossover

We defined the basic genetic operation, crossover, that splits a pair of binary integers at a random position and combines the head of one with the tail of the other and vice versa.

### Mutation

Additional operations, such as inverting a section of the binary representation (inversion) or randomly changing the state (0 or 1) of individual bits (mutation), also transform the population (Fig. 1).

### Selection and Reproduction

Before each such cycle (generation), population members are selected on the basis of their fitness (the value of the objective function for that solution) to be the “parents” of the new generation.

### Termination Condition

of the genetic algorithm was set a predefined number of evolved generations. This number of generations was proportional to the size of the response surface design that the genetic algorithm was searching for in each case. Thus the GA required only a few generations to find a small sized optimal response surface design, while a larger

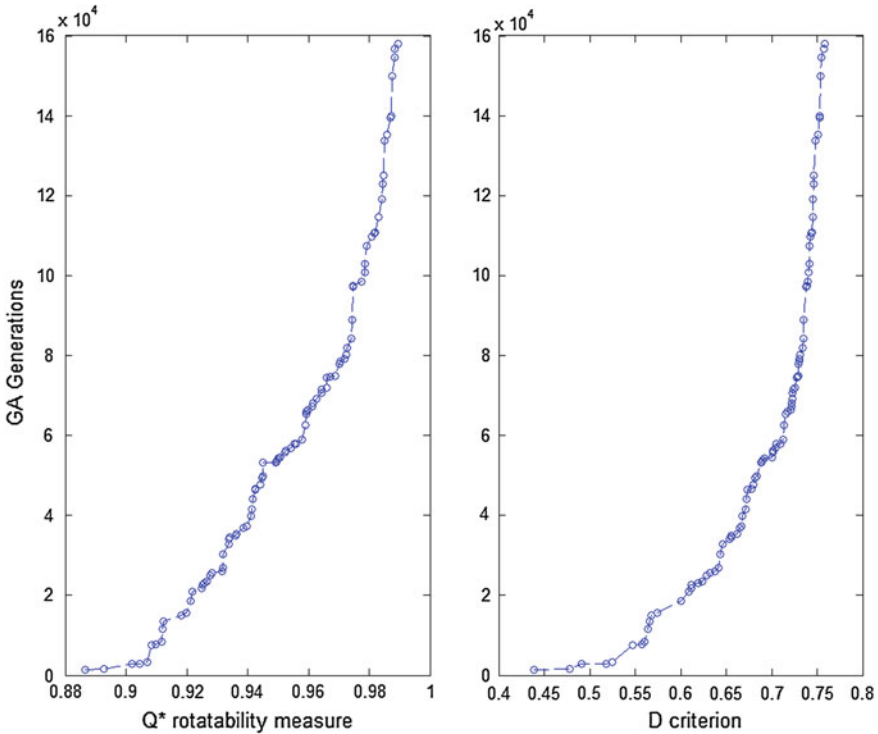


Fig. 1 Symmetries on objectives functions for optimization of response surface designs

design required additional generations to be evolved. Since GA is a heuristic process, the time complexity of the algorithm was relatively small compared to exhaustive search algorithms.

### 4 New Four and Eight-Level Response Surface Designs

In this Section we present the results of the construction method for four and eight-level response surface designs as described previously.

#### 4.1 New Four-Level Response Surface Designs

In Table 1,  $k$  stands for the number of the experimental factors and  $n$  for the number of the performed runs, while in the next two columns the achieved values for the  $Q^*$  and the  $D$  criterion are listed.

**Table 1** Some new four-level response surface designs with  $k$  factors

$k$	$n$	$Q^*$	$D$	$k$	$n$	$Q^*$	$D$
2	48	0.989625	0.730767	3	76	0.990628	0.751091
2	52	0.989279	0.730478	3	80	0.988196	0.748842
2	56	0.989625	0.730767	3	84	0.986588	0.749383
2	60	0.989282	0.730437	4	72	0.979375	0.753859
2	64	0.989625	0.730767	4	76	0.982954	0.757085
2	68	0.989051	0.730598	4	80	0.979180	0.759259
2	72	0.989625	0.730767	4	84	0.978854	0.759373
2	76	0.989166	0.730632	4	88	0.983253	0.759436
2	80	0.989539	0.730718	5	84	0.972752	0.752474
3	52	0.986202	0.746648	5	88	0.959241	0.754575
3	56	0.988977	0.748505	5	92	0.966673	0.757510
3	60	0.987961	0.748139	6	92	0.965235	0.748283
3	64	0.989528	0.747663	6	96	0.973121	0.745234
3	68	0.988830	0.748171	7	96	0.956693	0.720159
3	72	0.987170	0.748501	7	100	0.962628	0.726570

From the above results we note that the  $Q^*$  values fluctuate between 95.67 % and 99.06 % and the arithmetical mean equals to 98.20 %, while the maximum and the minimum values of the  $D$ -criterion are 75.94 % and 72.02 %, respectively, with the arithmetical mean equal to 74.36 %. Also, Koshal’s designs (see [2, 20]) are occasionally of use in response surface work. For the third-order Koshal design in 3 four-level predictor variables with 20 runs, given in page 504 of [2], we calculate the corresponding values of  $Q^*$  and  $D$ -criterion, which are equal to 0.3150 and 0.2613, respectively. In general, high values of the two criteria,  $Q^*$  and  $D$ , ensure that the designs are near-rotatable and efficient for estimating the set of the effects.

### 4.2 New Eight-Level Response Surface Designs

In this section we present the results of the construction method for eight-level response surface designs. In Table 2,  $k$  stands for the number of the experimental factors and  $n$  for the number of the performed runs, while in the next two columns the achieved values for the  $Q^*$  and the  $D$  criterion are listed.

From the above results we note that the  $Q^*$  values fluctuate between 95.30 % and 99.99 % and the arithmetical mean equals to 98.13 %, while the maximum and the minimum values of the  $D$ -criterion are 87.95 % and 65.28 %, respectively, with the arithmetical mean equal to 78.06 %.

As a conclusion, our construction method manages to generate near-rotatable and efficient response surface designs with a small number of required runs for both

**Table 2** Some new eight-level response surface designs with  $k$  factors

$k$	$n$	$Q^*$	$D$	$k$	$n$	$Q^*$	$D$
2	8	0.984915	0.680165	5	32	0.962000	0.652791
2	16	0.996432	0.867841	5	40	0.969055	0.727856
2	24	0.996035	0.867885	5	48	0.972596	0.758961
2	32	0.998573	0.874306	5	56	0.978982	0.740121
2	40	0.999523	0.875252	5	64	0.978354	0.784608
2	48	0.999696	0.875406	5	72	0.981888	0.803481
2	56	0.999929	0.876042	5	80	0.984097	0.819177
2	64	0.999974	0.876227	5	88	0.985177	0.821409
3	16	0.990309	0.673609	6	40	0.952993	0.676931
3	24	0.991651	0.827829	6	48	0.963848	0.688760
3	32	0.990673	0.815758	6	56	0.967265	0.682539
3	40	0.995118	0.865237	6	64	0.972076	0.717860
3	48	0.994795	0.844095	6	72	0.972703	0.740062
3	56	0.995701	0.871523	6	80	0.975458	0.766400
3	64	0.997662	0.867527	6	88	0.975989	0.748631
3	72	0.998499	0.879501	6	96	0.976775	0.801824
4	24	0.970440	0.727679	7	48	0.956076	0.658748
4	32	0.980760	0.773002	7	56	0.958585	0.690182
4	40	0.984617	0.791023	7	64	0.957366	0.659179
4	48	0.987015	0.821650	7	72	0.959647	0.694141
4	56	0.990165	0.827778	7	80	0.966887	0.712859
4	64	0.991205	0.831423	7	88	0.970750	0.724768
4	72	0.992321	0.852958	7	96	0.969300	0.726233
4	80	0.993720	0.860543	7	104	0.973606	0.746514

cases of four and eight-level designs thus demonstrating the efficiency of the genetic algorithm used. From these experimental results, it is anticipated that the proposed formulation for  $2^q$ -level response surface designs should produce similar results for higher number of levels when combined with a genetic algorithm utilized with the aid of a  $q$ -bit Gray Code.

**Acknowledgments** This work was carried out during the tenure of an ERCIM “Alain Bensoussan” Fellowship Programme. This Programme is supported by the Marie Curie Co-funding of Regional, National and International Programmes (COFUND) of the European Commission. This work was partly funded by COMET K1, FFG—Austrian Research Promotion Agency.

## References

1. Box GEP, Draper NR (1987) Empirical model building and response surfaces. Wiley, New York
2. Box GEP, Draper NR (2007) Response surfaces, mixtures, and ridge analyses. Wiley, New York
3. Box GEP, Hunter JS (1957) Multifactor experimental designs for exploring response surfaces. *Ann Math Stat* 28:195–241
4. Box MJ, Draper NR (1971) Factorial designs, the  $|X'X|$  criterion and some related matters. *Technometrics* 13:731–742
5. Carla S (1997) A survey of combinatorial gray codes. *Soc Ind Appl Math Rev* 39:605–629
6. Davis L (1991) Handbook of genetic algorithms. Van Nostrand, Reinhold
7. Drain D, Carlyle WM, Montgomery DC, Borror C, Anderson-Cook C (2004) A genetic algorithm hybrid for constructing optimal response surface designs. *Qual Reliab Eng Int* 20:637–650
8. Draper NR, Guttman I (1988) An index of rotatability. *Technometrics* 30:105–111
9. Draper NR, Pukelsheim F (1990) Another look at rotatability. *Technometrics* 32:195–202
10. Edmondson RN (1991) Agricultural response surface experiments based on four-level factorial designs. *Biometrics* 47:1435–1448
11. Forrest S (1993) Genetic algorithms: principles of natural selection applied to computation. *Science* 261:872–878
12. Gilmour SG (2004) Irregular four-level response surface designs. *J Appl Stat* 31:1043–1048
13. Gilmour SG (2006) Response surface designs for experiments in bioprocessing. *Biometrics* 62:323–331
14. Goldberg DE (1989) Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, Reading
15. Gupta TK, Dey A (1975) On some new second order rotatable designs. *Ann Inst Stat Math* 27:167–175
16. Holland JH (1975) Adaptation in natural and artificial systems, an introductory analysis with applications to biology, control and artificial intelligence. University of Michigan Press, Ann Arbor
17. Khuri AI (1988) A measure of rotatability for response surface designs. *Technometrics* 30:95–104
18. Khuri AI, Cornell JA (1996) Response surfaces, 2nd edn. Dekker, New York
19. Knuth DE (2004) Generating all n-tuples. *The Art Computer Programming, Volume 4A: Enumeration and Backtracking, pre-fascicle 2a*
20. Koshal RS (1933) Application of the method of maximum likelihood to the improvement of curves fitted by the method of moments. *J Roy Stat Soc Ser A* 96:303–313
21. Koukouvinos C, Mylona K, Simos DE (2007) Exploring k-circulant supersaturated designs via genetic algorithms. *Comput Stat Data Anal* 51:2958–2968
22. Koukouvinos C, Mylona K, Simos DE (2008)  $E(s^2)$ -optimal and minimax-optimal cyclic supersaturated designs via multi-objective simulated annealing. *J Stat Plann Infer* 138:1639–1646
23. Koukouvinos C, Mylona K, Simos DE (2009) A hybrid SAGA algorithm for the construction of  $E(s^2)$ -optimal cyclic supersaturated designs. *J Stat Plann Infer* 139:478–485
24. Koukouvinos C, Mylona K, Simos DE, Skountzou A (2009). An algorithmic construction of four-level response surface designs. *Comm Stat Simul Comput* 38:2152–2160
25. Myers RH, Montgomery DC (2002) Response surface methodology, 2nd edn. Wiley, New York
26. Wang JC, Wu CFJ (1995) A hidden projection property of Plackett-Burman and related designs. *Stat Sin* 5:235–250

# Reactive Power and Voltage Control Based on Mesh Adaptive Direct Search Algorithm

Seyyed Soheil Sadat Hosseini, Amir H. Gandomi, Alireza Nemati  
and Seyed Hamidreza Sadat Hosseini

**Abstract** This is a pioneer study that presents a new optimization algorithm called mesh adaptive direct search (MADS) to solve optimal steady-state performance of power systems. MADS is utilized to specify the optimal settings of control variables, i.e. transformer taps and generator voltages for optimal reactive power and voltage control of IEEE 30-bus system. Covariance matrix adaptation evolution strategy (CMAES) algorithm is utilized as a strong search strategy in the MADS technique to enhance its effectiveness. The results acquired by the hybrid search algorithm coupling MADS and CMAES, called MADS-CMAES, and the MADS algorithm itself without any search method are compared with multi-objective evolutionary and particle swarm optimization algorithms, demonstrating the superiority of MADS. The proposed MADS-based techniques are very robust against their parameters and changing the search space because of their inherent adaptive tuning.

**Keywords** Mesh adaptive direct search algorithm · Covariance matrix adaptation evolution strategy · Reactive power control · Voltage control

## 1 Introduction

A number of optimization problems have to be solved in order to the economic and secure operation of large-scale power systems. The optimal power flow (OPF) problem, introduced in 1960s by Carpentier, [1] is a powerful and critical tool in

---

S.S. Sadat Hosseini (✉) · A. Nemati  
Department of Electrical Engineering and Computer Science, University of Toledo,  
Toledo, OH 43606-3390, USA  
e-mail: s.sadathosseini@gmail.com

A.H. Gandomi  
BEACON Center for Study in Action, Michigan State University, East Lansing, MI 48824, USA  
e-mail: a.h.gandomi@gmail.com

S.H. Sadat Hosseini  
Department of Civil Engineering, Shomal University, Amol, Iran

power system operation and planning field. The reactive power dispatch (RPD) is a sub-problem of optimal power flow (OPF) calculation, which is generally a highly constrained non-linear non-convex optimization problem. The reactive power dispatch problem has a considerable effect on economic and secure operation of power systems. Maintaining the load bus voltages within the limits for high quality consumer services is a major task of a power system. Since the electric power loads vary from time to time, any change in the power demand results in lower or higher voltages [2]. This sub-problem specify all kinds of controllable variables, such as tap ratios of transformers and generator voltages. RPD optimizes transmission losses or other appropriate fitness functions, while satisfying a given set of operating and physical constraints. Another objective function with the same physical and operating constraints as RPD is voltage profile of the power system. In this problem, the main aim is to optimize the voltage deviations (VD) at load buses. Several conventional optimization algorithms such as linear programming, Newton method, quadratic programming, dynamic programming and interior point methods [3–7] have been developed to solve the reactive power dispatch problem. In general, most of these methods suffer from insecure convergence, algorithmic complexity, sensitivity to initial search point, etc. [8].

In recent years, global optimization algorithms (e.g. [9–12]) and specially a particular family introduced and developed by researchers in 1960 (e.g. [13]), has received great attentions. This family of techniques is named direct search techniques. Direct search algorithms search a set of point, around the current point, looking for a point that has less fitness function value than the current one does. This family contains powell optimization (PO), simplex methods (SM) (different form the simplex used in linear programming), pattern search (PS) techniques, and others [14]. Many interesting results come from the usage of PS methods in the optimization area [15]. Direct search algorithms are called derivative-free optimization techniques, where they do not need any further information about the gradient or even higher derivative of the fitness function to search for an optimal solution. Therefore, direct search methods may very well be used to solve non-continues, non-differentiable and multimodal, i.e. multiple local optima, optimization problem.

At this stage, mesh adaptive direct search (MADS) [16, 17] is one of the most powerful optimization techniques that has recently developed. MADS is supported by a thorough convergence analysis [18]. The reason for the using of MADS is that mathematical optimization algorithms, such as quadratic programming, nonlinear programming, Newton-based techniques, sequential unconstrained minimization, and interior point methods, have failed in handling non-smoothness and non-convexities in engineering optimization problems. Despite significant advantages of MADS over other optimization methods, there have been some little scientific efforts directed at applying it to practical and academic problems [18, 19].

The main goal of this study is to introduce the MADS method to find out the optimal settings of control variables, such as transformer taps and voltage magnitudes for two optimization problems, namely optimization of (a) real power losses in transmission lines and (b) sum of voltage deviations on load busses. Results of the MADS algorithm without any search strategy (MADS-N) and with covariance



matrix adaptation evolution strategy (MADS-CMAES) on the networks of IEEE 30-bus are presented. The results are compared with by other evolutionary computational algorithms such as multi-objective EA [20], global variant (PSO-PC) based on passive congregation [21] and local variant (CLONEPAC) PSO based on passive congregation [21]. The comparison shows the better performance of MADS-CMAES in locating optimal solutions. This paper is organized as follows: the problems of reactive power and voltage control are formulated in Sect. 2. Section 3 deals with the MADS method, which is effectively utilized in power engineering problems. Section 4 presents performance evaluation of MADS in comparison with the evolutionary computational techniques. Finally, Sect. 5 concludes this paper.

## 2 Optimal Power Flow

Optimal power flow is a static constrained nonlinear optimization problem, the solution of which specifies the optimal settings of control variables in a network respecting several constraints. Therefore, the goal is to determine a set of nonlinear equations illustrating the optimal solution of power system. It is described as:

$$\begin{aligned} \min & f(x, u) \\ \text{s.t.} & h(x, u) = 0 \\ & g(x, u) \leq 0 \end{aligned} \tag{1}$$

where,  $f$  is the fitness function that usually contains total generation cost, losses in transmission system etc. In general,  $h(x, u)$  represents the nonlinear power flow equations and  $g(x, u)$  represents transmission line limits and other security limits. The dependent and control variables vectors are respectively denoted by  $x$  and  $u$ . Generally, the dependent vector contains load bus voltage magnitudes  $V_L$ , bus voltage angles  $\theta$  and generator reactive power outputs  $Q_g$ , i.e.,  $x = [\theta, V_L, Q_g]^T$ . The control variable vector includes real power generation  $P_g$ , generator voltage  $V_g$ , transformer tap settings  $t$  and shunt VAR compensations  $Q_c$ , i.e.,  $u = [P_g, V_g, t, Q_c]^T$ . Of the mentioned control variable,  $P_g$  and  $V_g$  are continuous variables, while tap ratio,  $t$ , of tap changing transformers and reactive power output of compensation devices,  $Q_c$ , are discrete in nature.

Minimization of loss is required when cost minimization is the main aim with active power generation. A subsequent loss minimization will not yield enhancements, when all control variables are utilized in a cost optimization. Therefore, active power generation of all generators is fixed during the optimization procedure in reactive power dispatch problem, such as loss minimization, except slack generator.

### 3 Problem Formulation

MADS is tested and compared with the evolutionary computational techniques on optimal steady state performance in terms of optimization of (a) losses in transmission lines and (b) sum of voltage deviations on load busses while meeting various inequality and equality constraints. Since the main goal of this paper is the performance evaluation of MADS algorithm, two nonlinear optimization problems are individually studied. The first objective is to minimize the active power loss in the transmission network given as below:

$$f_1 = P_{loss}(x, u) = \sum_{l=1}^{N_l} P_l \quad (2)$$

In which  $u$  is the vector of control variables,  $x$  is the vector of dependent variables,  $P_l$  is the real power losses at line- $l$  and  $N_l$  is the number of transmission lines. The second fitness function is to minimize the voltage profile of the power system. The objective is to optimize the voltage deviations at load buses that can be described as:

$$f_2 = VD(x, u) = \sum_{i=1}^{N_d} |V_i - V_i^{SP}| \quad (3)$$

where  $V_i^{SP}$  is the pre-specified reference value at load bus- $i$ , which is usually set at the value of 1.0 pu, and is the number of load buses. The equality constraints of both optimization problems are load flow equations as follows:

$$P_{gi} - P_{di} - V_i \sum_{j=1}^{N_b} V_j (g_{ij} \cos \theta_{ij} + b_{ij} \sin \theta_{ij}) = 0 \quad (4)$$

$$Q_{gi} - Q_{di} + Q_{ci} - V_i \sum_{j=1}^{N_b} V_j (g_{ij} \sin \theta_{ij} - b_{ij} \cos \theta_{ij}) = 0 \quad (5)$$

where,  $g_{ij}$ ,  $b_{ij}$  are real and imaginary parts of  $(i, j)$  element of bus admittance matrix;  $P_{gi}$  and  $Q_{gi}$  are, respectively, the generator real and reactive power at bus  $i$ ;  $P_{di}$  and  $Q_{di}$  are, respectively, the load real and reactive power at bus  $i$ , respectively.  $Q_{ci}$  is the reactive power compensation source. The system operating constraints the inequality constraints for both of these problems are as follows.

- Generation constraints: Generators voltages and reactive power outputs are limited by their lower and upper limits as:

$$V_{gi}^{\min} \leq V_g \leq V_{gi}^{\max} \quad (6)$$

$$Q_{gi}^{\min} \leq Q_g \leq Q_{gi}^{\max} \quad (7)$$

- Shunt VAR constraints: Shunt VAR compensations are limited by their limits given as below:

$$Q_{ci}^{\min} \leq Q_c \leq Q_{ci}^{\max} \quad (8)$$

- Transformer constraints: Transformer tap settings  $t$  are limited given as follows:

$$t_i^{\min} \leq t_i \leq t_i^{\max} \quad (9)$$

- Functional operating constraints: This term refers to the constraints of load voltages at load buses  $V_L$  and transmission line loadings  $S_l$  as follows:

$$V_{L_i}^{\min} \leq V_L \leq V_{L_i}^{\max} \quad (10)$$

$$S_l \leq S_l^{\max} \quad (11)$$

## 4 Mesh Adaptive Direct Search Algorithm

The mesh adaptive direct search method for nonlinear optimization develops the generalized pattern search (GPS) [22, 23] methods. A major advantage of MADS over the GPS technique for both linearly constrained optimization and unconstrained is that the space of variables is not limited to a finite number of directions, named POLL directions. This is the prime drawback of the GPS methods, and the crucial motivation in developing MADS was to conquer this limitation [16].

**Notation.**  $\mathbf{R}$ ,  $\mathbf{Z}$ , and  $\mathbf{N}$  indicate the sets of real numbers, integers, and nonnegative integers respectively. For a matrix  $D$ , the notation  $d \in D$  shows that  $d$  is a column of  $D$ . Index  $i$  is denoted the iteration numbers.

### 4.1 Features of the MADS Method

The MADS technique is an iterative one. The MADS method begins with an initial point with finite objective value. The MADS method is derivative-free. This is crucial when  $\nabla f$  is inaccessible, either it does not exist, or it cannot be precisely computed because of noise in  $f$  or other reasons. A finite number of trial points are made at each iteration to determine an enhanced mesh point, that is, one with a lower fitness function value than that of the current incumbent.

Each iteration is split into two steps. The first step is named the search step. Any finite set of mesh points can be calculated in this step. This step lets great flexibility to choose strategies. When no trial points are considered, the search step is expressed to be empty. The search step adds nothing to the convergence theory, except to provide counterexamples as in [24]. It is notable that well-chosen search techniques can improve method performance (see [25–28]). The aim of iteration is to locate unfiltered points in  $X$ . If search fails to done an unfiltered point, then the second step, POLL is applied, and if POLL fails, then the mesh is modified.

## 4.2 Description of MADS Algorithm

The steps in the procedure of MADS are displayed in Fig. 2. These steps are explained in the next five subsections.

### 4.2.1 Initialization

In initialization step, a starting point  $x_0$  is chosen, and an initial mesh size parameter  $\Delta_0^m$ . The superscripts  $m$  and  $p$  stand for mesh and poll, respectively.

The algorithm parameters are defined as follows.

- $\Delta_i^p = \sqrt{\Delta_i^m}$  : the poll size parameter;
- $D = \{\pm e_k, k = 1, 2, \dots, n\}$  : to create polling direction, the basis is used, where  $e_k$  is the  $k$ th coordinate direction;
- $M_i = \{x \in S_i\} \cup \{x_i + \Delta_i^m Dz : z \in N^{2n}\}$ ; the mesh is defined with  $S_i$  which is the set of points, and the fitness function had been computed by the initialization. This mesh is to be generated by adding the current point to a set of independent vectors forming a certain pattern for the direction of the search towards optimality.

### 4.2.2 Search Step

The search step can be empty. This means that the algorithm can be implemented as a sequence of poll steps only. The search strategy utilized in the present study is briefly described in the following subsections.

## Covariance Matrix Adaptation Evolution Strategy

The first evolutionary methods come from the mid-60s with genetic algorithms (GAs) [29], evolutionary programming (EP) from Fogel et al. [30] and independently evolution strategy (ES) from Rechenberg [31]. Their studies brought a variety class of optimization techniques for difficult problems where few are famous about the

underlying search space. Facing a multitude of methods, the attention is focused on the evolution techniques branch of evolutionary methods. First, ES was advanced by Rechenberg [31] and Schwefel [32] and have developed into the cumulative step-path adaptation algorithm (CSAES) [33, 34] and the CMAES [35, 36]. CMAES [35, 37, 38] is an evolution technique that adapts the full covariance matrix of a normal search distribution.

Similar to quasi-Newton techniques, the CMAES computes the inverse Hessian matrix. As opposed to quasi-Newton methods the CMAES does neither calculate nor utilize gradients. The former makes the method feasible on non-separable and/or ill conditioned problems. The latter method makes the technique feasible on multimodal and/or noisy problems. The CMAES efficiently minimizes unimodal objective functions and is particularly superior on ill-conditioned and non-separable problems [35, 39]. CMAES has been utilized to solve many optimization problems [36] and successfully applied to a number of real world problems [40].

### 4.2.3 The POLL Step

Whenever the search step fails to make an improved mesh point, then the poll step is implemented before terminating the iteration. The poll step consists of a local exploration of the space of optimization variables near the current incumbent solution  $x_i$  (called the frame center). The set of trial points considered during the poll step is called a frame. If the poll step fails to produce an improved mesh point,  $P_i$  is expressed as a minimal frame with minimal frame center  $x_i$ . If both the search and poll steps are successful in finding an enhanced mesh point, the improved mesh point turns into the new current iterate  $x_{i+1}$  and the mesh is either retained or coarsened. If none of the steps are successful, then the minimal frame center is retained as the current iterate ( i.e.,  $x_{i+1} = x_i$ ) and the method continues to the parameters update step.

For MADS, the poll size parameter  $\Delta_i^p \in R_+$  for iteration  $i$  is presented. This new parameter regulates the magnitude of the distance from the trial points made by the poll step to the current incumbent solution  $x_i$ . The MADS frame is built by the usage of a current incumbent solution  $x_i$  and the poll and mesh size parameters  $\Delta_i^p$  and  $\Delta_i^m$  to obtain a positive spanning set of directions  $D_i$ . Generally, the MADS set of directions  $D_i$  is not a subset of  $D$ .

At iteration  $i$ , the MADS frame is described to be the set

$$P_i = \{x_i + \Delta_i^m d : d \in D_i\} \subset M_i, \tag{12}$$

where  $D_i$  is defined as a positive spanning set such that  $0 \notin D_i$  and for each  $d \in D_i$ ,

- each direction,  $d$ , can be computed using a nonnegative integer combination of the directions in  $D$ :  
 $d = Du$  for some vector  $u \in N^{n_{D_i}}$  that may change in some iteration

- the distance from the current point  $x_i$  to a frame point  $x_i + \Delta_i^m d \in P_i$  is limited and it is determined by a constant times the poll size parameter:  $\Delta_i^m \|d\| \leq \Delta_i^p \max\{\|d'\| : d' \in D\}$

#### 4.2.4 Parameters Update

The mesh size parameter  $\Delta_i^m$  is updated based on given a fixed rational number  $\tau > 1$  and two integers  $w^- \leq -1$  and  $w^+ \geq 0$  as follows.

$$\Delta_{i+1}^m = r^{w_i} \Delta_i^m \text{ for some } w_i \in \begin{cases} \{0, 1, \dots, w^+\} & \text{if an improved mesh point is found} \\ \{w^-, w^- + 1, \dots, -1\} & \text{otherwise} \end{cases} \quad (13)$$

#### 4.2.5 Termination

Some termination criteria must be specified, such as a minimal value on the mesh size parameter  $\Delta_i^m$ , a maximal number of fitness function evaluations, or a maximal number of consecutive unsuccessful function evaluations. It is described in [41] that the mesh size parameter is a measure of first-order stationary for GPS in the unconstrained case. The method ends, as soon as one termination criterion is attained. Otherwise, it returns to step 2.

## 5 Numerical Results and Performance Evaluation

In this study, MADS is utilized to solve the nonlinear optimization problems. Two approaches were considered for the implementation of the MADS algorithm. First, the MADS algorithm itself without any search method (MADS-N) was applied to the problems. Thereafter, CMAES was utilized as a strong search strategy in the MADS method (MADS-CMAES) to enhance its effectiveness. MADS-N and MADS-CMAES were applied on the standard IEEE 30-bus 6-generator test system to determine their effectiveness. The topology of the IEEE 30-bus test is given in Fig. 2 and the detailed information is provided in [20]. The network includes 48 branches, six generator-buses, and 22 load-buses. Four branches, 6–9, 6–10, 4–12 and 27–28, are under load tap setting transformer branches. The system has 6 generators at buses 1, 2, 5, 8, 11, and 13. The others are load-buses. The lower and upper limits of the transformer tapings are respectively 0.9 and 1.1 pu. The capacitor banks are connected to buses 10 and 24 respectively and fixed at 19.0 and 4.3 MVar. All bus voltages are needed to be maintained within the range of 0.95–1.1 pu. The initial values of the control variables and the fitness functions are provided in Table 2. In the test system, the base MVA is taken as 100.

**Table 1** Parameter settings for the MADS-N and MADS-CMAES algorithms

Parameters	Setting
<i>Termination parameters</i>	
Mesh size tolerance	$10^{-4}$
Maximum number of iterations	1000
Maximum number of function evaluations	1000
Maximum number of consecutive POLL failures	50
<i>Mesh parameters</i>	
Initial mesh size	1
Mesh refinement factor	0.5
Mesh coarsening factor	1
Cache tolerance	$10^{-4}$
<i>CMAES parameters</i>	
Population size ( $PS$ )*	$4 + \lfloor 3 * \log(N) \rfloor$
Parents	$\lfloor PS/2 \rfloor$
Recombination weighting	Superlinear

\* $N$  Number of objective variables/problem dimension

For the MADS analysis, the POLL directions can be chosen either MADS Positive basis  $N + 1$  or  $2N$  points, where  $N$  is the number of independent variables for the fitness function. At each iteration, the MADS Positive basis  $2N$  directions explore more points around the current point. In order to avoid finding a local minimum rather than the global minimum, the Positive basis  $2N$  directions are considered as the POLL technique in the present study. Consecutive order is the natural order that the points are stored in. Consecutive polling is also considered to solve the problems. Table 1 shows the other parameter settings for the MADS-N and MADS-CMAES algorithms. The MADS algorithm was implemented on a Dell Inspiron 6400 with a 2.00 GHz processor and 2 GB of RAM using Nomadm optimization software [42].

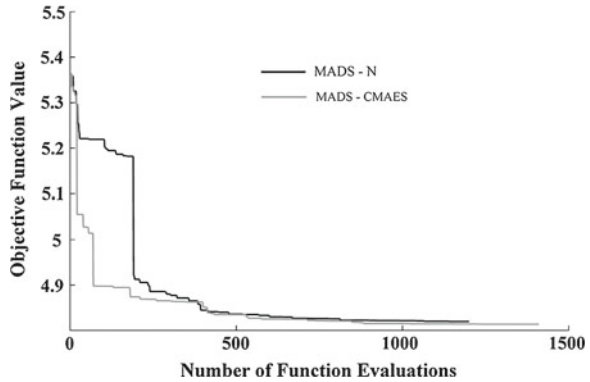
In this case, the MADS-N algorithm converges after 1189 function evaluations and in 138 iterations and the final minimum value of power losses is 4.8128 MW. The MADS-CMAES algorithm converges after 1420 function evaluations and in 150 iterations with the final minimum value of power losses equal to 4.8119 MW. The CLONEPAC-PSO [21] algorithm converges in 13 iterations and the final minimum value of power losses is 5.0949 MW. The PSOPC [43] converges in 48 iterations and its final optimum value is also the global best of 5.0960 MW. The EA [17] converges in about 70 iterations and its optimum value is 5.1065 MW. Table 2 summarizes the results of the optimal settings obtained by different methods. This table also shows the optimal settings and initial settings of decision variables for the same case study as proposed by PSOPC [43] and EA [20]. These results show that the minimum dispatch solutions found out by MADS-CMAES lead to lower active power loss than that found by other techniques. These results display that maximum saving is also

**Table 2** Settings of variables and real power losses in reactive power control of IEEE 30-bus test system

Decision variable/objective	Initial settings	CLONEPAC-PSO	PSOPC	EA	MADS-N	MADS-CMAES
$V_{G1}$	1.0500	0.9701	0.9900	1.0500	1.0394	1.0391
$V_{G2}$	1.0450	1.0175	1.0274	1.0440	1.0330	1.0318
$V_{G5}$	1.0100	1.0284	1.0500	1.0240	1.0168	1.0168
$V_{G8}$	1.0100	1.0361	1.0500	1.0260	1.0183	1.0180
$V_{G11}$	1.0500	0.9944	0.9500	1.0930	1.0400	1.0469
$V_{G13}$	1.0500	1.0445	1.0500	1.0850	1.0343	1.0347
$T_{6-9}$	0.9780	1.0471	1.1000	1.0780	0.9785	1.0200
$T_{6-10}$	0.9690	0.9910	0.9895	0.9060	1.0461	0.9900
$T_{4-12}$	0.9320	0.9992	1.0008	1.0070	1.0122	1.0060
$T_{27-28}$	0.9680	1.0091	1.0115	0.9590	1.0118	1.0131
$Q_{e1}(MVAr)(fixed)$	19	19	19	19	19	19
$Q_{e2}(MVAr)(fixed)$	4.3	4.3	4.3	4.3	4.3	4.3
$P_{Loss}$	5.3786	5.0949	5.0960	5.1065	4.8198	4.8121



**Fig. 1** Convergence of MADS-N and MADS-CMAES in reactive power control



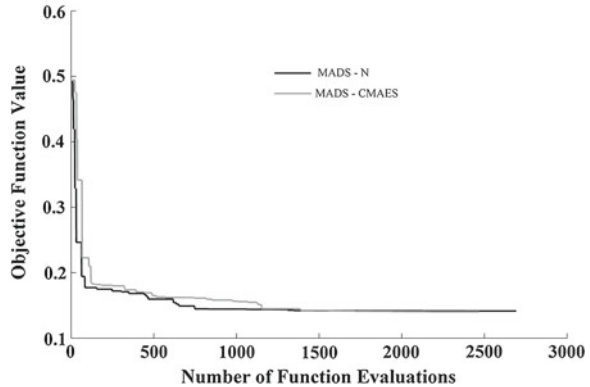
acquired by the MADS-CMAES algorithm. According to the results, the MADS-N algorithm has provided reasonable outcomes. The comparisons confirm that MADS is capable of locating the near-global or global optimum dispatch solution. The proposed algorithms simultaneously succeed in maintaining the dependent variables within their limits. It is notable that the MADS-N and MADS-CMAES results were acquired with the default parameters. While most of the previously presented optimization algorithms are significantly sensitive to their parameter settings changes, the proposed MADS-based algorithms are very robust against their parameters and changing the search space. Convergence nature of the MADS-N and MADS-CMAES algorithms are shown in Fig. 1.

The feasibility of MADS in voltage control of IEEE 30-bus system is also studied. In this case, the MADS-N converges after 2715 function evaluations and 205 iterations achieving the voltage deviation (VD) of 0.1397 pu. The MADS-CMAES converges after 2480 function evaluations and 172 iterations achieving the lowest sum of VD of 0.1389 pu. CLONEPAC-PSO [21] converges in 49 iterations, and its final optimum value of VD is 0.1400 pu. The PSOPC [43] converges in 27 iterations and its minimum fitness value is 0.1410 pu. The EA [20] converges in about 110 iterations and its optimum fitness value is 0.1477 pu. Table 3 presents the best solution of this problem acquired utilizing the MADS method and those given by other researchers. It can be seen from Table 3 that the result obtained using the MADS-CMAES algorithm is better than the best known solution reported previously in the literature [21]. This table shows the initial settings of decision variables for comparison purposes. Convergence nature of the MADS-N and MADS-CMAES algorithms are shown in Fig. 2.

**Table 3** Settings of variables and voltage deviations in voltage control of IEEE 30-bus test system

Decision variable/objective	Initial settings	CLONEPAC-PSO	PSOPC	EA	MADS-N	MADS-CMAES
$V_{G1}$	1.0500	0.9966	0.9895	1.037	1.0104	1.0074
$V_{G2}$	1.0450	0.9969	0.9883	1.027	1.0065	1.0023
$V_{G5}$	1.0100	1.0363	1.05	1.013	1.0184	1.0152
$V_{G8}$	1.0100	1.0287	1.0372	1.008	1.0091	1.0105
$V_{G11}$	1.0500	1.0046	1.0176	1.03	1.0191	1.0467
$V_{G13}$	1.0500	1.0149	1.0189	1.007	1.0288	1.0305
$T_{6-9}$	0.9780	1.0375	1.036	1.054	1.0413	1.0747
$T_{6-10}$	0.9690	1.0111	1.0167	0.907	0.9053	0.9034
$T_{4-12}$	0.9320	0.9985	0.9982	0.928	0.9610	0.9719
$T_{27-28}$	0.9680	1.0154	1.0172	0.945	1.0650	1.0628
$Q_{e1}(MVAr)(fixed)$	19	19	19	19	19	19
$Q_{e2}(MVAr)(fixed)$	4.3	4.3	4.3	4.3	4.3	4.3
$VD$	0.4993	0.14	0.141	0.1477	0.1397	0.1389

**Fig. 2** Convergence of MADS-N and MADS-CMAES in voltage control



## 6 Conclusion

This paper has studied a new optimization technique called MADS for the optimal steady-state performance of power systems. The MADS-based techniques were applied to the voltage control and reactive power problems of IEEE 30-bus systems. The CMAES algorithm is employed as an effective search strategy in MADS to improve its performance. The results show that the MADS-based algorithms are practical and valid for the investigated problems. The MADS-CMAES and MADS-N algorithms also provide superior compared with the other techniques reported in the literature. The presented numerical results were obtained with the standard default algorithmic parameters. Unlike many of the metaheuristic optimization algorithms that are significantly sensitive to their parameter settings changes, the proposed MADS-based algorithms are very robust against their parameters and changing the search space due to their inherent adaptive tuning. The comparisons confirm the efficiency of MADS for its future applications to the real problems with non-convex decision and space more hard constraints.

## References

1. Dommel HW, Tinney WF (1968) Optimal power flow solutions. *IEEE Trans Power Ap Syst* 87(10):1866–1876
2. Mamundur KRC, Chenoweth RD (1981) Optimal control of reactive power flow for improvements in voltage profiles and for real power loss minimization. *IEEE Trans Power Ap Syst* 100:3185–3194
3. Sun DI, Ashley B, Brewar B, Hughes A, Tinny WF (1984) Optimal power flow by Newton approach. *IEEE Trans Power Ap Syst* 103:2864–2880
4. Alsac O, Bright J, Prais M, Scott B, Marinho JL (1990) Further developments in LP based optimal power flow. *IEEE Trans Power Syst* 5:697–711
5. Lu FC, Hsu YY (1995) Reactive power/voltage control in a distribution substation using dynamic programming. *IEEE Proc Gener Transm Distrib* 142:639–645

6. Quintana VH, Santos-Nieto M (1989) Reactive-power dispatch by successive quadratic programming. *IEEE Trans Energy Convers* 4:425–435
7. Granville S (1994) Optimal reactive power dispatch through interior point methods. *IEEE Trans Power Sys* 9:136–146
8. Subbaraj P, Rajnarayanan PN (2009) Optimal reactive power dispatch using self-adaptive real coded genetic algorithm. *Electr Power Syst Res* 79(2):374–381
9. Hooke R, Jeeves TA (1961) Direct search solution of numerical and statistical problems. *J Assoc Comput Mach* 8:212–229
10. Wang GG, Gandomi AH, Alavi AH, An effective krill herd algorithm with migration operator in biogeography-based optimization. *Appl Math Model* (in Press)
11. Yang XS, Gandomi AH, Sadat Hosseini SS (2012) firefly algorithm for solving non-convex economic dispatch problems with valve loading effect. *Appl Soft Comput, Elsevier* 12(3):1180–1186
12. Sadat Hosseini SS, Yang XS, Gandomi AH, Nemati A (2015) Solutions of non-smooth economic dispatch problems by swarm intelligence. *Adapt Hybr Comput Intell* 18:129–146
13. Doganis P, Sarimveis H, Optimization of power production through coordinated use of hydro-electric and conventional power units. *Appl Math Model* (in Press)
14. Lewis RM, Torczon V, Trosset MW (2000) Direct search methods: then and now. *J Comput Appl Math* 124:191–207
15. Dennis JE Jr, Torczon V (1991) Direct search methods on parallel machines. *SIAM J Optim* 1(4):448–474
16. Audet C, Dennis JE Jr (2006) Mesh adaptive direct search algorithms for constrained optimization. *SIAM J Optim* 17(1):188–217
17. Zakerifar R, Sadat Hosseini SS, Jafarnejad A (2011) Application of mesh adaptive direct search method to power system valve-point. In: *Proceedings of the international conference on energy and electrical systems, Kuala Lumpur, Malaysia, ASME Digital Library* (in press)
18. Audet C, Bechard V, Chaouki J (2008) Spent potliner treatment process optimization using a MADS algorithm. *Optim Eng* 9:143–160
19. Sadat Hosseini SS, Jafarnejad A, Behrooz AH, Gandomi AH (2011) Combined heat and power economic dispatch by mesh adaptive direct search algorithm. *Expert Syst Appl* 38(6):6556–6564
20. Abido MA, Bakhshwain JM (2005) Optimal VAR dispatch using a multiobjective evolutionary algorithm. *Int J Electr Power* 27(1):13–20
21. Vlachogiannis JG (2006) Constricted local neighborhood particle swarm optimization with passive congregation for reactive power and voltage control. *Electr Pow Compon Syst* 34(6):509–520
22. Kolda TG, Lewis RM, Torczon V (2004) Optimization by direct search: new perspectives on some classical and modern methods. *SIAM Rev* 45(3):385–482
23. Lewis RM, Torczon V (2000) Pattern search algorithms for linearly constrained minimization. *SIAM J Optim* 10(3):917–941
24. Audet C (2004) Convergence results for pattern search algorithms are tight. *Optim Eng* 5:101–122
25. Abramson MA, Audet C, Dennis JE Jr (2004) Generalized pattern searches with derivative information. *Math Program* 100:3–25
26. Audet C, Orban D (2006) Finding optimal algorithmic parameters using the mesh adaptive direct search algorithm. *SIAM J Optim* 17(3):642–664
27. Booker AJ, Dennis JE Jr, Frank PD, Serafini DB, Torczon V, Trosset MW (1999) A rigorous framework for optimization of expensive functions by surrogates. *Struct Optim* 17:1–13
28. McKay MD, Conover WJ, Beckman RJ (1979) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 21:239–245
29. Holland JH (1975) *Adaptation in natural and artificial systems*. The University of Michigan Press, Ann Arbor

30. Fogel LJ, Owens AJ, Walsh MJ (1966) Artificial intelligence through simulated evolution. Wiley, New York
31. Rechenberg I (1973) Evolutionsstrategie: optimierung technischer systeme nach principen der biologischen evolution. Fromman-Holzboog, Stuttgart
32. Schwefel HP (1995) Evolution and optimum seeking. Wiley-Interscience, New York
33. Beyer HG, Schwefel HP (2002) Evolution strategies: a comprehensive introduction. *Nat Comput* 1:3–52
34. Dirk VA, Alexander M (2006) Hierarchically organized evolution strategies on the parabolic ridge. In: GECCO'06, Seattle, WA, USA
35. Hansen N, Ostermeier A (2001) Completely derandomized self-adaptation in evolution strategies. *Evol Comput* 9(2):159–195
36. Hansen N, Kern S (2004) Evaluating the CMA evolution strategy on multimodal test functions. In: *Parallel problem solving from nature*, Springer
37. Hansen N, Ostermeier A (1996) Adapting arbitrary normal mutation distributions in evolution strategies: the covariance matrix adaptation. In: *Proceedings of the 1996 IEEE conference on evolutionary computation (ICEC '96)*, pp 312–317
38. Hansen N, Müller SD, Koumoutsakos P (2003) Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evol Comput* 11(1):1–18
39. Hansen N, Ostermeier A (1997) Convergence properties of evolution strategies with the derandomized covariance matrix adaptation: the CMA-ES. In: *Proceedings of the 5th European congress on intelligent techniques and soft computing*, pp 650–654
40. [www.icos.ethz.ch/software/evolutionarycomputation/cmaapplications.pdf](http://www.icos.ethz.ch/software/evolutionarycomputation/cmaapplications.pdf)
41. Dolan ED, Lewis RM, Torczon V (2003) On the local convergence properties of pattern search. *SIAM J Optim* 14(2):567–583
42. <http://www.gerad.ca/NOMAD/Abramson/nomad.html>
43. He S, Wu QH, Wen JY, Saunders JR, Patton PC (2004) A particle swarm optimizer with passive congregation. *Biosystems* 78:135–147

# Optimal Placement of Hysteretic Dampers via Adaptive Sensitivity-Smoothing Algorithm

Yu Murakami, Katsuya Noshi, Kohei Fujita, Masaaki Tsuji and Izuru Takewaki

**Abstract** Since hysteretic dampers have nonlinear restoring-force characteristics with sensitive plastic flow and input earthquake ground motions propagating random media are extremely random in time and frequency domains, the seismic response of a building structure with hysteretic dampers deviates greatly depending on the installed quantity and location of dampers. This characteristic could become a barrier and difficulty to the reliable formulation of optimal placement problems of such dampers. In order to overcome such difficulty, a new optimization method including a variable adaptive step length is proposed. The proposed method to solve the optimum design problem is a successive procedure which consists of two steps. The first step is a sensitivity analysis by using nonlinear time-history response analyses, and the second step is a modification of the set of damper quantities based upon the sensitivity analysis. Numerical examples are presented to demonstrate the effectiveness and validity of the proposed design method.

## 1 Introduction

The concept of performance-based design is becoming popular worldwide and plays a key role in the current structural design practice of buildings. In earthquake-prone countries, the philosophy of earthquake-resistant design to resist ground shaking with sufficient stiffness and strength of a building itself has also been accepted as a relevant structural design concept for many years. On the other hand, a new strategy based on the concept of active and passive structural control including base-isolation has been introduced rather recently in order to provide structural designers with powerful tools for performance-based design.

While active control has some issues to be resolved from the viewpoint of reliability, feasibility and cost during severe earthquake ground motions, passive control is being widely accepted and used for building structures under earthquake

---

Y. Murakami · K. Noshi · K. Fujita · M. Tsuji · I. Takewaki (✉)  
Department of Architecture and Architectural Engineering, Kyoto University,  
Kyotodaigaku-Katsura, Nishikyo-ku, Kyoto 615-8540, Japan  
e-mail: takewaki@archi.kyoto-u.ac.jp

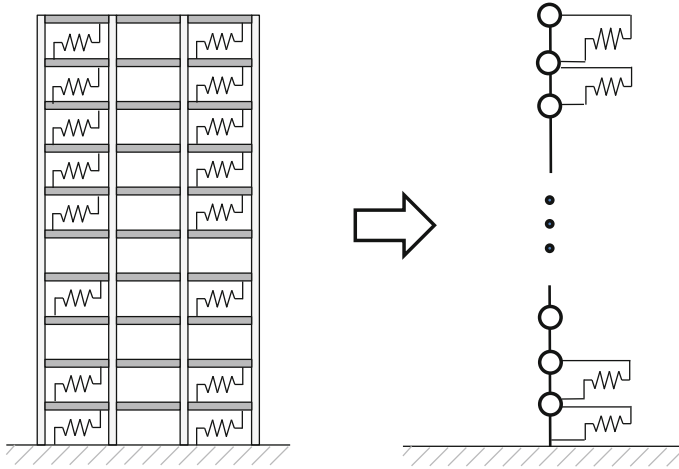
ground motions [9, 15, 17, 29, 32, 35]. Hysteretic steel dampers (shear deformation type, buckling restrained type), viscous wall-type dampers, viscous oil dampers [40], visco-elastic dampers, friction dampers, tuned mass dampers, inertial mass dampers [37] are representative ones. Recently viscous oil dampers (called oil dampers hereafter) are often used based on their stable mechanical properties, low frequency and temperature dependencies and cost effectiveness, etc. together with low cost hysteretic steel dampers. Compared to oil dampers, hysteretic steel dampers suit the strength-type performance check and are often preferred in the retrofit of buildings. It should be emphasized that, during the 2011 Tohoku (Japan) earthquake, the Osaka WTC building of 256 (m) high was shaken so hard irrespective of its long distance (800km) from the epicenter [36]. It is said that this results from the resonance of the building with the so-called long-period ground motion [33, 36]. To respond to this unfavorable situation, the retrofit of this building is being conducted with oil dampers and hysteretic steel dampers. It should be remembered that the oil dampers and inertial mass dampers do not change the natural period of a building which may cause a resonance with the long-period ground motion stated above. On the other hand, the hysteretic steel dampers can change the natural period of a building by yielding even in the early vibration process.

Many research works have been accumulated so far on the damper optimization [5, 6, 8, 10–12, 14, 16, 17, 20, 22, 24, 27, 28, 30, 31, 34, 38, 39, 41, 43]. While most of them deal with linear responses, quite a few treat non-linear responses in building structures or dampers [1, 2, 5, 13, 17–19, 21, 23, 25, 26, 41, 42]. However, there is no research on the optimization of location and quantity of dampers which deals with non-linear responses and includes simple and systematic algorithms.

The purpose of this paper is to propose a new optimization method including a variable adaptive step length for shear buildings with hysteretic dampers subjected to a set of design earthquake ground motions under a constraint on total cost. The response sensitivity of buildings including hysteretic dampers is high and a devised algorithm of adaptive step-length is useful to obtain a smooth and reliable response sensitivity. The high response sensitivity of buildings including hysteretic dampers may result from the timing of fast plastic flow and random process of input and the change of the natural period of a building depending on the installed quantity and location of hysteretic dampers. The proposed procedure enables structural designers to derive a series of optimal distribution of damper quantities with respect to the level of the total cost of dampers which is useful in seeking for the relation between the optimal response level and the quantity and placement of passive dampers. Numerical examples reveal some features of the optimal distribution of various passive dampers.

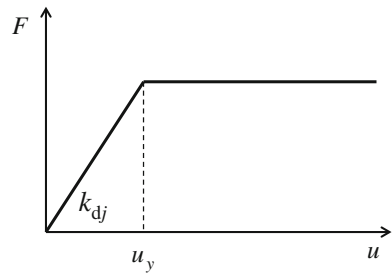
## 2 Optimal Hysteretic Damper Placement Problem

Consider an  $N$ -story shear building model with interstory-type hysteretic steel dampers as shown in Fig.1. A stiffness proportional viscous damping is employed here in the main frame (damping ratio = 0.02).



**Fig. 1** *N*-story planar frame with hysteretic steel dampers and its modeling into shear building model with hysteretic springs

**Fig. 2** Force-deformation relation of hysteretic damper



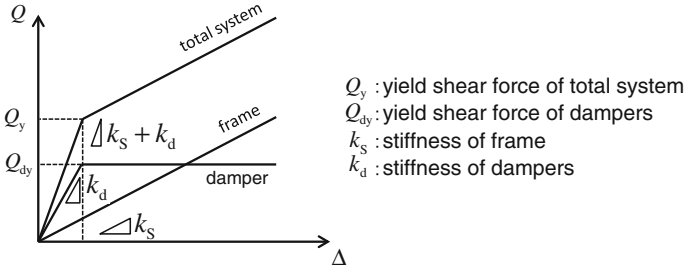
### 2.1 Modeling of Hysteretic Dampers

Steel hysteretic dampers are used in this paper. The initial stiffness  $k_{dj}$  and the yield displacement  $u_y$  are the major parameters to characterize the present steel hysteretic dampers. An elastic-perfectly plastic restoring force characteristic as shown in Fig. 2 is assumed. Figure 3 shows the story shear force with respect to interstory drift in which the part of the total system and those of frame and damper systems are illustrated. An example of hysteresis loop in the story shear force-interstory drift relation under an earthquake ground motion is presented in Fig. 4.

### 2.2 Design Earthquake Ground Motions and Envelope Response

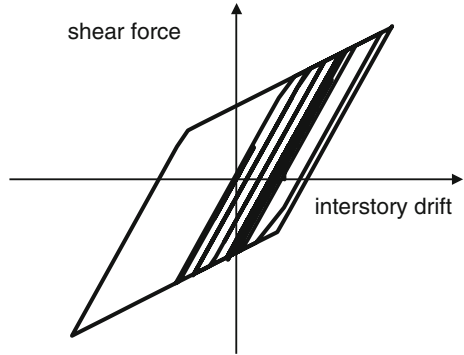
Two representative recorded ground motions, i.e. El Centro NS 1940 (maximum velocity = 0.5m/s; impulsive type ground motion) and Hachinohe NS 1968



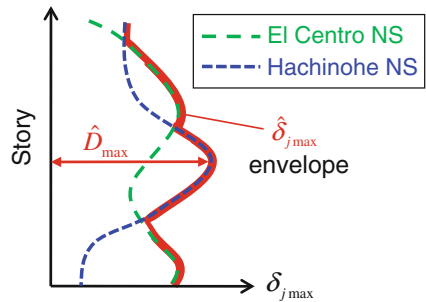


**Fig. 3** Story shear force with respect to interstory drift

**Fig. 4** Example of hysteresis loop in story shear force-interstory drift relation



**Fig. 5** Envelope response for demand



(maximum velocity = 0.5 m/s; slightly long-period motion), are employed as the design earthquake ground motions. The maximum value  $\hat{D}_{max}$  in the envelope response  $\hat{\Delta}_{jmax}$  of the maximum interstory drift for multiple candidate ground motions as shown in Fig. 5 is used for the demand in this paper. Although an example for two ground motions is presented here, this is applicable to a more general case for multiple ground motions without difficulty.

### 2.3 Optimal Damper Placement Problem

The design problem of hysteretic dampers may be stated as follows.

[Problem] Find  $\mathbf{k}_d = \{k_{dj}\}$  so as to minimize the selected seismic response  $F$  subject to

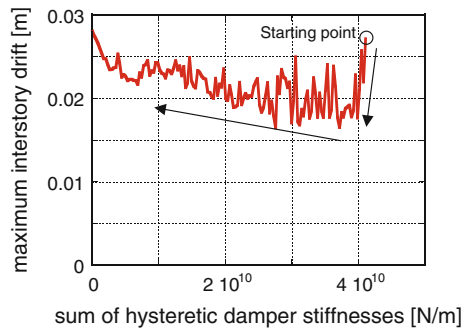
$$\sum_{j=1}^N k_{dj} = \bar{C}_d \tag{1}$$

In this problem,  $\bar{C}_d$  is the specified sum of stiffnesses of hysteretic dampers. It can be shown after some examination that the initial stiffness of hysteretic dampers is directly related to the quantity (and cost) of hysteretic dampers irrespective of its installation type (axial-type or shear type).  $\hat{D}_{max}$  is employed here as  $F$ . For simplicity of expression,  $\hat{D}_{max}$  is expressed simply as  $D_{max}$  later.

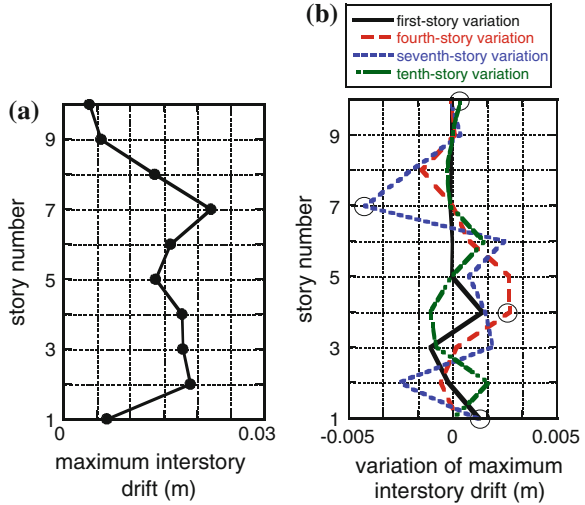
Since hysteretic dampers have nonlinear restoring-force characteristics with sudden, large stiffness change and input earthquake ground motions are random (because of propagation in random media), the seismic response of a building with hysteretic dampers deviates greatly depending on the installed quantity and location of dampers. The timing of fast plastic flow and random process of input may be the main reason of the response randomness. This characteristic disturbs a reliable formulation of the optimal damper placement different from other dampers [1, 2, 32, 35].

Figure 6 shows an example of variation of the maximum interstory drift with respect to the sum of hysteretic damper stiffnesses. The initial design of the hysteretic damper stiffness is proportional to the main frame stiffness and the stiffness ratio to the main frame is 5. In Fig. 6, the hysteretic damper stiffnesses (distribution with respect to height) have been changed keeping the hysteretic damper stiffness proportional to the main frame stiffness. The main causes of response irregularity may be (i) irregularity of ground motions, (ii) sudden change of stiffness due to yielding, (iii) irregularity of maximum response (change of story number, time of the maximum response and direction of the maximum interstory drift).

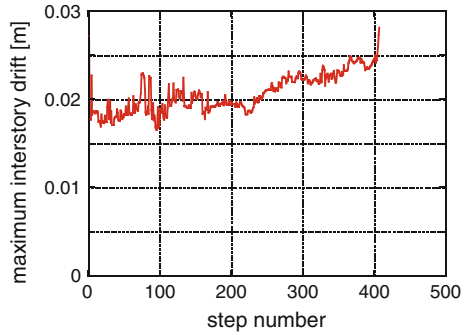
**Fig. 6** Maximum interstory drift with respect to sum of hysteretic damper stiffness



**Fig. 7** Maximum interstory drift and variation of maximum interstory drift to change (decrease) of damper stiffness in story marked by circle



**Fig. 8** Maximum interstory drift with respect to step number



In order to overcome such difficulty, a new optimization method including a variable adaptive step length for sensitivity smoothing is proposed. Although a constraint on accumulated plastic deformation ratio is sometimes required in hysteretic dampers for long-duration earthquake ground motions [3, 4, 7, 36], this is not taken into account here because of a simple, essential presentation of a new optimization procedure.

Figure 7 shows the maximum interstory drift and the variation of the maximum interstory drift to the change (decrease) of damper stiffness in the story marked by circle. The first, fourth, seventh and tenth-story damper stiffnesses have been varied. It can be observed that it is difficult to predict the variation of the maximum interstory drifts from the story number with the stiffness variation.

Figure 8 illustrates the maximum interstory drift with respect to step number. It can also be confirmed that the seismic response of a building with hysteretic dampers deviates greatly depending on the installed quantity and location of dampers.

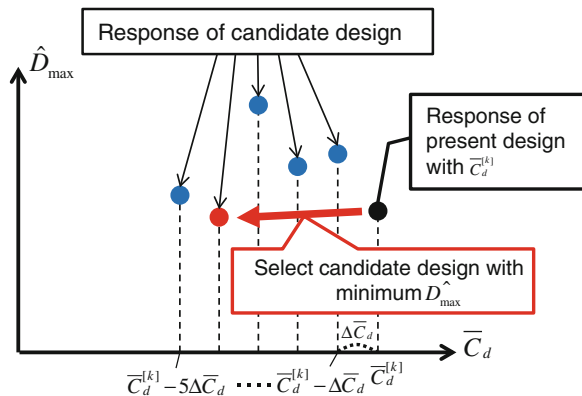
### 2.4 Optimization Algorithm Including Variable Adaptive Step Length

Figure 9 shows a schematic diagram of the proposed sensitivity evaluation algorithm including variable adaptive step length. The response of candidate design in Fig. 9 is obtained by minimizing the maximum interstory drift for variation of damper stiffness in respective story. This procedure is aimed at finding the most inactive damper and reducing the quantity of such damper. Among several candidates of the decreased hysteretic damper cost, the decreased hysteretic damper cost attaining the lowest value of the maximum interstory drift is employed as the next-step sensitivity (also next-step design). Figure 10 presents the flowchart of hysteretic damper optimization-1. Although the minimum value is used in this example, the maximum value (with respect to several step lengths) among the candidate designs for the maximum interstory drift can be employed alternatively in consideration of the safety level of the passively controlled buildings (worst-case scenario). An example using this maximum value of the maximum interstory drifts will be shown later. The average value of the maximum interstory drift may be another possibility.

A practical procedure for optimal oil damper design without laborious mathematical programming techniques has been proposed for reducing the computational load [2]. A similar procedure can be developed for hysteretic dampers. There are two practical aspects: (1) use of a reduced model (static condensation) from a frame model for computational efficiency, (2) search of a series of optimal damper distribution for different total damper quantities. Although a shear building model is used here, the reduced model (static condensation) developed by [2] can be used if desired. Figure 11 illustrates the conceptual approximate solution procedure. The design algorithm may be summarized as follows:

- Step 1 The along-height sum of hysteretic damper stiffnesses is determined (as the stiffness ratio to the main frame stiffness).

**Fig. 9** Sensitivity evaluation algorithm including variable adaptive step length-1 (Selection of candidate design with minimum among minimums for respective reduced damper levels)



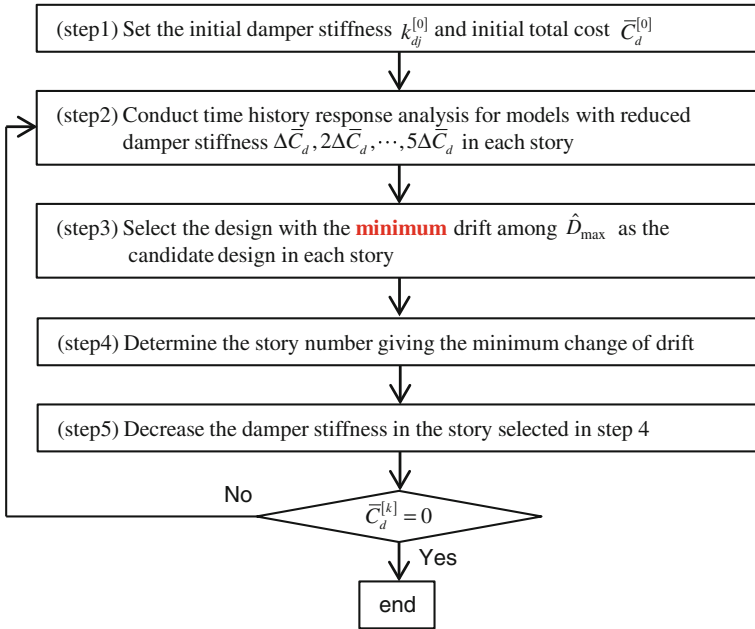


Fig. 10 Flowchart of hysteretic damper optimization-1

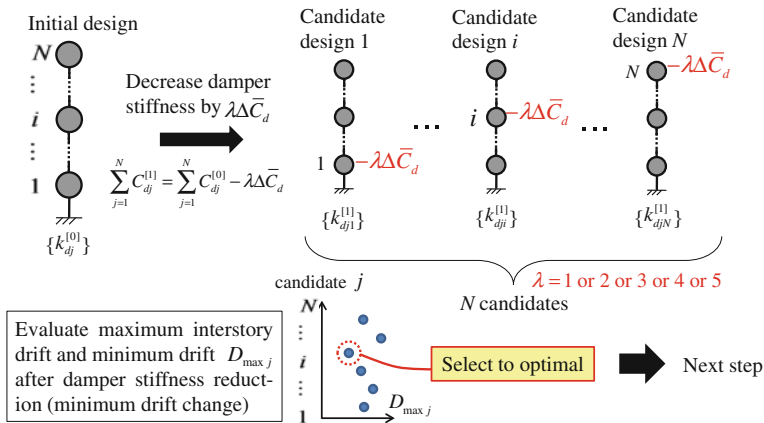


Fig. 11 Conceptual diagram of hysteretic damper optimization

Step 2 Produce  $N \times 5$  candidates in which damper stiffnesses  $\Delta C_d, 2\Delta C_d, 3\Delta C_d, 4\Delta C_d, 5\Delta C_d$  are reduced from the present hysteretic damper stiffness in each story. Compute the objective function for each model constructed in Step 2 through nonlinear time-history response analysis.

- Step 3 Select the design with the minimum drift as the candidate design in each story.
- Step 4 Select the best candidate with the minimum objective function (drift change) from the candidates produced in Step 3.
- Step 5 Decrease the damper stiffness in the story selected in Step 4. Then go to Step 2.

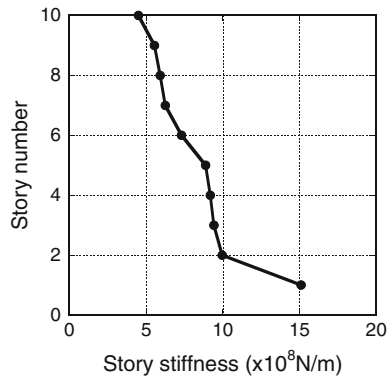
### 3 Numerical Examples

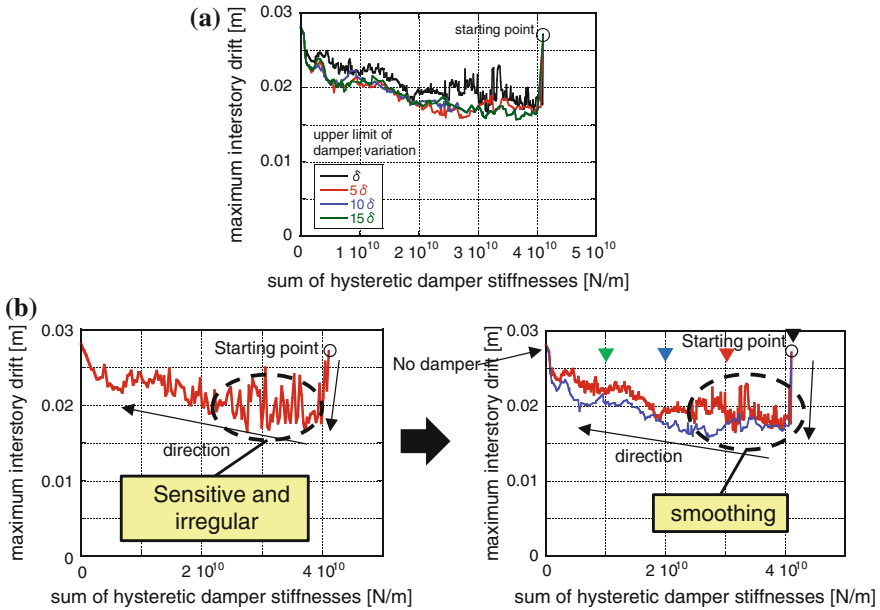
The main structure has been designed so that it has a fundamental natural period = 1.05 (s) and a realistic stiffness distribution as shown in Fig. 12. The constant mass is  $1.0 \times 10^6$  kg which corresponds approximately to 30 m  $\times$  30 m floor plan and the structural damping ratio (stiffness-proportional viscous damping) is assumed to be 0.02. The yield displacement of hysteretic dampers is 0.005 m and the stiffness ratio of hysteretic dampers to the main frame stiffness in the initial design is 5.

#### 3.1 Example 1 (Employment of Minimum Value Among Maximum Interstory Drifts in Algorithm of Variable Adaptive Step Length)

An example using the algorithm explained in Sect. 2.4 is presented here. Figure 13a shows the plot of the maximum interstory drift with respect to the sum of hysteretic damper stiffnesses. The sum of hysteretic damper stiffnesses is decreased gradually. Figure 13a indicates clearly the effect of upper limit of damper variation. The smoothing process due to change of upper limit of damper variation is illustrated in Fig. 13b.

**Fig. 12** Story stiffness of ten-story main frame





**Fig. 13** a Effect of upper limit of damper variation on damper optimization. b Smoothing process due to change of upper limit of damper variation

Figure 14a illustrates the distribution of hysteretic damper stiffnesses and Fig. 14b shows the distributions of the maximum interstory drifts. It can be understood from Fig. 14a that the first-story damper is reduced fast in the increasing step. This may result from the fact that the interstory drift in the first story is smaller compared to other stories and the installation of hysteretic dampers in the first story is not effective in this example. It can also be observed from Fig. 14b that the maximum ductility factor of hysteretic dampers is about 4–5 in later steps. Figure 15 shows the characteristics of optimal damper variation. It should be noted that Fig. 15 is a little bit different from Fig. 14 because of the difference of pick-up points. The early removal of top-story damper and the reduction of the first-story damper can be observed.

### 3.2 Example 2 (Employment of Maximum Value Among Maximum Interstory Drifts in Algorithm of Variable Adaptive Step Length)

In place of the algorithm in Sect. 2.4, another one is employed here, i.e. the selection of the design with the maximum drift among minimum interstory drifts as the candidate design in each story in Step 3. Figure 16 presents the sensitivity evaluation

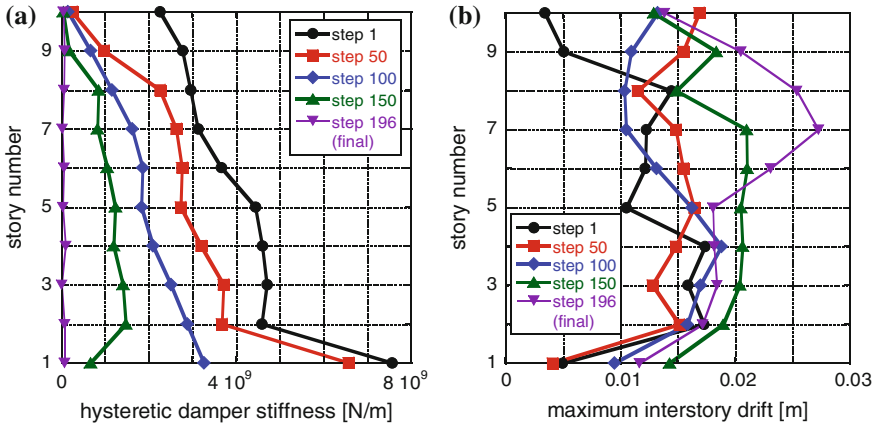


Fig. 14 Optimal design (Example 1: Minimum drift-sensitivity criterion). a Distribution of hysteretic damper stiffness, b Maximum interstory drift

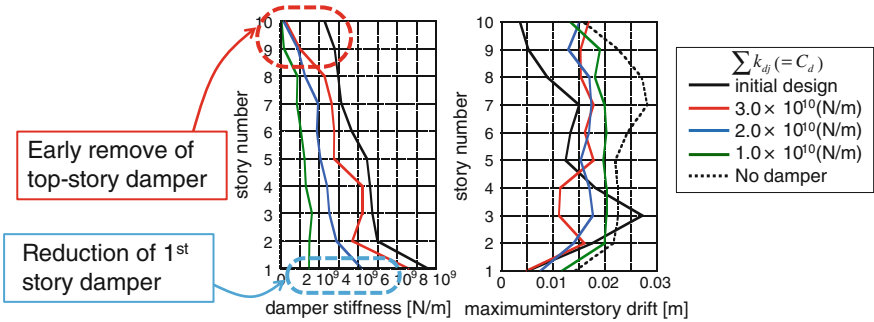
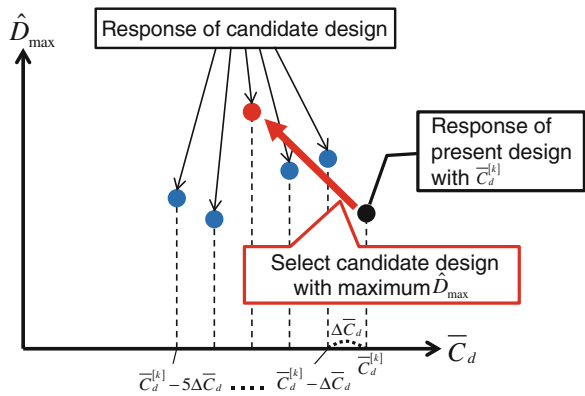


Fig. 15 Characteristics of optimal damper variation

Fig. 16 Sensitivity evaluation algorithm including variable adaptive step length-2 (Selection of candidate design with maximum among minimums for respective reduced damper levels)





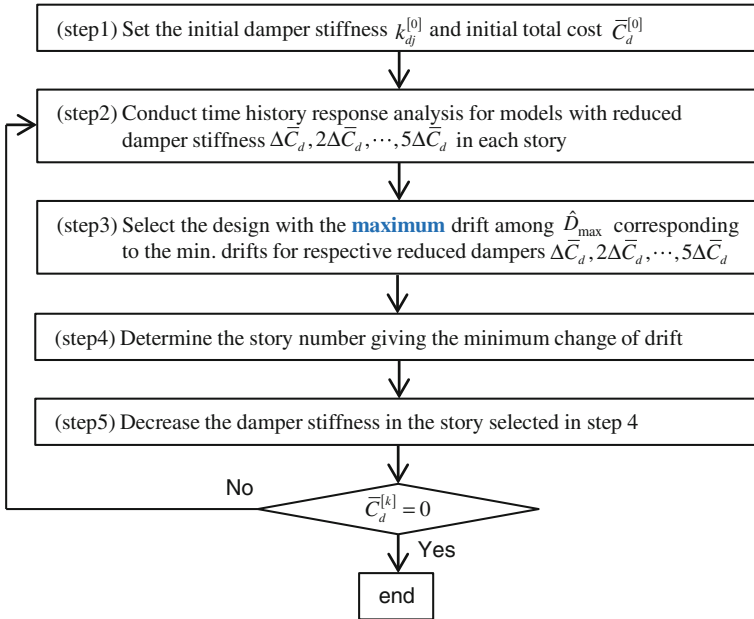
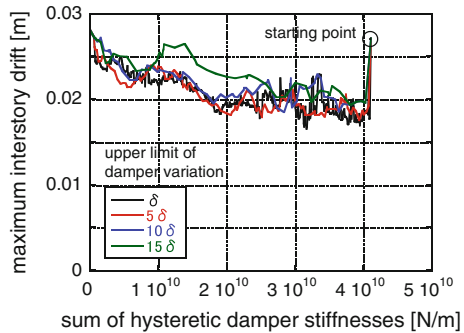


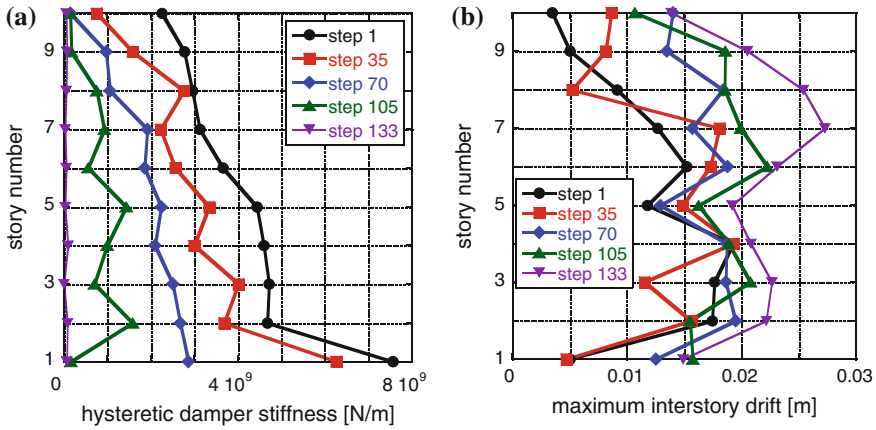
Fig. 17 Flowchart of hysteretic damper optimization-2

Fig. 18 Effect of upper limit of damper variation on damper optimization



algorithm including variable adaptive step length (Selection of candidate design with the maximum among minimum interstory drifts for design variation). This procedure can be conducted by changing ‘minimum’ to ‘maximum’ in Step 4 in Sect. 2.4. Figure 17 illustrates the flowchart of hysteretic damper optimization-2.

Figure 18 shows the plot of the maximum interstory drift with respect to the sum of hysteretic damper stiffnesses. The sum of hysteretic damper stiffnesses is decreased gradually. The plot for  $\delta$  is the same as that for  $\delta$  in Fig. 13a because the minimization or maximization procedure is not applied to the case for  $\delta$  (only one case for  $\delta$ ). Figure 18 indicates clearly the effect of upper limit of damper variation, i.e. smoothing of variation. While Fig. 19a illustrates the distribution of hysteretic



**Fig. 19** Optimal design (Example 2: Maximum drift-sensitivity criterion). **a** Distribution of hysteretic damper stiffness, **b** Maximum interstory drift

damper stiffnesses, Fig. 19b shows the distributions of the maximum interstory drifts. It can be observed from Fig. 19a that the maximum interstory drift distributions of the models obtained in this new algorithm are not different so much from those in Fig. 14a. Although Fig. 19b is also similar to Fig. 14b, a slight change can be observed in upper stories.

## 4 Conclusions

The following conclusions have been derived.

- (1) The proposed method for optimal placement of hysteretic dampers takes full advantage of a sensitivity-based redesign algorithm including nonlinear time-history response analysis in the optimization process. The method enables structural designers to find an optimal quantity and location of hysteretic dampers in each design step. The method is general and applicable to any type of passive dampers and any classes of design earthquake ground motions.
- (2) The response sensitivity of buildings including hysteretic dampers is high because of the timing of fast plastic flow and random process of input and the change of the natural period of a building depending on the installed quantity and location of hysteretic dampers. A devised algorithm of adaptive step-length in the response sensitivity computation is useful to obtain a smooth and reliable response sensitivity.
- (3) Employment of the *minimum or maximum* value of the maximum interstory drift can be used in the algorithm of variable adaptive step length. First select the design with the minimum drift as the candidate design in each story and select the

best candidate with the *minimum or maximum* objective function (drift change) from the candidates produced in this step. Both algorithms provide similar results on the optimal damper placement.

**Acknowledgments** Part of the present work is supported by the Grant-in-Aid for Scientific Research of Japan Society for the Promotion of Science (No. 24246095). This support is greatly appreciated.

## References

1. Adachi F, Fujita K, Tsuji M, Takewaki I (2013) Importance of interstory velocity on optimal along-height allocation of viscous oil dampers in super high-rise buildings. *Eng Struct* 56:489–500
2. Adachi F, Yoshitomi S, Tsuji M, Takewaki I (2013) Nonlinear optimal oil damper design in seismically controlled multi-story building frame. *Soil Dyn Earthq Eng* 44(1):1–13
3. Architectural Institute of Japan (AIJ) (2011) Preliminary reconnaissance report of the 2011 Tohoku-Chiho Taiheiyo-Oki, earthquake, July 2011 (in Japanese)
4. Architectural Institute of Japan (AIJ) (2012) Preliminary reconnaissance report of the 2011 Tohoku-Chiho Taiheiyo-Oki, earthquake, Springer
5. Attard TL (2007) Controlling all interstory displacements in highly nonlinear steel buildings using optimal viscous damping. *J Struct Eng ASCE* 133(9):1331–1340
6. Aydin E, Boduroglu MH, Guney D (2007) Optimal damper distribution for seismic rehabilitation of planar building structures. *Eng Struct* 29:176–185
7. Celebi M, Okawa I, Kashima T, Koyama S, Iiba M (2014) Response of a tall building far from the epicenter of the 11 March 2011 M9.0 Great East Japan earthquake and aftershocks. *Struct Des Tall Spec Build* 23:427–441
8. Cimellaro GP (2007) Simultaneous stiffness-damping optimization of structures with respect to acceleration, displacement and base shear. *Eng Struct* 29:2853–2870
9. de Silva CW (ed) (2007) *Vibration damping, control, and design*. CRC Press, Boca Raton
10. Fujita K, Takewaki I (2012) Robust passive damper design for building structures under uncertain structural parameter environments. *Earthq Struct* 3(6):805–820
11. Fujita K, Moustafa A, Takewaki I (2010) Optimal placement of viscoelastic dampers and supporting members under variable critical excitations. *Earthq Struct* 1(1):43–67
12. Fujita K, Yamamoto K, Takewaki I (2010) An evolutionary algorithm for optimal damper placement to minimize interstorey-drift transfer function in shear building. *Earthq Struct* 1(3):289–306
13. Fujita K, Kasagi M, Lang ZQ, Guo PF, Takewaki I (2014) Optimal placement and design of nonlinear dampers for building structures in the frequency domain. *Earthq Struct* (accepted for publication)
14. Garcia DL (2001) A simple method for the design of optimal damper configurations in MDOF structures. *Earthq Spectra* 17:387–398
15. Hanson RD, Soong TT (2001) *Seismic design with supplemental energy dissipation devices*. EERI, Oakland
16. Hwang JS, Lin WC, Wu NJ (2013) Comparison of distribution methods for viscous damping coefficients to buildings. *Struct Infrastruct Eng* 9(1):28–41
17. Lagaros ND, Plevris V, Mitropoulou CC (eds) (2012) *Design optimization of active and passive structural control systems*. IGI Global, Hershey
18. Lang ZQ, Guo PF, Takewaki I (2013) Output frequency response function based design of additional nonlinear viscous dampers for vibration control of multi-degree-of-freedom systems. *J Sound Vib* 332(19):4461–4481

19. Lavan O (2014) A methodology for the integrated seismic design of nonlinear buildings with supplemental damping. *Control Health Monit Struct* (published online)
20. Lavan O, Dargush GF (2009) Multi-objective evolutionary seismic design with passive energy dissipation systems. *J Earthq Eng* 13(6):758–790
21. Lavan O, Levy R (2005) Optimal design of supplemental viscous dampers for irregular shear-frames in the presence of yielding. *Earthq Eng Struct Dyn* 34(8):889–907
22. Lavan O, Levy R (2006) Optimal design of supplemental viscous dampers for linear framed structures. *Earthq Eng Struct Dyn* 35:337–356
23. Lavan O, Levy R (2010) Performance based optimal seismic retrofitting of yielding plane frames using added viscous damping. *Earthq Struct* 1(3):307–326
24. Liu W, Tong M, Lee G (2005) Optimization methodology for damper configuration based on building performance indices. *J Struct Eng ASCE* 131(11):1746–1756
25. Martinez CA, Curadelli O, Compagnoni ME (2014) Optimal placement of nonlinear hysteretic dampers on planar structures under seismic excitation. *Eng Struct* 65:89–98
26. Noshi K, Yoshitomi S, Tsuji M, Takewaki I (2013) Optimal nonlinear oil damper design in seismically controlled multi-story buildings for relief forces and damping coefficients. *J Struct Eng Archit Inst Jpn* 59B (in Japanese)
27. Silvestri S, Trombetti T (2007) Physical and numerical approaches for the optimal insertion of seismic viscous dampers in shear-type structures. *J Earthq Eng* 11:787–828
28. Singh MP, Moreschi LM (2001) Optimal seismic response control with dampers. *Earthq Eng Struct Dyn* 2001(30):553–572
29. Soong TT, Dargush GF (1997) *Passive energy dissipation systems in structural engineering*. Wiley, Chichester
30. Takewaki I (1997) Optimal damper placement for minimum transfer functions. *Earthq Eng Struct Dyn* 26:1113–1124
31. Takewaki I (2000) Optimal damper placement for planar building frames using transfer functions. *Struct Multidiscip Optim* 20(4):280–287
32. Takewaki I (2009) *Building control with passive dampers: optimal performance-based design for earthquakes*. Wiley, Chichester (Asia)
33. Takewaki I (2013) Smart system identification of super high-rise buildings using limited vibration data during the 2011 Tohoku earthquake. In: *Keynote lecture at ICEAS13 in ASEM13, 8–12 Sept. Jeju, Korea*, pp 118–145
34. Takewaki I, Yoshitomi S (1998) Effects of support stiffnesses on optimal damper placement for a planar building frame. *J Struct Des Tall Build* 7(4):323–336
35. Takewaki I, Fujita K, Yamamoto K, Takabatake H (2011) Smart passive damper control for greater building earthquake resilience in sustainable cities. *Sustain Cities Soc* 1(1):3–15
36. Takewaki I, Murakami S, Fujita K, Yoshitomi S, Tsuji M (2011) The 2011 off the Pacific coast of Tohoku earthquake and response of high-rise buildings under long-period ground motions. *Soil Dyn Earthq Eng* 31(11):1511–1528
37. Takewaki I, Murakami S, Yoshitomi S, Tsuji M (2012) Fundamental mechanism of earthquake response reduction in building structures with inertial dampers. *Struct Control Health Monit* 19(6):590–608
38. Trombetti T, Silvestri S (2004) Added viscous dampers in shear-type structures: the effectiveness of mass proportional damping. *J Earthq Eng* 8(2):275–313
39. Tsuji M, Nakamura T (1996) Optimum viscous dampers for stiffness design of shear buildings. *J Struct Des Tall Build* 5:217–234
40. Tsuji M, Tanaka H, Yoshitomi S, Takewaki I (2011) Model reduction method for buildings with viscous dampers under earthquake loading. *J Struct Constr Eng Archit Inst Jpn* 76(665):1281–1290 (in Japanese)
41. Uetani K, Tsuji M, Takewaki I (2003) Application of optimum design method to practical building frames with viscous dampers and hysteretic dampers. *Eng Struct* 25:579–592
42. Whittle JK, Williams MS, Karavasilis TL, Blakeborough A (2012) A comparison of viscous damper placement methods for improving seismic building design. *J Earthq Eng* 16(4):540–560
43. Zhang RH, Soong TT (1992) Seismic design of viscoelastic dampers for structural applications. *J Struct Eng ASCE* 118:1375–1392

# Design of Tuned Mass Dampers via Harmony Search for Different Optimization Objectives of Structures

Sinan Melih Nigdeli and Gebrail Bekdaş

**Abstract** In this chapter, an optimization methodology for tuning of tuned mass dampers (TMDs) on seismic structures was presented for two different objectives such as reducing the displacement of first story and absolute acceleration of top story of the structure. A metaheuristic method; harmony search (HS) was employed for optimization according to the time history analyses of structure under several earthquake excitations. Harmony search inspires musical performances in order to find optimum design variables according to optimization objective. Step by step, the methodology of the optimization process is explained in the chapter. The method was applied to find an optimum TMD for a seven story shear building and the optimum results were compared for the two cases considering displacement objective and acceleration objective. According to the results, optimum TMDs for both objectives are effective on both displacements and accelerations. But for acceleration objective, a small benefit for accelerations can be seen although the optimum mass of TMD is very heavy according to displacement objective.

**Keywords** Tuned mass dampers · Earthquake · Harmony search · Optimization · Time domain analyses · Metaheuristic methods

## 1 Introduction

Tuned mass dampers (TMDs) are vibration absorber devices used in all types of mechanic systems. The first type of this device was a mass connected with springs to the main system needs to be stabilized. The invention of this vibration absorber device was done by Frahm [1]. Since this device designed without inherent damping, absorbing of vibrations resulting from random excitations was not possible. Ormon-

---

S.M. Nigdeli (✉) · G. Bekdaş  
Department of Civil Engineering, Istanbul University, 34320 Istanbul, Avcılar, Turkey  
e-mail: melihnig@istanbul.edu.tr

G. Bekdaş  
e-mail: bekdas@istanbul.edu.tr

**Fig. 1** Berlin TV tower

droyd and Den Hartog implemented dampers to the absorber device [2]. Thus, the new form of the absorber device was regarded as tuned mass damper (TMD) and it is also effective on structures subjected to earthquake and wind loads with changing frequency, but an effective tuning of TMDs must be done for the best performance. Optimization methods are an important issue for the TMD tuning problem.

In practice, several high-rise structures and bridges were designed by including TMDs. Also, TMDs were installed after the construction of the structures after several negative experiences resulting from the disturbing sway of the structures. The sway of the structure may be a treat to the security of the structure or prevent people to live in comfort. One Wall Centre in Vancouver, Shanghai World Financial Center in Shanghai, Berlin TV Tower in Berlin, Dublin Spire in Dublin, Akashi-Kaikyō Bridge in Japan, Tokyo Skytree in Tokyo, Yokohama Landmark Tower in Yokohama, Sakhalin-I offshore drilling platform in Russia, Taipei 101 in Taipei, Burj al-Arab in Dubai, 731 Lexington, Citigroup Center, Trump World Tower and Random House Tower in New York, Comcast Center in Philadelphia, Grand Canyon Skywalk in Arizona, John Hancock Tower in Boston, One Rincon Hill Tower in San Francisco, Park Tower in Chicago, Theme Building in Los Angeles and Millennium Bridge in London are the example structures including a type of TMD. Berlin TV tower shown in Fig. 1 was renovated by installing a TMD because of strong and disturbing vibrations resulting from wind forces. In the seismic retrofit of Theme Building in



**Fig. 2** Theme building in Los Angeles under renovation for seismic retrofit [3]

**Table 1** The frequency and damping ratio expressions of TMDs

Method	$f_{opt} = \frac{w_{d,opt}}{w_s}$	$\xi_{d,opt} = \frac{c_{d,opt}}{2m_d w_{d,opt}}$
Den Hartog [4]	$\frac{1}{1+\mu}$	$\sqrt{\frac{3\mu}{8(1+\mu)}}$
Warburton [5]	$\frac{\sqrt{1-(\mu/2)}}{1+\mu}$	$\sqrt{\frac{\mu(1-\mu/4)}{4(1+\mu)(1-\mu/2)}}$
Sadek et al. [6]	$\frac{1}{1+\mu} \left[ 1 - \xi \sqrt{\frac{\mu}{1+\mu}} \right]$	$\frac{\xi}{1+\mu} + \sqrt{\frac{\mu}{1+\mu}}$

Los Angeles (seen in Fig. 2), a TMD with 20% mass ratio and supported by 8 rubber bearings was installed on the main core of the structure in order to obtain 30–40% response reduction [3].

Several closed form expressions were proposed for optimum frequency and damping ratio of TMDs [4–8]. Several expressions were given in Table 1. The expressions of Den Hartog [4] are theoretically derived for undamped main system under harmonic excitation. Warburton [5] proposed several expressions for different excitations. In Table 1, the optimum expressions for white noise base excitation were given for Warburton [5]. The expression of Den Hartog and Warburton are dependent to a preselected mass ratio ( $\mu$ ) of TMD and main structure. Sadek et al. [6] performed numerically searched optimum TMD values for damped main system and obtained the expression depending on the damping of the main system ( $\xi$ ) by using curve fitting. The optimum frequency ratio ( $f_{opt}$ ) is defined as the ratio of the optimum frequency of TMD ( $w_{d,opt}$ ) and the frequency of SDOF structure ( $w_s$ ). The optimum damping coefficient of TMD ( $c_{d,opt}$ ) is formulated by the multiplication of the mass



of TMD ( $m_d$ ), optimum frequency of TMD ( $w_{d,opt}$ ) and damping ratio of TMD ( $\xi_{d,opt}$ ). The study of Sadek et al. [6] also contains several modifications for multiple degree of freedom structures.

By using curve fitting and modification for multiple degree of freedom systems, optimum TMD parameters cannot be exactly found. For that reason, numerical search algorithms have been employed for the optimization problem. Thus, all design variables including the mass of TMD are optimized by considering all modes of structures. Metaheuristic algorithms, which are developed by the inspiration of natural phenomena, have been employed in the optimization methods for TMD optimization problem of civil structures [8–23].

In this chapter, the TMD optimization method employing harmony search (HS) algorithm is explained. The methodology is demonstrated by optimizing TMDs for a seven-story structure for two cases considering different objectives such as minimization of maximum first story displacement and absolute acceleration of top story of the structure. The optimization process considers time domain analyses for several earthquake excitations in the presented method.

## 2 Metaheuristic Algorithms in TMD Optimization

In the optimization of TMDs, different metaheuristic algorithms such as Particle Swarm Optimization (PSO) [8, 9], Genetic Algorithm (GA) [10–14], Bionic Algorithm (BA) [15], Harmony Search (HS) [16–20], Ant Colony Optimization (ACO) [21], Artificial Bee Colony Optimization (ABC) [22] and Shuffled Complex Evolution (SCE) [23] have been employed. Generally, different design variables were optimized by considering different objectives depending to time domain or frequency domain responses.

PSO was developed by Kennedy and Eberhart formulated the movement behavior of organisms in a swarm [24]. Leung developed an optimization approach based on PSO for TMD tuning for structures under non-stationary base excitation [9]. Also, PSO was employed by Leung and Zang in development of TMD design formulas [8].

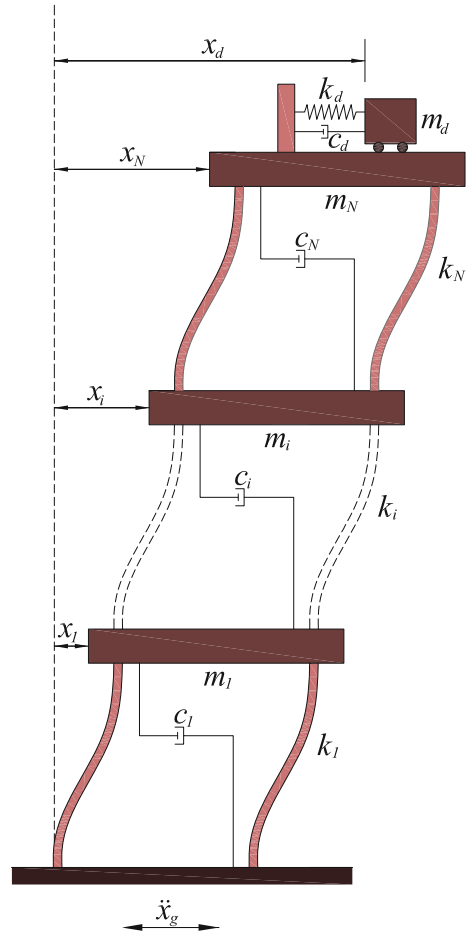
GA is the most known metaheuristic algorithm and it also belongs to evolutionary algorithm class. GA is inspired from the process of natural selection [25, 26]. Hadi and Arfiadi employed GA in search of stiffness and damping properties of TMD for seismic structures [10]. The optimization of the mass of TMDs were taken into consideration by Marano et al. in the study employing GA [11]. The response of torsionally irregular structure were reduced by TMDs optimized a methodology employing GA [12, 13]. Together by Fuzzy logic, GA is employed for active tuned mass damper optimization problem [14]. BA, which is also belongs to evolutionary algorithm class, was used in the optimization of high-rise structures excited by earthquakes [15].

HS imitates musical performances in search of optimum design variables [27]. In search of optimum mass, stiffness and damping coefficient of TMDs, HS based methodology was developed for the seismic structures [16, 17]. Mass ratio factor and comparison of HS with closed form TMD formulas were done by Bekdaş and





**Fig. 3** Physical model of N-story shear building including a TMD



mass ( $m_d$ ), damping coefficient ( $c_d$ ) and stiffness coefficient ( $k_d$ ). The displacement of the TMD is shown as  $x_d$ .

In the optimization methodology, the equations of motion given in Eq. (1) is solved for all iterations of the optimization process. Details of the optimization process are explained in Sect. 4.

### 4 Harmony Search Based TMD Optimization

The Harmony Search (HS) algorithm developed by Geem et al. [27] is a memory based random search method. It imitates the music performance process in which a musician tries to find a pleasing harmony that is a perfect state for appreciation of the audience. Like musicians, researchers try to find a global solution as a perfect state for maximum performance with a low cost.

Comparing to other metaheuristic algorithms, the usage of HS is not complex because a stochastic random search is used instead of a gradient search. Also, it is not a hill-climbing algorithm. For that reason, the local optima problem does not occur in solving problems. HS is suitable to solve problems with discrete and continuous variables [29, 30]. By using stochastic derivatives, the number of iteration can be reduced in the HS algorithm. When the function's mathematical derivative cannot be analytically obtained or function's type is step-wise or condition-wise, the usage of stochastic derivatives is important [31].

A musician can choose three possible options during a performance in order to gain the admiration of audience. The first option is to play a famous part of music from their memory. This option can be simulated as the usage of harmony memory (HM) in the HS algorithm. The HM may be constructed as a matrix for engineering problem searching for optimum design variables. This matrix contains harmony vectors and the number of these vectors are known as Harmony Memory Size (HMS). In that matrix, possible design variables will be stored to reach the optimum. In generation of a new harmony after the HM matrix initially constructed, a special parameter is used. This parameter is Harmony Memory Considering Rate (HMCR) and by using this rate parameter, it is possible to control the acceptance of the new harmonies.

A new harmony is generated according to the other two options of the musician. The second option is to play something similar to a famous part of music. By imitating this option, a new harmony can be generated from the HM. The HMCR is the possibility of a vector being selected from the existing HM. The third option is to compose a new or random note. According to this option, a new harmony can be randomly generated. If the value of HMCR is near to 1, a new harmony is strongly generated from the HM. In that case, the search will cover only the specific part of the range. The other parts in the solution domain can be missed. If the HMCR is too small, the optimization process will be long.

The randomization is the main source of the harmony search when generating a new harmony vector. As shown in Eq. (6). A random solution ( $S_r$ ) can be generated within a selected range defined by lower and upper limits named with  $S_{lower}$  and  $S_{upper}$ , respectively.  $Rand$  is a random number which is generated between 0 and 1.

$$S_r = S_{lower} + Rand(S_{upper} - S_{lower}) \quad (6)$$

Also, the adjusting of the pitches is related with the second option of musician. A parameter called Pitch Adjusting Rate (PAR) is utilized to adjust the range when the HM is chosen as the source of generation. The new harmony is searched in a smaller range around the values of HM. PAR can be accepted as the ratio between the smaller range around the stored values in HM and the whole range. Thus, it is possible to check the values which are close to existing ones in HM to find exact optimum values of design variables.

HS has several similarities with GA. As the usage of harmony memory, the least fit individuals are chosen in GA. The parameter, PAR is similar to mutation operator used in GA [32].

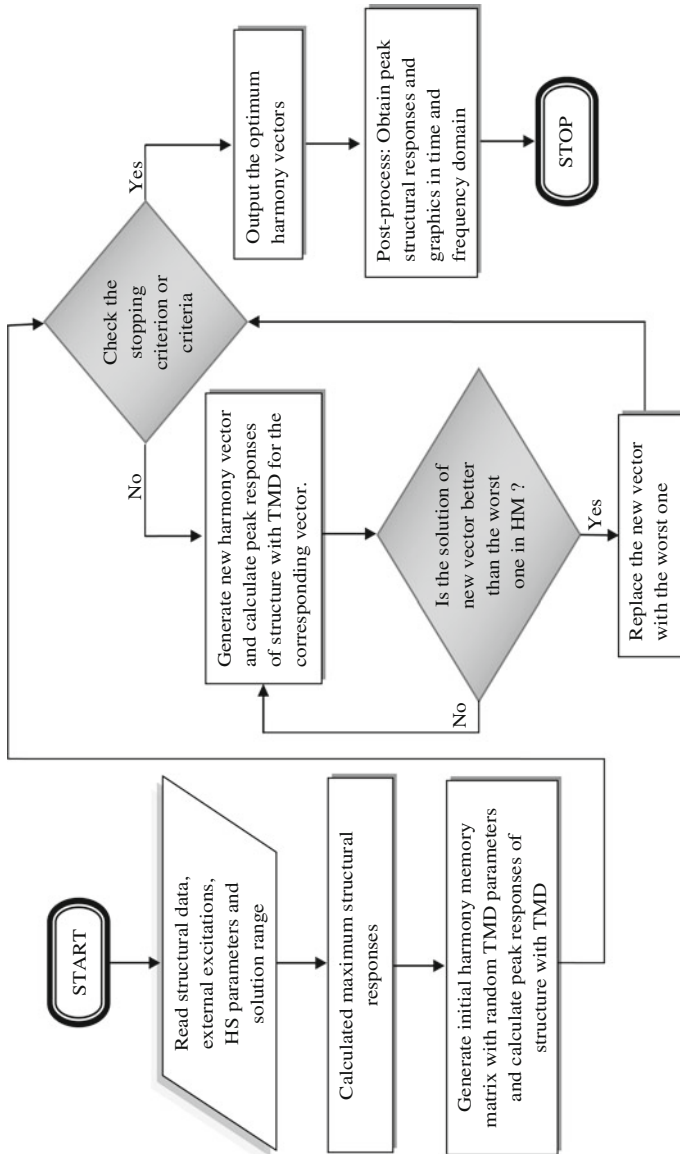


Fig. 4 The flowchart of the optimization process

The flowchart of the optimization process is given in Fig. 4. The process of TMD optimization employing HS can be summarized in six steps including the parameter setting procedure.

- i. In the first step, HS algorithm parameters; HMS, HMCR and PAR are defined. Also, the solution ranges for the design variables are defined. Selecting a wide range can increase the optimization time and a tight range may prevent to find best optimum solution. Termination criterion or criteria must be selected in this step according to the main purpose of the optimization problem. The main structure properties (design constants) must be also defined in this step.
- ii. The main purpose of using a TMD on a structure is to reduce an objective response. In order to make comparisons of optimization objective and termination criterion or criteria, the maximum objective response of the structure without TMD must be found. In the optimization process, several earthquake records are used in the dynamic analyses at the same run of the optimization code. In the dynamic analyses, the equations of motion given in Eq. (1) is solved by using Matlab with Simulink [33]. Runge-Kutta method with 0.001 s was chosen as a solver in the numerical analyses.
- iii. After the definition of known properties and dynamic analyses of the structure without TMD, the initial harmony memory (HM) matrix is generated with the combination of harmony vectors containing the unknown design variables. The number of the harmony vectors (HV<sub>1</sub> to HV<sub>HMS</sub>) stored in HM matrix is defined with HMS. The harmony vectors contain random numbers selected by the defined ranges. These possible optimum values are for mass (*m<sub>d</sub>*), period (*T<sub>d</sub>*) and damping ratio (*ξ<sub>d</sub>*) of the TMD. HM matrix and HV are defined in Eqs. (7) and (8), respectively. For each set of design variables, dynamic analyses are done for the structure including a TMD on the top of the structure. The value of objective function (OF) of the optimization is also stored in a vector for each set of design variables. The objective functions are given for displacement and acceleration objectives in Eqs. (9) and (10), respectively. Equation (9) is the ratio of maximum first story displacements of the structure with and without TMD and Eq. (10) is ratio of maximum absolute acceleration of the top story. The aim of the optimization is to minimize these objective functions.

$$HM = [HV_1 \ HV_2 \ \dots \ HV_{HMS}] \tag{7}$$

$$HV = \begin{bmatrix} m_{di} \\ T_{di} \\ \xi_{di} \end{bmatrix} \tag{8}$$

$$OF = \frac{\max(x_1)_{withTMD}}{\max(x_1)_{withoutTMD}} \tag{9}$$

$$OF = \frac{\max(\ddot{x}_N + \ddot{x}_g)_{withTMD}}{\max(\ddot{x}_N + \ddot{x}_g)_{withoutTMD}} \tag{10}$$

- iv. In the fourth step, a new harmony vector is randomly generated according to the rules of HS. This vector can be generated in two ways as explained before. It can be created around a randomly chosen existing vector, which is stored in the HM, or randomly generated within the initial solution range. In order to find best optimum solutions and escape local optima problem, the new vector is generated from the neighbouring values of a chosen vector in HM. The algorithm generates neighbouring values with a defined parameter called PAR. The parameter; PAR defines the ratio of the small range (the range modified according to the results of an existing HV in HM matrix) and initial range.
- v. In this step, HM matrix is modified. If the new harmony vector has lower OF value than the worst harmony vector in the HM, the worst one is replaced with the newly generated HV.
- vi. In the last step, if corresponding OF results of HVs in HM satisfy the termination criterion or criteria, the optimization process is ended. If not, iterations must continue from the fourth step where a new harmony vector is generated. The optimization process continue until termination criterion or criteria are satisfied. Two different criteria were used in this study. One of the criteria is to reduce the OF values under a user defined value. If the solution range is not suitable to reduce the OF value below the value entered by the user, the user defined value is iteratively increased after several attempts. For the minimization of the objective function, the user defined value may be entered as zero and the value of this value may be updated according to the values of the best harmony vector after several iterations. The other criterion is related with frequency domain results. For both objectives, acceleration transfer function of the first story must be smaller than the uncontrolled structure.

## 5 Numerical Example

The numerical example is a seven story building. The properties of all stories are the same. The mass, rigidity and damping coefficient of a story are 180t, 400 MN/m and 3 MNs/m, respectively. The period at the first mode is 0.64 s. The natural frequencies of the seven degrees of freedom structure are 1.57, 4.64, 7.5, 10.04, 12.14, 13.71 and 14.68 Hz.

During the optimization process, six different earthquake record were used. The optimization earthquakes have different characteristics and these earthquake records were downloaded from Pacific Earthquake Engineering Resource Center (PEER) NGA database [34]. Date, station and component information of earthquake records are given in Table 2.

In order to check the robustness of the optimum results for different excitations, the optimum TMD parameter were also tested on different records which were not considered in the optimization process. These earthquake records are BOL090 component of Bolu record of 1999 Duzce earthquake, PET090 component of Petrolia

**Table 2** Earthquake records used in the HS optimization [34]

Earthquake	Date	Station	Component
Loma Prieta	1989	16 LGPC	LGP000
Gazli	1976	9201Karaky	GAZ090
Erzincan	1992	95 Erzincan	ERZ-NS
Imperial valley	1940	El Centro Array #9	I-ELC180
Northridge	1994	24514 Sylmar	SYL360
Kobe	1995	0 KJMA	KJM000

record of 1992 Cape Mendocino earthquake and LCN000 component of Lucerne record of 1992 Landers earthquake.

In optimization of numerical example, mass, damping ratio and period of TMD are optimized. The range of the mass ratio ( $\mu$ ) of TMD and the total mass of the structure is between 1 and 5 %. The period of TMD is search between 0.8–1.2 times of the superstructure critical period. The lower and upper damping ratio limits are taken as 5 and 40 %, respectively.

The best and the worst harmony vectors are chosen according to ratios of maximum first story displacement in Case 1 and maximum top story acceleration in Case 2 between TMD controlled and uncontrolled structure. For all cases, frequency domain criterion defined in Sect. 4 was also used. The user defined value used for OF was taken as a small value in order to minimize the objective function. HS parameters; HMS, HMCR and PAR are taken as 5, 0.5 and 0.2, respectively.

For the displacement objective (Case 1), the optimum TMD parameters were found as 29.5 t ( $\mu = 2.34\%$ ), 0.666 s ( $k_d = 2625.63$  kN/m) and 0.398 ( $c_d = 221.53$  kNs/m) for the mass, period and damping ratio of TMD, respectively. The performance of TMD on reducing structural displacements is between 22 and 72 % for the optimization earthquakes and Case 1 results. The optimum results are also effective for the benchmark earthquakes. The maximum and minimum reductions are 58 and 27 %, respectively for these earthquake records.

The TMD is predominantly effective under Loma Prieta excitation. This excitation is the record under that the maximum structural displacements are observed. The first story displacement plots are given in Fig. 5. According to these plots, the TMD is optimally effective on reducing structural vibrations occurred under optimization earthquakes.

For the acceleration objective (Case 2), the optimum TMD parameters were found as 62.9 t ( $\mu = 4.99\%$ ), 0.687 s ( $k_d = 561.35$  kN/m) and 0.4 ( $c_d = 460.22$  kNs/m) for the mass, period and damping ratio of TMD, respectively. The performance of TMD on reducing top story accelerations is between 22 and 75 % for the optimization earthquakes and between 25 and 59 % for the benchmark earthquakes.

The maximum structural responses for uncontrolled structure and both cases are given and discussed in the last section of the chapter.

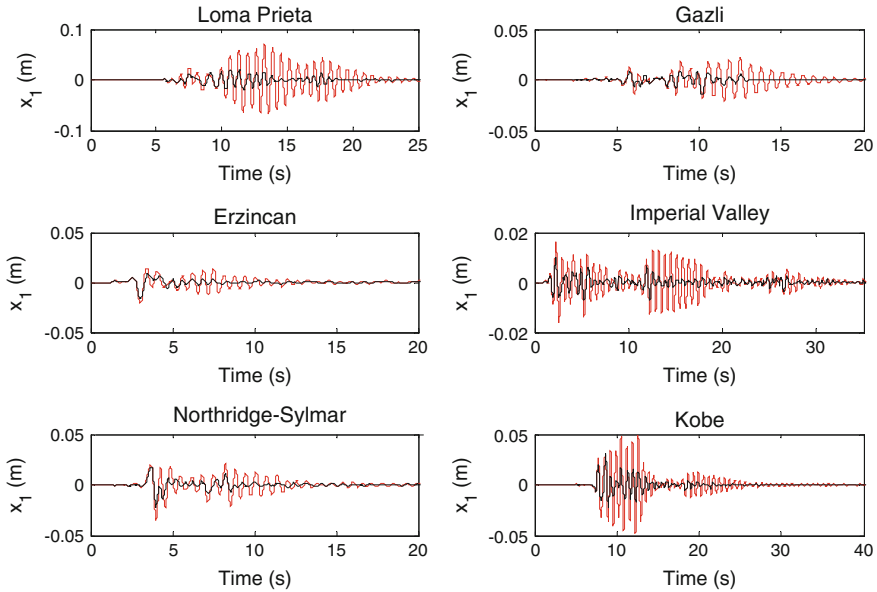


Fig. 5 The first story displacement plots under earthquakes (Case 1)

## 6 RESULTS and Conclusions

The maximum values of the displacements can be seen in Table 3 for both cases. Table 4 shows the maximum absolute accelerations for all stories.

According to the maximum displacement results the optimum TMD is effective to reduce all story displacement under all optimization and benchmark earthquakes for all cases. Another important factor is the maximum displacement of the TMD in design. The stroke of the TMD must be suitable to sustain the maximum displacement value without limiting the displacement of TMD. If the displacement of TMD is limited, the damping and tuning frequency of TMD is affected. In that reason, the optimum effectiveness of the TMD is lost.

For the maximum displacements, the most critical excitation is Loma Prieta record. The optimum TMDs are very effective to reduce the maximum displacement for that excitation, but the critical excitation is different for the TMD controlled structure. For the structures with optimum TMD, the most critical excitation is Kobe. For that excitation, case 1 in which the displacement objective is employed, the maximum displacement is lower than the results of case 2, although case 2 using an heavy mass comparing to case 1 for the other excitations including optimization and benchmark ones. These results show us the importance of using several records and a time-domain based optimization technique, since the critical excitation may differ according to the randomization of TMD parameters. Another conclusion of the numerical example is related with the robustness of the TMD for different exci-



**Table 3** Maximum displacements respect to ground under different earthquake record (m)

Earthquake record	Story	1	2	3	4	5	6	7	TMD
Loma Prieta, 1989	Without TMD	0.071	0.139	0.200	0.253	0.295	0.323	0.338	–
	Case 1	0.021	0.040	0.057	0.071	0.082	0.090	0.094	0.270
	Case 2	0.020	0.037	0.053	0.066	0.077	0.085	0.090	0.268
Gazli, 1976	Without TMD	0.022	0.043	0.062	0.078	0.092	0.101	0.106	–
	Case 1	0.016	0.032	0.046	0.058	0.067	0.073	0.076	0.126
	Case 2	0.016	0.031	0.045	0.056	0.065	0.071	0.074	0.129
Erzincan, 1992	Without TMD	0.021	0.040	0.057	0.072	0.084	0.093	0.097	–
	Case 1	0.016	0.031	0.044	0.056	0.065	0.072	0.075	0.154
	Case 2	0.017	0.032	0.045	0.057	0.066	0.073	0.077	0.162
Imperial valley 1940	Without TMD	0.016	0.032	0.047	0.060	0.071	0.078	0.082	–
	Case 1	0.010	0.019	0.028	0.035	0.041	0.045	0.047	0.087
	Case 2	0.009	0.018	0.026	0.033	0.039	0.043	0.045	0.086
Northridge, 1994	Without TMD	0.035	0.069	0.102	0.130	0.153	0.169	0.178	–
	Case 1	0.022	0.043	0.064	0.082	0.096	0.107	0.112	0.200
	Case 2	0.021	0.042	0.061	0.079	0.093	0.103	0.108	0.203
Kobe, 1995	Without TMD	0.049	0.097	0.141	0.180	0.211	0.233	0.244	–
	Case 1	0.031	0.059	0.084	0.105	0.122	0.133	0.140	0.285
	Case 2	0.031	0.061	0.087	0.109	0.127	0.139	0.147	0.311

(continued)

Table 3 (continued)

Earthquake record	Story	1	2	3	4	5	6	7	TMD
Duzce, 1999	Without TMD	0.033	0.063	0.091	0.114	0.132	0.144	0.151	–
	Case 1	0.022	0.042	0.060	0.075	0.087	0.095	0.099	0.235
	Case 2	0.021	0.041	0.058	0.073	0.085	0.094	0.099	0.245
Cape Mendocino, 1992	Without TMD	0.054	0.106	0.154	0.194	0.226	0.248	0.259	–
	Case 1	0.023	0.044	0.064	0.081	0.095	0.105	0.110	0.261
Landers, 1992	Case 2	0.023	0.044	0.063	0.079	0.093	0.102	0.108	0.249
	Without TMD	0.011	0.021	0.029	0.036	0.041	0.044	0.046	–
	Case 1	0.008	0.015	0.020	0.025	0.028	0.030	0.031	0.057
	Case 2	0.008	0.014	0.020	0.024	0.027	0.029	0.031	0.059

**Table 4** Maximum total acceleration respect to ground under different earthquake record (m/s<sup>2</sup>)

Earthquake record	Story	1	2	3	4	5	6	7	TMD
Loma Prieta, 1989	Without TMD	8.751	13.996	19.839	24.804	28.681	31.602	33.516	—
	Case 1	5.190	6.337	7.795	8.472	8.378	8.542	9.537	19.127
	Case 2	5.125	5.913	7.149	7.682	7.511	7.369	8.231	17.218
Gazli, 1976	Without TMD	4.860	5.445	6.789	8.457	9.116	10.075	10.708	—
	Case 1	5.927	5.448	5.843	6.853	7.304	7.530	8.262	9.258
	Case 2	6.018	5.575	5.718	6.601	6.985	7.295	8.008	8.950
Erzincan, 1992	Without TMD	5.290	5.577	6.004	6.795	7.947	9.181	9.870	—
	Case 1	5.287	5.515	5.636	5.742	6.316	7.269	7.834	8.635
	Case 2	5.263	5.465	5.566	5.660	6.220	7.132	7.666	8.681
Imperial valley 1940	Without TMD	3.933	5.476	6.238	6.529	7.425	8.278	8.793	—
	Case 1	3.224	3.806	3.760	4.129	4.448	4.647	4.723	6.305
	Case 2	3.182	3.581	3.657	3.933	4.213	4.367	4.420	5.975
Northridge, 1994	Without TMD	6.497	6.960	9.651	13.063	15.879	17.723	18.625	—
	Case 1	6.890	6.948	7.462	8.705	10.359	11.690	12.409	15.030
	Case 2	6.999	6.851	7.301	8.395	9.952	11.189	11.830	14.836
Kobe, 1995	Without TMD	8.481	11.536	14.062	17.521	21.023	23.667	25.221	—
	Case 1	8.562	9.733	10.947	11.657	11.881	11.927	12.124	19.621
	Case 2	8.576	9.746	10.967	11.709	11.982	12.092	12.330	19.695

(continued)

Table 4 (continued)

Earthquake record	Story	1	2	3	4	5	6	7	TMD
Duzce, 1999	Without TMD	7.328	7.621	9.647	11.463	13.169	14.102	14.520	—
	Case 1	7.715	7.586	8.054	8.477	8.917	9.858	10.936	15.202
	Case 2	7.732	7.610	8.064	8.472	8.886	9.759	10.780	15.235
Cape Mendocino, 1992	Without TMD	7.563	11.778	16.061	19.583	22.243	24.052	25.170	—
	Case 1	6.187	7.109	8.158	9.570	10.557	10.961	11.075	20.381
Landers, 1992	Case 2	6.157	6.995	7.806	8.908	9.749	10.065	10.272	18.400
	Without TMD	5.100	5.436	5.285	4.495	4.354	5.036	5.868	—
	Case 1	4.982	4.608	3.949	3.295	3.075	2.757	3.910	3.903
	Case 2	4.950	4.401	3.695	3.187	2.954	2.619	3.613	3.909

tations. The TMD with the heavy mass (case 2) is more effective on reducing of maximum responses including displacement and acceleration, but another factor; economy is in progress of the optimum designs. By the increase of the mass, the damping coefficient of the TMD is also increased. The damper of the TMD is the most expensive component of TMD. Design engineers must find a balance between performance and economy according to the requirements and opportunities in their hands. Also, the balance between performance and economy can be formulated to use in the optimization process.

Another important conclusion is the acceleration increase for low stories for TMD controlled structure. This situation is observed under several excitations such as Gazli, Northridge, Kobe and Düzce. For that reason, first story acceleration at which the absolute acceleration value is minimum cannot be taken as an optimization objective since the accelerations of the top stories are critical. In displacement objective, the drift of the first stories are generally critical and the displacement of the top stories are generally related with the displacement of lower stories. Thus, first story displacements and top story accelerations were considered as objective function in this study. Consequently, both approaches for different objective are suitable for tuning of TMDs. Both approach have negatives and positives as explained in the previous paragraphs of the Sect. 6. Metaheuristic methods and harmony search algorithm is a feasible approach for employing in optimization approaches of tuning of TMDs for seismic structures and similar problems.

## References

1. Frahm H (1911) Device for damping of bodies. U.S. Patent No: 989,958
2. Ormondroyd J, Den Hartog JP (1928) The theory of dynamic vibration absorber. *T. ASME* 50:9–22
3. Miyamoto HK, Gilani ASJ, Gündoğdu YZG (2011) Innovative seismic retrofit of an iconic building with mass damper. Seventh national conference on earthquake engineering, Istanbul, Turkey, 30 May–3 June
4. Den Hartog JP (1947) *Mechanical vibrations*. McGraw-Hill, New York
5. Warburton GB (1982) Optimum absorber parameters for various combinations of response and excitation parameters. *Earthq Eng Struct Dyn* 10:381–401
6. Sadek F, Mohraz B, Taylor AW, Chung RM (1997) A method of estimating the parameters of tuned mass dampers for seismic applications. *Earthq Eng Struct Dyn* 26:617–635
7. Chang CC (1999) Mass dampers and their optimal designs for building vibration control. *Eng Struct* 21:454–463
8. Leung AYT, Zhang H (2009) Particle swarm optimization of tuned mass dampers. *Eng Struct* 31:715–728
9. Leung AYT, Zhang H, Cheng CC, Lee YY (2008) Particle swarm optimization of TMD by non-stationary base excitation during earthquake. *Earthq Eng Struct Dyn* 37:1223–1246
10. Hadi MNS, Arfiadi Y (1998) Optimum design of absorber for MDOF structures. *J Struct Eng ASCE* 124:1272–1280
11. Marano GC, Greco R, Chiaia B (2010) A comparison between different optimization criteria for tuned mass dampers design. *J Sound Vib* 329:4880–4890
12. Singh MP, Singh S, Moreschi LM (2002) Tuned mass dampers for response control of torsional buildings. *Earthq Eng Struct Dyn* 31:749–769

13. Desu NB, Deb SK, Dutta A (2006) Coupled tuned mass dampers for control of coupled vibrations in asymmetric buildings. *Struct Control Health Monit* 13:897–916
14. Pourzeynali S, Lavasani HH, Modarayi AH (2007) Active control of high rise building structures using fuzzy logic and genetic algorithms. *Eng Struct* 29:346–357
15. Steinbuch R (2011) Bionic optimisation of the earthquake resistance of high buildings by tuned mass dampers. *J Bionic Eng* 8:335–344
16. Bekdaş G, Nigdeli SM (2011) Estimating optimum parameters of tuned mass dampers using harmony search. *Eng Struct* 33:2716–2723
17. Bekdaş G, Nigdeli SM (2013) Optimization of tuned mass damper with harmony search. In: Gandomi AH, Yang X-S, Alavi AH, Talatahari S (eds) *Metaheuristic applications in structures and infrastructures*. Elsevier, Chapter 14
18. Bekdaş G, Nigdeli SM (2013) Mass ratio factor for optimum tuned mass damper strategies. *Int J Mech Sci* 71:68–84
19. Nigdeli SM, Bekdaş G (2013) Optimum tuned mass damper design for preventing brittle fracture of RC buildings. *Smart Struct Syst* 12(2):137–155
20. Nigdeli SM and Bekdaş G (2014) Optimization of TMDs for different objectives. In: *An international conference on engineering and applied sciences optimization*, Kos Island, Greece, 4–6 June
21. Farshidianfar A, Soheili S (2013) Ant colony optimization of tuned mass dampers for earthquake oscillations of high-rise structures including soil-structure interaction. *Soil Dyn Earthq Eng* 51:14–22
22. Farshidianfar A, Soheili S (2013) ABC optimization of TMD parameters for tall buildings with soil structure interaction. *Interact Multiscale Mech* 6:339–356
23. Farshidianfar A, Soheili S (2013) Optimization of TMD parameters for earthquake vibrations of tall buildings including soil structure interaction. *Int J Optim Civil Eng* 3:409–429
24. Kennedy J, Eberhart RC (1995) Particle swarm optimization. In: *Proceedings of IEEE international conference on neural networks*. Perth, Nov 27–Dec 1, pp 1942–1948
25. Holland JH (1975) *Adaptat Nat Artif Syst*. University of Michigan Press, Ann Arbor MI
26. Goldberg DE (1989) *Genetic algorithms in search. Optimization and machine learning*. Addison Wesley, Boston
27. Geem ZW, Kim JH, Loganathan GV (2001) A new heuristic optimization algorithm: harmony search. *Simulation* 76:60–68
28. Dorigo M, Maniezzo V, Colomi A (1996) The ant system: optimization by a colony of cooperating agents. *IEEE Trans Syst Man Cybernet B* 26:29–41
29. Lee KS, Geem ZW (2005) A new meta-heuristic algorithm for continuous engineering optimization: harmony search theory and practice. *Comput Methods Appl Mech Eng* 194:3902–3933
30. Lee KS, Geem ZW, Lee SH, Bae KW (2005) The harmony search heuristic algorithm for discrete structural optimization. *Eng Optim* 37:663–684
31. Geem ZW (2008) Novel derivative of harmony search algorithm for discrete design variables. *Appl Math Comput* 199:223–230
32. Yang X-S (2008) *Nature-inspired metaheuristic algorithms*. Luniver Press, Bristol
33. The MathWorks Inc (2010) *MATLAB R2010a*. Natick, MA, USA
34. Peer (2005) Pacific earthquake engineering resource center: NGA database. University of California, Berkeley. <http://peer.berkeley.edu/nga>

# Tailoring Macroscale Response of Mechanical and Heat Transfer Systems by Topology Optimization of Microstructural Details

Joe Alexandersen and Boyan Stefanov Lazarov

**Abstract** The aim of this book chapter is to demonstrate a methodology for tailoring macroscale response by topology optimizing microstructural details. The microscale and macroscale response are completely coupled by treating the full model. The multiscale finite element method (MsFEM) for high-contrast material parameters is proposed to alleviate the high computational cost associated with solving the discrete systems arising during the topology optimization process. Problems within important engineering areas, heat transfer and linear elasticity, are considered for exemplifying the approach. It is demonstrated that it is important to account for the boundary effects to ensure prescribed behavior of the macrostructure. The obtained microstructures are designed for specific applications, in contrast to more traditional homogenization approaches where the microstructure is designed for specific material properties.

## 1 Introduction

The focus of this book chapter is on the topology optimization of microstructural details for tailoring the macroscale response of mechanical and heat transfer systems. Topology optimization [7] is an iterative design process which distributes material in a design domain by optimizing a prescribed objective and satisfying a set of constraints. In mechanical and structural engineering applications, the typical objective is to maximize structural stiffness subjected to material constraints, or minimize material volume subjected to stiffness constraints. Over the last decade topology optimization has become one of the preferred design tools in the automotive and aerospace industries. In addition, the method has spread to other disciplines for design of optical crystals and circuits, antennas and fluid mechanics systems [14, 30].

---

J. Alexandersen (✉) · B.S. Lazarov  
Department of Mechanical Engineering, Technical University of Denmark,  
Kongens Lyngby, Denmark  
e-mail: joealex@mek.dtu.dk

B.S. Lazarov  
e-mail: bsl@mek.dtu.dk

© Springer International Publishing Switzerland 2015  
N.D. Lagaros and M. Papadrakakis (eds.), *Engineering and Applied Sciences Optimization*, Computational Methods in Applied Sciences 38,  
DOI 10.1007/978-3-319-18320-6\_15

The main burden in topology optimization is the computational cost associated with modeling the physical behavior of the optimized system. The system response is evaluated for each optimization iteration. Relatively coarse discretizations are utilized in order to save computational time. Refining the discretization improves the physical model and provides a larger solution space for the optimization process. Therefore, one of the main goals in the development of the methodology is to reduce the computational complexity without restricting the design freedom. Several approaches like material homogenization, coupled and decoupled multiscale models and efficient state solvers, discussed below, are suggested in the literature.

The systematic design of novel materials with extremal properties using topology optimization has been demonstrated in several papers starting with the pioneering work for 2D designs presented in [28, 29] to the recent manufacturable 2/3D material designs with negative Poisson's ratio [5, 35]. The optimization is performed on periodic microstructures with the aim to achieve prescribed effective properties. Such optimization affects indirectly the macroscopic response and an alternative multiscale approach to the topological design is to introduce homogenized microstructural properties in the optimization of a macrostructural response. This coincides with the original homogenization approach to topology optimization presented in [6]. The macroscale design is realized with homogenized material properties without the need to precisely specify the unit cell topology. Later a hierarchical optimization strategy has been applied to bone modeling [12, 13] where the microscopic structure and the macroscopic density are designed simultaneously. The macroscale response is decoupled from the microscale and the microstructural details affect the macroscale response through the homogenized material properties. The scale separation reduces the computational cost, however, the design often lacks connectivity between the varying microstructural details. Furthermore, practical realizations of such designs with modern manufacturable technologies (e.g. [5]) lead to finite size periodic cells, which contradict the infinite periodicity assumption applied in the homogenization process.

Here, the macroscale response of the system is completely coupled to the structural response at the microscale. The fine discretization of the physical system requires the solution of large linear systems of equations. The system response can be obtained using direct or iterative solvers. Direct solvers are often preferable due to their robust behavior, however, for large 3D problems, the computational time becomes prohibitive even on large parallel systems. On the other hand, even though they lack the robustness of direct solvers, iterative solvers provide scalable and easy to implement parallel solutions. Their convergence is improved by utilizing preconditioning techniques [26] which in the context of topology optimization are discussed in [1, 2, 4].

Here iterative solvers with preconditioning using the multiscale finite element method (MsFEM) for high-contrast media are utilized, in order to speed up the design process and to allow the optimization of large scale problems without compromising the resolution. The original MsFEM [19] represents the system behavior by constructing basis functions on a coarse grid. The coarse basis functions provide a good approximation to the system response and reduce significantly the problem



size. The method has been applied mainly to scalar problems, and recently extensions to elasticity [11] and problems modeled by positive definite bilinear forms [17], have been demonstrated as well. The MsFEM for high-contrast media [18] constructs several basis functions per coarse node, which represents well the important features of the solution with a convergence rate independent of the contrast. The method has been extended and applied to topology optimization problems in linear elasticity in [3, 22] and is presented in details in Sect. 5.

## 2 Physical Models

The partial differential equations (PDEs) governing the physical behavior for heat transfer and linear elasticity are introduced for 2D in the following subsections. The presented examples follow this simplification. However, the approach considered in this book chapter can be extended to 3D without any significant modifications, which will be demonstrated in following works.

### 2.1 Heat Transfer

The system response for heat transfer problems in a conductive medium distributed in a given domain  $\Omega$  is governed by the following PDE

$$-\nabla^T \mathbf{q} + p(\mathbf{x}) = 0 \quad \mathbf{x} \in \Omega \quad (1)$$

where  $\mathbf{q}$  is the heat flux per unit area and  $p(\mathbf{x})$  is a source term. The conductive heat flux  $\mathbf{q}$  is obtained from Fourier's law as

$$\mathbf{q} = -\kappa \nabla \theta \quad (2)$$

where  $\kappa_{\min} \leq \kappa(\mathbf{x}) \leq \kappa_{\max}$  is a spatially-varying conduction coefficient and  $\theta$  is a scalar temperature field defined over the domain  $\Omega$ . The boundary  $\Gamma = \partial\Omega$  is decomposed into disjoint subsets  $\Gamma = \overline{\Gamma_D} \cup \overline{\Gamma_N}$ . The following boundary conditions are prescribed on the different subsets

$$\theta = 0 \quad \text{on } \Gamma_D \quad (3)$$

$$q_n = g \quad \text{on } \Gamma_N \quad (4)$$

where  $q_n = \mathbf{q}^T \mathbf{n}$ .

The variational formulation [9] of the above problem is to find  $u \in H_0(\Omega)$  such that

$$a(u, v) = l(v) \quad \text{for all } v \in H_0^1(\Omega) \quad (5)$$

where the bilinear form  $a$  and the linear functional  $l$  are defined as

$$a(u, v) = \int_{\Omega} \kappa(\mathbf{x}) \nabla u(\mathbf{x}) \nabla v(\mathbf{x}) \, d\mathbf{x} \quad \text{for all } u, v \in H_0^1(\Omega) \quad (6)$$

$$l(v) = \int_{\Omega} p(\mathbf{x}) v(\mathbf{x}) \, d\mathbf{x} + \int_{\Gamma_N} \kappa(\mathbf{x}) g v(\mathbf{x}) \, d\mathbf{x} \quad \text{for all } v \in H_0^1(\Omega) \quad (7)$$

and  $H_0^1(\Omega)$  is defined as

$$H_0^1(\Omega) = \left\{ v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D \right\} \quad (8)$$

$H^1(\Omega)$  is a standard Sobolev space on  $\Omega$ . The Galerkin formulation of Eq. 5 is obtained using the finite element space  $V_h(\Omega) \subset H_0^1(\Omega)$  with test and trial functions  $u, v \in V_h(\Omega)$ . The space  $V_h(\Omega)$  consists of standard Lagrange shape functions defined on a uniform rectangular mesh  $\mathcal{T}^h$  with characteristic length  $h$ . The Galerkin formulation leads to a linear system of equations of the form

$$\mathbf{K}\mathbf{u} = \mathbf{f} \quad (9)$$

where the vector  $\mathbf{u}$  consists of all nodal values of the temperature field  $\theta$  and  $\mathbf{f}$  is a vector with the supplied input to the system.

## 2.2 Linear Elasticity

The response of a linear elastic system is governed by the Navier-Cauchy partial differential equation, e.g. [9], given as

$$\nabla \cdot \boldsymbol{\sigma}(\mathbf{u}) + \mathbf{f}(\mathbf{x}) = 0, \quad \mathbf{x} \in \Omega \quad (10)$$

$$\boldsymbol{\sigma}(\mathbf{u}) = \mathbf{C} : \boldsymbol{\varepsilon}(\mathbf{u}) \quad (11)$$

where  $\boldsymbol{\sigma}$  is the stress tensor,  $\boldsymbol{\varepsilon}$  is the linearized strain tensor, the vector  $\mathbf{u}$  consists of the displacements in the coordinate directions and  $\mathbf{C}$  is the linear elastic stiffness tensor. The vector function  $\mathbf{f}(\mathbf{x})$  represents the system input. The mechanical system occupies the bounded domain  $\Omega$ , where the boundary  $\Gamma = \overline{\Gamma_{D_i}} \cup \overline{\Gamma_{N_i}}$  is decomposed into two disjoint subsets for each component  $u_i, i = 1, 2$ .  $\Gamma_{D_i}$  is the part of the boundary where  $u_i = 0$  and  $\Gamma_{N_i}$  denotes the part with prescribed traction  $t_i$ . The stiffness tensor is isotropic with predefined Poisson's ratio  $\nu < 0.5$  and spatially-varying Young's modulus  $E_{\min} \leq E(\mathbf{x}) \leq E_{\max}$ .

The weak formulation of the linear elasticity problem is to find  $\mathbf{u} \in V_0$  such that

$$a(\mathbf{u}, \mathbf{v}) = l(\mathbf{v}) \quad \text{for all } \mathbf{v} \in V_0 \quad (12)$$

with bilinear form  $a$  and linear functional  $l$  defined as

$$a(\mathbf{u}, \mathbf{v}) = \int_{\Omega} (\mathbf{C} : \varepsilon(\mathbf{u})) : \varepsilon(\mathbf{v}) \, d\mathbf{x} \quad \text{for all } \mathbf{v} \in V_0 \quad (13)$$

$$l(\mathbf{v}) = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \, d\mathbf{x} + \int_{\Gamma_N} \mathbf{t} \cdot \mathbf{v} \, d\mathbf{x} \quad \text{for all } \mathbf{v} \in V_0 \quad (14)$$

where the space  $V_0$  is defined as

$$V_0 = \left\{ \mathbf{v} \in \left[ H^1(\Omega) \right]^2 : v_i = 0 \text{ on } \Gamma_{D_i}, i = 1, 2 \right\} \quad (15)$$

The weak formulation is discretized using standard finite element functions defined on uniform rectangular mesh  $\mathcal{T}^h$ . Similar to the heat transfer case, the discrete formulation results in a linear system of the form given by Eq. 9 with vector  $\mathbf{u}$  consisting of all nodal displacements.

### 3 Topology Optimization Formulation

Topology optimization is an iterative method that seeks to distribute material in a given design domain by optimizing an objective functional and fulfilling a set of design constraints [7]. The material distribution is represented by a density field  $0 \leq \rho(\mathbf{x}) \leq 1$ . The density field takes values one for all points in the design domain  $\Omega$  occupied with material and zero for the void regions. In order to utilize gradient-based optimization techniques, the density field is allowed to take intermediate values.

The main steps in the topology optimization algorithm will be demonstrated first for thermal compliance minimization, which coincides with the first example in Sect. 6. The optimization problem is defined as

$$\begin{aligned} \min_{\rho \in \mathcal{Q}_{ad}} : c(\rho, u) &= \int_{\Omega} \kappa(\rho(\mathbf{x})) \nabla u(\mathbf{x}) \nabla u(\mathbf{x}) \, d\mathbf{x} \\ \text{s.t. } a(\rho; u, v) &= l(v) \\ \int_{\Omega} \rho \, d\mathbf{x} &\leq V^* \end{aligned} \quad (16)$$

where  $\mathcal{Q}_{ad}$  is the space of admissible density material distributions,  $V^*$  is the allowed volume of material and  $a(\rho; u, v)$  is the bilinear form given by Eq. 6. In the optimization problem, the bilinear form Eq. 6 depends on the density field  $\rho$ . The heat conduction coefficient in Eq. 2 is interpolated between  $\kappa_{\min}$  and  $\kappa_{\max}$  using the modified SIMP scheme [7] given as

$$\kappa = \kappa_{\min} + (\kappa_{\max} - \kappa_{\min}) \rho^p \quad (17)$$

where  $p$  is the penalization parameter,  $\kappa_{\max}$  is the conduction coefficient of the solid material, and  $\kappa_{\min}$  is set to be a very small number in order to ensure that the bilinear form is coercive. The above optimization problem can be written in discrete form using the finite element discretization given by Eq. 9. The design field  $\rho$  is represented using independent design variables associated to each element. The discrete problem is given as

$$\begin{aligned} \min_{\rho} : c &= \mathbf{f}^T \mathbf{u} & (18) \\ \text{s.t. } \mathbf{K} \mathbf{u} &= \mathbf{f} \\ \rho^T \mathbf{v} &\leq V^* \\ 0 \leq \rho_i &\leq 1 \quad i = 1, \dots, n_{\text{el}} \end{aligned}$$

where the vector  $\rho$  consists of all design variables and  $\mathbf{v}$  is a vector with element  $v_i$  equal to the volume of the  $i$ th finite element.

The optimization problem is solved using the so-called nested formulation, where the discrete system of equations for the state problem is solved during each optimization step. The gradients of the objective with respect to the design variables are computed using adjoint sensitivity analysis [7] and are given as

$$\frac{\partial c}{\partial \rho_e} = -p \rho_e^{p-1} (\kappa_{\max} - \kappa_{\min}) \mathbf{u}_e^T \mathbf{K}_{0,e} \mathbf{u}_e, \quad e = 1, \dots, N_{\text{el}} \quad (19)$$

The design update is performed using the method of moving asymptotes (MMA) [31].

The optimization problem defined by Eq. 18 is mesh dependent. Instead of obtaining a better representation of a coarse optimized topology, the optimization might result in a completely different topology by refining the mesh. Such behavior is avoided here by utilizing density filtering [8, 10]. The filtered density  $\rho_f(\mathbf{x})$  at a point  $\mathbf{x}$  in the design domain is obtained using convolution of the original design field  $\rho$  and a filter function

$$\rho_f(\mathbf{x}) = \int_{\Omega} F(\mathbf{x} - \mathbf{y}) \rho(\mathbf{y}) \, d\mathbf{y} \quad (20)$$

The filter function is chosen to be

$$F(x) = \frac{1}{R} \left( 1 - \frac{|x|}{R} \right), \quad x \in [-R, R] \quad (21)$$

where  $R$  is the filter radius, which controls the length scale. Instead of using an explicit weighting function  $F(\cdot)$ , the filtered field can be obtained as a solution of a PDE [24] given as

$$-r^2 \Delta \rho_f + \rho_f = \rho \quad (22)$$

with  $r = R/(2\sqrt{3})$ . The PDE filter simplifies the enforcement of different boundary conditions on the density field, reutilizes the already developed discretization framework for solving the state problem, simplifies large scale parallel implementations of the topology optimization process, and reduces the computational cost in 3D [1, 2, 24]. The classical filter is utilized for the heat transfer example and the PDE filter is utilized for the linear elastic designs.

## 4 Robust Design

The filtered field consists of large gray regions which require post-processing of the optimized results. Such a transformation can affect the optimality of the solution and in many cases [34] completely destroy the performance of the optimized design. These post-processing effects are alleviated here by using projection and introducing a requirement on the performance to be insensitive with respect to uncertainties in the geometry [23, 34]. The physical density in this case is represented by a projected density field obtained as

$$\rho_p = \frac{\tanh(\beta\eta) + \tanh(\beta(\rho_f - \eta))}{\tanh(\beta\eta) + \tanh(\beta(1 - \eta))} \quad (23)$$

where  $\eta$  is a selected threshold and  $\beta$  controls the sharpness of the projections. For  $\beta \rightarrow \infty$  the above expression approaches a Heaviside function. The gradients of the objective functional and the constraints with respect to the original design field  $\rho$  are obtained by the chain rule.

The projection improves the contrast in the design, however, the length scale imposed from the filter is lost. All manufacturing processes introduce uncertainties in the realizations of the optimized designs, which might result in complete loss of the performance [21, 34, 36]. Imperfections along the design perimeter can be modeled by varying the threshold  $\eta$  in Eq. 23, and for cases with non-uniform uncertainties the threshold can be replaced with spatially-varying random field [27].

Here the threshold is assumed to be a random variable with uniform distribution  $\eta \in [\eta_d; \eta_e]$ , where the threshold  $\eta_d$  corresponds to the most dilated design and  $\eta_e$  corresponds to the most eroded case. The optimization problem is posed as follows

$$\begin{aligned} \min_{\rho} : \quad & c = \mathbb{E}[\mathbf{f}^T \mathbf{u}] + w\sqrt{\text{Var}[\mathbf{f}^T \mathbf{u}]} \\ \text{s.t.} \quad & \mathbf{K}\mathbf{u} = \mathbf{f} \\ & \mathbb{E}[\rho^T \mathbf{v}] \leq V^* \\ & 0 \leq \rho_i \leq 1 \quad i = 1, \dots, n_{el} \end{aligned} \quad (24)$$

where  $E[\cdot]$  and  $\text{Var}[\cdot]$  denote the expected value and the variance of a given quantity, and  $w$  is a weight factor. The state problem in the above formulation becomes stochastic and approximations to expectation and the variance are obtained using Stochastic collocation and Monte Carlo sampling [25]. The gradients are computed as described in [23].

## 5 Multiscale Finite Element Method

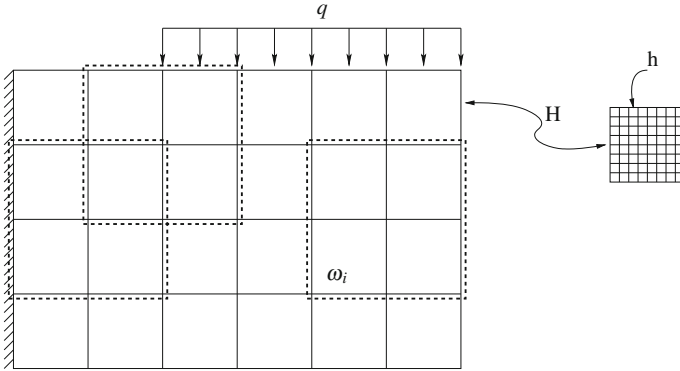
Topology optimization is an iterative approach which requires the computation of the state solution, and possibly adjoint solution also, at every design iteration. Often, the required state and adjoint field computations account for more than 95–99 % of the total computational time [2]. The solution for small problems is usually obtained using direct solvers due to their robustness. Realistic 3D and large 2D designs with fine details require fine resolutions, which makes the computational cost prohibitive. An alternative is to use iterative solution techniques [26] also known as Krylov iterative methods. Iterative solvers alleviate some of the issues observed with direct solvers in terms of memory utilization and parallel scalability. However, their convergence speed is determined by the condition number of the system matrix, which can be improved by preconditioning.

Classical preconditioners such as incomplete factorization, diagonal scaling and successive over-relaxation, cannot provide mesh independent convergence. Furthermore, for problems with high contrast between material parameters, as the ones arising in topology optimization, the number of iterations increases with increasing contrast [2, 4]. Mesh independent convergence can be obtained using geometric multigrid (MG) [33], if the coarse grid is capable of resolving the fine scale details. Such a condition cannot be guaranteed in the topology optimization process which results in deteriorated convergence. A compelling alternative demonstrated in [15, 18, 20] is the multiscale finite element method (MsFEM) with spectral basis functions.

MsFEM with spectral bases has initially been developed for diffusion type problems [18, 20], for general bilinear forms [17], and extended later for topology optimization problems in linear elasticity [3, 22]. Here the method is presented for heat transfer problems and follows closely [18]. The idea is to construct a coarse space capable of representing the important features of the solution.

The fine mesh  $\mathcal{T}^h$  utilized for the discretization of Eq. 1 and 2 is assumed to be obtained by a refinement of a coarser one  $\mathcal{T}^H = \{K_j\}_{j=1}^{N_{cc}}$ , where  $K_j$  denotes a coarse mesh cell and  $N_{cc}$  the number of coarse cells (e.g. Fig. 1). The nodes of the coarse mesh are denoted as  $\{\mathbf{y}_i\}_{i=1}^{N_c}$ , where  $N_c$  denotes the number of coarse nodes. The neighborhood of node  $\mathbf{y}_i$  is defined as

$$\omega_i = \bigcup \left\{ \bar{K}_j \in \mathcal{T}^H : \mathbf{y}_i \in \bar{K}_j \right\} \quad (25)$$



**Fig. 1** Illustration of fine, coarse mesh and several agglomerates for cantilever beam subjected to distributed load  $q$

The neighborhoods  $\omega_i, i = 1, \dots, N_c$ , will be called agglomerates as they can be viewed as a group of coarse elements agglomerated together.

A set of coarse basis functions  $\{\phi_{i,j}, j = 1, \dots, N_c\}$ , defined with respect to  $\mathcal{T}^h$ , is introduced for each coarse node  $y_j$ . An approximation to the solution in the coarse space is sought as  $u_c = \sum_{i,j} c_{i,j} \phi_{i,j}$ . The coefficients  $c_{i,j}$  are determined by solving the coarse problem  $\mathbf{K}_c \mathbf{u}_c = \mathbf{f}_c$ , with

$$\mathbf{K}_c = \mathbf{R}_c \mathbf{K} \mathbf{R}_c^T \tag{26}$$

$$\mathbf{f}_c = \mathbf{R}_c \mathbf{f} \tag{27}$$

where  $\mathbf{R}_c = [\phi_{i,1}, \phi_{i,2}, \dots, \phi_{N_c,1}, \phi_{N_c,2}, \dots]$  consists of all coarse basis functions defined on the fine scale grid, and  $\mathbf{u}_c$  consists of all coefficients  $c_{i,j}$ . The matrix  $\mathbf{R}_c$  provides a map between temperature fields defined on the fine and the coarse grids. An approximation to the nodal solution in the fine space can be obtained as  $\mathbf{u}_a = \mathbf{R}_c^T \mathbf{u}_c$ .

The set of coarse basis functions is built using the set of eigenmodes of local eigenvalue problems [18] defined on each agglomerate  $\omega_i$ . The eigenvalue problem for agglomerate  $\omega_i$  is given as

$$-\nabla^T \kappa(\mathbf{x}) \nabla u = \lambda \kappa(\mathbf{x}) u, \quad \mathbf{x} \in \omega_i \tag{28}$$

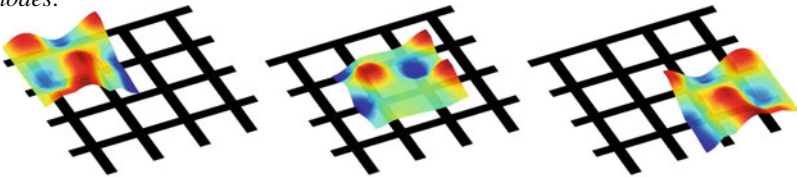
with homogeneous Neumann boundary conditions on the agglomerate boundary if  $\partial \omega_i \cap \Gamma = \emptyset$ , and boundary conditions applied to Eq. 1 on  $\partial \omega_i \cap \Gamma \neq \emptyset$ , where  $\Gamma$  is the boundary of the design domain  $\Omega$  and  $\partial \omega_i$  is the boundary of the agglomerate  $\omega_i$ . The eigenvalue problem is discretized using  $V_h(\omega_i) = \{v_h \in V_h : \text{supp } v_h \subset \omega_i\}$  and in matrix vector form is given as

$$\mathbf{K}_{\omega_i} \psi_j^{\omega_i} = \lambda_j^{\omega_i} \mathbf{M}_{\omega_i} \psi_j^{\omega_i} \tag{29}$$

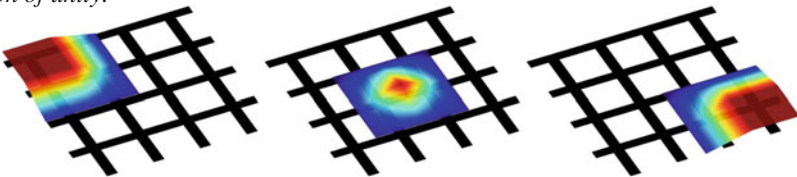
where  $\mathbf{K}_{\omega_i}$  is the stiffness matrix,  $\mathbf{M}_{\omega_i}$  is a mass matrix,  $\psi_j^{\omega_i}$  is the  $j$ th eigenvector, and  $\lambda_j^{\omega_i}$  is the  $j$ th eigenvalue. The eigenvalues are ordered as  $\lambda_1^{\omega_i} \leq \lambda_2^{\omega_i} \leq \dots \leq \lambda_j^{\omega_i} \leq \dots$ , and the first eigenvectors corresponding to eigenvalues smaller than a selected threshold  $\lambda_\Omega$  are selected to form the coarse basis. The coarse basis functions, represented on the fine grid, are defined as  $\phi_{i,j} = \xi_i \psi_j^{\omega_i}$ , i.e., they are constructed by multiplication of the eigenfunctions  $\psi_j^{\omega_i}$  with a partition of unity  $\{\xi_i\}_{i=1}^{N_c}$  subordinated to  $\omega_i$  such that  $\xi_i \in H^1(\Omega)$  and  $|\nabla \xi_i| \leq 1/H, i = 1, \dots, N_c$ , where  $H$  is the characteristic length of a coarse element  $K$ . Hence, for each coarse node, the basis functions  $\{\phi_{i,j}\}$  are defined as the fine space finite element interpolants of  $\xi_i \psi_j^{\omega_i}, j = 1, \dots, N_i$ , where  $N_i$  is determined as the number of eigenvalues smaller than the globally selected threshold  $\lambda_\Omega$ . It is important to note that, since the eigenvalue problem defined on agglomerate  $\omega_i$  and the full problem share the same boundary conditions on the common boundaries, the eigenfunctions and hence the coarse basis functions automatically fulfill the boundary conditions of the global problem. The construction process of several coarse basis functions is exemplified in Fig. 2.

In [18] the coarse system is utilized as a solver, where the accuracy of the coarse approximation depends on the global threshold  $\lambda_\Omega$ , which controls the number of the basis functions and the computational cost. For topology optimization problems

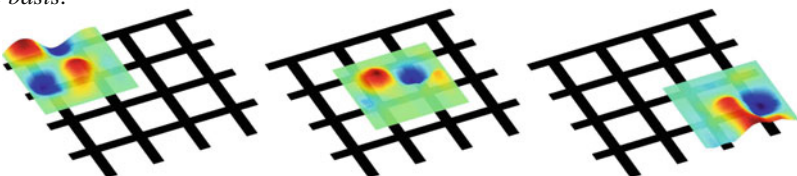
*Eigenmodes:*



*Partition of unity:*



*Coarse basis:*



**Fig. 2** Illustration of spectral basis construction



using the nested formulation, the optimizer can take advantage of the approximation error. As discussed in [22], the optimization for linear elasticity might result in isolated islands of material. Such topologies are not optimal and appear due to the homogenization effect of the approximation. Therefore, here the coarse solver is utilized as a preconditioner for iterative solvers applied to the fine-scale problem. Using the coarse system as a preconditioner results in mesh- and contrast-independent number of iterations for the Preconditioned Conjugate Gradient (PCG) and the Generalized Minimal Residual Method (GMRES). In [20] the coarse space is utilized in a two-level additive Schwarz preconditioner. Instead of implementing local sub-domain solvers for the Schwarz preconditioner, here, the coarse space is utilized as a coarse-level in a two-level multigrid preconditioner for GMRES (e.g. [33]). The smoothing is performed by a single symmetric Gauss-Seidel step.

The time consuming part of the MsFEM algorithm is the construction of the coarse basis and the projection given by Eq. 26. Several strategies for reducing the computational cost are discussed in [3, 22]. The main idea utilizes the fact that the design changes during the optimization process are relatively slow and hence consecutive design realizations can share the same coarse basis. When the difference in the topologies for the reference and the current design becomes large, the basis is updated. A heuristic rule is suggested in [3] where the basis is updated when the solver iterations exceed the previous iteration number by more than a given limit. More rigorous criteria is a subject of future research. In the stochastic case, the basis is constructed for the most dilated design and utilized for all realizations which further reduces the computational cost. For linear elastic problems, the MsFEM coarse basis algorithm follows the same steps and is demonstrated in [3, 22].

## 6 Numerical Examples

### 6.1 Heat Sink Design

The first example is the topology optimization of thermal compliance. The design domain is shown in Fig. 3. The temperature  $T_0$  is set to zero. The conduction coefficient of the solid material is set to one and the conduction of the void region is  $10^{-6}$ . The volume occupied with solid material is restricted to be 50 % of the total volume. Uniform heat flux is supplied over the design domain. The penalization factor  $p$  is increased from 1.2 to 3.0 after the first 100 iterations. The projection coefficient  $\beta$  is increased from 8 to 32 after the first 200 iterations. The optimization is performed with three realizations  $\eta_e = 0.7$ ,  $\eta_i = 0.5$  and  $\eta_d = 0.3$  of the threshold projection  $\eta \in [0.3, 0.7]$  and are verified by Monte Carlo simulations. Four coarse cell configurations with  $4 \times 4$ ,  $8 \times 8$ ,  $16 \times 16$  and  $32 \times 32$  coarse cells, are selected. Each coarse cell consists of  $40 \times 40$  elements. The filtering step is performed with standard hat filter function with radius  $R = 3h$ . All coarse cells are kept identical. The optimization problem in discrete form is given by Eq. 24 with  $w = 1$ .

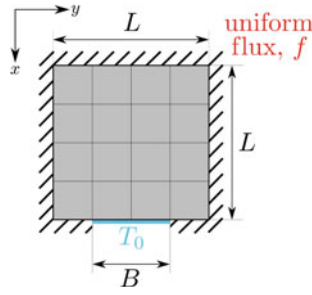


Fig. 3 Heat sink optimization problem—design domain with dimensions  $L$  and  $B = L/2$ . Unit heat flux is applied uniformly over the design domain

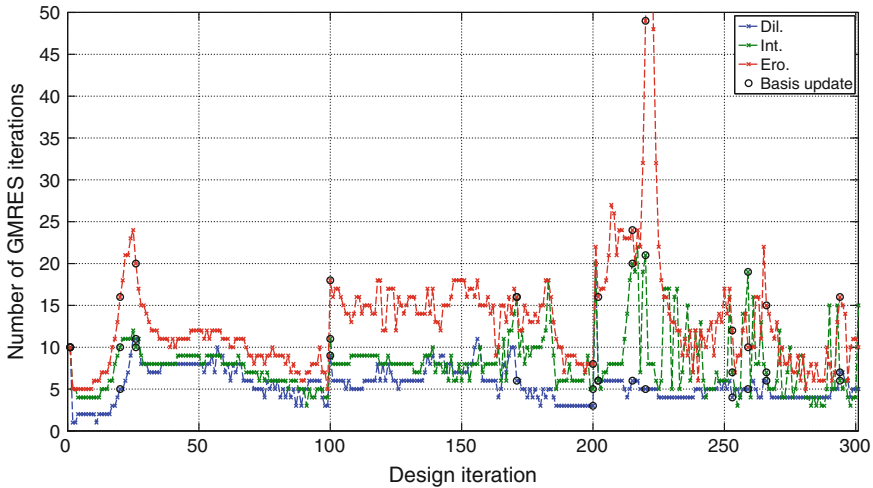


Fig. 4 Optimized heat sink topology for  $4 \times 4$ ,  $8 \times 8$ ,  $16 \times 16$  coarse cells



Fig. 5 Optimized heat sink coarse cell topology for  $4 \times 4$ ,  $8 \times 8$ ,  $16 \times 16$  and  $32 \times 32$  coarse cells

Optimized topologies for the heat sink design problem are shown in Fig. 4 and enlarged cell designs are shown in Fig. 5. The mean objective values for the four considered cases are 46.3; 12.8; 7.3; 3.7. The decrease in the compliance is due to the nature of the problem. The optimal design will consist of smaller and smaller features covering more uniformly the design domain due to the distributed flux. Increasing the number of coarse cells with a constant relative length scale at the microscale, results in a smaller overall design length scale which improves the objective. The length scale is imposed with respect to the cell characteristic length and is not related to the global macroscale. It can be observed that the cell topology is not preserved during

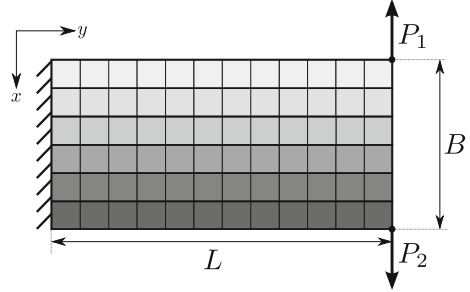


**Fig. 6** GMRES iterations for relative tolerance  $10^{-5}$ . The circles denote the basis updates

the refinement. The GMRES iteration number is kept under 20 with the selected eigenvalue threshold. Increasing the threshold, decreases the iteration number. However, as the cost for computing the basis increases, it also results in an increase of the total computational time [3]. The basis is obtained for the dilated realization, and thus the GMRES iterations differ between the realizations, as seen in Fig. 6. An alternative is the reduced basis approach as suggested in [16], which provides good coarse space for all realizations. However, this increases the computational cost related to the eigenproblems and results in longer optimization time.

The selection of the eigenvalue threshold is a non-trivial task and needs further investigations. The value and the computational time depend on the CPU architecture, the implementation of the preconditioner, the eigenvalue solver and the number of unique agglomerates in the design. For the selected example only ten unique agglomerates can be identified. As demonstrated in [22], MsFEM can also be applied to general problems without microstructure. Such an approach removes the restrictions on the design space and the design performance is expected to improve further. However, the design freedom comes at higher computational cost due to the large number of local eigenvalue problems. All of them are completely independent. Thus, the MsFEM preconditioner will excel in parallel implementations which are subject to future research. It should be noted that for the small 2D problems, the total computational time becomes larger compared to the total time with direct solvers. However, increasing the problems size leads to shorter computational times for the proposed approach [3].

**Fig. 7** Boundary conditions and design domain of layered cantilever beam problem with multiple load cases. The vertical dimension is  $B = L/2$



## 6.2 Linear Elastic Designs with Multiple Load Cases

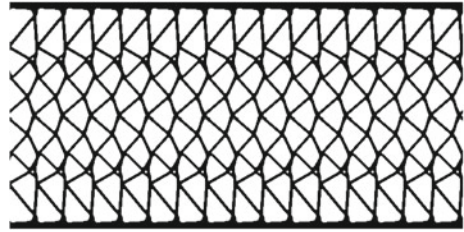
The second example, shown in Fig. 7, is the design of a cantilever beam with two load cases. For the first load case the only active force is  $P_1$  and for the second case  $P_2$ . The coarse mesh consists of  $16 \times 8$  coarse cells and periodicity is enforced only in the horizontal direction. The filtering step is performed using the PDE filter with filter parameter  $R = 4h$ , where  $h$  is the characteristic length of the fine mesh. Each coarse cell consists of  $40 \times 40$  elements. The weight coefficient in the stochastic formulation is set to 1.0. The volume fraction is 30% of the design domain volume. The Poisson's ratio is set to  $\nu = 0.3$ , the modulus of elasticity for the solid is set to  $E_{\max} = 1$ , and the modulus of elasticity for the void material is set to  $E_{\min} = 10^{-9}$ . The rest of the parameters are set to be the same as for the thermal case, except that the final value of  $p$  is set to 5.0. The optimization problem is given as

$$\begin{aligned} \min_{\rho} : c &= \sum_{i=1}^{n_l} E[\mathbf{f}_i^T \mathbf{u}_i] + w\sqrt{\text{Var}[\mathbf{f}_i^T \mathbf{u}_i]} & (30) \\ \text{s.t. } \mathbf{K}\mathbf{u}_i &= \mathbf{f}_i, \quad i = 1, \dots, n_l \\ E[\rho^T \mathbf{v}] &\leq V^* \\ 0 \leq \rho_i &\leq 1 \quad i = 1, \dots, n_{\text{el}} \end{aligned}$$

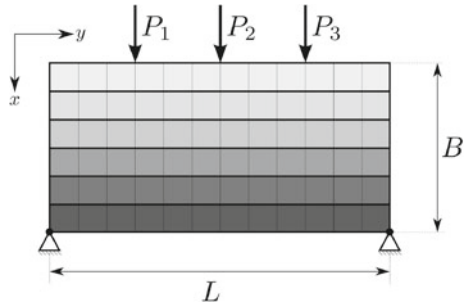
where  $n_l$  is the number of load cases.

The optimized design is shown in Fig. 8. In contrast to the designs obtained for a single active load presented in [3], the obtained design is symmetric with respect to the horizontal mid-axis and the microstructural details closely resembles triangular truss structures. Triangular truss-like structures are optimal for problems with changes of the principal stress orientation for the different load cases. For a single load case without any restrictions on the design pattern, the optimal design will follow the principal stress trajectories. The mean compliance is 2.9 for both load cases. Optimization for a single load case resulted in a mean compliance of 2.5 which as expected is better for that particular load case, and worse for the load in the other direction yielding a compliance of 3.8.

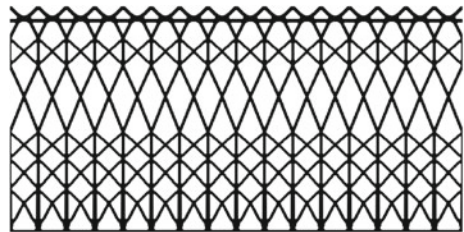
**Fig. 8** Intermediate design realization  $\eta = 0.5$  for optimized multiple load cases cantilever beam problem



**Fig. 9** Boundary conditions and design domain of layered beam problem with multiple load cases

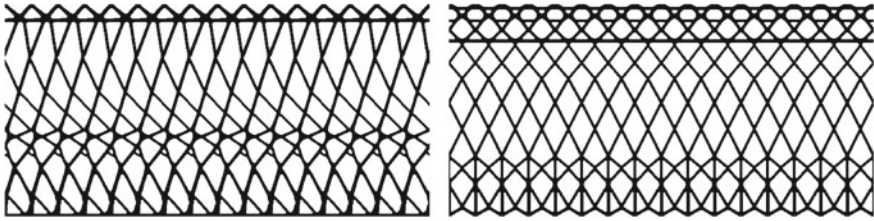


**Fig. 10** Intermediate design realization  $\eta = 0.5$  for optimized multiple load case beam problem



The third example, shown in Fig. 9, is the design of a simply supported beam with three load cases. The optimization setup parameters are the same as for the previous example. The optimized intermediate design realization is shown in Fig. 10. The multiload case design shares some similarities to the single load case with central active force  $P_2 = 1$  shown in Fig. 11. However, a cross-check of the designs show that it performs better for all three cases in contrast to the single load designs which perform well only for the corresponding design case. Requiring periodicity in the horizontal direction implicitly ensures some robustness of the  $P_2$  single load case with respect to a shift of the applied load with a single or multiple coarse cells. This property is not shared for the single load designs obtained for  $P_1$  or  $P_3$ . The microstructural details vary along in the vertical direction, however, some of the layers show similar topology with small variations.

The periodicity requirement implicitly imposes a maximum length scale on the design [3] as it requires the material to be distributed regularly along the design domain. Removing the periodicity requirement in the horizontal direction would provide additional freedom to the optimizer and would allow more material to be



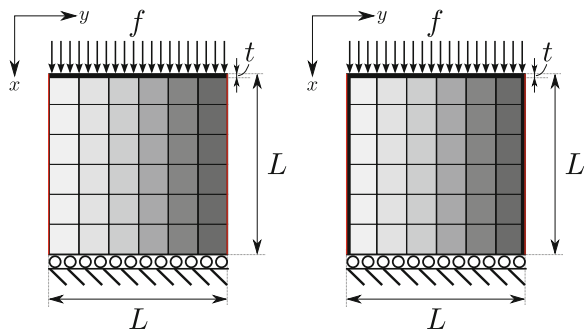
**Fig. 11** Intermediate design realization  $\eta = 0.5$  for optimized single load case beam problems— $P_1$  is active on the first design (left) and  $P_2$  is active on the second design (right)

concentrated in the central areas of the beam, which will result in better performance. Therefore, such restrictions on the design space should be imposed only for manufacturing, aesthetic or other reasons not related directly to the optimality of the design. As mentioned earlier, the computational cost of the coarse basis increases with increasing the design freedom. However, for multiple load cases the basis is utilized for multiple solutions which makes the approach even more competitive compared to the alternatives.

### 6.3 Linear Elastic Designs with Zero and Negative Expansion

The final example is topology optimization of a linear elastic compression test with restrictions on the horizontal displacements. The boundary conditions and the design domains are shown in Fig. 12. Two cases are considered: for the first, a solid region of thickness  $t = 0.0125$  is enforced only on the top of the design domain, and for the second, the solid region is enforced on the horizontal edges as well. The design domain is partitioned using  $8 \times 8$  coarse cells with design symmetry with respect to the vertical axis. Each coarse cell is discretized using  $40 \times 40$  finite elements. The filtering is performed by the PDE filter with parameter  $R = 5h$ . The dilated, intermediate and eroded design thresholds are set to 0.4, 0.5 and 0.6, respectively.

**Fig. 12** Boundary conditions and design domains for compression tests with restrictions on the horizontal displacements. Solid regions marked with thick black line are enforced on the top edge in the first case (left) and also on the horizontal edges in the second case (right)





The dimension of the design domain is set to  $L = 2$ . Distributed load of total size  $10^{-3}$  is applied on the upper edge of the design for the two cases. The material volume is restricted to be 50% of the design domain volume. The penalization is set to  $p = 5$  and the projection parameter  $\beta$  is increased from 8 to 32 after the first 150 iterations. The rest of the optimization parameters are the same as for the previous example.

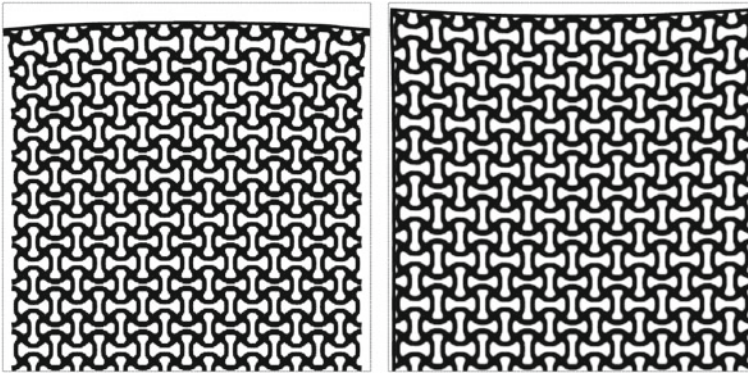
The optimization formulation in discrete form is given as

$$\begin{aligned} \min_{\rho} : c &= E \left[ \mathbf{f}^T \mathbf{u} \right] + w \sqrt{\text{Var} \left[ \mathbf{f}^T \mathbf{u} \right]} & (31) \\ \text{s.t. } \mathbf{K} \mathbf{u} &= \mathbf{f} \\ E \left[ \rho^T \mathbf{v} \right] &\leq V^* \\ \bar{\mathbf{u}}_j + \sigma_j - u_{\text{ref}} &\leq \varepsilon_{\text{con}}, j = e, i, d \\ \bar{\mathbf{u}}_j - \sigma_j - u_{\text{ref}} &\geq \varepsilon_{\text{con}}, j = e, i, d \\ 0 \leq \rho_i &\leq 1 \quad i = 1, \dots, n_{\text{el}} \end{aligned}$$

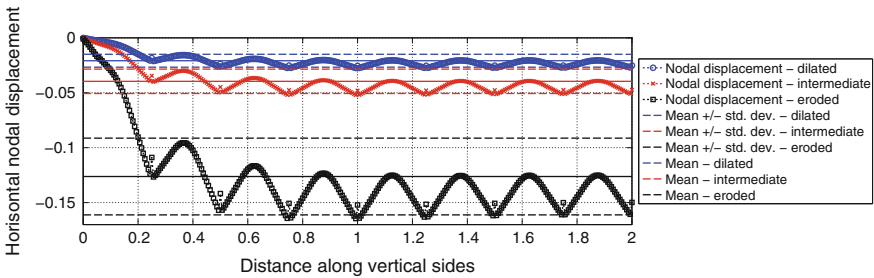
where the expectation and the variance in the objective are approximated using only three realizations: the most eroded case  $\eta_e = 0.6$ , the most dilated case  $\eta_d = 0.4$  and the intermediate case for  $\eta_i = 0.5$ . The final results are verified using Monte Carlo simulations. The objective is to minimize the compliance of the system with constraints on the horizontal displacements along the vertical edges, where  $\bar{u}_j$ ,  $j = e, i, d$ , is the average displacement along the horizontal edge for the eroded, intermediate and dilated realizations, respectively,  $\sigma_j$  is the standard deviation of the horizontal displacements along the edge for realization  $j$ , and  $\varepsilon_{\text{con}}$  is a prescribed tolerance.

The initial design is obtained by repetition of a unit cell negative Poisson's ratio design from [35]. The unit cell design is robust with respect to uniform erosion and dilation. Deformed structures for the considered cases are shown in Fig. 13. The global behavior of the two structures differs significantly due to the difference in the boundaries. For the first case of unframed design, the bulk material is free to contract and the negative Poisson's effect can be clearly seen. In the second case, the stiff frame around the bulk material restrains the horizontal movement which lowers the Poisson's effect and adds additional stiffness to the structure in the vertical direction. This results in lower vertical displacements of the upper edge. The displacements along the vertical edge for framed and unframed designs are shown in Figs. 14 and 15. For the unframed design, the horizontal displacements for the three realizations are large and negative as expected from the homogenized material properties. However, for the framed design, shown in Fig. 15, the horizontal displacements for the eroded and dilated cases are significantly smaller.

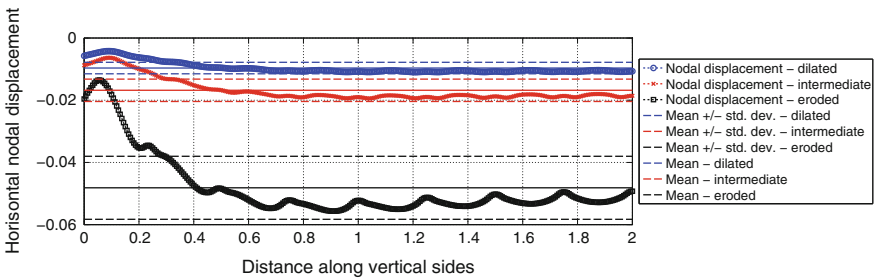
As demonstrated above, optimized microstructure designs for a selected material property might lead to different global responses for equivalent load patterns due to small differences in cells close to boundaries of the design. Classical homogenization theory [32] does not take into account the boundary conditions and localized effects.



**Fig. 13** Deformed structures for unframed (*left*) and boxed (*right*) design domains with microstructural pattern optimized for negative Poisson’s ratio



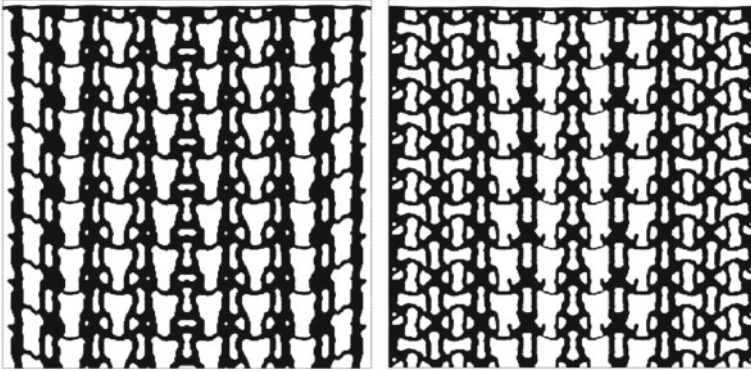
**Fig. 14** Horizontal displacements along the *vertical edge* of the design domain with unframed boundaries for design realizations with thresholds 0.4, 0.5 and 0.6 (dilated, intermediate, and eroded)



**Fig. 15** Horizontal displacements along the *vertical edge* of the design domain with framed boundaries for design realizations with thresholds 0.4, 0.5 and 0.6 (dilated, intermediate, and eroded)

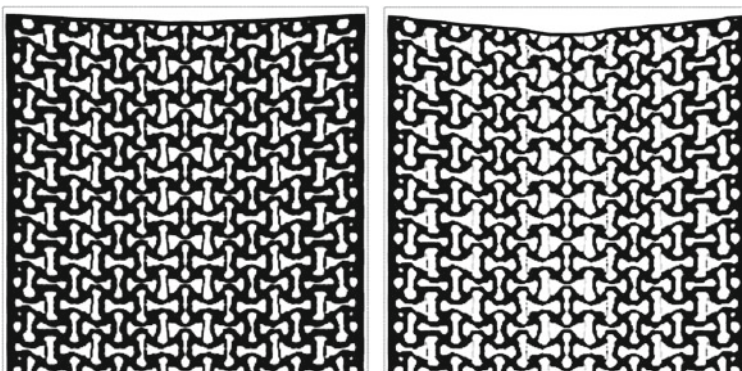
Hence, in all cases where the global structural response is of interest, the boundary effects should be taken into account during the optimization process. As demonstrated here, the proposed MsFEM methodology provides such solution at a relatively low computational cost.





**Fig. 16** Deformed structures for unframed (*left*) and framed (*right*) design domains with microstructural patterns optimized for tailoring macroscale response. The reference horizontal displacement along the *vertical edge* is zero

Topology optimized design, using the formulation given by Eq. 31, with zero reference displacement  $u_{\text{ref}} = 0$  and  $\varepsilon_{\text{con}} = 10^{-4}$ , are shown in Fig. 16. The microstructures differs significantly close to the vertical edges, which demonstrates the need to account for boundary effects in the design process. In the first case, the optimization utilizes the fact that solid material is not required along the vertical edge and shifts the force transmitting structure from the boundary. In the second case, a complex microstructure is designed around the solid frame in order to avoid displacements in the horizontal direction. Another important feature observed during the design process, is that the eroded, dilated and intermediate designs might not share the same topology. In such cases length scale cannot be guaranteed on the intermedi-



**Fig. 17** Dilated (*left*) and intermediate (*right*) deformed structures realizations for framed design domain with microstructural patterns optimized for tailoring macroscale response. The reference horizontal displacement along the *vertical edge* is  $u_{\text{ref}} = -0.01$  and  $\varepsilon_{\text{con}} = 10^{-3}$

ate design [34], however, since the design performance is insensitive with respect to small imperfections, removing or adding small features along the perimeter will not change significantly the optimized performance. This property can be clearly observed for the case with negative reference displacement shown in Fig. 17.

## 7 Conclusions

In this book chapter, a methodology has been demonstrated for tailoring macroscale responses of mechanical and heat transfer systems by topology optimization of microstructural details. These details are herein restricted to full periodicity or grading in a single direction. For a heat transfer problem, increased periodicity is shown to aid the optimization objective, and for certain elastic structures with multiple load cases it is shown that partial periodicity can provide an implicit robustness to load position. Finally, it has been demonstrated that it is important to take the boundary effects and finite size microstructural details into account during the optimization process in order to tailor the macroscopic response. These details can be easily accounted for by the proposed multiscale approach. The spectral MsFEM for high-contrast problems reduces the computational cost and allows for the optimization of large resolution models within a reasonable amount of time.

**Acknowledgments** Both authors were funded by Villum Fonden through the NextTop project, as well as the EU FP7-MC-IAPP programme LaScISO. The authors would like to thank Dr. Fengwen Wang for providing them with optimized periodic microstructural design of negative Poisson's ratio material utilized as initial guess in the last example.

## References

1. Aage N, Andreassen E, Lazarov BS (2014) Topology optimization using PETSc: an easy-to-use, fully parallel, open source topology optimization framework. *Struct Multi Optim* 1–8. doi:[10.1007/s00158-014-1157-0](https://doi.org/10.1007/s00158-014-1157-0)
2. Aage N, Lazarov B (2013) Parallel framework for topology optimization using the method of moving asymptotes. *Struct Multi Optim* 47(4):493–505. doi:[10.1007/s00158-012-0869-2](https://doi.org/10.1007/s00158-012-0869-2)
3. Alexandersen J, Lazarov BS (2015) Topology optimisation of manufacturable microstructural details without length scale separation using a spectral coarse basis preconditioner. *Comput Methods Appl Mech Eng* 290(1):156–182. doi:[10.1016/j.cma.2015.02.028](https://doi.org/10.1016/j.cma.2015.02.028)
4. Amir O, Aage N, Lazarov BS (2014) On multigrid-CG for efficient topology optimization. *Struct Multi Optim* 49(5):815–829 (2014). doi:[10.1007/s00158-013-1015-5](https://doi.org/10.1007/s00158-013-1015-5)
5. Andreassen E, Lazarov BS, Sigmund O (2014) Design of manufacturable 3d extremal elastic microstructure. *Mech Mater* 69(1):1–10. doi:[10.1016/j.mechmat.2013.09.018](https://doi.org/10.1016/j.mechmat.2013.09.018)
6. Bendsoe MP, Kikuchi N (1988) Generating optimal topologies in structural design using a homogenization method. *Comput Methods Appl Mech Eng* 71(2):197–224. doi:[10.1016/0045-7825\(88\)90086-2](https://doi.org/10.1016/0045-7825(88)90086-2)
7. Bendsoe MP, Sigmund O (2003) *Topology optimization—Theory, methods and applications*. Springer, Berlin

8. Bourdin B (2001) Filters in topology optimization. *Int J Numer Methods Eng* 50:2143–2158
9. Braess D (2007) *Finite elements: theory, fast solvers, and applications in solid mechanics*. Cambridge University Press, Cambridge
10. Bruns TE, Tortorelli DA (2001) Topology optimization of non-linear elastic structures and compliant mechanisms. *Comput Methods Appl Mech Eng* 190:3443–3459
11. Buck M, Iliev O, Andrä H (2013) Multiscale finite element coarse spaces for the application to linear elasticity. *Central European Journal of Mathematics* 11(4):680–701. doi:[10.2478/s11533-012-0166-8](https://doi.org/10.2478/s11533-012-0166-8)
12. Coelho P, Fernandes P, Guedes J, Rodrigues H (2008) A hierarchical model for concurrent material and topology optimisation of three-dimensional structures. *Struct Multi Optim* 35:107–115. doi:[10.1007/s00158-007-0141-3](https://doi.org/10.1007/s00158-007-0141-3)
13. Coelho P, Fernandes P, Rodrigues H, Cardoso J, Guedes J (2009) Numerical modeling of bone tissue adaptation: a hierarchical approach for bone apparent density and trabecular structure. *J Biomech* 42(7):830–837. doi:[10.1016/j.jbiomech.2009.01.020](https://doi.org/10.1016/j.jbiomech.2009.01.020)
14. Deaton J, Grandhi R (2014) A survey of structural and multidisciplinary continuum topology optimization: post 2000. *Struct Multi Optim* 49(1):1–38. doi:[10.1007/s00158-013-0956-z](https://doi.org/10.1007/s00158-013-0956-z)
15. Efendiev Y, Galvis J (2011) A domain decomposition preconditioner for multiscale high-contrast problems. In: Huang Y, Kornhuber R, Widlund O, Xu J, Barth TJ, Griebel M, Keyes DE, Nieminen RM, Roose D, Schlick T (eds) *Domain decomposition methods in science and engineering XIX*, lecture notes in computational science and engineering. Springer, Berlin, pp 189–196. doi:[10.1007/978-3-642-11304-820](https://doi.org/10.1007/978-3-642-11304-820)
16. Efendiev Y, Galvis J, Hou TY (2013) Generalized multiscale finite element methods (gmsfem). *J Comput Phys* 251(0):116–135. doi:[10.1016/j.jcp.2013.04.045](https://doi.org/10.1016/j.jcp.2013.04.045)
17. Efendiev Y, Galvis J, Lazarov R, Willems J (2012) Robust domain decomposition preconditioners for abstract symmetric positive definite bilinear forms. *ESAIM: Math Model Numer Anal* 46:1175–1199
18. Efendiev Y, Galvis J, Wu XH (2011) Multiscale finite element methods for high-contrast problems using local spectral basis functions. *J Comput Phys* 230(4):937–955. doi:[10.1016/j.jcp.2010.09.026](https://doi.org/10.1016/j.jcp.2010.09.026)
19. Efendiev Y, Hou TY (2009) *Multiscale finite element methods: theory and applications*. Springer, Berlin
20. Galvis J, Efendiev Y (2010) Domain decomposition preconditioners for multiscale flows in high-contrast media. *Multiscale Model Simul* 8(4):1461–1483. doi:[10.1137/090751190](https://doi.org/10.1137/090751190)
21. Jansen M, Lazarov B, Schevenels M, Sigmund O (2013) On the similarities between micro/nano lithography and topology optimization projection methods. *Struct Multi Optim* 48(4):717–730. doi:[10.1007/s00158-013-0941-6](https://doi.org/10.1007/s00158-013-0941-6)
22. Lazarov B (2014) Topology optimization using multiscale finite element method for high-contrast media In: Lirkov I, Margenov S, Waniowski J (eds) *Large-scale scientific computing, lecture notes in computer science*, pp 339–346. Springer, Berlin. doi:[10.1007/978-3-662-43880-038](https://doi.org/10.1007/978-3-662-43880-038)
23. Lazarov BS, Schevenels M, Sigmund O (2012) Topology optimization considering material and geometric uncertainties using stochastic collocation methods. *Struct Multi Optim* 46:597–612. doi:[10.1007/s00158-012-0791-7](https://doi.org/10.1007/s00158-012-0791-7)
24. Lazarov BS, Sigmund O (2011) Filters in topology optimization based on Helmholtz-type differential equations. *Int J Numer Meth Eng* 86(6):765–781. doi:[10.1002/nme.3072](https://doi.org/10.1002/nme.3072)
25. Maitre OPL, Knio OM (2010) *Spectral Methods for uncertainty quantification: with applications to computational fluid dynamics*. Springer, Berlin
26. Saad Y (2003) *Iterative methods for sparse linear systems*. SIAM, Philadelphia
27. Schevenels M, Lazarov B, Sigmund O (2011) Robust topology optimization accounting for spatially varying manufacturing errors. *Comput Meth Appl Mech Eng* 200(49–52):3613–3627. doi:[10.1016/j.cma.2011.08.006](https://doi.org/10.1016/j.cma.2011.08.006)
28. Sigmund O (1994) Materials with prescribed constitutive parameters: an inverse homogenization problem. *Int J Sol Struct* 31(17):2313–2329. doi:[10.1016/0020-7683\(94\)90154-6](https://doi.org/10.1016/0020-7683(94)90154-6)

29. Sigmund O (1995) Tailoring materials with prescribed elastic properties. *Mech Mater* 20(4):351–368. doi:[10.1016/0167-6636\(94\)00069-7](https://doi.org/10.1016/0167-6636(94)00069-7)
30. Sigmund O, Maute K (2013) Topology optimization approaches. *Struct Multi Optim* 48(6):1031–1055. doi:[10.1007/s00158-013-0978-6](https://doi.org/10.1007/s00158-013-0978-6)
31. Svanberg K (1987) The method of moving asymptotes - a new method for structural optimization. *Int J Numer Meth Eng* 24:359–373
32. Torquato S (2002) *Random heterogeneous materials*. Springer, Berlin
33. Vassilevski PS (2008) *Multilevel block factorization preconditioners: matrix-based analysis and algorithms for solving finite element equations*. Springer, New York
34. Wang F, Lazarov B, Sigmund O (2011) On projection methods, convergence and robust formulations in topology optimization. *Struct Multidi Optim* 43(6):767–784. doi:[10.1007/s00158-010-0602-y](https://doi.org/10.1007/s00158-010-0602-y)
35. Wang F, Sigmund O, Jensen JS (2014) Design of materials with prescribed nonlinear properties. *J Mech Phys Sol* 69(1):156–174. doi:[10.1016/j.jmps.2014.05.003](https://doi.org/10.1016/j.jmps.2014.05.003)
36. Zhou M, Lazarov BS, Sigmund O (2014) Topology optimization for optical projection lithography with manufacturing uncertainties. *Appl Opt* 53(12):2720–2729. doi:[10.1364/AO.53.002720](https://doi.org/10.1364/AO.53.002720)

# Aerodynamic Shape Optimization Using “Turbulent” Adjoint And Robust Design in Fluid Mechanics

Kyriakos C. Giannakoglou, Dimitrios I. Papadimitriou,  
Evangelos M. Papoutsis-Kiachagias and Ioannis S. Kavvadias

**Abstract** This article presents adjoint methods for the computation of the first- and higher-order derivatives of objective functions  $F$  used in optimization problems governed by the Navier–Stokes equations in aero/hydrodynamics. The first part of the chapter summarizes developments and findings related to the application of the continuous adjoint method to turbulence models, such as the Spalart–Allmaras and  $k$ - $\varepsilon$  ones, in either their low- or high-Reynolds number (with wall functions) variants. Differentiating the turbulence model, over and above to the differentiation of the mean–flow equations, leads to the computation of the exact gradient of  $F$ , by overcoming the frequently made assumption of neglecting turbulence variations. The second part deals with higher-order sensitivity analysis based on the combined use of the adjoint approach and the direct differentiation of the governing PDEs. In robust design problems, the so-called second-moment approach requires the computation of second-order derivatives of  $F$  with respect to (w.r.t.) the environmental or uncertain variables; in addition, any gradient-based optimization algorithm requires third-order mixed derivatives w.r.t. both the environmental and design variables; various ways to compute them are discussed and the most efficient is adopted. The equivalence of the continuous and discrete adjoint for this type of computations is demonstrated. In the last part, some other relevant recent achievements regarding the adjoint approach are discussed. Finally, using the aforementioned adjoint methods, industrial geometries are optimized. The application domain includes both incompressible or compressible fluid flow applications.

---

K.C. Giannakoglou (✉) · D.I. Papadimitriou · E.M. Papoutsis-Kiachagias · I.S. Kavvadias  
National Technical University of Athens, Athens, Greece  
e-mail: kgianna@central.ntua.gr

D.I. Papadimitriou  
e-mail: dpapadim@mail.ntua.gr

E.M. Papoutsis-Kiachagias  
e-mail: vaggelisp@gmail.com

I.S. Kavvadias  
e-mail: kavvadiasj@hotmail.com

## 1 Flow Equations and Objective Function

In this section, the equations governing the state (i.e. flow) problem for incompressible fluid flows, using either the one-equation Spalart-Allmaras [1] or the two equation Launder-Sharma  $k - \varepsilon$  [2] turbulence models, are briefly presented. The mean-flow state equations are

$$R^p = -\frac{\partial v_j}{\partial x_j} = 0 \quad (1)$$

$$R^{v_i} = v_j \frac{\partial v_i}{\partial x_j} + \frac{\partial p}{\partial x_i} - \frac{\partial}{\partial x_j} \left[ (v + \nu_t) \left( \frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) \right] = 0 \quad (2)$$

where  $v_i$  are the velocity components,  $p$  the static pressure divided by the density,  $\nu$  and  $\nu_t$  the bulk and turbulent viscosities. The turbulence model (TM) equation(s) is/are

$$R^{\tilde{v}} = v_j \frac{\partial \tilde{v}}{\partial x_j} - \frac{\partial}{\partial x_j} \left[ \left( \nu + \frac{\tilde{v}}{\sigma} \right) \frac{\partial \tilde{v}}{\partial x_j} \right] - \frac{c_{b2}}{\sigma} \left( \frac{\partial \tilde{v}}{\partial x_j} \right)^2 - \tilde{v} P(\tilde{v}) + \tilde{v} D(\tilde{v}) = 0 \quad (3)$$

for the Spalart-Allmaras model (TM=SA) and

$$\begin{aligned} R^k &= v_j \frac{\partial k}{\partial x_j} - \frac{\partial}{\partial x_j} \left[ \left( \nu + \frac{\nu_t}{Pr_k} \right) \frac{\partial k}{\partial x_j} \right] - P_k + \varepsilon + D = 0 \\ R^\varepsilon &= v_j \frac{\partial \varepsilon}{\partial x_j} - \frac{\partial}{\partial x_j} \left[ \left( \nu + \frac{\nu_t}{Pr_\varepsilon} \right) \frac{\partial \varepsilon}{\partial x_j} \right] - c_1 P_k \frac{\varepsilon}{k} + c_2 f_2 \frac{\varepsilon^2}{k} - E = 0 \end{aligned} \quad (4)$$

for the Launder-Sharma  $k - \varepsilon$  (TM=KE) one.  $\tilde{v}$  is the turbulence state variable if TM=SA ( $\nu_t = \tilde{v} f_{\nu_t}$ ) and  $k$ ,  $\varepsilon$  are the corresponding quantities (turbulent kinetic energy and turbulent energy dissipation) if TM=KE ( $\nu_t = c_\mu \frac{k^2}{\varepsilon}$ ). In both cases, the boundary conditions and the model constant values are omitted in the interest of space; see [1] and [2].

In general, the objective function may comprise both surface ( $S$ ) and volume ( $\Omega$ ) integrals, as follows

$$F = \int_S F_S dS + \int_\Omega F_\Omega d\Omega = \int_S F_{S_i} n_i dS + \int_\Omega F_\Omega d\Omega \quad (5)$$

where  $n_i$  are the components of the normal to the boundary outward unit vector.

## 2 The Adjoint Method for Shape Optimization in Turbulent Flows

In discrete adjoint, the differentiation of the turbulence model equations is straightforward and can be found in several published works, [3, 4]. In contrast, the majority of existing continuous adjoint methods/codes rely on the so-called “frozen turbulence” assumption, in which the sensitivities of the turbulence quantities w.r.t. the design variables are neglected [5–9]. The first continuous adjoint to the Spalart-Allmaras model, for incompressible flows, was presented by the current group of authors in [10] and was extended to compressible flows in [11]. Regarding the adjoint approach to high-Reynolds turbulence models, the continuous adjoint to the  $k-\varepsilon$  model with wall functions has recently been published, [12], whereas the continuous adjoint to the low-Reynolds Launder-Sharma  $k-\varepsilon$  model can be found in [13]. All these adjoint approaches which rely upon the differentiated turbulence model will hereafter be referred to as “turbulent” adjoint, to distinguish it from the “frozen turbulence” approach.

### 2.1 Continuous Adjoint to Low-Re Turbulence Models

In the continuous adjoint approach for shape optimization problems, the total derivative (symbol  $\delta$ ) of any function  $\Phi$  w.r.t. the design variables  $b_n$  must be distinguished from the corresponding partial sensitivity (symbol  $\partial$ ) since

$$\frac{\delta\Phi}{\delta b_n} = \frac{\partial\Phi}{\partial b_n} + \frac{\partial\Phi}{\partial x_l} \frac{\delta x_l}{\delta b_n} \quad (6)$$

where  $\frac{\delta x_l}{\delta b_n}$  are the sensitivities of nodal coordinates. In case  $\Phi$  is defined along a surface, Eq. 6 becomes  $\frac{\delta_s\Phi}{\delta b_n} = \frac{\partial\Phi}{\partial b_n} + \frac{\partial\Phi}{\partial x_k} n_k \frac{\delta x_m}{\delta b_n} n_m$ . Since any sufficiently small surface deformation can be seen as a normal perturbation, only the normal part of the surface deformation velocity  $\delta x_m/\delta b_n n_m$  contributes to changes in  $\Phi$ .

In order to formulate the adjoint method, the augmented objective function  $F_{aug}$  is defined as the sum of  $F$  and the field integrals of the products of the adjoint variable fields and the state equations, as follows

$$F_{aug} = F + \int_{\Omega} u_i R_i^u d\Omega + \int_{\Omega} q R^p d\Omega + E_{TM} \quad (7)$$

where  $u_i$  are the adjoint velocity components,  $q$  the adjoint pressure and the extra terms  $E_{TM}$  depend on the turbulence model ( $TM$ ). If  $TM = SA$ ,

$$E_{SA} = \int_{\Omega} \tilde{v}_a R^{\tilde{v}} d\Omega \quad (8)$$

whereas if  $TM = KE$ ,

$$E_{KE} = \int_{\Omega} \left( k_a R^k + \varepsilon_a R^\varepsilon \right) d\Omega \quad (9)$$

where  $\tilde{v}_a$ ,  $k_a$  and  $\varepsilon_a$  are the adjoints to  $\tilde{v}$ ,  $k$  and  $\varepsilon$ , respectively.

Based on the Leibniz theorem, the derivative of  $F_{aug}$  w.r.t.  $b_n$  is

$$\begin{aligned} \frac{\delta F_{aug}}{\delta b_n} &= \frac{\delta F}{\delta b_n} + \int_{\Omega} u_i \frac{\partial R_i^v}{\partial b_n} d\Omega + \int_{\Omega} q \frac{\partial R^p}{\partial b_n} d\Omega \\ &+ \int_{S_{W_p}} (u_i R_i^v + q R^p) \frac{\delta x_k}{\delta b_n} n_k dS + \frac{\delta(E_{TM})}{\delta b_n} \end{aligned} \quad (10a)$$

$$\frac{\delta(E_{SA})}{\delta b_n} = \int_{\Omega} \tilde{v}_a \frac{\partial R^{\tilde{v}}}{\partial b_n} d\Omega + \int_{S_{W_p}} \tilde{v}_a R^{\tilde{v}} \frac{\delta x_k}{\delta b_n} n_k dS \quad (10b)$$

$$\frac{\delta(E_{KE})}{\delta b_n} = \int_{\Omega} k_a \frac{\partial R^k}{\partial b_n} d\Omega + \int_{\Omega} \varepsilon_a \frac{\partial R^{\varepsilon_a}}{\partial b_n} d\Omega + \int_{S_{W_p}} (k_a R^k + \varepsilon_a R^\varepsilon) \frac{\delta x_k}{\delta b_n} n_k dS \quad (10c)$$

where  $S_{W_p}$  is the parameterized (in terms of  $b_n$ ) part of the solid wall. The development of the volume integrals in Eqs. 10a, b, based on the Green-Gauss theorem and the elimination of terms depending on the variations of the mean-flow and turbulence model variables, lead to the adjoint mean-flow equations

$$R^q = \frac{\partial u_j}{\partial x_j} = 0 \quad (11)$$

$$R_i^u = u_j \frac{\partial v_j}{\partial x_i} - \frac{\partial(v_j u_i)}{\partial x_j} - \frac{\partial}{\partial x_j} \left[ (v + v_t) \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \right] + \frac{\partial q}{\partial x_i} + AMS_i = 0 \quad (12)$$

The extra terms  $AMS_i$  arise from the differentiation of the turbulence model, see [10, 13]. The adjoint turbulence model variables fields  $\tilde{v}_a$ ,  $k_a$  and  $\varepsilon_a$  are governed by the ‘‘turbulent’’ adjoint PDEs, which are

$$\begin{aligned} R^{\tilde{v}_a} &= -\frac{\partial(v_j \tilde{v}_a)}{\partial x_j} - \frac{\partial}{\partial x_j} \left[ \left( v + \frac{\tilde{v}}{\sigma} \right) \frac{\partial \tilde{v}_a}{\partial x_j} \right] + \frac{1}{\sigma} \frac{\partial \tilde{v}_a}{\partial x_j} \frac{\partial \tilde{v}}{\partial x_j} + 2 \frac{cb_2}{\sigma} \frac{\partial}{\partial x_j} \left( \tilde{v}_a \frac{\partial \tilde{v}}{\partial x_j} \right) \\ &+ \tilde{v}_a \tilde{v} C_{\tilde{v}} + \frac{\partial v_t}{\partial \tilde{v}} \frac{\partial u_i}{\partial x_j} \left( \frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) + (-P + D) \tilde{v}_a = 0 \end{aligned} \quad (13a)$$

$$\begin{aligned} R^{k_a} &= -\frac{\partial(v_j k_a)}{\partial x_j} - \frac{\partial}{\partial x_j} \left[ \left( v + \frac{v_t}{Pr_k} \right) \frac{\partial k_a}{\partial x_j} \right] \\ &+ \left( \frac{B_1}{Pr_k} - \frac{v}{k} \right) \frac{\partial k}{\partial x_j} \frac{\partial k_a}{\partial x_j} + \frac{B_1}{Pr_\varepsilon} \frac{\partial \varepsilon}{\partial x_j} \frac{\partial \varepsilon_a}{\partial x_j} + B_1 \left( \frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) \frac{\partial u_i}{\partial x_j} \end{aligned}$$



$$\begin{aligned}
& + \left[ \frac{\nu}{2k^2} \left( \frac{\partial k}{\partial x_j} \right)^2 - \frac{\nu}{k} \frac{\partial^2 k}{\partial x_j^2} - PB_1 \right] k_a \\
& - \left[ c_1 \frac{\varepsilon}{k} PB_1 + 2\nu \left( \frac{\partial^2 \nu k}{\partial x_i \partial x_j} \right)^2 B_1 + c_2 f_2 \frac{\varepsilon^2}{k^2} - 1.2c_2 \frac{k^2}{\nu^2} e^{-Re_t^2} - c_1 P k \frac{\varepsilon}{k^2} \right] \varepsilon_a = 0
\end{aligned} \tag{13b}$$

$$\begin{aligned}
R^{\varepsilon_a} = & - \frac{\partial(\nu_j \varepsilon_a)}{\partial x_j} - \frac{\partial}{\partial x_j} \left[ \left( \nu + \frac{\nu_t}{Pr_\varepsilon} \right) \frac{\partial \varepsilon_a}{\partial x_j} \right] \\
& + \frac{B_2}{Pr_\varepsilon} \frac{\partial \varepsilon}{\partial x_j} \frac{\partial \varepsilon_a}{\partial x_j} + \frac{B_2}{Pr_k} \frac{\partial k}{\partial x_j} \frac{\partial k_a}{\partial x_j} + B_2 \left( \frac{\partial \nu_i}{\partial x_j} + \frac{\partial \nu_j}{\partial x_i} \right) \frac{\partial u_i}{\partial x_j} + (1 - PB_2) k_a \\
& + \left[ -2\nu \left( \frac{\partial^2 \nu k}{\partial x_i \partial x_j} \right)^2 B_2 - c_1 \frac{\varepsilon}{k} PB_2 + 2c_2 f_2 \frac{\varepsilon}{k} - 0.6c_2 \frac{k^3}{\nu^2 \varepsilon} e^{-Re_t^2} - c_1 P k \frac{1}{k} \right] \varepsilon_a = 0
\end{aligned} \tag{13c}$$

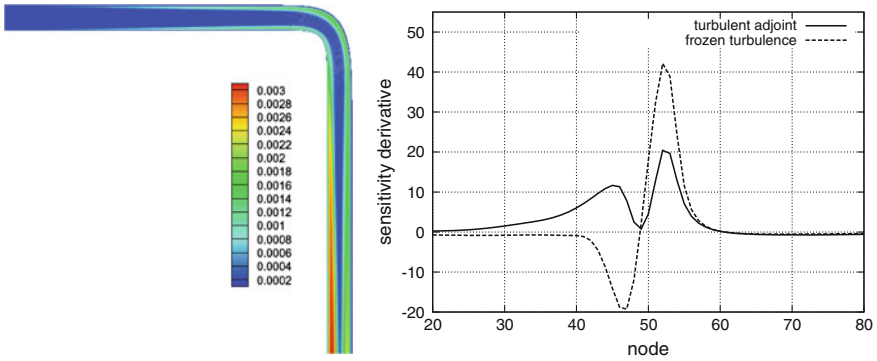
The detailed derivation of the adjoint PDEs, the various terms or constants in Eqs. 13a, b, c and the corresponding adjoint boundary conditions can be found in [10] or [13].

After satisfying the field adjoint equations, the sensitivity derivatives of  $F_{aug}$  are given by

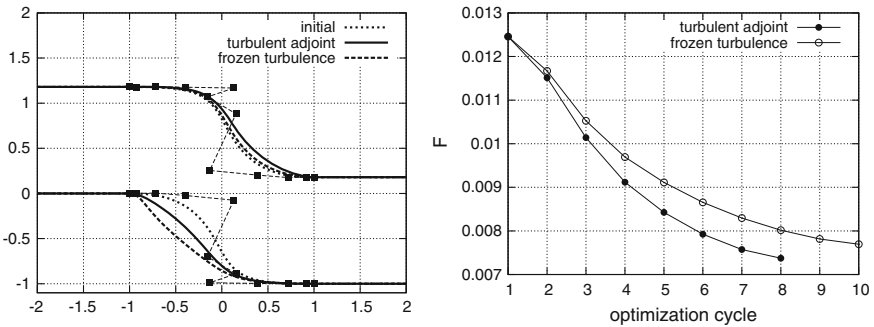
$$\begin{aligned}
\frac{\delta F_{aug}}{\delta b_n} = & \int_S BC_i^u \frac{\partial \nu_i}{\partial b_n} dS + \int_S (u_j n_j + \frac{\partial F_{S_i}}{\partial p} n_i) \frac{\partial p}{\partial b_n} dS + \int_S (-u_i n_j + \frac{\partial F_{S_k}}{\partial \tau_{ij}} n_k) \frac{\partial \tau_{ij}}{\partial b_n} dS \\
& + \int_{S_{W_p}} n_i \frac{\partial F_{S_{W_p,i}}}{\partial x_m} n_m \frac{\delta x_k}{\delta b_n} n_k dS + \int_{S_{W_p}} F_{S_{W_p,i}} \frac{\delta n_i}{\delta b_n} dS + \int_{S_{W_p}} F_{S_{W_p,i}} n_i \frac{\delta(dS)}{\delta b_n} \\
& + \int_{S_{W_p}} (u_i R_i^v + q R^p) \frac{\delta x_k}{\delta b_n} n_k dS + SD
\end{aligned} \tag{14}$$

where, depending on the turbulence model, terms  $BC_i^u$  and  $SD$  can be found in [10] or [13]. The gain from using the “turbulent” adjoint approach and overcoming the “frozen turbulence” assumption, at the expense of additionally solving the adjoint to the turbulence model PDEs, is demonstrated below in a few selected cases. The “frozen turbulence” assumption may lead to wrongly signed sensitivities, misleading or delaying the optimization process. As an example, the optimization of a 90° elbow duct, targeting minimum total pressure losses,  $min F = - \int_{S_I} (p + \frac{1}{2} v^2) v_i n_i dS - \int_{S_O} (p + \frac{1}{2} v^2) v_i n_i dS$ , where  $S_I$  and  $S_O$  are the inlet to and outlet from the flow domain, with a Reynolds number equal to  $3.5 \times 10^4$ , modeled using TM=SA model is demonstrated in Fig. 1, [10]. Comments can be found in the caption.

The shape optimization of an S-shaped duct, with the same target as before, is demonstrated in Fig. 2. The flow Reynolds number based on the inlet height is  $Re = 1.2 \times 10^5$  and TM=KE is used. The upper and lower duct contours are parameterized using Bézier–Bernstein polynomials with 12 control points each. The Fletcher-Reeves Conjugate Gradient (CG) method is used. The gradients used by

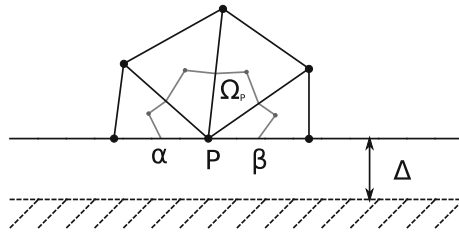


**Fig. 1** Adjoint to the low-Re Spalart–Allmaras model: *Left* adjoint pressure field in a 90° elbow duct with constant cross-section. *Right* sensitivity derivatives of the total pressure losses function ( $\delta F/\delta b_n$ ), where  $b_n$  are the normal displacements of the solid wall grid nodes. Two sensitivity distributions, close to the 90° bend, are compared. The abscissa stands for the nodal numbers of the wall nodes. By making the “frozen turbulence” assumption, wrongly signed sensitivities between nodes 20 and 50 are computed. Extensive validation of the adjoint solver against direct differentiation is conducted in [10]



**Fig. 2** Adjoint to the low-Re Launder–Sharma  $k-\varepsilon$  model: Shape optimization of an S-shaped duct targeting minimum total pressure losses. *Left* Starting duct shape compared to the optimal shapes resulting from **a** “turbulent” adjoint and **b** adjoint based on the “frozen turbulence” assumption; axes not in scale. *Right* Convergence history of the CG algorithm driven by the two different adjoint methods. From [13]

each method to update the design variables are based on (a) “turbulent” adjoint and (b) adjoint with the “frozen turbulence” assumption. The starting duct shape along with the optimal ones computed by CG, based on the two variants of the adjoint formulation, are presented in Fig. 2. The shape resulting from (a) has an  $F$  value by about 3 % lower than that of (b) and reaches the optimal solution after ~20 % less cycles.



**Fig. 3** The adjoint technique with wall functions: A vertex-centered finite volume  $\Omega_P$  associated with the “wall” (horizontal line) at node  $P$ . The real solid wall lies underneath  $P$ , at a distance  $\Delta$

### 2.2 Continuous Adjoint to High-Re Turbulence Models

In industrial projects, many analysis codes rely on the use of the wall function (WF) technique, due to the less stretched and generally coarser meshes required close to the walls and the resulting economy in the overall CPU cost. The development of the adjoint approach to the wall function model is, thus, necessary. This is briefly presented below for the  $k-\varepsilon$  and the Spalart-Allmaras models. The two developments differ since the first was based on the in-house GPU-enabled RANS solver, [14], with slip velocity at the wall, [15], while the second on the OpenFOAM code with a no-slip condition at the wall. Note that these differences in the primal boundary conditions at the wall cause differences in the corresponding adjoint boundary conditions.

Regarding the  $k-\varepsilon$  model, the development, which was carried out by the authors’ group [12] was based on vertex-centered finite volumes with non-zero slip velocity at the wall. The real solid wall is assumed to lie at a distance  $\Delta$  underneath  $S_W$ . Integrating the state equations over the finite volume of Fig. 3, the diffusive flux across segment  $\alpha\beta$  depends on the friction velocity  $v_\tau$ ,

$$v_\tau^2 = (v + v_t) \left( \frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) n_j t_i \tag{15}$$

and  $v_t = v_t i_i$ ;  $v_t$  computed via the local application of the law of the wall.

With known  $v_\tau$ , the  $k$  and  $\varepsilon$  values at  $P$  are

$$\begin{aligned} k_P &= \frac{v_\tau^2}{\sqrt{c_\mu}}, & \varepsilon_P &= \frac{v_\tau^3}{\kappa \Delta}, & \text{if } y^+ &\geq y_c^+ \\ k_P &= \frac{v_\tau^2}{\sqrt{c_\mu}} \left( \frac{y^+}{y_c^+} \right)^2, & \varepsilon_P &= k_P^{\frac{3}{2}} \frac{1 + \frac{5.3v}{\sqrt{k_P \Delta}}}{\kappa c_\mu^{-\frac{3}{4}} \Delta}, & \text{if } y^+ &< y_c^+ \end{aligned} \tag{16}$$

where  $y^+ = \frac{v_\tau \Delta}{\nu}$ ,  $v^+ = \frac{v_t}{\nu}$ , which result from the expressions  $v^+ = \frac{1}{\kappa} \ln y^+ + B$ , with  $\kappa = 0.41$  and  $B = 5.5$ , if  $y^+ \geq y_c^+$  or  $v^+ = y^+$  if  $y^+ < y_c^+$ .

Similar to the definition of  $v_\tau$ , Eq. 15, the development of the adjoint equations introduces the adjoint friction velocity  $u_\tau$  at each “wall” node (such as  $P$ ) defined by, [12],

$$u_\tau^2 = (v + v_i) \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) n_j t_i \tag{17}$$

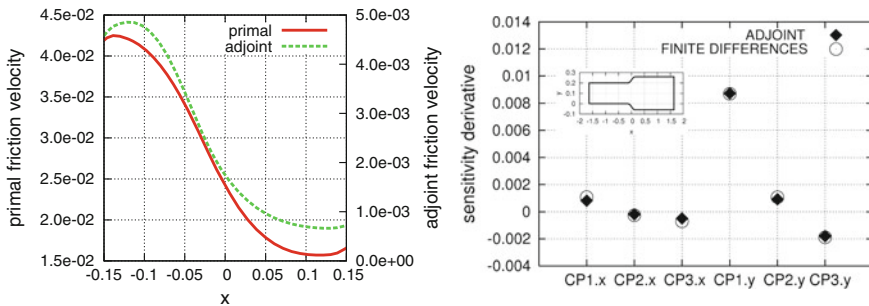
Attention should be paid to the close similarity of Eqs. 17 and 15. During the solution of the adjoint PDEs (TM=KE), the value of  $u_\tau$ , which contributes to the adjoint viscous fluxes at the “wall” nodes, is expressed in terms of the gradients of  $k$ ,  $k_a$ ,  $\varepsilon$  and  $\varepsilon_a$ , as follows

$$u_\tau^2 = \frac{1}{c_v} \left[ 2u_k t_k v_\tau - \left( v + \frac{v_i}{Pr_k} \right) \frac{\partial k_a}{\partial x_j} n_j \frac{\delta k}{\delta v_\tau} - \left( v + \frac{v_i}{Pr_\varepsilon} \right) \frac{\partial \varepsilon_a}{\partial x_j} n_j \frac{\delta \varepsilon}{\delta v_\tau} \right] \tag{18}$$

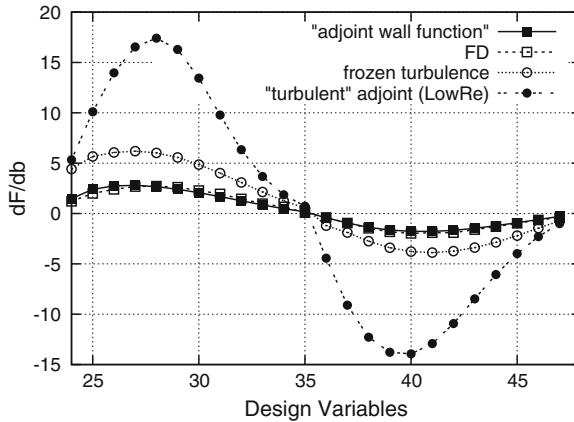
On the other hand, if TM=SA (based on a cell-centered finite-volume scheme with a no-slip condition at the solid wall boundary faces), the wall function technique is based on a single formula modeling both the viscous sublayer and the logarithmic region of the boundary layer

$$f_{WF} = y^+ - v^+ - e^{-\kappa B} \left[ e^{\kappa v^+} - 1 - \kappa v^+ - \frac{(\kappa v^+)^2}{2} - \frac{(\kappa v^+)^3}{6} \right] = 0 \tag{19}$$

In this case, the adjoint friction velocity must be zeroed. This is the major difference between the two finite-volume approaches (cell- and vertex-centered); despite this difference and any difference in the interpretation of the adjoint friction velocity, both will be referred to as “adjoint wall function” technique. Here, also, the role of (zero)  $u_\tau$  is to complete the adjoint momentum equilibrium at the first cell adjacent to the wall. The development is omitted in the interest of space. Applications, including validation, of the “adjoint wall function” technique are shown in Figs. 4 and 5, with comments in the caption.



**Fig. 4** Adjoint to the high-Re  $k$ - $\varepsilon$  model: Optimization of an axial diffuser, for minimum total pressure losses, using the “adjoint wall function” technique. *Left* Friction velocity  $v_\tau$  and adjoint friction velocity  $u_\tau$  distributions along its lower wall. *Right* Sensitivity derivatives of  $F$  w.r.t. the design variables, i.e. the coordinates of Bézier control points parameterizing the side walls. The “adjoint wall function” method perfectly matches the sensitivity derivatives computed by finite differences. From [12]



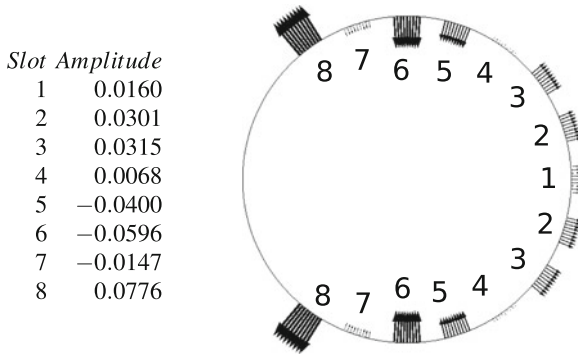
**Fig. 5** Adjoint to the high-Re Spalart–Allmaras model, flow around the isolated NACA0012 airfoil,  $\alpha_\infty = 3^\circ$ ,  $Re = 6 \times 10^6$ : Drag sensitivities computed using the “adjoint wall function” method are compared to finite-differences (FD), the adjoint method using the “frozen turbulence” assumption and the adjoint method with the “low-Reynolds” approach. The latter implies that the turbulence model is differentiated but the differentiation of the wall functions is disregarded. Only the latest 24 design variables, namely the  $y$  coordinates of the control points, where the magnitude of the computed derivatives is greater, are considered; the first 12 correspond to the suction side and the other to the pressure side. It is interesting to note that the “low-Re” adjoint approach performs even worse than the “frozen turbulence” one. In other words, the incomplete differentiation of the turbulence model produces worse results than its complete omission!

Recently, the “turbulent adjoint” method for the  $k - \omega$  SST turbulence model with wall functions was published by the same group, [16].

### 2.3 Other Applications of the Continuous Adjoint Methods

The continuous adjoint method is a low-cost tool to derive information regarding the optimal location and type of steady suction/blowing jets, used to control flow separation, [13]. In unsteady flows, the adjoint method, [17], can also be used to compute the optimal characteristics of unsteady jets, such as pulsating or oscillating ones. Such an application, where the optimal amplitudes of pulsating jets have been computed using the unsteady continuous adjoint method is presented in Fig. 6.

An inherent difficulty of the adjoint method, applied to unsteady flows, is the need of having the primal solution field available for the solution of the adjoint equations in each time-step. Since the adjoint solution evolves backwards in time, the need to store every primal solution arises. Since storing everything is expensive memory-wise, some turnaround is often used instead. A very common approach is the checkpointing technique, [18], where selected primal flow fields are stored (checkpoints) and the rest are recomputed starting from the checkpoints. Checkpointing is much cheaper memory-wise, at the expense of extra CPU time, needed for the re-computations of the primal fields.



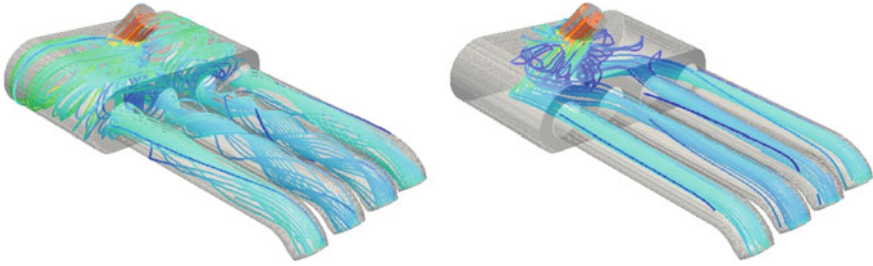
**Fig. 6** Time-averaged drag minimization of the unsteady flow around a cylinder at  $Re = 100$ , using pulsating jets: Optimal amplitudes for the symmetrically placed jets computed by the continuous adjoint method

Another way to overcome this is to use an approximation to the time evolution of the primal fields, such as the proper orthogonal decomposition (POD) technique, [19]. Approximating the primal field of each iteration bears no extra CPU cost.

On the other hand, topology optimization in fluid mechanics exclusively relies upon the adjoint method. In these problems, a real-valued porosity ( $\alpha$ ) dependent term is introduced into the flow equations. Based on the local porosity values, domain areas corresponding to the fluid flow are identified as those with nodal values  $\alpha \leq \varepsilon$ , where  $\varepsilon$  is a user-defined infinitesimally small positive number. All the remaining areas where  $\alpha > \varepsilon$  define the part of the domain to be solidified. The goal of topology optimization is to compute the optimal  $\alpha$  field in order to minimize the objective function under consideration. Since the number of the design variables is equal to the number of mesh cells (and thus, very high), the adjoint method is the perfect choice for computing  $\delta F / \delta \alpha$ , as its cost is independent of the number of design variables. Continuous adjoint methods for solving topology optimization problems for laminar and turbulent ducted flows, with or without heat transfer, are described in [20]. For turbulent flows, the adjoint approach is exact, i.e. includes the differentiation of the turbulence model (“turbulent” adjoint), (Fig. 7).

### 3 Robust Design Using High-Order Sensitivity Analysis

In aerodynamics, robust design methods aim at optimizing a shape in a range of possible operating conditions or by considering environmental uncertainties, such as manufacturing imprecisions or fluctuations of flow conditions, etc. The latter depend on the so-called environmental variables  $\mathbf{c}$  ( $c_i, i \in [1, M]$ ). In robust design problems, the function to be minimized can be expressed as  $\widehat{F} = \widehat{F}(\mathbf{b}, \mathbf{c}, \mathbf{U}(\mathbf{b}, \mathbf{c}))$ , to denote the dependency of  $\widehat{F}$  on the flow variables  $\mathbf{U}$ , the design variables  $\mathbf{b}$  ( $b_l, l \in [1, N]$ )



**Fig. 7** Topology optimization of a plenum chamber targeting minimum fluid power losses ( $F$ ), subject to a constraint requiring half of the plenum chamber volume to be filled by fluid. Primal velocity streamlines computed in the starting (*left*) and optimized geometries (*right*). Streamlines are colored based on the the primal velocity magnitude. A 29% reduction in  $F$  was achieved after a 12 hour computation on 40 cores of 5 Intel Xeon E5620 CPUs (2.40 GHz)

which parameterize the aerodynamic shape and the uncertain environmental variables  $\mathbf{c}$ . A probability density function  $g(\mathbf{c})$  can be associated with  $\mathbf{c}$ . In the so-called Second-Order Second-Moment (SOSM) approach,  $\widehat{F}$  combines the mean value  $\mu_F$  and variance  $\sigma_F^2$  of  $F$

$$\mu_F(\mathbf{b}, \mathbf{c}) = \int Fg(\mathbf{c})d\mathbf{c} \simeq F + \frac{1}{2} \left[ \frac{\delta^2 F}{\delta c_i^2} \right]_{\bar{\mathbf{c}}} \sigma_i^2 \tag{20}$$

$$\sigma_F^2(\mathbf{b}, \mathbf{c}) = \int (F - \mu_F)^2 g(\mathbf{c})d\mathbf{c} \simeq \left[ \frac{\delta F}{\delta c_i} \right]_{\bar{\mathbf{c}}}^2 \sigma_i^2 + \frac{1}{2} \left[ \frac{\delta^2 F}{\delta c_i \delta c_j} \right]_{\bar{\mathbf{c}}}^2 \sigma_i^2 \sigma_j^2 \tag{21}$$

where the gradients are evaluated at the mean values  $\bar{\mathbf{c}}$  of the environmental variables.

Based on the previous definitions, in robust design,  $\widehat{F}$  becomes

$$\widehat{F}(\mathbf{b}, \mathbf{c}) = w\mu_F + (1 - w)\sigma_F^2 \tag{22}$$

where  $w$  is a user-defined weight. To compute  $\widehat{F}$ , efficient and accurate methods for first- and second-order derivatives of  $F$  w.r.t. the environmental variables are needed.

### 3.1 Computation of Second-Order Moments

In aerodynamic optimization, the computation of the Hessian of  $F$ , subject to the constraint of satisfying the flow equations, can be conducted in at least four different ways. All of them can be set up in either discrete or continuous form [21–23]. The presentation is always much more synoptic in the discrete sense. In this case, the first-order variation rate of  $F$  w.r.t.  $c_i, i = 1, \dots, M$  is given by

$$\frac{dF}{dc_i} = \frac{\partial F}{\partial c_i} + \frac{\partial F}{\partial U_k} \frac{dU_k}{dc_i} \quad (23)$$

whereas the sensitivities of the discretized residuals  $R_m$  of the flow equations w.r.t.  $c_i$  are given by

$$\frac{dR_m}{dc_i} = \frac{\partial R_m}{\partial c_i} + \frac{\partial R_m}{\partial U_k} \frac{dU_k}{dc_i} = 0 \quad (24)$$

where  $U_k$  are the discretized field of the flow variables. Solving Eq. 24 for  $\frac{dU_k}{dc_i}$ , at the cost of  $M$  equivalent flow solutions (EFS; this is approximately the cost of solving the primal equations) and, then, computing  $\frac{dF}{dc_i}$  from Eq. 23 is straightforward but costly and will be referred to as Direct Differentiation (DD). Since the cost to compute the gradient of  $F$  using DD scales with  $M$ , the Adjoint Variable (AV) method can be used instead. The adjoint equations to be solved for the adjoint variables  $\Psi_m$  are

$$R_k^\Psi = \frac{\partial F}{\partial U_k} + \Psi_m \frac{\partial R_m}{\partial U_k} = 0 \quad (25)$$

and  $\frac{dF}{dc_i}$  are computed as

$$\frac{dF}{dc_i} = \frac{\partial F}{\partial c_i} + \Psi_m \frac{\partial R_m}{\partial c_i} \quad (26)$$

To compute the Hessian of  $F$ , starting from Eq. 23, the so-called DD-DD approach is set up, so that

$$\begin{aligned} \frac{d^2F}{dc_i dc_j} &= \frac{\partial^2 F}{\partial c_i \partial c_j} + \frac{\partial^2 F}{\partial c_i \partial U_k} \frac{dU_k}{dc_j} + \frac{\partial^2 F}{\partial U_k \partial c_j} \frac{dU_k}{dc_i} \\ &+ \frac{\partial^2 F}{\partial U_k \partial U_m} \frac{dU_k}{dc_i} \frac{dU_m}{dc_j} + \frac{\partial F}{\partial U_k} \frac{d^2 U_k}{dc_i dc_j} \end{aligned} \quad (27)$$

where  $\frac{d^2 U_k}{dc_i dc_j}$  is computed by first solving the following DD equations

$$\begin{aligned} \frac{d^2 R_n}{dc_i dc_j} &= \frac{\partial^2 R_n}{\partial c_i \partial c_j} + \frac{\partial^2 R_n}{\partial c_i \partial U_k} \frac{dU_k}{dc_j} + \frac{\partial^2 R_n}{\partial U_k \partial c_j} \frac{dU_k}{dc_i} \\ &+ \frac{\partial^2 R_n}{\partial U_k \partial U_m} \frac{dU_k}{dc_i} \frac{dU_m}{dc_j} + \frac{\partial R_n}{\partial U_k} \frac{d^2 U_k}{dc_i dc_j} = 0 \end{aligned} \quad (28)$$

Note that  $\frac{dU_k}{dc_i}$  are already known from the solution of Eq. 24.



The DD-DD approach requires upon the computation of  $\frac{dU_k}{dc_i}$  and  $\frac{d^2U_k}{dc_idc_j}$  and its CPU cost is  $M + \frac{M(M+1)}{2}$  EFS in total (excluding the cost of solving the flow equations). So, the overall CPU cost scales with  $M^2$  and becomes too expensive for use in real-world optimization.

Two less expensive approaches to compute the Hessian of  $F$  are the AV-DD (AV for the gradient and DD for the Hessian) and AV-AV ones. As shown in [23], both cost an many as  $2M+1$  EFS. It can be shown that the fourth alternative way, i.e. the DD-AV approach (DD for the gradient and AV for the Hessian), is the most efficient one to compute the Hessian matrix. In DD-AV, the Hessian matrix is computed by

$$\begin{aligned} \frac{d^2F}{dc_idc_j} &= \frac{\partial^2F}{\partial c_i\partial c_j} + \Psi_n \frac{\partial^2R_n}{\partial c_i\partial c_j} + \left( \frac{\partial^2F}{\partial U_k\partial U_m} + \Psi_n \frac{\partial^2R_n}{\partial U_k\partial U_m} \right) \frac{dU_k}{dc_i} \frac{dU_m}{dc_j} \\ &+ \left( \frac{\partial^2F}{\partial c_i\partial U_k} + \Psi_n \frac{\partial^2R_n}{\partial c_i\partial U_k} \right) \frac{dU_k}{dc_j} + \left( \frac{\partial^2F}{\partial U_k\partial c_j} + \Psi_n \frac{\partial^2R_n}{\partial U_k\partial c_j} \right) \frac{dU_k}{dc_i} \end{aligned} \quad (29)$$

where  $\frac{dU_k}{dc_i}$  result from Eq. 24 and  $\Psi_m$  is computed by solving the adjoint equation, Eq. 25 (same as before). The total CPU cost of DD-AV is equal to  $M+1$  EFS being, thus, the most economical approach.

### 3.2 Robust Shape Optimization Using Third-Order Sensitivities

If the problem of minimizing the combination of the two first statistical moments is to be solved using a stochastic method such as an evolutionary algorithm, the methods presented in Sect. 3.1 serve to provide  $\mu_F$  and  $\sigma_F^2$ ; no other derivation is required. However, if a gradient-based method is used, the gradient of  $\widehat{F}$  Eq. 22 must be differentiated w.r.t.  $b_q$ ,

$$\frac{\delta \widehat{F}}{\delta b_q} = w \left( \frac{\delta F}{\delta b_q} + \frac{1}{2} \frac{\delta^3 F}{\delta c_i^2 \delta b_q} \sigma_i^2 \right) + (1-w) \frac{2 \frac{\delta F}{\delta c_i} \frac{\delta^2 F}{\delta c_i \delta b_q} \sigma_i^2 + \frac{\delta^2 F}{\delta c_i \delta c_j} \frac{\delta^3 F}{\delta c_i \delta c_j \delta b_q} \sigma_i^2 \sigma_j^2}{2 \sqrt{\left[ \frac{\delta F}{\delta c_i} \right]^2 \sigma_i^2 + \frac{1}{2} \left[ \frac{\delta^2 F}{\delta c_i \delta c_j} \right]^2 \sigma_i^2 \sigma_j^2}} \quad (30)$$

From Eq. 30, the computation of  $\frac{\delta \widehat{F}}{\delta b_q}$  involves up to third-order mixed sensitivities w.r.t.  $c_i$  and  $b_q$ , such as  $\frac{\delta^3 F}{\delta c_i \delta c_j \delta b_q}$ . The computation of the second and third-order sensitivity derivatives is presented in detail in [24–26]. For instance, in the discrete sense, the highest-order derivative  $\frac{d^2F}{dc_idc_jdb_q}$  is computed using the expression

$$\begin{aligned}
\frac{d^3 F}{dc_i dc_j db_l} &= \frac{\partial^3 F}{\partial c_i \partial c_j \partial b_l} + \frac{\partial^3 F}{\partial c_i \partial b_l \partial U_k} \cdot \frac{dU_k}{dc_j} + \frac{\partial^3 F}{\partial c_j \partial b_l \partial U_k} \cdot \frac{dU_k}{dc_i} \\
&+ \frac{\partial^3 F}{\partial b_l \partial U_k \partial U_m} \cdot \frac{dU_k}{dc_i} \cdot \frac{dU_m}{dc_j} + \frac{\partial^2 F}{\partial b_l \partial U_k} \cdot \frac{d^2 U_k}{dc_i dc_j} \\
&+ \mathcal{K}_n^{i,j} \frac{\partial R_n}{\partial b_l} + \mathcal{L}_n^j \left( \frac{\partial^2 R_n}{\partial c_i \partial b_l} + \frac{\partial^2 R_n}{\partial b_l \partial U_k} \cdot \frac{dU_k}{dc_i} \right) \\
&+ \mathcal{M}_n^i \left( \frac{\partial^2 R_n}{\partial c_j \partial b_l} + \frac{\partial^2 R_n}{\partial b_l \partial U_k} \cdot \frac{dU_k}{dc_j} \right) \\
&+ \mathcal{N}_n \left( \frac{\partial^3 R_n}{\partial c_i \partial c_j \partial b_l} + \frac{\partial^3 R_n}{\partial c_i \partial b_l \partial U_k} \cdot \frac{dU_k}{dc_j} + \frac{\partial^3 R_n}{\partial c_j \partial b_l \partial U_k} \cdot \frac{dU_k}{dc_i} \right. \\
&\left. + \frac{\partial^3 R_n}{\partial b_l \partial U_k \partial U_m} \cdot \frac{dU_k}{dc_i} \cdot \frac{dU_m}{dc_j} + \frac{\partial^2 R_n}{\partial b_l \partial U_k} \cdot \frac{d^2 U_k}{dc_i dc_j} \right) \quad (31)
\end{aligned}$$

where the adjoint variables  $\mathcal{N}_n$  satisfy the equation

$$\frac{\partial F}{\partial U_k} + \mathcal{N}_n \frac{\partial R_n}{\partial U_k} = 0 \quad (32)$$

and the equations to be solved for  $\mathcal{L}_n^j$  and  $\mathcal{M}_n^i$  can be found in [24].

According to Eq. 31,  $\frac{dU_k}{dc_j}$  and  $\frac{d^2 U_k}{dc_i dc_j}$  must be available. These are computed by twice applying the DD technique, practically by solving Eqs. 24 and 28. This is the costly part of the algorithm, since it costs as many as  $M + \frac{M(M+1)}{2}$  EFS. However, in the majority of cases, the environmental variables are much less than the design ones,  $M \ll N$ . The computation of  $\mathcal{K}_N^{i,j}$ ,  $i, j \in [1, M]$  costs  $M + \frac{M(M+1)}{2}$  EFS and that of  $\mathcal{M}_n^i$ ,  $i \in [1, M]$   $M$  EFS. The overall cost per optimization cycle becomes  $M^2 + 3M + 2$  EFS; where the last two EFS correspond to the solution of the primal and adjoint (i.e. Eq. 32 for  $\mathcal{N}$ ) equations. The aforementioned technique, which is referred to as DD<sub>c</sub>-DD<sub>c</sub>-AV<sub>b</sub> (subscripts denote whether the differentiation is made w.r.t. **c** or **b**) has the minimum computational cost, provided that  $M < N$ .

### 3.3 Robust Design Using Continuous Adjoint

This section aims at briefly demonstrating that the material presented in Sects. 3.1 and 3.2 can also be based on the continuous, rather than the discrete, adjoint. Without loss in generality, this will be demonstrated in an inverse design problem, by assuming inviscid flow of a compressible fluid.

The steady-state 2D Euler equations of a compressible fluid are given by

$$\frac{\partial f_{nk}}{\partial x_k} = 0 \quad (33)$$

where  $k = 1, 2$  (for the Cartesian components) and  $n = 1, \dots, 4$  (four equations in 2D). The inviscid fluxes  $f_{nk}$  are

$$[f_{1k}, f_{2k}, f_{3k}, f_{4k}] = [\rho v_k, \rho v_k v_1 + p \delta_{k1}, \rho v_k v_2 + p \delta_{k2}, v_k(E + p)]$$

where  $\rho, p, v_k$  and  $E$  stand for the density, pressure, Cartesian velocity components and total energy per unit volume, respectively. The array of conservative flow variables is  $[U_1, U_2, U_3, U_4] = [\rho, \rho v_1, \rho v_2, E]$ . For the inverse design problem, the objective function is

$$F = \frac{1}{2} \int_{S_w} (p - p_{tar})^2 dS \quad (34)$$

where  $p_{tar}$  is the target pressure distribution along the solid wall.

In this problem, it is straightforward to derive the continuous adjoint PDEs which take the form

$$-A_{nmk} \frac{\partial \mathcal{N}_n}{\partial x_k} = 0, \quad m = 1, \dots, 4 \quad (35)$$

where  $A_{nmk} = \frac{\partial f_{nk}}{\partial U_m}$  ( $n = 1, 4, m = 1, 4, k = 1, 2$ ) are the Jacobian matrices of the inviscid fluxes. Eq. 35 is equivalent to Eq. 32 in the continuous sense, considering that, in continuous adjoint,  $\frac{\partial F}{\partial U_k}$  appears in the application of boundary conditions.

In the continuous approach, the DD<sub>c</sub>-DD<sub>c</sub> approach can also be formulated by setting up, discretizing and numerically solving PDEs for  $\frac{\delta U_m}{\delta c_i}$  and  $\frac{\delta^2 U_m}{\delta c_i \delta c_j}$ . The  $M$  systems of PDEs, to be solved for  $\frac{\delta U_m}{\delta c_i}$ , result from the first-order sensitivities of the Euler equations w.r.t. the environmental variables,

$$\frac{\partial}{\partial x_k} \left( A_{nmk} \frac{\delta U_m}{\delta c_i} \right) = 0, \quad n = 1, \dots, 4 \quad i = 1, \dots, M \quad (36)$$

along with appropriate boundary conditions. For the  $\frac{M(M+1)}{2}$  systems of equations, to be solved for  $\frac{\delta^2 U_m}{\delta c_i \delta c_j}$ ,  $i = 1, M, j = 1, M$ , Eq. 36 are differentiated once more to give

$$\frac{\partial}{\partial x_k} \left( A_{nmk} \frac{\delta^2 U_m}{\delta c_i \delta c_j} + \frac{\delta A_{nmk}}{\delta c_j} \frac{\delta U_m}{\delta c_i} \right) = 0, \quad n = 1, \dots, 4; \quad i, j = 1, \dots, M \quad (37)$$

With known  $\frac{\delta U_m}{\delta c_i}$  and  $\frac{\delta^2 U_m}{\delta c_i \delta c_j}$  fields, the first- and second-order sensitivities of  $F$  w.r.t. the environmental variables are given by

$$\frac{\delta F}{\delta c_i} = \int_{S_w} (p - p_{tar}) \frac{\delta p}{\delta c_i} dS \quad (38)$$

and

$$\frac{\delta^2 F}{\delta c_i \delta c_j} = \int_{S_w} \left[ \frac{\delta p}{\delta c_i} \frac{\delta p}{\delta c_j} + (p - p_{tar}) \frac{\delta^2 p}{\delta c_i \delta c_j} \right] dS \quad (39)$$

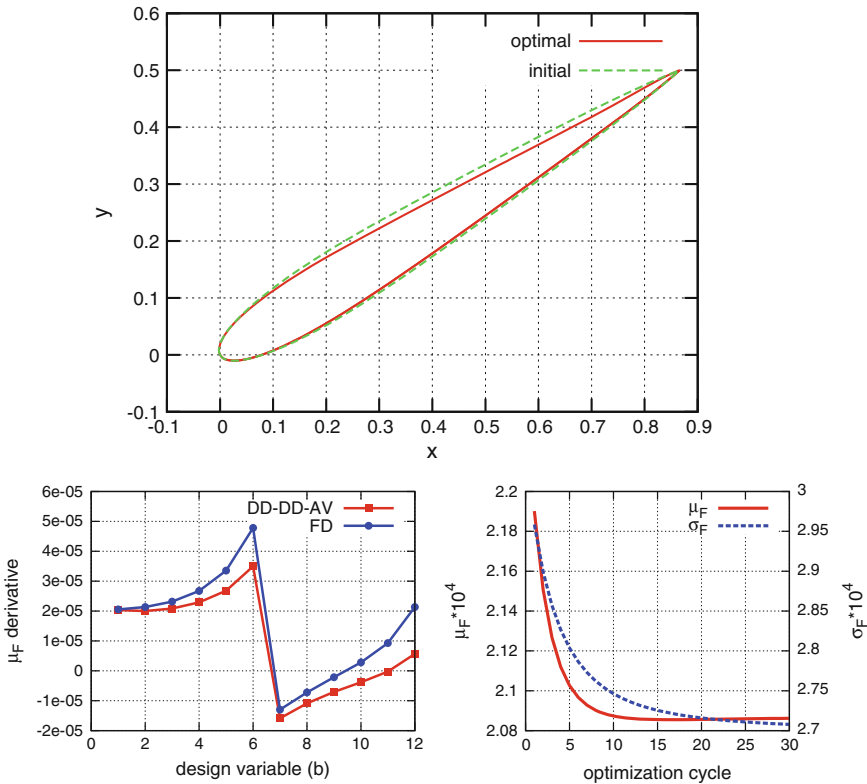
where  $\frac{\delta p}{\delta c_i}$  and  $\frac{\delta^2 p}{\delta c_i \delta c_j}$  can be expressed in terms of the corresponding derivatives of the conservative flow variables  $U_m$ .

The highest-order mixed derivatives are computed through the solution of additional adjoint PDEs, similar to the corresponding discrete equations. For instance, the third-order mixed sensitivity derivatives of  $F$ , required in Eq. 30, are given by

$$\begin{aligned} \frac{\delta^3 F}{\delta c_i \delta c_j \delta b_q} &= \int_{S_w} \frac{\delta p}{\delta c_i} \frac{\delta p}{\delta c_j} \frac{\delta(dS)}{\delta b_q} + \int_{S_w} (p - p_{tar}) \frac{\delta^2 p}{\delta c_i \delta c_j} \frac{\delta(dS)}{\delta b_q} \\ &+ \int_{S_w} \left( \mathcal{K}_{k+1}^{i,j} p - \mathcal{K}_n^{i,j} f_{nk} \right) \frac{\delta(n_k dS)}{\delta b_q} \\ &+ \int_{S_w} \left( \mathcal{L}_{k+1}^i \frac{\delta p}{\delta c_j} - \mathcal{L}_n^i \frac{\delta f_{nk}}{\delta c_j} \right) \frac{\delta(n_k dS)}{\delta b_q} \\ &+ \int_{S_w} \left( \mathcal{L}_{k+1}^j \frac{\delta p}{\delta c_i} - \mathcal{L}_n^j \frac{\delta f_{nk}}{\delta c_i} \right) \frac{\delta(n_k dS)}{\delta b_q} \\ &+ \int_{S_w} \left( \mathcal{N}_{k+1} \frac{\delta^2 p}{\delta c_i \delta c_j} - \mathcal{N}_n \frac{\delta^2 f_{nk}}{\delta c_i \delta c_j} \right) \frac{\delta(n_k dS)}{\delta b_q} \\ &- \int_{S_w} \mathcal{K}_n^{i,j} \frac{\partial f_{nk}}{\partial x_l} \frac{\delta x_l}{\delta b_q} n_k dS - \int_{S_w} \mathcal{L}_n^i \frac{\partial}{\partial x_l} \left( \frac{\delta f_{nk}}{\delta c_j} \right) \frac{\delta x_l}{\delta b_q} n_k dS \\ &- \int_{S_w} \mathcal{L}_n^j \frac{\partial}{\partial x_l} \left( \frac{\delta f_{nk}}{\delta c_i} \right) \frac{\delta x_l}{\delta b_q} n_k dS - \int_{S_w} \mathcal{N}_n \frac{\partial}{\partial x_l} \left( \frac{\delta^2 f_{nk}}{\delta c_i \delta c_j} \right) \frac{\delta x_l}{\delta b_q} n_k dS \\ &+ \int_{S_w} \mathcal{L}_n^j \frac{\partial}{\partial x_k} \left( \frac{\delta f_{nk}}{\delta c_i} \right) \frac{\delta x_l}{\delta b_q} n_l dS + \int_{S_w} \mathcal{L}_n^i \frac{\partial}{\partial x_k} \left( \frac{\delta f_{nk}}{\delta c_j} \right) \frac{\delta x_l}{\delta b_q} n_l dS \\ &+ \int_{S_w} \mathcal{K}_n^{i,j} \frac{\partial f_{nk}}{\partial x_k} \frac{\delta x_l}{\delta b_q} n_l dS + \int_{S_w} \mathcal{N}_n \frac{\partial}{\partial x_k} \left( \frac{\delta^2 f_{nk}}{\delta c_i \delta c_j} \right) \frac{\delta x_l}{\delta b_q} n_l dS \quad (40) \end{aligned}$$

where the additional adjoint fields  $\mathcal{L}_n^i = \frac{\delta \mathcal{A}_n}{\delta c_i}$  and  $\mathcal{K}_n^{i,j} = \frac{\delta \mathcal{L}_n^i}{\delta c_j}$  are computed by solving the adjoint equations

$$-A_{nmk} \frac{\partial \mathcal{L}_n^i}{\partial x_k} - \frac{\delta A_{nmk}}{\delta c_i} \frac{\partial \mathcal{A}_n}{\partial x_k} = 0 \quad (41)$$



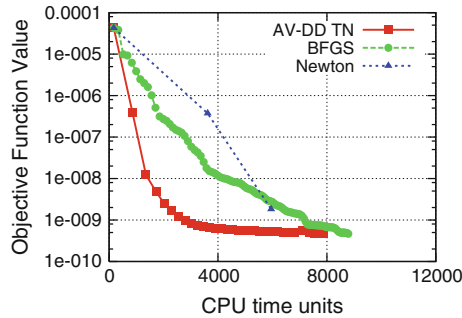
**Fig. 8** Robust inverse design of a 2D cascade for a given target pressure distribution. *Top* Optimal cascade geometry compared to the initial one; the initial airfoil is symmetric. *Bottom-left* Comparison of sensitivities  $\frac{\delta \mu_F}{\delta b_q}$  ( $b_q$  are the coordinates of the Bézier control points) computed using the  $DD_c-DD_c-AV_b$  method and FD. The proposed method matches the third derivatives captured by FD. *Bottom-right* Convergence of the mean value and standard deviation of  $F$  using  $w = 0.7$ , see Eq. 22. From [26]

and

$$-A_{nmk} \frac{\partial \mathcal{N}_n^{i,j}}{\partial x_k} - \frac{\delta A_{nmk}}{\delta c_j} \frac{\partial \mathcal{L}_n^i}{\partial x_k} - \frac{\delta A_{nmk}}{\delta c_i} \frac{\partial \mathcal{L}_n^j}{\partial x_k} - \frac{\delta^2 A_{nmk}}{\delta c_i \delta c_j} \frac{\partial \mathcal{N}_n}{\partial x_k} = 0 \quad (42)$$

as explained in [25]. Similarities between the discrete and continuous variants of the  $DD_c-DD_c-AV_b$  method can easily be identified.

An application of the robust design algorithm is illustrated in Fig. 8; it is related to the inverse design of a 2D cascade, [25, 26]. The airfoil shape controlling parameters are the design variables and the inlet/outlet flow conditions are the environmental ones.



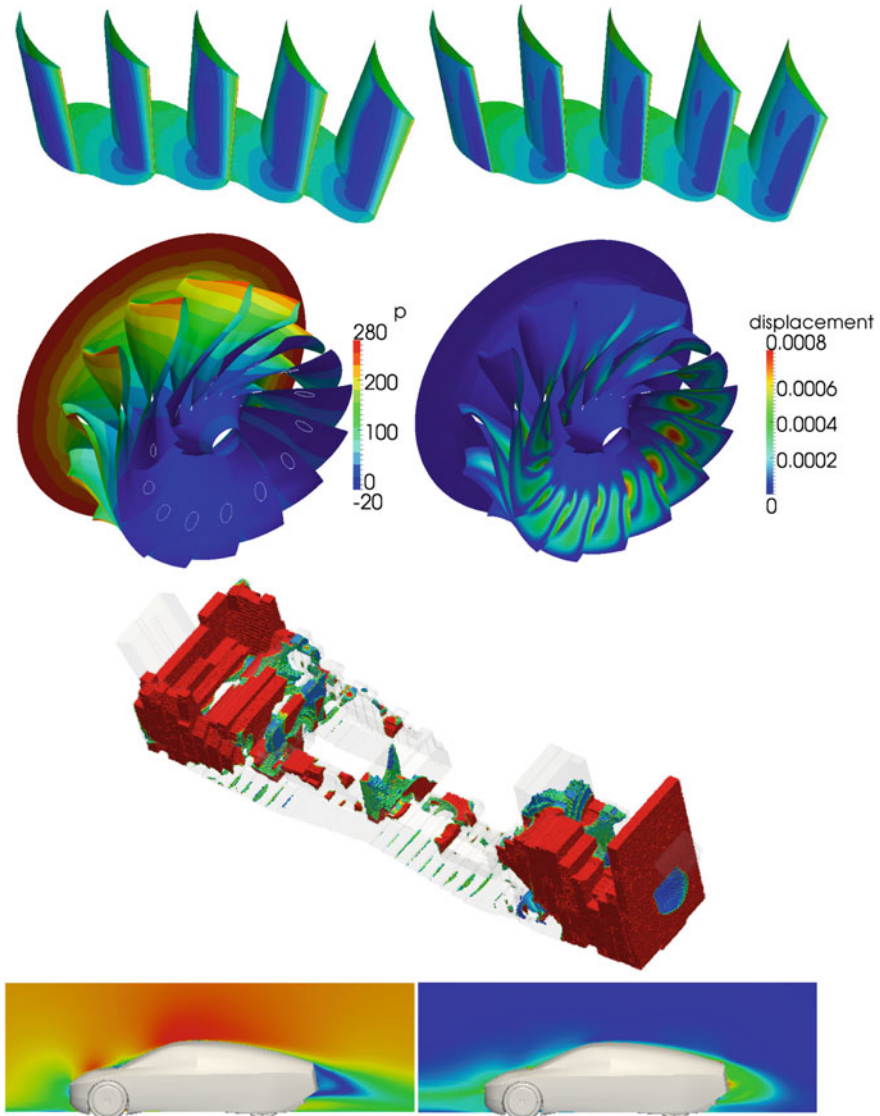
**Fig. 9** Inverse design of a 2D airfoil cascade (42 design variables) using the truncated Newton method: Comparison of the convergence rates of the AV-DD truncated Newton method (with 4 CG steps per cycle) with other second-order methods (BFGS and exact Newton). From [27]

### 3.4 Other Usage of the DD and AV Method

Apart from robust design applications, the developed methods for the computation of higher-order derivatives of  $F$  can also be used to support more efficient optimization methods, such as the (exact) Newton method. In such a case, however, the cost per optimization cycle depends on  $N$  and this may seriously hinder the use of such a method in industry. To cope with large scale optimization problems, the Newton equations can be solved through the CG method with truncation. By doing so, the Hessian matrix itself is not needed anymore, [27]. The adjoint approach followed by the DD of both the flow and adjoint equations (AV-DD) is the most efficient way to compute the product of the Hessian matrix with any vector required by the truncated Newton algorithm. The cost per Newton iteration scales linearly with the number of CG steps, rather than the much higher number of the design variables (if the Hessian itself was computed in the “exact” Newton method). The efficiency of the truncated Newton method is demonstrated in Fig. 9, in a problem with 42 design variables.

## 4 Industrial Applications

In Fig. 10, the application of the developed adjoint-based software to four industrial problems is presented. The first case deals with the blade optimization of a 3D peripheral compressor cascade in which the objective is the minimization of entropy losses within the flow passage, [28]. The second case is concerned with the shape optimization of a Francis turbine runner targeting cavitation suppression, [29], the third one with an air-conditioning duct targeting minimum total pressure losses and the last one with the shape optimization of a Volkswagen concept car, targeting minimum drag force, [30]. More comments can be found in the caption [31].



**Fig. 10** Row 1 Shape optimization of a 3D peripheral compressor cascade, targeting minimum entropy generation rate within the flow passage with constraints on the blade thickness. Pressure distributions over the initial (*left*) and optimal (*right*) blade geometries; from [28]. Row 2 Optimization of a Francis runner blade for cavitation suppression. Pressure distribution over the initial blading (*left*); areas within the white isolines are considered to be cavitated; surface deformation magnitude over the optimized blading (*right*), after eliminating cavitation; from [29]. Row 3 Topology optimization of an air-conditioning duct, used in a passenger car, targeting minimum total pressure losses. Porosity field at the last optimization cycle. The topology optimization led to the solidification of areas (in *red*) where, in the starting geometry, intense flow recirculation appeared. Row 4 Optimization of the VW L1 concept car targeting minimum drag force. Primal velocity field calculated using the RANS equations along with the low-Re Spalart–Allmaras model (*left*) and adjoint velocity field calculated by using the “turbulent” adjoint method (*right*); from [30]

**Acknowledgments** Parts of the research related to the exact differentiation of the turbulence models were funded by Volkswagen AG (Group Research, K-EFFG/V, Wolfsburg, Germany) and Icon Technology and Process Consulting Ltd. The support of Volkswagen AG for the development of the adjoint method for flow control applications is also acknowledged.

Research related to topology optimization was partially supported by a Basic Research Project funded by the National Technical University of Athens.

The authors would like to acknowledge contributions to topology optimization from Dr. E Kontoleonos and to “turbulent” adjoint by Dr. A. Zymaris. They are also indebted to Dr. Carsten Othmer, Volkswagen AG (Group Research, K-EFFG/V), for his support, some interesting discussions on the continuous adjoint method and his contributions in several parts of this work.

## References

1. P. Spalart and S. Allmaras: A one-equation turbulence model for aerodynamic flows. AIAA Paper, 04(39), 1992
2. Launder BE, Sharma BI (1974) Application of the energy-dissipation model of turbulence to the calculation of flow near a spinning disc. *Lett Heat Mass Transf* 1:131–137
3. Lee BJ, Kim C (2007) Automated design methodology of turbulent internal flow using discrete adjoint formulation. *Aerosp Sci Technol* 11:163–173
4. Mavriplis DJ (2007) Discrete adjoint-based approach for optimization problems on three-dimensional unstructured meshes. *AIAA J* 45:740–750
5. Pironneau O (1974) On optimum design in fluid mechanics. *J Fluid Mech* 64:97–110
6. Jameson A (1988) Aerodynamic design via control theory. *J Sci Comput* 3:233–260
7. Anderson WK, Venkatakrisnan V (1997) Aerodynamic design optimization on unstructured grids with a continuous adjoint formulation. AIAA Paper 97–0643
8. Papadimitriou DI, Giannakoglou KC (2007) A continuous adjoint method with objective function derivatives based on boundary integrals for inviscid and viscous flows. *J Comput Fluids* 36:325–341
9. Othmer C (2008) A continuous adjoint formulation for the computation of topological and surface sensitivities of ducted flows. *Int J Numer Meth Fluids* 58:861–877
10. Zymaris AS, Papadimitriou DI, Giannakoglou KC, Othmer C (2009) Continuous adjoint approach to the Spalart-Allmaras turbulence model for incompressible flows. *Comput Fluids* 38:1528–1538
11. Bueno-Orovio A, Castro C, Palacios F, Zuazua E (2012) Continuous adjoint approach for the SpalartAllmaras model in aerodynamic optimization. *AIAA J* 50:631–646
12. Zymaris AS, Papadimitriou DI, Giannakoglou KC, Othmer C (2010) Adjoint wall functions: a new concept for use in aerodynamic shape optimization. *J Comput Phys* 229:5228–5245
13. Papoutsis-Kiachagias EM, Zymaris AS, Kavvadias IS, Papadimitriou DI, Giannakoglou KC (2014) The continuous adjoint approach to the  $k-\varepsilon$  turbulence model for shape optimization and optimal active control of turbulent flows. *Eng Optim*. doi:10.1080/0305215X.2014.892595
14. Asouti VG, Trompoukis XS, Kampolis IC, Giannakoglou C (2011) Unsteady CFD computations using vertex-centered finite volumes for unstructured grids on graphics processing units. *Int J Numer Meth Fluids* 67(2):232–246
15. Trompoukis XS, Tsiakas KT, Ghavami Nejad M, Asouti VG Giannakoglou KC (2014) The continuous adjoint method on graphics processing units for compressible flows. In: OPT-i, international conference on engineering and applied sciences optimization, Kos Island, Greece, 4–6 June 2014
16. Kavvadias IS, Papoutsis-Kiachagias EM, Dimitrakopoulos G, Giannakoglou KC (2014) The continuous adjoint approach to the  $k\omega$  SST turbulence model with applications in shape optimization. *Eng Optim*. doi:10.1080/0305215X.2014.979816



17. Kavvadias IS, Karpouzas GK, Papoutsis-Kiachagias EM, Papadimitriou DI, Giannakoglou KC (2013) Optimal flow control and topology optimization using the continuous adjoint method in unsteady flows. In: EUROGEN conference 2013, Las Palmas de Gran Canaria, Spain
18. Griewank A, Walther A (2000) Algorithm 799: revolve: an implementation of checkpointing for the reverse or adjoint mode of computational differentiation. *ACM Trans Math Softw* 26(1):19–45
19. Vezyris CK, Kavvadias IS, Papoutsis-Kiachagias EM, Giannakoglou KC (2014) Unsteady continuous adjoint method using POD for jet-based flow control. In: ECCOMAS conference 2014, Barcelona, Spain
20. Papoutsis-Kiachagias EM, Kontoleontos EA, Zymaris AS, Papadimitriou DI, Giannakoglou KC (2011) Constrained topology optimization for laminar and turbulent flows, including heat transfer. In: EUROGEN Conference 2011, Capua, Italy
21. Papadimitriou DI, Giannakoglou KC (2008) Computation of the Hessian matrix in aerodynamic inverse design using continuous adjoint formulations. *Comput Fluids* 37:1029–1039
22. Papadimitriou DI, Giannakoglou KC (2008) The continuous direct-adjoint approach for second-order sensitivities in viscous aerodynamic inverse design problems. *Comput Fluids* 38:1539–1548
23. Papadimitriou DI, Giannakoglou KC (2008) Aerodynamic shape optimization using adjoint and direct approaches. *Arch Comput Meth Eng* 15:447–488
24. Papoutsis-Kiachagias EM, Papadimitriou DI, Giannakoglou KC (2012) Robust design in aerodynamics using third-order sensitivity analysis based on discrete adjoint. Application to quasi-1D flows. *Int J Numer Meth Fluids* 69:691–709
25. Papadimitriou DI, Giannakoglou KC (2013) Third-order sensitivity analysis for robust aerodynamic design using continuous adjoint. *Int J Numer Meth Fluids* 71:652–670
26. Papoutsis-Kiachagias EM, Papadimitriou DI, Giannakoglou KC (2011) On the optimal use of adjoint methods in aerodynamic robust design problems. In: CFD and OPTIMIZATION, (2011) ECCOMAS thematic conference. Antalya, Turkey 2011
27. Papadimitriou DI, Giannakoglou KC (2012) Aerodynamic design using the truncated Newton algorithm and the continuous adjoint approach. *Int J Numer Meth Fluids* 68:724–739
28. Papadimitriou DI, Giannakoglou KC (2006) Compressor blade optimization using a continuous adjoint formulation. In: ASME TURBO EXPO, GT2006/90466, Barcelona, Spain, 8–11 May 2006
29. Papoutsis-Kiachagias EM, Kyriacou SA, Giannakoglou KC (2014) The continuous adjoint method for the design of hydraulic turbomachines. *Comput Meth Appl Mech Eng* 276:621–639
30. Othmer C, Papoutsis-Kiachagias E, Haliskos K (2011) CFD Optimization via sensitivity-based shape morphing. In: 4th ANSA &  $\mu$ ETA international conference, Thessaloniki, Greece, 2011
31. Menter FR (1994) Two-equation eddy-viscosity turbulence models for engineering applications. *AIAA* 32(8):269–289

# Hierarchical Topology Optimization for Bone Tissue Scaffold: Preliminary Results on the Design of a Fracture Fixation Plate

Emily Gogarty and Damiano Pasini

**Abstract** A porous material can be designed to promote tissue regeneration as well as satisfy mechanical and biological requirements. The porous microarchitecture can be specifically tailored to locally match the specific properties of the host tissue resulting in a biologically fixed implant. A 2D hierarchical topology optimization scheme is presented here to design a cellular scaffold that optimally reconciles bone resorption and permeability, two antagonist objectives of bone tissue scaffolds. The implant is tailored to reproduce the variable stiffness properties of the surrounding bone while maximizing its permeability for bone ingrowth. The procedure integrates multi-objective optimization with multi-scale topology optimization. In particular, the material layout is sequentially optimized at two length scales: (1) the property distribution varying throughout the implant body, and (2) the topology of each pore of the scaffold. In the first stage, an optimal material distribution is obtained to generate a stiffness match between implant and bone tissue. In the second stage, the optimal relative density distribution is used to interpolate target material properties at each location of the implant domain. Target matching topology optimization is used to obtain unit cells with desired stiffness and maximum permeability throughout the implant. The procedure currently developed in 2D can be extended to produce clinically relevant 3D implant models. As a case study, a 2D bone fracture fixation plate under in-plane load is optimized at both the implant and cellular material level. While the preliminary results presented here need further refinement, such as on the filtering method and the calculation of permeability, the paper contributes to the development of a method to design engineered scaffolds that are both mechanically optimal and conducive to bone tissue regeneration.

---

E. Gogarty (✉) and D. Pasini  
Department of Mechanical Engineering, McGill University, Montreal, Canada  
e-mail: Emily.gogarty@gmail.com

D. Pasini  
e-mail: Damiano.pasini@mcgill.ca

# 1 Introduction

Bone tissue serves four important functions in the body: (1) protection of organs, (2) structural support of muscle attachment for locomotion, (3) generation of red and white blood cells, and (4) calcium and other ion storage. It is apparent that any damage to the skeletal system has widespread effects [1]. Damage due to disease, abnormal development, or trauma can be addressed through artificial bone tissue scaffolds. The design of such scaffolds is a multidisciplinary area with much research potential. Currently, technological and scientific advances in areas such as additive manufacturing and biomaterials enable the design and manufacturing of bone tissue scaffolds with tunable properties. The scope of this paper is to present a design methodology to improve upon the current state of the art of bone scaffold design. It is therefore essential to first review the biology of bone before describing the design requirements for bone tissue scaffolds.

## 1.1 Bone Tissue

On the macroscopic level, the human skeleton consists of long bones, flat bones, and cuboid bones (femur, skull, and vertebrae respectively). The structure of bone can be divided into two categories: cortical and trabecular (also referred to as cancellous). In adults, approximately 80 % of bone is cortical and 20 % is trabecular, percentages that vary throughout the body. Cortical bone is mainly located in the shafts of long bones and peripheral linings of flat bones [2]. The structural arrangement of a bone can be described as a hollow tube or bilaminar plate of cortical bone, with trabecular “struts” reinforcing the architecture. The presence of cancellous bone allows for a reduced-weight structure that contributes to satisfying mechanical requirements [3].

Bone has a hierarchical structure. On the microscopic level, bone structure can be either woven or lamellar. Woven bone is immature and unorganized. A remodelling process occurs to organize woven bone into a lamellar form, such as Haversian bone [2]. Adult cortical bone has a lamellar collagen fibre arrangement, which is densely packed and arranged both circumferentially and in a tubular formation. The tubular formation is made of concentric lamellae layers, forming an osteon. Each osteon surrounds a central Haversian canal, which contains blood vessels. The osteons are arranged around branching blood vessels, oriented along the long axis of the bone. Because of this orientation, the osteons act as fibres reinforcing the long bone and are essential in resisting deformation. This hierarchical structure is both mechanically and biologically significant [4]. It is thus critical to understand the mechanical response of bone for the design of an implant with properties closely matching those of the host bone tissue [3].

The biological composition of cortical and trabecular bone is very similar. However, the anisotropic structure of cortical bone consists of partial alignment of the mineral hydroxyapatite in the longitudinal direction (fibre-like), making that the

**Table 1** Mechanical properties of wet cortical bone [3]

Property	Value
Young's modulus (GPa)	
Longitudinal	17.0
Radial	11.5
Tangential	11.5
Compressive Strength (MPa)	
Along	193
Normal	133
Tensile strength (MPa)	
Along	148
Normal	49

stiffer and stronger axis. Table 1 summarizes the mechanical properties of wet cortical bone [3]. When bone is dried, its elastic moduli increase, whereas its strength and strain to failure decrease. In the case study examined in this work, we use the properties of wet cortical bone (Table 1).

Bone remodelling is a dynamic and lifelong process of bone resorption (removal from the skeleton) and ossification (formation of new bone). It has long been accepted that bone grows in response to stress, as stated most notably by Wolff et al. [1], but the mechanism is currently not fully understood. Wolff's law states the following: "Every change in the form and function of bone or of its function alone is followed by certain definite changes in the bone internal architecture, and equally definite alteration in its external conformation, in accordance with mathematical laws". Essentially, the skeleton adds or removes tissue in response to functional requirements with the purpose of reducing stress or strain.

The mechanical behaviour of trabecular bone is typical of a cellular material, as evident in the characteristic stress-strain plot of both materials. In addition, the Young's modulus, compressive and tensile strength of cancellous bone are highly dependent on relative density. The shape and density of the trabecular cells are biologically governed by the loads that bone must support. Cell walls of trabecular bone tend to align and thicken in the direction which will best support load, and the relative density of the cells depends on the load magnitude. The mechanical behaviour of trabecular bone has been proven to follow that of a cellular material [3].

### 1.2 Design Requirements of Bone Tissue Scaffolds

Bone tissue scaffolds provide temporary or permanent support as tissue regenerates and assumes primary function [5]. Scaffold architecture can be specifically tailored to match the mechanical environment of bone tissue while concurrently providing sufficient porosity for cell migration to achieve tissue regeneration [6]. Load bearing bone tissue scaffolds present design challenges that soft tissue scaffolds do not

necessarily need to address. In the case of a permanent, non-degradable bone tissue scaffold, the implant must withstand relatively high physiological loading conditions as well as promote tissue regeneration at the periphery of the implant for fixation [7]. At a given site, approximately 1 mm of bone ingrowth into the scaffold is required for biological fixation. Advances in additive manufacturing make it possible to fabricate scaffolds with such characteristics.

There are four fundamental requirements, 4F, that a bone tissue scaffold must satisfy [8]: Form, Function, Formation, and Fixation. Form refers to scaffold shape completely filling a complex 3D anatomical defect. Function means supporting mechanical demands, i.e. normal physiological loading conditions. Formation refers to the enhancing tissue regeneration by providing a sufficient mass transport environment for new tissue growth. Finally, fixation implies that the scaffold can be readily implanted and attached to tissue at the defect site. To design a tissue scaffold, these four requirements must be addressed in a quantitative manner [8].

With computer assisted design, the Form requirement can easily be fulfilled. For example, a scaffold can be specifically designed to fit an anatomical defect based on a patient CT scan [7]. Factors affecting Fixation include scaffold microarchitecture (pore shape, size, interconnectivity), cellular interaction with the scaffold surface, and release of growth factors. The basic requirements of a tissue scaffold often present a design trade-off between Function and Formation: a denser, mechanically suitable scaffold versus a more porous scaffold which would provide better mass transport [5]. The requirements governing Formation are difficult to specify quantitatively. Bone regeneration is influenced by mass transport and delivery of biologics. Mass transport can be quantitatively expressed as permeability and diffusivity, which are controlled by the scaffold microarchitecture.

It is therefore challenging to design a scaffold that addresses all four requirements for a number of reasons. Properties such as elasticity, permeability, and diffusion are related on a hierarchical scale (material and pore level) and require complex computational design and fabrication methods [9]. The main key players affecting the 4F requirements are described in the following sections.

### ***Microarchitecture Characteristics***

Scaffold microarchitecture refers to the microscopic features of the scaffold that can be tailored to achieve desired field properties. The need for a bone scaffold to match complex anatomic shapes and desired physical properties requires the separation of a scaffold into the microscopic (< 1 mm feature size) and macroscopic scales (> 1 mm). The division of feature sizes into two scales allows the utilization of areas of the design space which are not accessible with a solid material [7]. For example, the bounds on effective mechanical and mass transport properties are defined by properties of a completely solid material and no material at all. The range of possible effective properties fall within these bounds and can be achieved through the design of the cell topology. It has been proved that cell architecture tailoring permits the achievement of target properties, such as elasticity, diffusion, and permeability, which meet the 4F requirements [10]. Some microarchitecture characteristics that are reported to

influence tissue regeneration and vascularization include volume porosity, pore size, pore interconnectivity, and pore geometry [11].

Combining computer-aided design with additive manufacturing technology allows for a fine control of scaffold design at both the macro- and micro-architectural level. Recent work on the design of hip implants has demonstrated that multi-scale and multi-objective optimization allows to functionally grade a cellular domain to meet specific mechanical and biological requirements [12–17].

### ***Mechanical Properties***

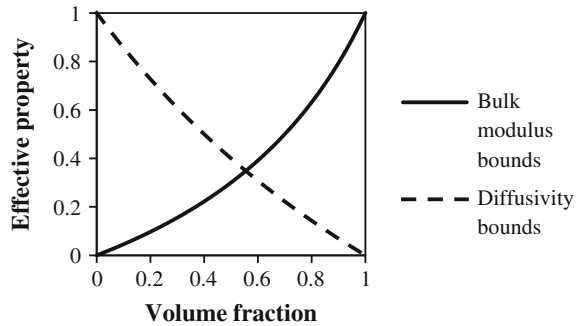
A bone implant must have appropriate mechanical properties to substitute for the loss of function of the replaced bone tissue [7, 18]. It is widely accepted that scaffolds should be designed to match healthy tissue stiffness and strength while providing a suitable network of pores to allow for cell migration and nutrient transport [11]. Over-designing for mechanical loading can result in a scaffold that is too stiff compared to the local tissue environment. In this case, a much stiffer scaffold can have adverse effects on local tissue, such as bone tissue resorption. Titanium and stainless steel are often used in orthopaedic implants because of their biocompatibility and superior mechanical properties compared to bone. However, a solid metal implant may absorb the forces that are required to stimulate bone remodelling, as discussed in Sect. 1. This phenomenon is known as stress shielding. The stress shielding effect can lead to bone resorption around the implant and prevent implant fixation [2]. Ideally, the mechanical properties would be similar to those of the local environment, so that scaffold failure would not occur, and structure stresses would be sufficiently low to avoid tissue resorption [8].

### ***Permeability/Diffusivity***

A scaffold should provide a suitable environment for mass transportation of nutrients and metabolic waste [7, 18]. Mass transport can be quantitatively expressed with permeability and diffusivity. The former relates fluid velocity in a porous medium to the pressure gradient, and the latter ion concentration to chemical concentration gradients.

Porosity is crucial to bone growth. Porosity is measured as the ratio of void space to the total bulk volume of a scaffold. High porosity increases surface area and allows for mass transport of nutrients and metabolic waste [2]. The surface of a bone tissue scaffold should be approximately 60–70% porous to effectively promote bone in growth. The optimal average size of a pore for bone tissue ingrowth in a scaffold material is in the range of 50–400  $\mu\text{m}$ . Studies examining the effect of pore size found that both amount and rate of bone growth increase with decreased pore size, when comparing pores in the pore range 200–500  $\mu\text{m}$  [19]. In addition, an interconnected porous architecture is essential to allow blood vessels and surrounding bone enter the scaffold [2].

**Fig. 1** Conflict between normalized diffusivity and bulk modulus bounds for an isotropic two-phase material (e.g. solid-void) as a function of volume fraction (relative density) [20]



### 1.2.1 Requirement Conflict: Mechanical Versus Mass Transport Function

Scaffold design is challenging because effective properties for mechanical support typically conflict with mass transport properties, which are essential for tissue regeneration. The increase in stiffness and strength of a scaffold comes at the expense of mass transport, and the opposite is also true (Fig. 1). The cross property bounds shown are those defined by Kang et al. [20].

Cellular material provides a unique advantage in addressing this trade-off because both macroscopic and microscopic features can be specifically tailored for mechanical and permeability properties [5]. Via multiobjective optimization we can determine the material layout of a scaffold that can achieve these target effective properties [10].

The objective of this paper is to illustrate that a cellular material can be designed to address the multi-scale requirements of a bone tissue scaffold. Via computational methods, both the geometry and microarchitecture of the scaffold can be designed to achieve specific effective mechanical and mass transport properties. A porous scaffold with extremely fine features on the micrometer scale can be manufactured using a biocompatible material (e.g. Ti-6Al-4V) by additive manufacturing, such as electron beam melting and selective laser sintering. The next and following sections present a hierarchical topology optimization scheme that uses multi-scale computations to design a bone tissue scaffold, and Sect. 4 illustrates its application to the design a bone fracture fixation plate.

## 2 Methods and Theory

The hierarchical structure of bone necessitates multi-scale design of a bone tissue scaffold. The hierarchical design of a cellular implant for bone tissue is treated here as a material distribution problem at two geometric scales: (1) macro-architecture: topology of the cellular implant, and (2) micro-architecture: topology of each unit cell. We first determine the optimal macro topology of the scaffold. Then the homog-

enized properties of each scaffold unit cell are tailored to locally yield the desired mechanical properties of the implant. Finally, the optimal property distribution of the implant is mapped with these unit cell geometries to create an optimal porous architecture of the scaffold.

There is extensive work on the characterization and structural analysis of cellular materials [6, 16, 18, 20–34]. The multi-objective optimization of both stiffness and permeability has also been explored for bone tissue scaffolds [5, 6, 11, 20, 31, 35] and methods have been developed to generate unit cells close to the theoretical cross property bounds [20, 36]. Hierarchical topology optimization methods [37–42] have been shown effective in problems involving target property matching. These works are the foundation of the method presented here, where we combine a hierarchical topology design with multi-objective optimization for bone tissue scaffold.

## ***2.1 Computational Mechanics for Scaffold Material***

### **2.1.1 Effective Properties**

The macroscopic behaviour of a composite material is largely dependent upon its microstructure. The analysis of the structural mechanics of a cellular material is challenging due to large geometric heterogeneity at the microscopic level [43]. The highly complex geometry requires a large computational effort. Finite element analysis of microscopic behaviour is generally unfeasible, as it would be very computationally expensive to create, mesh, and analyse each strut of a discrete lattice [44]. As such, homogenization methods are developed to accurately approximate the behaviour and properties of a composite cellular structure based on the smallest repeating element of the structure: the unit cell, or the RVE (representative volume element).

Asymptotic homogenization, which is used here, is one among several methods available in literature to determine an equivalent homogenous structure representing its detailed cellular counterpart [30]. Asymptotic homogenization is based on decoupling the analysis of a cellular material into analyses at the micro (local) level, and the macro (global) level. The method involves first analyzing one unit cell to determine its effective properties, by finding its unique behavioural response to a specified loading condition under periodic boundary conditions. To determine the effective stiffness matrix ( $C^H$ ) of a unit cell, a load is applied in each of the unique normal and shear directions, a process equivalent to imposing unit strains. The unit cell can then be treated as an equivalent homogeneous structure with the behavioural response equivalent to that of the detailed unit cell. Then, the entire macro structure can be mapped with equivalent homogeneous cells. This allows for a much simpler analysis of the structure at the global level. Homogenization can be used for any periodic physical property, and is used here to calculate the effective stiffness properties [44].



### 2.1.2 Effective Permeability

#### Weissberg's Approximation

A scaffold should provide an environment conducive to mass transportation of nutrients and metabolic waste [7, 18]. Mass transport can be quantitatively expressed as permeability and diffusivity. There are many ways to model the permeability of a scaffold, such as Stoke's flow homogenization, and various numerical approximations. Weissberg's formula is an approximation that was originally derived to determine the effective diffusion coefficient through a bed of randomly overlapping spheres of uniform or non-uniform shape, and is solely a function of porosity,  $\varepsilon$  [45]:

$$\frac{D_e}{D} = \frac{\varepsilon}{[1 - \frac{1}{2} \ln \varepsilon]} \quad (1)$$

It is valid to use Weissberg's formula to approximate effective diffusivity in 2D porous media if the following assumptions can be made: (1) the media is isotropic and (2) the media porosity is above the percolation threshold. A study by Trinh et al. shows that the effective diffusivity coefficient of a porous media can be computationally determined using Monte Carlo simulations [46]. The good agreement of the simulation results to the theoretical predictions using Weissberg's formula in the range of 60–80% porosity (in the required bone ingrowth range) leads to the choice of using the latter to estimate the effective diffusivity of the scaffold. If the unit cells of a designed scaffold are constrained to be isotropic then assumption (1) will hold. The desired outcome of this research is to optimize an overall anisotropic cellular medium with isotropic unit cells. Therefore, the Weissberg approximation will be used in the proposed optimization procedure to calculate unit cell diffusivity.

## 2.2 Inverse Homogenization

Homogenization theory is used to calculate the effective properties of a bulk material based on knowledge of the topology of a repeating unit cell. Recalling the goal of designing a bone tissue scaffold with specific effective properties, it is therefore necessary to solve an *inverse homogenization* problem [47]. Inverse topology optimization was originally formulated as a minimization of the difference between homogenized material properties and target material properties of a unit cell. The goal here is to seek a microstructural configuration that attains desired effective material properties [47], as described below.

At the microscopic level, the problem aims at determining the material distribution within the design domain of the unit cell. The unit cell is discretized into uniform mesh elements, each of which possesses a relative density: a fraction indicating how much solid material phase is present in that element. A relative density of 1 indicates

completely solid material, and a relative density of 0 indicates a void element where no material is present. The connection of solid elements defines the topology of the unit cell and consequently the effective properties. Relative densities of the elements of the unit cell are typically the design variables when topology optimization is used to solve the inverse homogenization problem.

There are two common approaches to defining objective functions for finite element based inverse homogenization. The first is minimizing or maximizing the critical components of a homogenized tensor. This is formulated for the stiffness tensor as minimizing these parameters (or their reciprocal). For example maximizing bulk modulus can be formulated as:

$$\min_{\rho} f(\rho) = \left( \frac{1}{9} \sum_{i,j=1}^3 C_{iijj}^H(\rho) \right)^{-1} \tag{2}$$

where  $\rho$  is relative density and  $C_{iijj}^H$  is the homogenized stiffness tensor. The second approach is to use the least squares formulation, where the square of the difference between homogenized tensor and target tensor is minimized. It is mathematically formulated in the following way:

$$\min_{\rho} f(\rho) = \sum_{i,j,k,l=1}^3 w_{ijkl} (C_{ijkl}^* - C_{ijkl}^H)^2 \tag{3}$$

where  $w_{ijkl}$  is a weighting factor to vary roles of different stiffness components and  $C_{ijkl}^*$  is a target stiffness tensor.

The method of inverse homogenization is well suited for the design of a bone tissue scaffold. The goal is typically to match the stiffness of the surrounding tissue while maximizing permeability. The hierarchical design of the bone tissue scaffold is addressed by using the multi-functional inverse homogenization at both the macro and micro scales, with topology optimization used to solve the inverse homogenization problem.

### 2.3 Topology Optimization for Bone Tissue Scaffold

Topology optimization refers to the determination of the connectivity of a design domain, through features such as number, location, and shape of holes in a structure. Topology optimization seeks to determine the optimal placement of an isotropic material in a given design space. A set of distributed functions defined on a fixed design domain are used to represent the topology, size, and shape of the structure. Although the base material in classical topology optimization problems is isotropic, this research deals with an anisotropic cellular material. Bendsoe and Sigmund describe in detail the theoretical basis for topology optimization [48]. The follow-

ing section and many published works in this area are based on their formulation of the topology optimization problem, which is also the foundation of the work presented here.

### 2.3.1 Definition and Derivation

The starting point for topology optimization is the formulation of a general shape optimization problem in terms of material distribution, for a minimum compliance structure with material constraints. The problem involves finding the optimal choice of the stiffness tensor  $E_{ijkl}(x)$  of the solid body, which is variable over the domain  $\Omega_{\text{mat}}$ , discretized using finite elements. To allow for the introduction of holes into the structure, a fixed mesh can be used where void elements are assigned very low stiffness properties, so that re-meshing is avoided [49]. Both the displacement and stiffness fields are discretized using identical prescribed mesh. The minimization problem is thus written as:

$$\begin{aligned} \min_{\mathbf{u}, E_e} \quad & \mathbf{f}^T \mathbf{u} \\ \text{s.t.} \quad & \mathbf{K}(E_e) \mathbf{u} = \mathbf{f} \\ & E_e \in E_{ad} \end{aligned} \quad (4)$$

where  $\mathbf{f}$  and  $\mathbf{u}$  are the load and displacement vectors respectively, the stiffness matrix  $\mathbf{K}$  is a function of  $E_e$  in an element  $e$ , and  $E_{ad}$  is the set of admissible stiffness tensors for the given problem. The stiffness matrix  $\mathbf{K}$  can be written as a sum of the stiffness of each element in the form:

$$\mathbf{K} = \sum_{e=1}^N \mathbf{K}_e(E_e) \quad (5)$$

where element  $e$  is numbered as  $e = 1, \dots, N$  and  $\mathbf{K}_e$  is the global level element stiffness matrix.

In a discretized design space, the topology of a structure can be visually represented as a black and white rendering of pixels (or voxels, in 3D). In these terms, the design problem involves finding the optimal subset  $\Omega_{\text{mat}}$  of material pixels. The set of admissible stiffness tensors consists of those tensors for which:

$$E_{ijkl} = 1_{\Omega^{\text{mat}}} E_{ijkl}^0, \quad 1_{\Omega^{\text{mat}}} = \begin{cases} 1 & \text{if } x \in \Omega^{\text{mat}} \\ 0 & \text{if } x \in \Omega \setminus \Omega^{\text{mat}} \end{cases} \quad (6)$$

$$\int_{\Omega} 1_{\Omega^{\text{mat}}} d\Omega = \text{Vol}(\Omega^{\text{mat}}) \leq V \quad (7)$$

The inequality expression [50] imposes a limit on the volume fraction  $V$  of material that can be used in the design, resulting in a minimum compliance design for a fixed

volume. The stiffness tensor  $E_{ijkl}^0$  is for a given isotropic material, which varies with point  $x$  over the domain. Solving this problem is most commonly achieved by replacing the integer variables ( $1_{\Omega^{mat}}$ ) with continuous variables, and applying a penalty that can direct the solution to have a binary 0–1, void-solid material distribution. With a fixed domain, the problem becomes a sizing problem by modifying the stiffness matrix to depend on a continuous function representing the density of the material. The function representing density is the design variable.

### 2.3.2 Solid Isotropic Material with Penalization Method

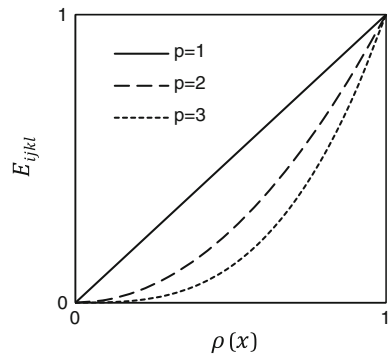
The introduction of a penalty allows for the design of a structure with regions of either solid material or void space, as opposed to an intermediate value. A popular and efficient penalization method is called “Solid Isotropic Material with Penalization”, i.e. SIMP [51]. Using SIMP, the sizing problem would be reformulated with a penalization factor  $p$  as:

$$E_{ijkl} = \rho(x)^p E_{ijkl}^0, \quad p > 1 \tag{8}$$

$$\int_{\Omega} \rho(x) d\Omega \leq V; \quad 0 \leq \rho(x) \leq 1, \quad x \in \Omega \tag{9}$$

The continuous density function  $\rho(x)$  is the design variable and  $E_{ijkl}^0$  is the isotropic base material stiffness. The stiffness tensor  $E_{ijkl}$  interpolates between 0 (void space) and  $E_{ijkl}^0$ . The penalization method is commonly used in structural optimization where intermediate values of material density do not have physical meaning, and a completely solid-void design is desired. With the exponent on the density function  $p \geq 1$ , values of density that are in the intermediate range are penalized because a smaller stiffness is obtained for a given material volume. Thus, it becomes uneconomical to use intermediate density values [51] (Fig. 2).

**Fig. 2** Effect of penalty factor  $p$  in SIMP method



Often,  $p \geq 3$  is used to obtain designs that are 0–1. Rietz shows that for a large enough  $p$ , there is a global optimum solution in 0–1 form, as long as the volume constraint is compatible [52]. Too severely penalizing the function, however, can result in a design that is a local minimum which is overly sensitive to the choice of the initial design. That is, the design skips too quickly to a 0–1 design. The choice of  $p$  is dependent on the design problem [51].

The physical interpretation of SIMP can be visualized using a composite or cellular material. If each pixel of a mesh is regarded as one unit cell, a design which has some grey regions can be achieved by designing the topology of each unit cell to match the required relative density.

### 2.3.3 Algorithms for Topology Optimization

Common algorithms specific to topology optimization include the Optimality Criteria (OC), the Method of Moving Asymptotes (MMA), and the Level Set Method (LSM) [53]. Evolutionary Structural Optimization (ESO) and Genetic Algorithms (GA) are alternatives to gradient-based methods, but are currently not as commonly used in topology optimization procedures for multi-functional bone tissue scaffolds. In this paper, we use two gradient based optimization procedures: OC and MMA, which are briefly explained below.

#### *Optimality Criteria Method*

In topology optimization, an iterative method is used to update relative density of each mesh element. In the method used here, the conditions of optimality are sought for the density of the minimum compliance design problem. For an iterative scheme integrated with the SIMP interpolation scheme, the optimality criteria are used to update the design variable  $\rho$  to achieve a stationary Lagrangian system. The relative densities are updated independently from the other elements and with respect to conditions of optimality, based on a previously computed design. The optimality criteria (OC) method is effective for large-scale topology optimization. However, the algorithm is not suitable for certain problems. For example, multiple objectives and constraints, and constraints of geometric nature, may require a more costly, and more robust mathematical programming method. The method of moving asymptotes is a versatile algorithm, well suited to address the limitations of the OC method.

#### *Method of Moving Asymptotes*

The method of moving asymptotes (MMA) is similar to Sequential Linear Programming and Sequential Quadratic Programming, and is well suited for topology optimization [48]. These methods solve smooth, non-linear optimization problems by using a sequence of subproblems which are simpler approximations. The sensitivity of a given design point and knowledge of previous iterations are used to decompose the problem into separable and convex subtasks. Subproblems can be solved each iteration using algorithms such as the dual method, or the interior point algorithm. The separable property of the approximation subproblems means that the

design variables are not coupled by the optimality conditions of the subproblems, and subproblems have a unique solution. The convexity property of the subproblems allows the use of dual methods or primal-dual methods. The combination of these properties generally results in a computationally efficient method commonly used in topology optimization [54].

While slower than OC, MMA offers an advantage because it can accommodate geometric scenarios with limited knowledge of the physical space. In addition, MMA can efficiently handle a large number of design variables and complex min-max functions. MMA has, however, one main disadvantage: convergence cannot be guaranteed [54]. A common experience with MMA is that if it converges, a solution is found quickly with steadily improving designs.

### *Numerical instabilities in topology optimization*

Topology optimization approaches often suffer from a variety of numerical instability problems, such as checkerboard patterns, mesh dependence, and computational inefficiency. Checkerboard patterns refer to the case where solid and void elements appear in alternating fashion, only connected by a corner, and create artificially high stiffness regions. To avoid this problem, higher order mesh elements can be used. Techniques such as local gradient constraints, filtering, and various material interpolation schemes, such as SIMP, are also used to eliminate the presence of checkerboard regions [53].

Mesh dependency is a numerical instability where increased mesh refinement results in a larger number of holes appearing in the optimal topology. Ideally, mesh refinement would result in improved boundaries of the optimal topology. One way to efficiently achieve mesh independent designs is to reduce the admissible design space with a global or local restriction on the variation of density. This can be achieved by either adding constraints to the optimization problem, reducing the parameter space directly, or applying filters in the optimization method. Convergence of finite element approximations can be found with the addition of one of these solutions [48].

One highly efficient filter to ensure mesh-independency is to modify design sensitivity, such that the sensitivity of an element is determined from a weighted average of the neighbouring element sensitivities. This filtering method is heuristic, but it is computationally efficient and simple to implement. It is reported that results are very similar to those obtained by a local gradient constraint [48].

## **3 Hierarchical Topology Optimization Algorithm**

This section describes each step of the algorithm and discusses the model assumptions. Section 4 reports the application of the methodology to the design of a bone fracture fixation plate with stiffness close to that of bone and maximum permeability.

### 3.1 Algorithm Structure

The procedure is divided into sequential material layout problems at two scales: (1) the topology of the implant, and (2) the topology of each unit cell to meet functional requirements at each location of the cellular material. Although some methods in literature perform these searches in parallel, here the design of each unit cell is obtained after the optimal topology of the scaffold is determined. The SIMP method is used to interpolate target material properties for each unit cell based on the optimal material distribution found at the implant optimization stage. The procedure is performed with in-house scripts implemented in MATLAB, in combination with ANSYS for finite element analysis. A detailed flowchart is shown in Fig. 3.

### 3.2 Implant Topology Optimization (I)

The first stage is to determine an optimal material layout for an implant with minimal compliance (or strain energy), to reduce the difference between the strain energy of the base material (in this case Ti-6Al-4V Ti6Al4V) and bone. Bone is much less stiff than titanium alloys used in implant design. The objective function  $c(\mathbf{x})$  to minimize compliance using a power based topology optimization is formulated as follows:

$$\begin{aligned} \min_{\mathbf{x}} : c(\mathbf{x}) &= \mathbf{U}^T \mathbf{K} \mathbf{U} = \sum_{e=1}^N (x_e)^p \mathbf{u}_e^T \mathbf{k}_e \mathbf{u}_e \\ \text{subject to} : & \frac{V(\mathbf{x})}{V_0} = V^* \\ & : \mathbf{K} \mathbf{U} = \mathbf{F} \\ & : \mathbf{0} < \mathbf{x}_{min} < \mathbf{x} \leq \mathbf{1} \end{aligned} \quad (10)$$

where  $\mathbf{x}$  is the design variable (relative density) of each element  $e$ ,  $\mathbf{U}$  is displacement and  $\mathbf{K}$  is the implant stiffness. The optimization scheme is governed by the structural elasticity equation, where  $\mathbf{x}_e$  is the vector of design variables, relative density,  $\mathbf{u}_e$  is the element displacement matrix, and  $\mathbf{k}_e$  is the element stiffness matrix [47]. A volume fraction constraint is applied as an equality constraint, with target volume  $V^*$  defined by the user. Volume fraction is measured as the ratio of solid material  $V(\mathbf{x})$  to the size of the design domain  $V_0$ . An inequality constraint on the design variable  $\mathbf{x}$  is imposed to restrict values of relative density to lie between a value close to zero and 1. A penalty factor  $p = 3$  is chosen to ensure convergence [48].

Figure 3 illustrates topology optimization at the implant level. A design space is first defined with both the displacement and stiffness fields discretized without mesh change. Identical four-node quadrilateral mesh elements are used, and a uniform relative density distribution is initially defined. Material properties of the solid phase

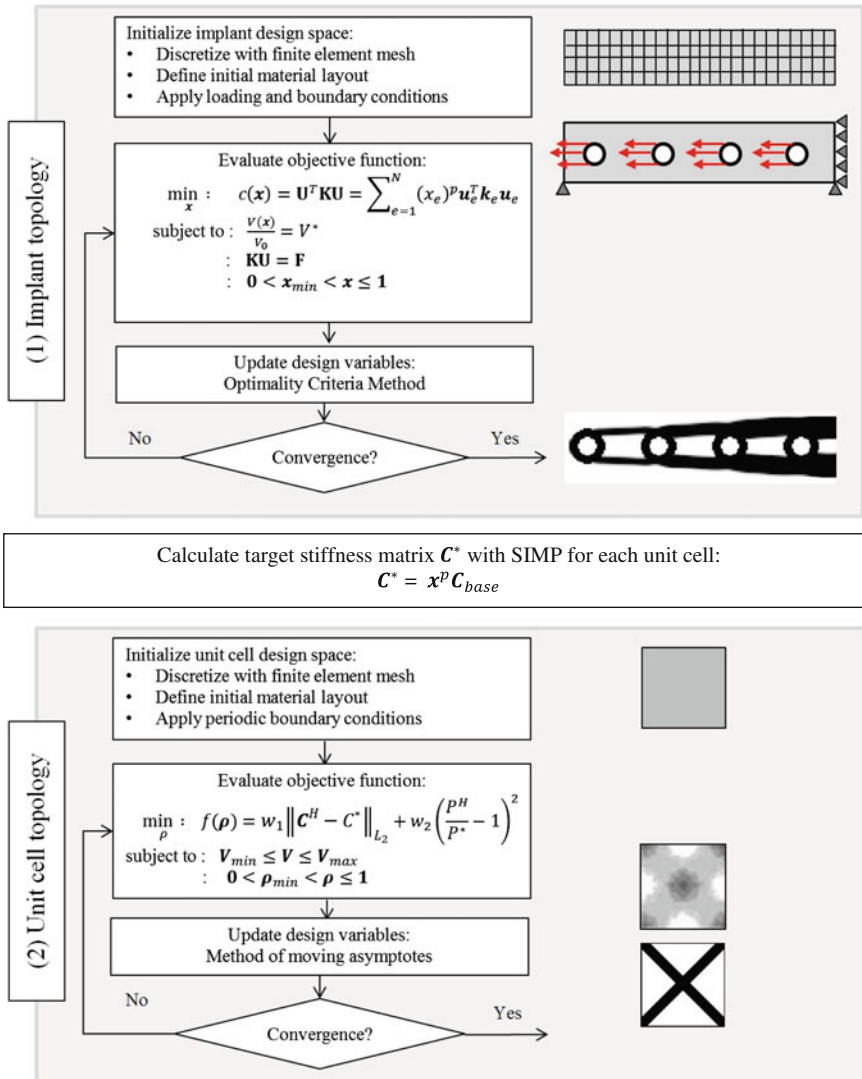


Fig. 3 Hierarchical topology optimization flowchart

are defined with Young’s modulus and Poisson’s ratio used to calculate the material stiffness. Plane stress is assumed in this two-dimensional analysis.

As shown in Fig. 3, the iterative procedure begins with the evaluation of the objective function for the initial material distribution within the design space. The sensitivity of the objective function is calculated, and a filtering technique is applied to smooth the gradients. The design variables are updated based on the filtered sensitivities, according to the Optimality Criteria (Sect. 2.3.3). A convergence check is



performed to evaluate the maximum change in relative density for each element from its current value to its value from the previous design. Convergence is found when the maximum change in relative density of all elements is below a defined threshold, in this case 0.02. When the design variables stabilize within the prescribed tolerance, the optimal material distribution is achieved.

With the optimal relative density for each unit cell determined, the target stiffness at each location of the implant is calculated. A relative density of 1 indicates that the desired stiffness of that unit cell is equal to the predefined base material stiffness. Similarly, a relative density near zero indicates that no stiffness is required at that location. The SIMP relation is used to interpret intermediate values of relative density as material properties for each unit cell:

$$\mathbf{C}^* = x^p \mathbf{C}_{base} \quad (11)$$

where the penalty factor is  $p = 3$ ,  $\mathbf{C}^*$  is the interpreted stiffness matrix of a unit cell with a given relative density  $x$ , and  $\mathbf{C}_{base}$  is the solid material stiffness. With target stiffness for each location of the implant defined, unit cells can be specifically designed to match the local requirements in the next stage of the optimization loop.

### 3.3 Unit Cell Topology Optimization (2)

The second stage involves the design of unit cells to achieve local target stiffness and permeability, based on bone ingrowth requirements. Each unit cell is designed independently from its surrounding unit cells. The objective function is formulated as follows:

$$\begin{aligned} \min_{\rho} : f(\rho) &= w_1 \|\mathbf{C}^H - \mathbf{C}^*\|_{L_2} + w_2 \left( \frac{P^H}{P^*} - 1 \right)^2 + w_3 \left( \frac{C_{22}^H}{C_{11}^H} - 1 \right)^2 \\ \text{subject to} : & \Phi_{min} \leq \Phi \leq \Phi_{max} \\ & : \mathbf{0} \leq \rho_{min} < \rho \leq \mathbf{1} \end{aligned} \quad (12)$$

where  $\Phi$  is the unit cell porosity and  $\rho$  is the relative density of each mesh element. The first term of the minimization problem represents the difference between target stiffness and effective stiffness of the unit cell. The  $L_2$  norm is calculated as the square of the difference between target ( $\mathbf{C}^*$ ) and homogenized ( $\mathbf{C}^H$ ) components. The target stiffness matrix is determined from Eq. 11. The effective stiffness terms are calculated using asymptotic homogenization (Sect. 2.1.1). The second term of the objective function is the squared difference between target and effective permeability of the unit cell. Effective permeability is calculated using Weissberg's formula (Eq. 1). The target permeability is also calculated using Weissberg's formula, which is solely a function of porosity,  $\Phi$ . The acceptable range of porosity for bone ingrowth is 60–80%, so the maximum porosity (80%) is used to determine the target permeability.

The third term in the objective function ensures square symmetry of the unit cell. The acceptable range of porosity is enforced using an inequality constraint, where  $\Phi_{min} = 60\%$  and  $\Phi_{max} = 80\%$ . An inequality constraint is also used to ensure that the relative density of each mesh element is between  $\rho_{min} = 0.001$  and  $1$ .

The design space is initialized with a uniform finite element mesh of four-node quadrilateral elements. In-plane stress is assumed. The iterative procedure begins with evaluating the objective function for the initial material distribution, based on the calculation of effective stiffness and permeability properties. The method of moving asymptotes is used to update the design variables (Sect. 2.3.3). A convergence check is performed to evaluate the maximum change in relative density for each element, from its current value to its value from the previous design. Convergence is reached when the maximum change in relative density of all elements is below a prescribed tolerance of  $1\%$ . This procedure is conducted for each unit cell, which can then be mapped to their respective locations within the implant structure. Further discussion on unit cell mapping is given in Sect. 4.

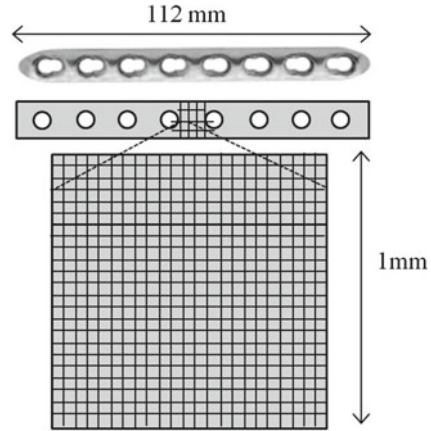
## 4 Application to Fracture Fixation Plate: Problem Definition

Before the problem definition describing the design requirements of fracture fixation plate, background information on the clinical aspects is provided. Internal fracture fixation plate and screw systems are a method of treating fractured long bones. The purpose of the mechanism is to provide necessary stabilization and a critical amount of compressive stress at the bone fracture site to facilitate healing. Additionally, the plate must minimize devascularisation, and allow early motion and partial loading to restore some load bearing capacity of the bone [55]. Compression also helps prevent transverse displacement of bone fragments and torque about the long axis of the bone [56]. The compression plate and screw components are typically made of solid, rigid, biocompatible materials such as stainless steel, cobalt chromium, titanium, and composites.

Ongoing concerns with fracture fixation plates are (1) excessive stiffness resulting in stress shielding, and (2) osteoporosis of underlying bone. The resulting decrease in bone mass and density increases the risk of re-fracture at the site [57]. One cause of osteoporosis beneath a fracture fixation plate is the disruption of the periosteal capillary network at the fracture site. Areas of bone in contact with the plate receive insufficient blood supply and necrosis follows. Low contact surface plates and limited contact dynamic compression plates have previously been designed to reduce the disruption of blood flow [57]. Stress shielding results from the mismatch of mechanical properties between bone tissue and plate, resulting in bone resorption. Lower stiffness and functionally graded bone plates have been investigated to address this issue with varying success in results [58, 59].

It is hypothesized in this paper that a hierarchically designed plate of cellular material can address both the stress shielding and osteoporosis problems currently occurring in fracture fixation plates. The porous nature of the cellular material reduces

**Fig. 4** Schematic of fracture fixation plate and a zoom of its mesh [60]

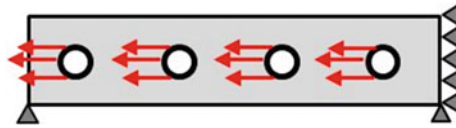


disruption of blood flow to the bone, while the mechanical properties are specifically tailored at the unit cell level to match that of the local tissue and reduce stress shielding. The design and optimization are performed in two dimensions, as the dominant forces acting on a fracture fixation plate are in-plane. First, the material distribution of the plate is determined. Secondly, a sample of unit cells for various locations throughout the plate are designed for target stiffness and permeability based on bone ingrowth requirements (Fig. 4). Loading and boundary conditions are specifically applied depending on the expected physiological loading of the implant. The topology optimization procedure is highly sensitive to loading and boundary conditions, so careful selection is essential. Finally, preliminary results on the mapping of unit cells into the plate structure are illustrated and discussed.

#### **4.1 Plate Topology Optimization**

The plate design space is initialized from the dimensions of a small fragment locking compression plate system by Synthes, Inc [60]. For an 8-screw plate with 3.5 mm hole diameter, the plate length is 112 mm and width is 12 mm. It is reported that a compression plate should provide approximately 600 N of compressive force [56]. In this analysis, a completely in-plane loading is assumed, and 600 N is distributed as a tensile force on the outer face of each screw hole, as shown in Fig. 5. Symmetry is exploited and only half of the plate is modeled, using a symmetric boundary condition (right side in Fig. 5).

The mesh resolution was varied and finally chosen as 48 elements by 448 elements. At higher mesh resolutions, no difference in topology was observed. One mesh element is equivalent to 0.25 mm in length and width. Each hole is prescribed to have void elements with a solid material boundary for screw threading. Elsewhere, the initial material distribution has uniformly 50% relative density. The target stiffness



**Fig. 5** Schematic of loading and boundary conditions for half of the fracture fixation plate

**Table 2** Optimal plate topologies with 50 % material fraction

Design	Optimal topology
(a) Solid	
(b) $p = 3$	
(c) $p = 1.25$	
(d) $p = 1.1$	
(e) $p = 1$	

for the implant is that of the in-plane tensile stiffness of cortical bone: approximately 20,092 N/mm. This was determined by a tensile stiffness analysis conducted in ANSYS for a solid plate with cortical bone properties ( $E = 20 \text{ GPa}$ ,  $\nu = 0.3$ ).






Table 2 shows the optimal Ti-6Al-4V plate topologies found using a 50 % material fraction equality constraint, with various penalty factors used to interpolate material properties. The first entry in the table shows a solid plate with 8 screw holes for comparison. Table 3 summarizes the strain energy and stiffness for the optimal plates. The solid plate has significantly higher stiffness than cortical bone, and is an order of magnitude greater than the target. Intuitively, lower stiffness is observed with a 50 % material fraction constraint. By using less material and allowing for intermediate values of relative density, a titanium plate can be designed with stiffness much closer to cortical bone.

For 50% volume fraction, it appears that the penalty factor should lie between 1 and 1.1 in order to achieve the target stiffness of cortical bone. A change in the material fraction can also result in plates with stiffness closer to the target. For example, the optimal plates with a 45 % material fraction constraint are shown in Tables 4 and 5, with the penalty factors of the 50 % volume fraction plates. Table 5 shows that with, the plate stiffness is 18,910 N/mm, which is approximately 6 % less

**Table 3** Strain energy and stiffness for 50 % material fraction plates

Design number and penalty factor, $p$	Material fraction (%)	Strain energy (N/mm)	Tensile stiffness (N/mm)
(a) Solid	94.11	10,974	204,857
(b) $p = 3$	50	4494	83,905
(c) $p = 1.25$	50	3215	60,013
(d) $p = 1.1$	50	1715	32,021
(e) $p = 1$	50	783	14,621

**Table 4** Optimal plate topologies with 45 % material fraction

Design	Optimal Topology
(a) Solid	
(b) $p = 3$	
(c) $p = 1.25$	
(d) $p = 1.1$	
(e) $p = 1$	

**Table 5** Strain energy and stiffness for 45 % material fraction plates

Design number and penalty factor, $p$	Material fraction (%)	Strain energy (N/mm)	Tensile stiffness (N/mm)
(a) Solid	94.11	10,974	204,857
(b) $p = 3$	45	4080	76,161
(c) $p = 1.25$	45	2654	49,555
(d) $p = 1.1$	45	1013	18,910
(e) $p = 1$	45	620	11,586

than that of cortical bone. The adjustment of input parameters allows for fine-tuning of the optimal results, to achieve target stiffness.

In general, a lower penalty factor yields higher gradients in material distribution, which can be realized by designing further at the unit cell level for each location, to achieve target material properties. The SIMP relation is used to interpolate effective material properties from intermediate values of relative density. This allows for new areas of the design space to be explored, beyond what is achievable with a solid

material. Design (d) in Table 4 is chosen as the optimal plate topology, and is further designed at the unit cell level in Sect. 4.2.

## 4.2 Unit Cell Topology Optimization

The optimization of the unit cell was achieved for a range of input parameters. Target stiffness properties were used for relative densities ranging from 10 to 90 %, with increments of 10 %. A penalty factor  $p$  was also varied between 1 and 3 in increments of 0.5, with the goal of achieving a completely solid void design. Each combination of penalty factor and relative density was repeated three times to ensure the repeatability of the procedure. Stiffness and permeability components of the object function were initially weighted equally. A mesh size of  $26 \times 26$  elements was used.

Convergence to an optimal topology was challenging to find. As reported in literature, convergence using the method of moving asymptotes is often not found [54]. Adjusting the MMA parameters, including step length for moving the asymptotes, were ineffective in yielding converging results. The best convergence was found with target stiffness properties determined by a relative density of 60 %. As shown in Table 6, stiffness can be found within the range of 3–13 % of this target, and permeability is found within 0.5–3.5 %. Porosity is also within the acceptable range of 60–80 %.

It was observed that higher penalty factors lead to more checkerboard patterns in the optimal design. The convergence to an optimal design with properties in the approximate range of the target was achieved; however regions of disconnected material within the unit cells and regions of intermediate material properties are present. This is a problem that needs to be addressed in a future study.



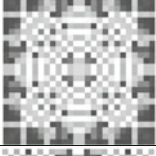
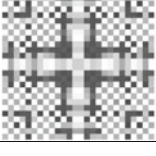
An optimal unit cell that has a maximum percent difference in stiffness of 7.43 % and a 1.43 % difference in permeability was compared to the Hashin-Shtrikman theoretical bounds of bulk modulus and diffusivity (Table 7).

Figure 6 shows that the effective bulk modulus for this unit cell falls beneath the theoretical bounds; however, the effective permeability lies well outside the predicted maximum value. This may be indicative of the inappropriate method chosen to calculate permeability solely as a function of porosity. The error may be reduced by using a more robust approach, which can account for the geometric features of the unit cell. Modifications to the permeability calculation method are further discussed in Sect. 5.

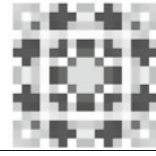
## 4.3 Unit Cell Mapping

This section presents preliminary results on the procedure to map unit cells into the optimal implant domain. As stated in Sect. 2, the optimal scaffold environment for bone ingrowth has pore sizes of 50–400  $\mu\text{m}$  and is 60–80 % porous, where the poros-

**Table 6** Optimal unit cells with target stiffness determined by a relative density of 60 %

	Porosity	p	% Difference in the stiffness matrix components			% Difference in permeability
	65.23	1	-5.52	4.80	0	-0.45
	65.77	1.5	-10.98	5.16	0	-1.42
	65.50	2	-13.72	-3.87	0	-0.91
	66.95	2.5	-13.17	-10.22	0	-3.46

**Table 7** Optimal unit cell compared against Hashin Shtrikman bounds

	Porosity	p	% Difference in the stiffness matrix components			% Difference in permeability
	65.79	2	-7.43	-1.44	0	-1.43

ity constraint is addressed in the unit cell optimization procedure. Theoretically, cell size can be tailored to meet pore size constraints during the procedure of mapping unit cells onto the optimal plate topology. Practically, manufacturing constraints on minimum allowable feature size govern the mapping procedure. Limits on the smallest possible pore and strut size determine the allowable unit cell dimensions. Currently, the nominal minimum strut size is approximately 200  $\mu\text{m}$  to be manufactured with additive processes, such as Electron Beam Melting. To address these constraints, instead of representing each mesh element as one unit cell, target material properties can be averaged over a region that has the size of the smallest manufacturable cell. For example, assuming the minimum cell length is 1 mm and the mesh element

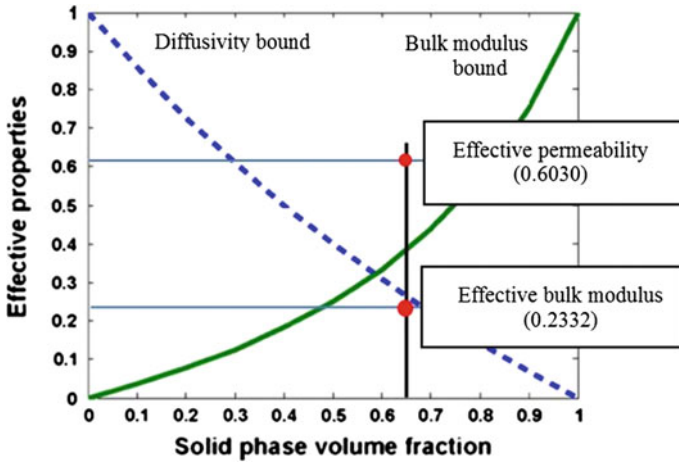


Fig. 6 Comparison of optimal unit cell effective properties to theoretical bounds [20]

Table 8 CAD representation of optimal unit cells with tessellation

	Optimized unit cell	CAD unit cell	Tessellation
1			
2			
3			

length is 0.25 mm, 1 mm square unit cells are mapped into the implant based on the average relative density over a  $4 \times 4$  element region.

To create implant models for manufacturing, computer-aided design (CAD) can be used to construct the optimized unit cells as solids. Table 8 shows three optimal unit cells that have been recreated using SolidWorks 3D CAD design software. Refined boundary interpretation methods, such as the level set method [61] can be used to translate the optimal material distribution to a solid-void unit cell topology. However, due to low mesh size and prevalence of a checkerboard pattern in the results shown,



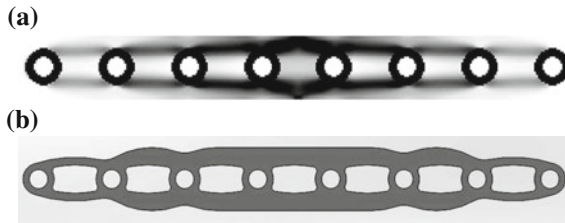


Fig. 7 CAD representation of optimized implant (material fraction 45 %,  $p = 1.1$ )

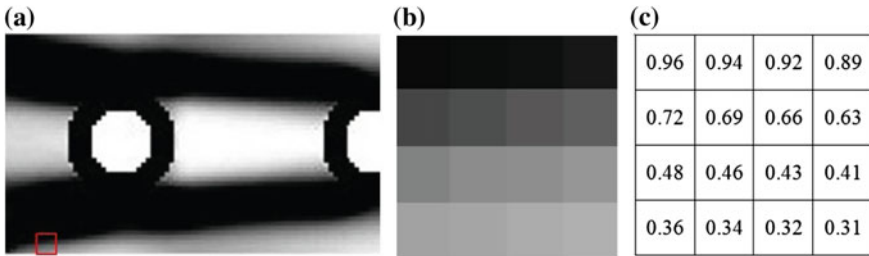


Fig. 8 a Optimized implant material layout where 1 mm  $\times$  1 mm region b is highlighted in red, with mesh element relative density shown in (c)

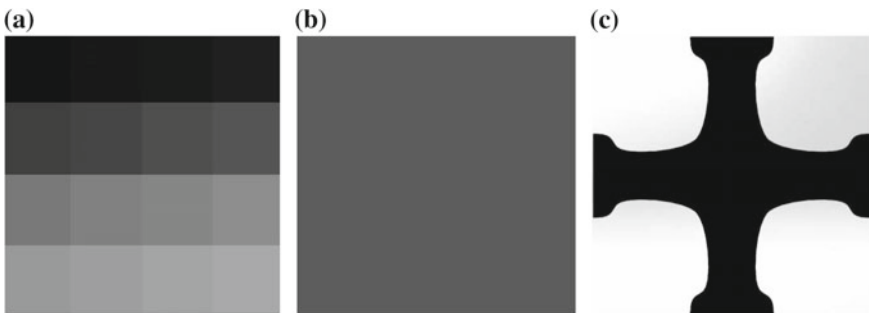
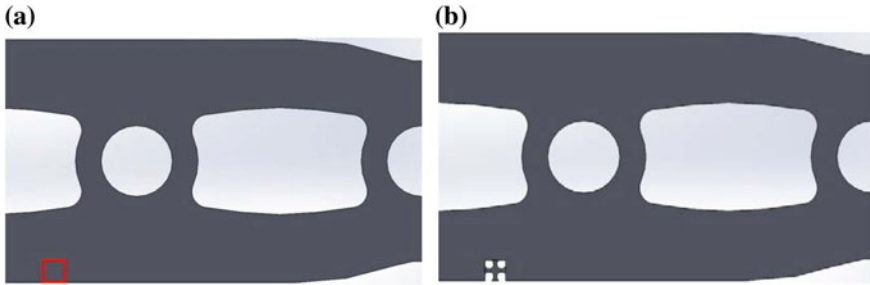


Fig. 9 a 1 mm  $\times$  1 mm region of optimal material distribution, with average relative density 60%. b 1 mm  $\times$  1 mm region of uniform relative density 60%. c Optimal unit cell with target material properties based on relative density of 60%

the implant topology is interpolated heuristically for simplicity. This poses major problems that need to be addressed in a further work. A tessellation of each unit cell is also shown to illustrate the respective scaffold topologies. In addition, it is important to note that the relevance of the CAD models is limited by the filtering technique used to interpolate the topology into a solid-void structure. In the results presented here, the filtering technique used is simple and must be improved.

Figure 7 shows a CAD model prior to the mapping, with solid material tentatively representing areas of relative density greater than  $\sim 20\%$ , and all other regions considered void. An example of the mapping procedure is shown in Figs. 8, 9 and 10.



**Fig. 10** **a** CAD model of Fig. 8a. **b** Mapping of optimal unit cell

Once the procedure is capable of optimizing unit cells to meet target properties based on the entire range of relative densities (from 0 to 100%), the entire implant can be mapped.

## 5 Discussion

### 5.1 Limitations and Outlook on Problem Resolution

This work is a preliminary step towards the development of a methodology for bone tissue scaffold. Substantial work and adjustments are required to address a number of issues, both computational and clinical, as described below.

**Effective permeability calculation.** As shown in Fig. 6, the method used to calculate permeability leads to values outside of theoretical bounds of permeability and stiffness. This indicates the Weissberg's formula chosen for the calculation, while simple as solely dependent on porosity, is not appropriate as expected. Alternative methods, such as Stokes flow homogenization, can be used to take into account the geometry of the unit cell. It is hypothesized that using Stokes flow homogenization may alter calculations for the permeability objective function and its sensitivity, thereby contributing to improve convergence.

**Filtering.** It is apparent that a more effective filtering technique is required at the unit cell level. The sensitivity filtering method is highly effective with the optimality criteria method used in the first stage of optimization. However, this filtering has been less effective at the unit cell level with unacceptable checkerboard patterns. There are other filtering techniques available that could be applied [47]. For example, limitations can be imposed on the allowable variation in density distribution. Restrictions to the gradient can be imposed with pointwise bounds on the derivatives of relative density with respect to mesh element location. Also, limits on the perimeter of mechanical elements in the design space can prevent solid material from appearing separately from the main structure [47]. While these techniques would have to be specifically catered to this design problem, they have been proven successful in

reducing checkerboard patterns [6, 20]. It is hypothesized that applying appropriate filtering at the unit cell level will reduce checkerboard patterns and a more distinct boundary between solid and void material will be produced.

**Sensitivity of final solution to initial design.** It is observed that the optimal material distribution in the unit cell design space is highly dependent on the initial material distribution at the beginning of the optimization. This dependency is not necessarily a drawback in this case, because many local minima of the function may exist that all exhibit target material properties. However, a search for a global minimum could be conducted with various initial designs.

**Connectivity of solid material within unit cell.** At the unit cell level, some optimal designs were achieved with disconnected solid material (Table 6), therefore unfeasible to manufacture. Connectivity of material within the unit cell can be imposed using available software packages on the market, such as the visualization toolkit image-processing library (*vtkPolyData-ConnectivityFilter*) by Kitware, Inc. This software is used in the iterative design of unit cells by Lin et al. [6] to ensure inner structure connectivity. One strategy is to identify the largest connected region in the design space and treat it as the main unit cell topology, disregarding unconnected material. FEA would be performed on this connected region and the material fraction constraint would be modified to apply to only the identified region. Connectivity between cells can also be enforced with prescribed regions of solid and void material, which are maintained throughout the iterative procedure. However, this limits the available design space and convergence to a minimum might be more challenging. Alternatively, a unit cell library with prescribed connectivity and optimized inner structures could be compiled, as proposed by Hollister and Lin [10].

**Translation of theoretical scaffold to CAD model.** The translation of theoretical optimal unit cells to CAD models for manufacturing can be eased by improving filtering. An appropriate filter will reduce checkerboard patterns and result in a design with a distinct boundary between solid and void regions. It was also shown that B-Spline based parametric smoothing functions are effective filters in topology optimization to control the size of the voids throughout the design domain, avoiding sharp changes in topology [62]. Furthermore, Sigmund et al. propose to perform a secondary shape optimization problem after topology optimization, with the optimal unit cell as an input. The shape optimization problem would smooth sharp corners in the optimal design, making manufacturing easier and eliminating stress concentrations [47].

**Clinical considerations.** There are many clinical considerations for the practical implementation of the proposed fracture fixation plate. The scope of this work is limited to two dimensions; hence comparing the stiffness and material properties to a 3D plate is not suitable. The thickness of the plate is an important contributing factor to stiffness, and bending and torsion should be accounted for too. However, the benefit of a cellular material design is supported by showing an increase in the accessible design space, allowing for specific tailoring of mechanical and permeability properties. It is important to note that the optimal implant topology resulting from this procedure is highly sensitive to the applied loads during optimization; hence a more realistic loading scenario may improve the final design of the implant.

## 5.2 Concluding Remarks

In conclusion, cellular materials provide a unique advantage in bone tissue scaffold design because material properties can be tailored to match non-homogeneous properties of bone. Inspired by the natural hierarchy of bone structure, a multi-scale designed scaffold can be proposed to closely mimic the physiological and mechanical response of bone tissue. Thus, controlling the hierarchical features of an artificial scaffold can allow for the optimization of function and tissue regeneration [11].

Hierarchical topology optimization can be used to design a scaffold with stiffness properties close to those of bone and high permeability for mass transport requirements. This was illustrated through the design of a fracture fixation plate in two dimensions. The optimized cellular design can reduce stress shielding by tailoring mechanical properties to match bone, and reduce the occurrence of osteoporosis by minimizing disruption of blood flow. In the proposed method, optimization convergence was not always found at the unit cell level, and several necessary improvements to the procedure have been suggested for future work. Nevertheless, with the target properties based on a relative density of 60%, optimal unit cells have been found within the range of 3–13% of desired stiffness and within 0.5–3.5% of desired permeability.

## References

1. Sikavitsas VI, Temenoff JS, Mikos AG (Oct 2001) Biomaterials and bone mechanotransduction. *Biomaterials* 22:2581–2593
2. van Gaalen S, Kruyt M, Meijer G, Mistry A, Mikos A, van den Beucken J et al (2008) Chapter 19—Tissue engineering of bone. In: van Blitterswijk C, Thomsen P, Lindahl A, Hubbell J, Williams DF, Cancedda R et al (eds) *Tissue engineering*. Academic Press, Burlington, pp 559–610
3. Lorna MFA, Gibson J (1997) *Cellular solids: structure and properties*. Cambridge University Press, Cambridge
4. Khurana JS (2009) Bone pathology. doi:10.1007/978-1-59745-347-9
5. Hollister SJ (2005) Porous scaffold design for tissue engineering. *Nat Mater* 4:518–524
6. Lin CY, Kikuchi N, Hollister SJ (2004) A novel method for biomaterial scaffold internal architecture design to match bone elastic properties with desired porosity. *J Biomech* 37:36–623
7. Woodruff MA, Lange C, Reichert J, Berner A, Chen F, Fratzl P et al (2012) Bone tissue engineering: from bench to bedside. *Mater Today* 15:430–435
8. Hollister SJ (2009) Scaffold design and manufacturing: from concept to clinic. *Adv Mater* 21:3330–3342
9. Hollister SJ, Murphy WL (2011) Scaffold translation: barriers between concept and clinic. *Tissue Eng Part B: Rev* 17:459–474
10. Hollister SJ, Lin CY (2007) Computational design of tissue engineering scaffolds. *Comput Methods Appl Mech Eng* 196:2991–2998
11. Hollister SJ, Maddox RD, Taboas JM (2002) Optimal design and fabrication of scaffolds to mimic tissue properties and satisfy biological constraints. *Biomaterials* 23:4095–4103
12. Khanoki SA, Pasini D (2013) The fatigue design of a bone preserving hip implant with functionally graded cellular material. *Trans ASME J Med Devices* 7:020908

13. Khanoki AS, Pasini D (2013) Fatigue design of a mechanically biocompatible lattice for a proof-of-concept femoral stem. *J Mech Behav Biomed Mater* 22:65–83
14. Abad EMK, Khanoki SA, Pasini D (2013) Fatigue design of lattice materials via computational mechanics: application to lattices with smooth transitions in cell geometry. *Int J Fatigue* 47: 126–136
15. Khanoki SA, Pasini D (2011) Multiscale design and multiobjective optimization of orthopaedic cellular hip implants. In: International design engineering technical conferences and computers and information in engineering conference, IDETC/CIE 2011, 28–31 Aug 2011, vol 2011, pp 935–944
16. Khanoki AS, Pasini D (2012) Multiscale design and multiobjective optimization of orthopedic hip implants with functionally graded cellular material. *J Biomech Eng* 134:031004
17. Abad EMK, Khanoki SA, Pasini D (2011) Shape design of periodic cellular materials under cyclic loading. In: ASME international design engineering technical conferences and computers and information in engineering conference, IDETC/CIE 2011, 28–31 Aug 2011. Washington, DC, United States, vol 2011, pp 945–954
18. Chen Y, Schellekens M, Zhou S, Cadman J, Li W, Appleyard R et al (2011) Design optimization of scaffold microstructures using wall shear stress criterion towards regulated flow-induced erosion. *J Biomech Eng* 133:081008
19. Knychala J, Bouropoulos N, Catt CJ, Katsamenis OL, Please CP, Sengers BG (2013) Pore geometry regulates early stage human bone marrow cell tissue formation and organisation. *Ann Biomed Eng* 41:917–930
20. Kang H, Lin C-Y, Hollister SJ (2010) Topology optimization of three dimensional tissue engineering scaffold architectures for prescribed bulk modulus and diffusivity. *Struct Multidiscip Optim* 42:633–644
21. Elsayed MSA, Pasini D (2010) Analysis of the elastostatic specific stiffness of 2D stretching-dominated lattice materials. *Mech Mater* 42:709–725
22. Elsayed MSA, Pasini D (2009) Characterization and performance optimization of 2D lattice materials with hexagonal Bravais lattice symmetry. In: ASME international design engineering technical conferences and computers and information in engineering conference, DETC2009, 30 Aug–2 Sept 2009. San Diego, CA, United States, vol 2010, pp 1315–1323
23. Elsayed MSA, Pasini D (2008) Multiscale model of the effective properties of the octet-truss lattice material. In: 12th AIAA/ISSMO multidisciplinary analysis and optimization conference, MAO, 10–12 Sept 2008, Victoria, BC, Canada
24. Elsayed MSA, Pasini D (2010) Multiscale structural design of columns made of regular octet-truss lattice material. *Int J Solids Struct* 47:1764–1774
25. Elsayed MSA, Clement H, Pasini D (2009) Structural analysis of pin jointed lattice structures. In: 3rd international conference on advances and trends in engineering materials and their applications, AES-ATEMA'2009, 6–10 July 2009. Montreal, QC, Canada, pp 51–57
26. Elsayed MSA, Pasini D (2010) Theoretical and experimental characterization of the 34.6 2D lattice material. In: ASME international design engineering technical conferences and computers and information in engineering conference, IDETC/CIE2010, 15–18 Aug 2010. Montreal, QC, Canada, vol 2010, pp 11–20
27. Vigliotti A, Pasini D (2012) Linear multiscale analysis and finite element validation of stretching and bending dominated lattice materials. *Mech Mater* 46:57–68
28. Vigliotti A, Deshpande VS, Pasini D (2014) Non linear constitutive models for lattice materials. *J Mech Phys Solids* 64:44–60
29. Vigliotti A, Pasini D (2012) Stiffness and strength of tridimensional periodic lattices. *Comput Methods Appl Mech Eng* 229–232:27–43
30. Arabnejad S, Pasini D (2013) Mechanical properties of lattice materials via asymptotic homogenization and comparison with alternative homogenization methods. *Int J Mech Sci* 77:249–262
31. Challis VJ, Roberts AP, Wilkins AH (2008) Design of three dimensional isotropic microstructures for maximized stiffness and conductivity. *Int J Solids Struct* 45:4130–4146
32. Gibson MFALJ, Harley BA (2010) Cellular materials in nature and medicine. Cambridge University Press, Cambridge

33. Vigliotti A, Pasini D (2013) Mechanical properties of hierarchical lattices. *Mech Mater* 62: 32–43
34. Khanoki SA, Pasini D (2013) Fatigue design of a mechanically biocompatible lattice for a proof-of-concept femoral stem. *J Mech Behav Biomed Mater* 22:65–83
35. Guest JK, Prévost JH (2006) Optimizing multifunctional materials: Design of microstructures for maximized stiffness and fluid permeability. *Int J Solids Struct* 43:7028–7047
36. Challis VJ, Guest JK, Grotowski JF, Roberts AP (2012) Computationally generated cross-property bounds for stiffness and fluid permeability using topology optimization. *Int J Solids Struct* 49:3397–3408
37. Coelho PG, Fernandes PR, Guedes JM, Rodrigues HC (2007) A hierarchical model for concurrent material and topology optimisation of three-dimensional structures. *Struct Multidisc Optim* 35:107–115
38. Chen Q, Huang S (2013) Mechanical properties of a porous bioscaffold with hierarchy. *Mater Lett* 95:89–92
39. Rodrigues H, Guedes JM, Bendsoe MP (2002) Hierarchical optimization of material and structure. *Struct Multidisc Optim* 24:1–10
40. Fernandes PR, Coelho PG, Guedes JM, Rodrigues HC (2010) Hierarchical optimization of the structure and the material used in its manufacture. *Mecánica Computacional XXIX*:405–415
41. Coelho ALV, Fernandes E, Faceli K (2011) Multi-objective design of hierarchical consensus functions for clustering ensembles via genetic programming. *Dec Support Syst* 51:794–809
42. Coelho PG, Fernandes PR, Rodrigues HC, Cardoso JB, Guedes JM (2009) Numerical modeling of bone tissue adaptation—a hierarchical approach for bone apparent density and trabecular structure. *J Biomech* 42:830–837
43. Hassani B (1996) A direct method to derive the boundary conditions of the homogenization equation for symmetric cells. *Commun Numer Methods Eng* 12:185–196
44. Hollister SJ, Kikuchi N (1992) Comparison of homogenization and standard mechanics analyses for periodic porous composites. *Comput Mech* 10:73–95
45. Weissberg HL (1963) Effective diffusion coefficient in porous media. *J Appl Phys* 34:2636–2639
46. Arce P, Locke BR, Trinh S (2000) Effective diffusivities of point-like molecules in isotropic porous media by Monte Carlo simulation. *Transp Porous Media* 38:241–259
47. Sigmund O (1994) Materials with prescribed constitutive parameters: an inverse homogenization problem. *Int J Solids Struct* 31:2313–2329
48. Bendsoe MP, Sigmund O (2003) *Topology optimization: theory, methods, and applications*. Springer, Berlin
49. Suzuki K, Kikuchi N (1991) A homogenization method for shape and topology optimization. *Comput Methods Appl Mech Eng* 93:291–318
50. Cowin SC (2007) The significance of bone microstructure in mechanotransduction. *J Biomech* 40(Suppl 1):S105–S109
51. Rozvany GIN (2001) Aims, scope, methods, history and unified terminology of computer-aided topology optimization in structural mechanics. *Struct Multidisc Optim* 21:90–108
52. Rietz A (2001) Sufficiency of a finite exponent in SIMP (power law) methods. *Struct Multidisc Optim* 21:159–163
53. Cadman JE, Zhou S, Chen Y, Li Q (2013) On design of multi-functional microstructural materials. *J Mater Sci* 48:51–66
54. Zilber C (1993) A globally convergent version of the method of moving asymptotes. *Struct Optim* 6:166–174
55. Tacvorian EK (2012) Evaluation of canine fracture fixation bone plates. Degree of Master of Science, Biomedical Engineering, Worcester Polytechnic Institute
56. Dwyer T (2009) The bone school. <http://www.boneschool.com/trauma/principles-internal-fixation>
57. Fouad H (2010) Effects of the bone-plate material and the presence of a gap between the fractured bone and plate on the predicted stresses at the fractured bone. *Med Eng Phys* 32:783–789

58. Ganesh VK, Ramakrishna K, Ghista D (2005) Biomechanics of bone-fracture fixation by stiffness-graded plates in comparison with stainless-steel plates. *BioMed Eng OnLine* 4:1–15
59. Perren KKSM, Pohler O, Predieri M, Steinemann S, Gautier E (1990) The limited contact dynamic compression plate (LC-DCP). *Arch Orthop Trauma Surg* 109:304–310
60. Synthes I (2002) Small fragment locking compression plate (LCP) system. In: Synthes D (ed) *Stainless steel and titanium*
61. Wang MY, Wang X, Guo D (2003) A level set method for structural topology optimization. *Comput Methods Appl Mech Eng* 192:227–246
62. Sundararajan VG (2010) Topology optimization for additive manufacturing of customized meso-structures using homogenization and parametric smoothing functions. *Engineering*, The University of Texas at Austin, Master of Science in Engineering

# Boundary Constraint Handling Affection on Slope Stability Analysis

Amir H. Gandomi, Ali R. Kashani and Mehdi Mousavi

**Abstract** In an engineering optimization problem such as soil slope problem, each design variable has permissible solution domain. Therefore, efficiency of an optimization algorithm may be affected by the method used for keeping the solutions within the defined boundaries or boundary constraint handling method. Despite importance of selecting constraint handling approach, there aren't adequate studies in this field. Heterogeneous slope stability optimization in the presence of a band of weak soil layer is considered as a complex geotechnical problem that requires satisfying boundary constraints. Evolutionary boundary constraint handling is one of the recently proposed methods that is very easy to implement and very effective. The present study intended to improve the optimization results by means of evolutionary boundary constraint handling scheme on slope stability optimization problem. In the current chapter five benchmark problems are analyzed using absorbing and evolutionary boundary constraint handling schemes and their results are compared to check the validity of this method. Based on achieved results optimization algorithm performance is improved by using the proposed boundary constraint handling method.

## 1 Introduction

Newly heuristic optimization methods have found a reliable position to solve geotechnical engineering problems. One of the most important geotechnical engineering problems is slope stability analysis. The consideration of non-circular slip surface has produced more efficient results in the heterogeneous soil slopes. The safety of slope is expressed in term of the factor of safety (FOS) and the limit equilibrium

---

A.H. Gandomi (✉)

BEACON Center for the Study of Evolution in Action, Michigan State University,  
East Lansing, MI 48824, USA  
e-mail: a.h.gandomi@gmail.com

A.R. Kashani · M. Mousavi

Department of Civil Engineering, Arak University, Arak, Iran



approach has been the most popular method in computing this factor. This method uses the plastic limit theorem of solid mechanics to analyze the stability of the potential slippery mass [33]. Large numbers of selective slip surfaces are required to be tested to find location of minimum factor of safety in order to use limit equilibrium method for slope stability analysis.

Slope stability analysis with non-circular slip surface is considered as a complicated optimization problem. As Chen and Shao [5] demonstrated, the objective function has a lot of local minimum within solution domain. Cheng et al. [8] also pointed out the objective function is usually non-convex and discontinuous over the search space. It is necessary to select a good initial failure surface to apply classical optimization techniques.

Recently by developing metaheuristic optimization algorithms, it is possible to overcome this issue. Several metaheuristic algorithms have been adopted to slope stability problems. Monte carlo random walk type was used by Greco [20]; genetic algorithm was applied by Goh [19], Das [12], McCombie and Wiklson [31], Zolfaghari et al. [45], Jianping et al. [22] and Sengupta and Upadhyay [36]; leap frog was used by Bolton et al. [3]; ant colony optimization selected by Kahatadenya et al. [23]; artificial neural network optimization technique was tried by Samui and Kumar [35]; fuzzy logic has also been adopted to find critical slip surface several simple slope stability problems by Mathada et al. [30], Rubio et al. [34] and Giasi et al. [18].

Cheng [6] and Cheng et al. [7–9] studied simulate annealing, harmony search, tabu search, particle swarm optimization and fish swarm for finding minimum FOS. Newly Cheng et al. [10] utilized a hybrid approach for locating the critical failure surface; Morgenstern and Price [32] used ant colony optimization for slope stability optimization; Khajehzadeh et al. [26, 27] used gravitational search algorithm and modified particle swarm optimization respectively; Zhao et al. [42] tried relevance vector machine in slope stability analysis and Kaveh and Talatahari [25] studied imperialistic competitive algorithm performance on 2-dimensional soil slopes.

Good optimization will be achieved by providing two requirements; a robust algorithm and proper handling of constraints. Boundary constraint handling is one of the most important parts of constraint handling that can affect power of algorithms. Unlike the importance of constraint handling method, there are limited studies in this area.

For the first requirement cuckoo search (CS) algorithm, proposed by Yang and Deb [41] is selected based on its satisfying records. CS is a new metaheuristic optimization technique inspired by reproduction strategy of some cuckoo species. The initial test of CS algorithm shows that this algorithm is very efficient for some benchmark optimization problems [41]. The CS algorithm has also been used to some structural and geotechnical engineering problems to reach optimum design by Gandomi et al. [15, 16], respectively.

Like most optimization algorithms, new produced solution of CS in each iteration may be gone beyond the boundaries. In this case traditional absorbing scheme was utilized by original CS. Recently Gandomi and Yang [17] developed a simple and effective method for boundary constraint handling that is so-called evolutionary boundary constraint handling (EBCH). This evolutionary scheme is also very easy

to implement for any optimization algorithm. The results showed that EBCH can outperform other existing methods. Therefore for the second requirement, EBCH is selected to handle boundary constraints.

The current study is allocated to contrast the location of critical slip surface using original CS and CS with EBCH called CS\_EB. Assuming non-circular slip surface for Morgenstern-Price [32] method, FOS is calculated. Five different case studies are evaluated here to show the efficiency of the proposed method. As a result in all cases, better results are gained using CS\_EB than CS. This fact is magnified in more complicated cases and CS\_EB are capable to evade local minima far better than CS.

## 2 Slope Stability Analyzing

### 2.1 Generation of Trial Slip Surface

An acceptable slip surface is required to be generated to find critical failure surface. A proper slip surface should be concave upward to be cinematically acceptable. Procedure proposed by Cheng [6] is used to shape slip surface. Slope geometry in Cartesian coordinate system  $XOY$  is shown as Fig. 1. Slope geometry and bedrock are defined by  $y = g(x)$  and  $y = B(x)$  mathematical functions, respectively.

By dividing slippery mass into  $n$  vertical slices,  $(n + 1)$  edge coordinates of each slice have to be determined. Therefore  $V$  vector, containing control variable, is defined for optimization as follows:

$$V = [x_1, y_1, x_2, y_2, \dots, x_n, y_n, x_{n+1}, y_{n+1}] \tag{1}$$

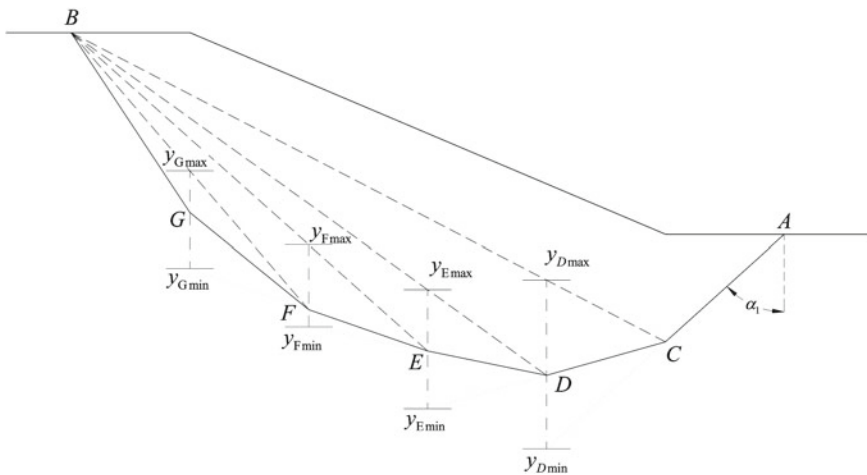


Fig. 1 Generation of Non-circular slip surface

The widths of all the slices are considered to be equal for simplicity. Then  $x_i$  can be computed as follows:

$$x_{i+1} = x_i + \frac{x_{n+1} - x_1}{n} \times (i - 1) \tag{2}$$

In this method upper and lower bound for  $y_i$  value,  $y_{imax}$  and  $y_{imin}$  are defined as slope geometry and bedrock, respectively. A random value between  $y_{imin}$  and  $y_{imax}$  is selected for  $y_i$  as Eq. (3).

$$y_i = rand \times (y_{i \max} - y_{i \min}) \tag{3}$$

Finally a trial slip surface will be defined by using above mentioned control variables.

### 2.2 Factor of Safety Calculation

A quantitative value is defined as FOS to explore the stability of a slope. In this study a concise procedure proposed by Zhu et al. [44] is used. By considering effective inter-slice forces as Fig. 2, FOS could be calculated by an iterative procedure as follows:

First, calculate  $R_i$  and  $T_i$  using Eqs. 4 and 5;

$$R_i = [W_i \cos \alpha_i - W_i \alpha_h \sin \alpha_i + Q_i \cos(\omega_i - \alpha_i) - U_i] \times \tan \phi'_i + c'_i b_i \sec \alpha_i \tag{4}$$

$$T_i = W_i \sin \alpha_i + W_i \alpha_h \cos \alpha_i - Q_i \sin(\omega_i - \alpha_i) \tag{5}$$

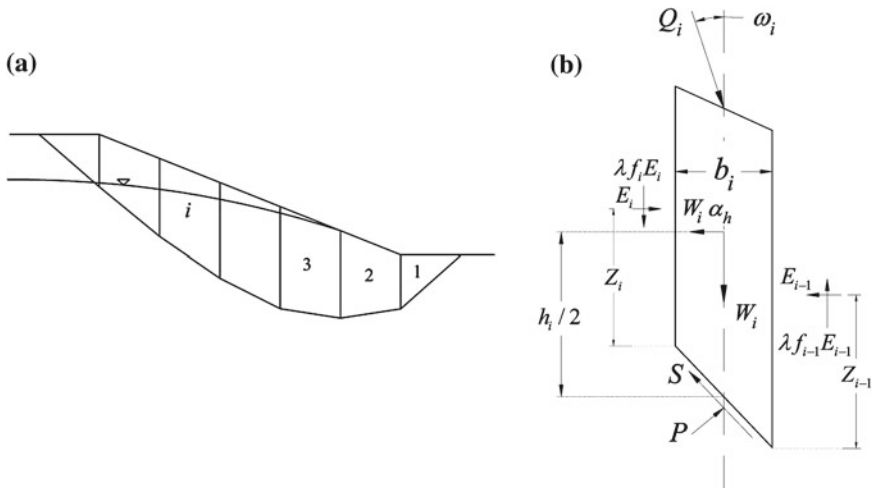


Fig. 2 a General failure surface, b Inter-slice forces in slice number  $i$

Second, specify inter slice force function,  $f(x)$  (it could be chosen constant, sine, half sine), as Eq. 6:

$$f(x) = \sin\left(\pi \times \frac{x-a}{b-a}\right) \quad (6)$$

where  $a$  and  $b$  are  $x$ -coordinates of two ends of slip surface. In this study constant function is selected.

Third, consider initial values of  $F_s$  and  $\lambda$  so that Eq. 7 will be satisfied.

$$F_s > -\frac{\sin \alpha_i - \lambda f_i \cos \alpha_i}{\cos \alpha_i + \lambda f_i \sin \alpha_i} \tan \phi' \quad (7)$$

Fourth, calculate  $\Phi_i$  and  $\Psi_{i-1}$  using Eqs. 8 and 9 for all the slices.

$$\Phi_i = (\sin \alpha_i - \lambda f_i \cos \alpha_i) \tan \phi'_i + (\cos \alpha_i + \lambda f_i \sin \alpha_i) F_s \quad (8)$$

$$\Psi_{i-1} = [(\sin \alpha_i - \lambda f_i \cos \alpha_i) \tan \phi'_i + (\cos \alpha_i + \lambda f_i \sin \alpha_i) F_s] / \Phi_{i-1} \quad (9)$$

Fifth, calculate  $F_s$  using Eq. 10.

$$F_s = \frac{\sum_{i=1}^{n-1} (R_i \prod_{j=i}^{n-1} \Psi_j) + R_n}{\sum_{i=1}^{n-1} (T_i \prod_{j=i}^{n-1} \Psi_j) + T_n} \quad (10)$$

Sixth, repeat forth step with new  $F_s$  and compute  $F_s$  again with new  $\Phi_i$  and  $\Psi_{i-1}$  values using Eq. 8.

Seventh, calculate  $E_i$  using Eq. 11 by updated  $F_s$  value for all the slices.

Finally, calculate  $\lambda$  using Eq. 12.

$$E_i \Phi_i = \Psi_{i-1} E_{i-1} \Phi_{i-1} + F_s T_i - R_i \quad (11)$$

$$\lambda = \frac{\sum [b_i (E_i + E_{i-1}) \tan \alpha_i + W_i \alpha_h h_i + 2Q \sin \omega_i h_i]}{\sum [b_i (f_i E_i + f_{i-1} E_{i-1})]} \quad (12)$$

Repeat all the eight above mentioned steps to  $F_s$  and  $\lambda$  converge to nearly constant values.

## 3 Optimization Techniques

### 3.1 Cuckoo Search

Cuckoo search (CS) algorithm is one of the swarm intelligence metaheuristic optimization algorithms. CS, inspired by cuckoo's life, has been proposed by Yang and Deb [41] recently. The cuckoos are fascinating because of their kind of reproduction strategy. Their eggs are laid in the nest of other host birds, nearly other species. At the same time they throw away host bird eggs to raise their egg hatching probability. Cuckoos are able to select recently spawned nests. Generally the cuckoo chicks are capable to hatch slightly earlier than host bird chickens. The cuckoo's hatchling will evict the other eggs by blindly propelling them instinctively. Also cuckoos are specialized to mimic the call of its host bird. In this way cuckoo chick can increase its share of food. However, some host birds can combat with infringing cuckoos. If these birds discover alien egg either throw this egg away or abandon the whole nest and build a new one.

In nature, animals and insects try to find food by following a random or quasi random prototype. Based on random walk which can model animals foraging path, the next move is derived from current position based on a probability which can be modeled mathematically.

In order to ease three idealized regulations proposed by Gandomi et al. [16]:

- Each cuckoo flyblows one egg at a time, and leaves it in an arbitrarily chosen nest.
- The best nests (solution) with highest quality of eggs will usable over the next generations.
- The number of available host nests is fixed, and a host can discover an alien egg with a probability  $P_a \in [0, 1]$ . If this encroachment has occurred, the host bird goes for either getting rid of the alien egg or leaving the nest and building a completely new one in a new location.

Inductively, each solution (considered as nest) will be replaced by a new one with a probability of  $P_a$ .

CS defines the main problem that optimization will be done for; completely similar to other popular optimization algorithms (i.e., GA, PSO and so on) as Objective Function.

By considering above three rules, CS conform the following procedures:

New solution using Levy-Flight is related to the current solution by Eq. 13.

$$x_i^{(t+1)} = x_i^{(t)} + \alpha \oplus Levy(\lambda) \quad (13)$$

where  $\alpha > 0$  is step size parameter which is supposed to be change with the scales of the problem. Mostly,  $\alpha$  sets to unity. The product  $\oplus$  means entry-wise multiplications.

The Levy term provides a random walk type search, and the probability distribution defined as Eq. 14 that has an infinite variance with an infinite mean.

$$Levy \sim u = t^{-\lambda}, (1 \leq \lambda \leq 3) \quad (14)$$

By implementing above procedure iteratively CS will approach the nearest best solution for minimization problems. For more detail refer to the main source (i.e., [41]).

### 3.2 Evolutionary Boundary Constraint Handling

In many optimization algorithms new solutions may be violated from allowable range of variables within reproduction procedures. There are proposed several methods to push the solution inside boundaries. The classical boundary constraint handling method is absorbing which is presented in Eq. 15. The other methods are random scheme as Eq. 16 and the toroidal space scheme as Eq. 17 or some schemes like replacing components with a mirror image relative to the boundary as Eq. 18.

$$f(z_i \rightarrow x_i) = \begin{cases} lb_i & \text{if } z_i < lb_i \\ ub_i & \text{if } z_i > ub_i \end{cases} \quad (15)$$

$$f(z_i \rightarrow x_i) = lb_i + rand \times (ub_i - lb_i) \quad \text{if } z_i < lb_i \text{ or } z_i > ub_i \quad (16)$$

$$f(z_i \rightarrow x_i) = \begin{cases} lb_i - (z_i - lb_i) & \text{if } z_i < lb_i \\ ub_i - (z_i - ub_i) & \text{if } z_i > ub_i \end{cases} \quad (17)$$

$$f(z_i \rightarrow x_i) = \begin{cases} ub_i + z_i - lb_i & \text{if } z_i < lb_i \\ lb_i + z_i - ub_i & \text{if } z_i > ub_i \end{cases} \quad (18)$$

where  $lb_i$  and  $ub_i$  are the  $i$ th lower bound and upper bound, by order,  $z_i$  and  $x_i$  are violated component and corrected component and  $rand$  is a random number between 0 and 1.

Also some literature devoted to examine certain methods for boundary constraint handling such as: Haung and Mohan [21] used damping scheme in particle swarm (PSO), Xu and Rahmat-Samii [39] proposed some hybrid methods in PSO, Chu et al. [4] proposed a method in PSO based on reducing velocity, Chu et al. [11] done a comparative study using various boundary constraint handling methods in PSO and Kaveh and Talatahari [25], proposed a harmony search-based method.

Recently Gandomi and Yang [17], developed an evolutionary boundary constraint handling (EBCH) in Differential Evolution (DE) algorithm that is examined on wide set of benchmark problems. Not only EBCH is simple and can be used in any optimization algorithm, but also it is efficient and can simply outperform the other existing methods. In the current study, EBCH method adopted on CS algorithm and the results are compared to original CS that is used classical absorbing scheme.

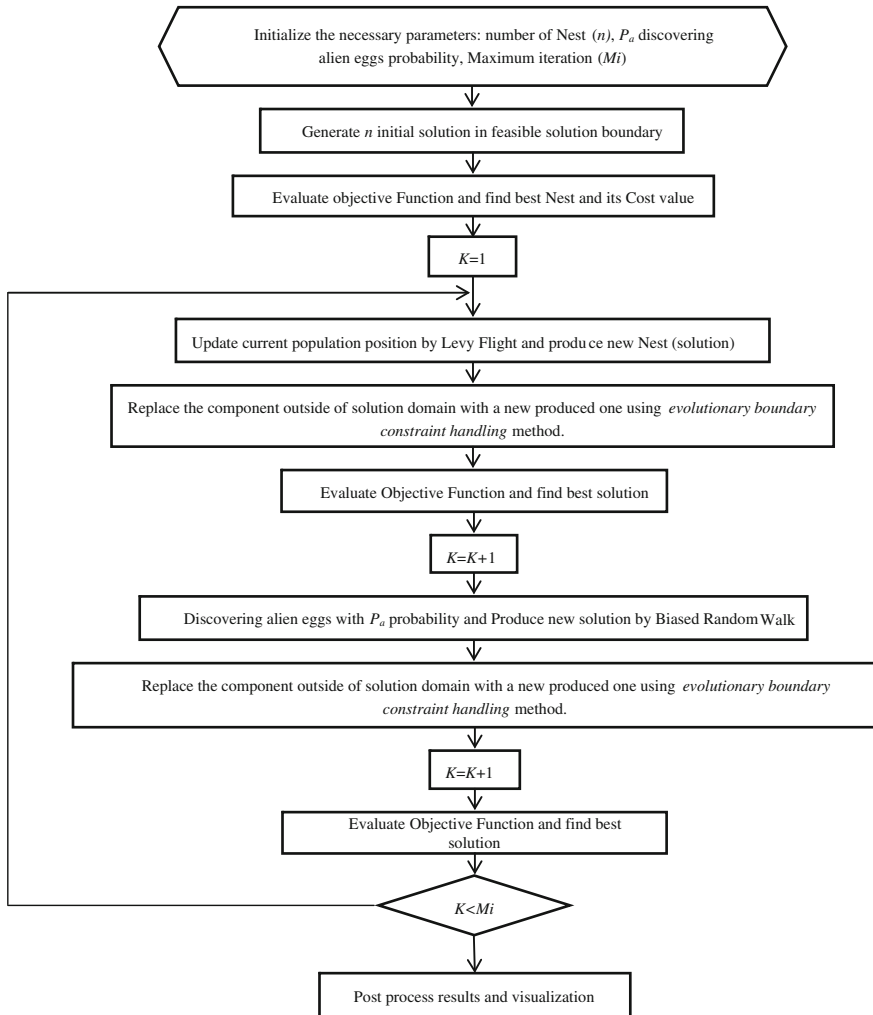


Fig. 3 The CS\_EB flowchart

In this method, if a component goes outside of boundaries, this component replace with new one produced using the following mutation operator:

$$f(z_i \rightarrow x_i) = \begin{cases} \alpha \times lb_i + (1 - \alpha)x_i^b & \text{if } z_i < lb_i \\ \beta \times ub_i + (1 - \beta)x_i^b & \text{if } z_i > ub_i \end{cases} \quad (19)$$

in which  $x_i^b$  is the related component of the best solution, and  $\alpha$  and  $\beta$  are random number between 0 and 1.

Representation of the CS\_EB algorithm is presented in Fig.3.

## 4 Numerical Simulation

In order to compare proposed algorithm efficiency, five soil slope examples are solved and final results are reported. In this chapter for evaluation of Factor of Safety number of slices is considered equal to 25. Furthermore because of the chaotic operation of optimization algorithm all the examples run about 20 times and results are reported by the best value, mean and standard deviation to illustrate the performance of every algorithm more efficient. The CS parameters used in this study are shown in Table 1. Number of nests is considered 50 and number of iteration is equal to 3000, therefore the number of function evaluation will be 15,000 times.

### 4.1 Case I

The first case is a homogeneous slope with an effective friction angle  $\phi$  of  $10^\circ$ , an effective cohesion intercept  $c$  of 9.8 kPa, a unit weight  $\gamma$  of  $17.64 \text{ kN/m}^3$  selected from the work by Yamagami and Ueta [40]. The geometry of slope and slip surfaces are as Fig. 4.

This example was analyzed by Yamagami and Ueta [40] for the first time, and then it was analyzed in the works of Greco [20] by pattern search and the Monte-Carlo methods, Solati and Habibagahi [37] by genetic algorithm, Kahatadeniya et al. [23] by ant colony optimization (ACO). In the current study this example solved once again by using CS and CS\_EB to explore these algorithms efficiency. As shown in Table 1, the resulted FOS values from CS and CS\_EB are equal, but the lower value of standard deviation of CS\_EB proves better performance of this algorithm respect to the original CS. Table 2 shows the previous studies in which FOS was computed.

### 4.2 Case II

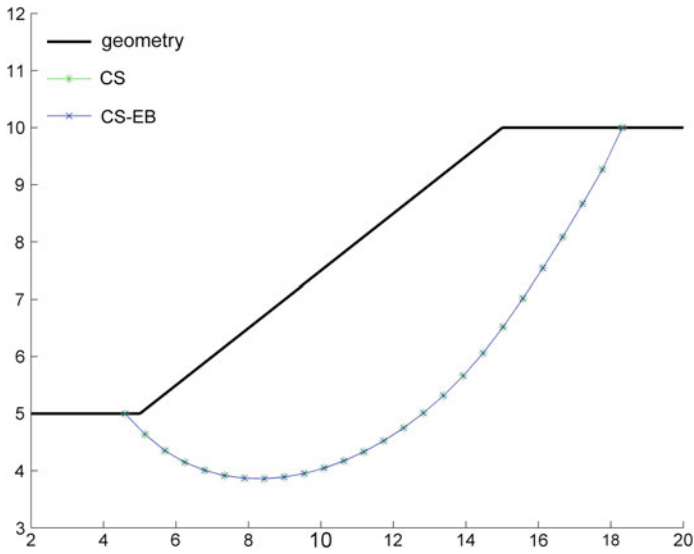
The second case is selected from the work by Arai and Tagyo [1]. In this case, a weak soil layer is stated between two stronger ones. The soil properties, geometry of slope and slip surfaces are as Table 3 and Fig. 5, respectively.

This example is surveyed in the literature, for example Arai and Tagyo [1] used Janbu's simplified method in combination with the conjugate gradient method,

**Table 1** Values of FOS comparison for Case I

Optimization algorithm	CS	CS_EB
Mean	1.3206	1.3206
Best	1.3206	1.3206
Standard deviation	2.08E-08	1.05E-08





**Fig. 4** Slope geometry and critical slip surface for Case I

**Table 2** Previous studies computed FOS for Case I

References	Optimization algorithm	FOS
Yamagami and Ueta [40]	Broyden–Fletcher–Goldfarb–Shanno (BFGS)	1.338
Yamagami and Ueta [40]	Davidon–Fletcher–Powell (DFP)	1.338
Greco [20]	Pattern search	1.326–1.330
Greco [20]	Monte Carlo	1.327–1.333
Malkawi et al. [29]	Monte Carlo	1.238
Solati and Habibagahi [37]	Genetic algorithm	1.380
Jianping et al. [22]	Genetic algorithm (spline slip surface)	1.321
Jianping et al. [22]	Genetic algorithm (line slip surface)	1.324
Kahatadeniya et al. [23]	Ant colony optimization	1.311
Kashani et al. [24]	Imperialistic competitive algorithm	1.3206

**Table 3** Soil Layers properties for Case II

Layer	$\gamma$ (kN/m <sup>3</sup> )	c(kPa)	$\phi$ (°)
1	18.82	29.4	12
2	18.82	9.8	5
3	18.82	294.0	40

Sridevi and Deep [38] and Malkawi et al. [29] applied the random search technique (RST-2) and Monte Carlo method and Khajezadeh et al. [27] utilized PSO and MPSO optimization algorithms, respectively.

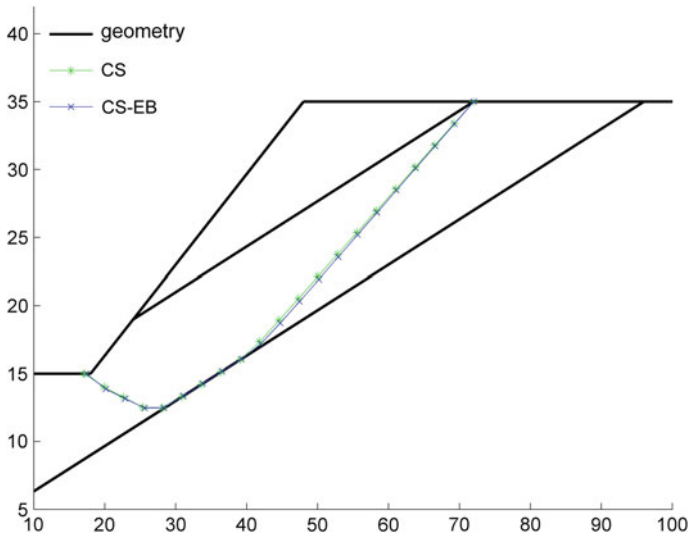


Fig. 5 Slope geometry and critical slip surface for Case II

Table 4 Values of FOS comparison for Case II

Optimization algorithm	CS	CS_EB
Mean	0.409	0.392
Best	0.391	0.391
Standard deviation	0.0319	0.00016

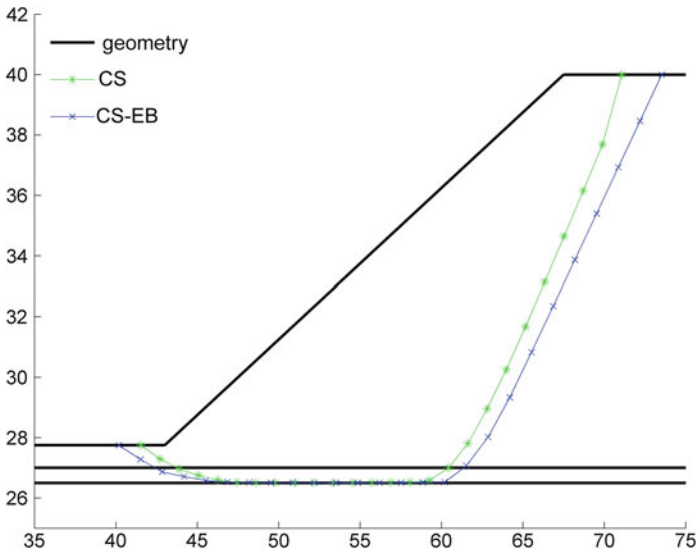
In the present study CS and CS\_EB are used to solve this problem and their results are presented in Table 4 by minimum FOS, mean and standard deviation. In order to compare these algorithms results with previous studies all the results are summarized in Table 5. In this case the values of FOS are equal again and from the SD it is concluded that the CS\_EB is the best algorithm on this case among all the past proposed ones.

### 4.3 Case III

The third case is a sample of more complicated slope geometry which a band of weak soil layer is sandwiched between two strong layers borrowed from SVSLOPE’s manual [13] as Fig. 6. Soil layers properties are, also presented in Table 6. In this case water table is at the base of the weak layer. As shown in Fig. 6 the slip surface is laid within weak layer. The factor of safety published by SVSLOPE’s manual was equal to 1.26 and the one calculated here are depicted in Table 7. Moreover this case was the

**Table 5** Previous studies computed FOS for Case II

References	Optimization algorithm	FOS
Arai and Tagyo [1]	Conjugate gradient	0.405
Sridevi and Deep [38]	Random search technique	0.401
Malkawi et al. [29]	Monte Carlo	0.401
Khajehzadeh et al. [27]	Particle swarm optimization	0.393
Khajehzadeh et al. [27]	Modified particle swarm optimization	0.391
Kashani et al. [24]	Imperialistic competitive algorithm	0.392
Gandomi et al. [15]	Particle swarm optimization	0.392
Gandomi et al. [15]	Firefly algorithm	0.392
Gandomi et al. [15]	Cuckoo search	0.391
Gandomi et al. [15]	Levy-Flight Krill Herd	0.391



**Fig. 6** Slope geometry and critical slip surface for case III

**Table 6** Soil layers properties for Case III

Layer	$\gamma$ (kN/m <sup>3</sup> )	$c$ (kPa)	$\phi$ (°)
1	18.84	28.5	20
2	18.84	0	10

aim of study in the work done by Gandomi et al. [15] and the results are summarized in Table 8. As results show, the performance of CS is benchmarked better in this case and CS\_EB obtained a lower value for FOS.

**Table 7** Values of FOS for Case III

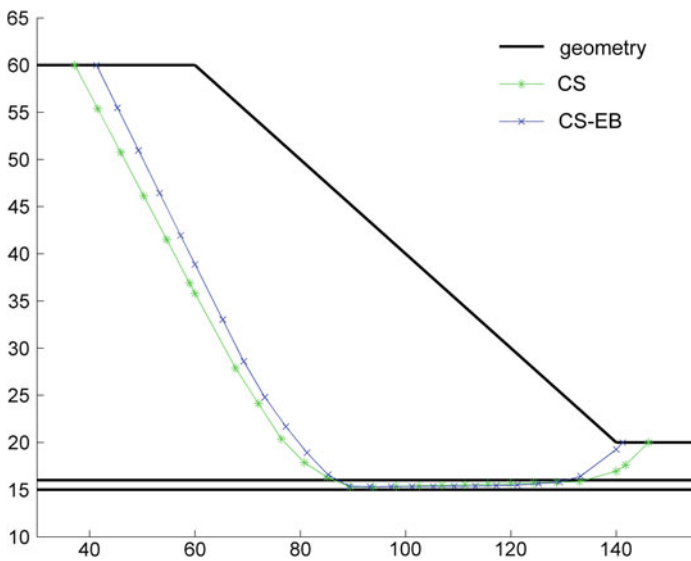
Optimization algorithm	CS	CS_EB
Mean	1.235279319	1.232020613
Best	1.226171806	1.223049725
Standard deviation	0.006388442	0.004708358

**Table 8** Previous studies computed FOS for Case III

References	Optimization algorithm	FOS
Gandomi et al. [15]	Particle swarm optimization	1.2462
Gandomi et al. [15]	Firefly algorithm	1.466
Gandomi et al. [15]	Cuckoo search	1.2261
Gandomi et al. [15]	Levy-Flight Krill Herd	1.2237

### 4.4 Case IV

In this example, the dry case of slope problem proposed by Fredlund and Krahn [14] is considered. Some researchers such as Kim et al. [28], Baker [2], and Zhu et al. [43] solved this problem in their studies. The slope geometry, location of slip surface and soil properties are shown in Fig. 7 and Table 9, respectively.



**Fig. 7** Slope geometry and critical slip surface for Case IV

**Table 9** Soil layers properties for Case IV

Layer	$\gamma$ (kN/m <sup>3</sup> )	$c$ (kPa)	$\phi$ (°)
1	19.22	28.73	20
2	19.22	0	10

**Table 10** Values of FOS comparison for Case IV

Optimization algorithm	CS	CS_EB
Mean	1.341315805	1.33212733
Best	1.323208687	1.308168866
Standard deviation	0.011051164	0.013925208

**Table 11** Previous studies computed FOS for Case IV

References	Optimization algorithm	FOS
Fredlund and Krahn [14]	–	1.373
Zhu et al. [43]	–	1.381
Kashani et al. [24]	Imperialistic competitive algorithm	1.3625

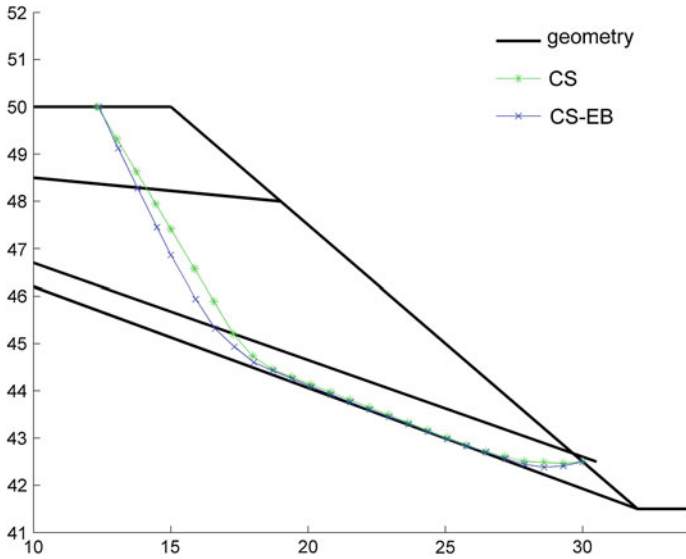
A brief comparison of present study and previous results are presented in Tables 10 and 11, respectively. From the results it is obvious that the CS and CS\_EB reach the best solution, and CS\_EB does even better than CS.

## 4.5 Case V

For the last case study, to investigate algorithms efficiency more accurately, more complicate example is selected from the literature of Zolfaghari et al. [45]. The soil parameters and slope geometry and slip surfaces are shown in Table 12 and Fig. 8, respectively.

**Table 12** Soil layers properties for Case V

Layer	$\gamma$ (kN/m <sup>3</sup> )	$c$ (kPa)	$\phi$ (°)
1	19.00	15.0	20
2	19.00	17.0	21
3	19.00	5.00	10
4	19.00	35.0	28



**Fig. 8** Slope geometry and critical slip surface for Case V

**Table 13** Values of FOS comparison for Case V

Optimization algorithm	CS	CS_EB
Mean	1.11585765	1.0731675
Best	1.0635	1.0502
Standard deviation	0.034966155	0.025047471

This problem is analyzed in various studies such as: Zolfaghari et al. [45] by using genetic algorithm, Cheng et al. [9] by using the artificial fish swarm algorithm (AFSA), Kahatadeniya et al. [23] by using the ant-colony method and Cheng et al. [10] by using HSPSO. The present study and latest studies results are summarized in Tables 13 and 14, respectively.

As a result, because of presence of thin weak soil layer between two strong ones multiple strong local minima have occurred and ACO and GA fail to converge to a very good solution. The computed FOS by CS and CS\_EB demonstrate that the present study provides a good solution in this example. Because of lower value of FOS by CS\_EB, it is concluded that CS\_EB is the best algorithm among other utilized algorithms.

**Table 14** Previous studies computed FOS for Case V

References	Optimization algorithm	FOS
Zolfaghari et al. [45]	Genetic algorithm	1.24
Cheng et al. [9, 10]	Simulated annealing	1.2813
Cheng et al. [9, 10]	Genetic algorithm	1.1440
Cheng et al. [9, 10]	Particle swarm optimization	1.1095
Cheng et al. [9, 10]	Simple harmony search	1.2068
Cheng et al. [9, 10]	Modified harmony search	1.1385
Cheng et al. [9, 10]	Tabu search	1.4650
Cheng et al. [9, 10]	Ant colony optimization	1.5817
Khajezadehet al. [26]	Gravitational search algorithm	1.0785
Kashani et al. [24]	Imperialistic competitive algorithm	1.0642
Gandomi et al. [15]	Particle swarm optimization	1.1148
Gandomi et al. [15]	Firefly algorithm	1.303
Gandomi et al. [15]	Cuckoo search	1.0635
Gandomi et al. [15]	Levy-Flight Krill Herd	1.0579

## 5 Conclusion

Effect of evolutionary boundary constraint handling scheme is assessed in complex geotechnical problems. This scheme is one of the recently proposed methods to implement boundary limitation on optimizations algorithms such as slope stability optimization problems. In this study a metaheuristic optimization algorithm that traditionally uses absorbing scheme is adopted to optimize five slope stability benchmark problems then their results are compared to the results with evolutionary boundary constraint handling scheme. The obtained results, such as best FOS values and standard deviation, using the classical and new proposed method prove the efficiency of the new method on making better the location of critical slip surface. Refer to the case studies; in the cases that obtained FOS are nearly equal, Case I and Case II, the lower value of standard deviation is belong to CS\_EB and from Case III to V, the lower values of FOS yield by CS\_EB. Altogether, the results declared the current proposed algorithm CS\_EB are capable to reach better solution than original CS. Not only this new boundary constraint handling method is easy to implement, but also it is efficient. This means evolutionary boundary constraint handling can make the optimization algorithm performance better without complex action like hybridizing.

## References

1. Arai K, Tagyo K (1986) Determination of noncircular slip surface giving the minimum factor of safety in slope stability analysis. *Soils Found* 26(3):152–154
2. Baker R (1980) Determination of the critical slip surface in slope stability computations. *Int J Numer Anal Methods Geomech* 4:333–359
3. Bolton HPJ, Heymann G, Groenwold AA (2003) Global search for critical failure surface in slope stability analysis. *Eng Optim* 35(1):51–65
4. Chen TY, Chi TM (2010) On the improvements of the particle swarm optimization algorithm. *Adv Eng Softw* 41:229–239
5. Chen Z, Shao C (1983) Evaluation of minimum factor of safety in slope stability analysis. *Can Geotech J* 25(4):735–748
6. Cheng YM (2003) Locations of critical failure surface and some further studies on slope stability analysis. *Comput Geotech* 30:255–267
7. Cheng YM, Li L, Ch SC (2007) Performance studies on six heuristic global optimization methods in the location of critical failure surface. *Comput Geotech* 34:462–484
8. Cheng YM, Li L, Chi SC, Wei WB (2007) Particle swarm optimization algorithm for the location of the critical non-circular failure surface in two-dimensional slope stability analysis. *Comput Geotech* 34(2):92–103
9. Cheng YM, Liang L, Chi SC, Wei WB (2008) Determination of the critical slip surface using artificial fish swarms algorithm. *J Geotech Geoenviron Eng* 134(2):244–251
10. Cheng YM, Li L, Sun YJ, Au SK (2012) A coupled particle swarm and harmony search optimization algorithm for difficult geotechnical problems. *Struct Multidisc Optim* 45:489–501
11. Chu W, Gao X, Sorooshian S (2011) Handling boundary constraints for particle swarm optimization in high-dimensional search space. *Inf Sci* 181(20):4569–4581
12. Das SK (2005) Slope stability analysis using genetic algorithm. *Electron J Geotech Eng* 10(A)
13. Feng T, Fredlund M (2012) SVSLOPE, Slope stability modeling software's verification manual
14. Fredlund DG, Krahn J (1977) Comparison of slope stability methods of analysis. *Can Geotech J* 14(3):429–439
15. Gandomi AH, Kashani AR, Mousavi M, Jalalvandi M (2014) Slope stability analyzing using recent swarm intelligence techniques. *Int J Numer Anal Methods Geomech* 39(3):295–309
16. Gandomi AH, Yang XS, Alavi AH (2013) Cuckoo search algorithm: a metaheuristic approach to solve structural optimization problems. *Eng Comput* 29(1):17–35
17. Gandomi AH, Yang XS (2012) Evolutionary boundary constraint handling scheme. *Neural Comput Appl* 21:1449–1462
18. Giasi CJ, Masi P, Cherubini C (2003) Probabilistic and fuzzy reliability analysis of a sample slope near Aliano. *Eng Geol* 67(3):391–402
19. Goh A (2000) Search for critical slip circle using genetic algorithms. *Civil Eng Environ Syst* 17(3):181–211
20. Greco YR (1996) Efficient Monte Carlo technique for locating critical slip surface. *J Geotech Eng ASCE* 122:517–525
21. Huang T, Mohan AS (2005) A hybrid boundary condition for robust particle swarm optimization. *IEEE Antennas Wirel Propag Lett* 4:112–117
22. Jianping S, Li J, Liu Q (2008) Search for critical slip surface in slope stability analysis by spline-based GA method. *J Geotech Geoenviron Eng* 134(2):252–256
23. Kahatadeniya KS, Nanakorn P, Neaupane KM (2009) Determination of the critical failure surface for slope stability analysis using ant colony optimization. *Eng Geol* 108:133–141
24. Kashani AR, Gandomi AH, Mousavi M (2014) Imperialistic competitive algorithm: a metaheuristic algorithm for locating the critical slip surface in 2-dimensional soil slopes. *Geosci Front* (in Press). doi:10.1016/j.gsf.2014.11.005
25. Kaveh A, Talatahari S (2009) Particle swarm optimizer, ant colony strategy and harmony search scheme hybridized for optimization of truss structures. *Comput Struct* 87(5–6):267–283



26. Khajehzadeh M, Taha MR, El-shafie A, Eslami M (2011) Search for critical failure surface in slope stability analysis by gravitational search algorithm. *Int J Physic Sci* 6(21): 5012–5021
27. Khajehzadeh M, Taha MR, El-Shafie A, Eslami M (2012) Locating the general failure surface of earth slope using particle swarm optimization. *Civil Eng Environ Syst* 29(1):41–57
28. Kim J, Salgado R, Lee J (2002) Stability analysis of complex soil slopes using limit analysis. *J Geotech Geoenviron Eng* 128(7):546–557
29. Malkawi AIH, Hassan WF, Sarma SK (2001) Global search method for locating general slip surface using Monte Carlo techniques. *J Geotech Geoenviron Eng* 127(8):688–698
30. Mathada VS, Venkatachalam G, Sridivya A (2007) Slope stability assessment—a comparison of probabilistic, possibilistic and hybrid approaches. *Int J Performability Eng* 3(2):11–21
31. McCombie P, Wilkinson P (2002) The use of the simple genetic algorithm in finding the critical factor of safety in slope stability analysis. *Comput Geotech* 29(8):699–714
32. Morgenstern NR, Price VE (1965) The analysis of the stability of general slip surfaces. *Géotechnique* 15:79–93
33. Rezaeean A, Noorzad R, Dankoub AKM (2011) Ant colony optimization for locating the critical failure surface in slope stability analysis. *World Appl Sci J* 13(7):1702–1711
34. Rubio E, Hall JW, Anderson MG (2004) Uncertainty analysis in a slope hydrology and stability model using probabilistic and imprecise information. *Comput Geotech* 31(7):529–536
35. Samui P, Kumar B (2006) Artificial neural network prediction of stability numbers for two-layered slopes with associated flow rule. *EJGE*
36. Sengupta A, Upadhyay A (2009) Locating the critical failure surface in a slope stability analysis by genetic algorithm. *Appl Soft Comput* 9(1):387–392
37. Solati S, Habibagahi G (2006) A genetic approach for determining the generalized interslice forces and the critical non-circular slip surface. *Iran J Sci Technol Trans B Eng* 30(1):1–20
38. Sridevi B, Deep K (1992) Application of global optimization technique to slope stability analysis. In: *Proceedings of the 6th international symposium on landslide*. Christchurch, New Zealand, pp 573–578
39. Xu S, Rahmat-Samii Y (2007) Boundary conditions in particle swarm optimization revisited. *IEEE Trans Antennas Propag* 55(3):112–117
40. Yamagami T, Ueta Y (1988) Search for noncircular slip surfaces by the Morgenstern-Price method. In: *The 6th international conference on numerical methods in geomechanics. Numerical methods in geomechanics (Innsbruck 1988)*. Balkema, Innsbruck, pp 1335–1340
41. Yang XS, Deb S (2009) Cuckoo search via Levy flights. In: *Proceedings of world congress on nature and biologically inspired computing*. IEEE Publications, USA, pp 210–214
42. Zhao H, Yin S, Ru Z (2012) Relevance vector machine applied to slope stability analysis. *Int J Numer Anal Methods Geomech* 36(5):643–652
43. Zhu D, Lee CF, Jiang HD (2003) Generalized framework of limit equilibrium methods for slope stability analysis. *Geotechnique* 4:337–395
44. Zhu DY, Lee CF, Qian QH, Chen GR (2005) A concise algorithm for computing the factor of safety using the Morgenstern-Price method. *Can Geotech J* 42(1):272–278
45. Zolfaghari AR, Heath AC, McCombie PF (2005) Simple genetic algorithm search for critical non-circular failure surface in slope stability analysis. *Comput Geotech* 32(3):139–152

# Blackbox Optimization in Engineering Design: Adaptive Statistical Surrogates and Direct Search Algorithms

Bastien Talgorn, Le Digabel Sébastien and Michael Kokkolaras

*We are honored to have this work appear in a book dedicated to the memory of Professor M.G. Karlaftis*

**Abstract** Simulation-based design optimization relies on computational models to evaluate objective and constraint functions. Typical challenges of solving simulation-based design optimization problems include unavailable gradients or unreliable approximations thereof, excessive computational cost, numerical noise, multimodality and even the models' failure to return a value. It has become common to use the term "blackbox" for a computational model that features any of these characteristics and/or is inaccessible by the design engineer (i.e., cannot be modified directly to address these issues). A possible remedy for dealing with blackboxes is to use surrogate-based derivative-free optimization methods. However, this has to be done carefully using appropriate formulations and algorithms. In this work, we use the R `dynaTree` package to build statistical surrogates of the blackboxes and the direct search method for derivative-free optimization. We present different formulations for the surrogate problem considered at each search step of the Mesh Adaptive

---

This material was presented by the third author in an invited semi-plenary lecture at OPT-i 2014.

---

B. Talgorn · M. Kokkolaras (✉)  
GERAD and Department of Mechanical Engineering, McGill University,  
Montréal, Québec, Canada  
e-mail: Michael.Kokkolaras@mcgill.ca

B. Talgorn  
e-mail: Bastien.Talgorn@gerad.ca

L.D. Sébastien  
GERAD and Département de Mathématiques Et Génie Industriel,  
École Polytechnique de Montréal, Montréal, Québec, Canada  
e-mail: Sebastien.Le.Digabel@gerad.ca

© Springer International Publishing Switzerland 2015  
N.D. Lagaros and M. Papadrakakis (eds.), *Engineering and Applied  
Sciences Optimization*, Computational Methods in Applied Sciences 38,  
DOI 10.1007/978-3-319-18320-6\_19

Direct Search (MADS) algorithm using a surrogate management framework. The proposed formulations are tested on two simulation-based multidisciplinary design optimization problems. Numerical results confirm that the use of statistical surrogates in MADS improves the efficiency of the optimization algorithm.

## 1 Introduction

We consider the nonlinear and constrained design optimization problem

$$\begin{aligned} & \min_{x \in \mathcal{X}} f(x) \\ & \text{subject to } c_j(x) \leq 0, \quad j \in J, \end{aligned} \tag{1}$$

where  $J = \{1, 2, \dots, m\}$  and  $\mathcal{X}$  is a subset of  $\mathbb{R}^n$  typically defined by bound constraints. The objective and constraint functions  $f$  and  $c_j, j \in J$ , map  $\mathbb{R}^n$  to  $\mathbb{R} \cup \{\infty\}$ , and in simulation-based engineering design most, if not all, of them are evaluated using *blackboxes*. A blackbox is a computational or simulation model whose internal structure is unknown and/or inaccessible. Typical challenges associated with blackboxes include numerical noise, multi-modality, high computational cost, and failure to return a value, e.g., when the simulation crashes. However, the most salient feature of blackboxes is that gradients are unavailable and their approximations are unreliable.

Derivative-free methods are developed to handle these issues [15]. Direct search algorithms such as GPS [3, 40] or MADS [4] rely on the *search-and-poll* paradigm. The *search* can implement any strategy (including none) to evaluate a finite number of trial points. It typically favors global exploration of the design space and allows users to implement any appropriate method that exploits their knowledge of the problem. If the search fails to find an improvement, the *poll* proposes trial points around the incumbent solution according to rigorous conditions. These points can then be evaluated by the blackbox in any order, and the evaluation can be interrupted if an improved solution is found. The poll ensures global convergence of the algorithm toward a local optimum.

Surrogate-based optimization methods construct approximations of the objective and constraint functions using the points evaluated by blackboxes during an iterative algorithm [7, 14, 20, 23, 26, 27, 34, 36, 38, 44]. These approximations are then used to propose promising candidates. This implies the formulation and solution of an optimization problem called the *surrogate problem*. Much effort can be devoted to solving this problem since the computational cost of a surrogate evaluation is negligible compared to the blackbox. These methods have proved to be efficient, but they have mostly been used on smooth or unconstrained problems.

Our work is based on the postulation that the use of surrogates would be more efficient if they were integrated within direct search algorithms as described in the *surrogate management framework* of Ref. [8]. The search would then involve build-

ing or updating the surrogate models and solving the surrogate problem to propose a candidate. Simple implementations of this framework have been presented for the unconstrained case in [8, 37] and for the constrained case in [16, 22]. However, only the simplest formulation of the surrogate problem was considered in these implementations, namely the optimization of a model of the objective function subject to models of the constraint functions.

The contribution of this work consists of new formulations of the surrogate problem that exploit the different capabilities of statistical surrogate modeling methods and in particular the `dynaTree` library [23, 39]. Our goal is to exploit the availability of statistical information such as the mean, the standard deviation and the cumulative density function of the `dynaTree` models, which are not restricted to be continuous or stationary Gaussian processes. Consequently, they are non-interpolating, and thus possibly better equipped to perform robust regression when using nonsmooth data.

Seven formulations are proposed. Three of them are constrained while the others quantify the relevance of a candidate via a single statistical criterion. Six of these formulations emphasize the exploration of the design space, and all of them handle nonconvex and nonsmooth constrained problems.

The paper is organized as follows. Section 2 describes the MADS algorithm, the implementation of the surrogate-based search, and the `dynaTree` models. Section 3 presents the seven new surrogate problem formulations. Section 4 compares the performance of the formulations by means of two simulation-based multidisciplinary design optimization (MDO) problems related to aircraft design. Section 5 provides concluding remarks.

## 2 Background: MADS and `dynaTree`

### 2.1 Mesh Adaptive Direct Search (MADS)

Mesh adaptive direct search (MADS) [4] is an algorithm for blackbox optimization that can handle nonlinear constraints. It ensures global convergence to a solution satisfying local optimality conditions based on the Clarke calculus for nonsmooth functions [13]. At each MADS iteration  $k$ , trial points are evaluated on the *mesh*  $M_k$  defined as

$$M_k = \{x + \Delta_k^m Dz : z \in \mathbb{N}^{n_D}, x \in \mathbf{X}_k\} \subset \mathbb{R}^n \quad (2)$$

where  $\Delta_k^m$  is a mesh size parameter, the columns of  $D \in \mathbb{R}^{n \times n_D}$  form a positive spanning set of  $n_D$  directions in  $\mathbb{R}^n$  [30], and  $\mathbf{X}_k = \{x^1, x^2, \dots\} \subset \mathbb{R}^n$  denotes the set of already evaluated points, called the *cache*. MADS relies on a search-and-poll paradigm, named after the two steps that constitute each iteration.

The search is an optional step during which several different methods can be used to propose candidates anywhere on the mesh  $M_k$ . These methods can be heuristic

in the sense that they can be guided by user insight into the problem at hand. Alternatively, more systematic methods, such as genetic algorithms [21], variable neighborhood search [5], particle swarms [42], or Latin hypercube-based design of experiments [32] can be used during the search step. In this work, we focus on surrogate-based search methods [8, 16, 18, 22].

The data  $[\mathbf{X}_k, f(\mathbf{X}_k), c_1(\mathbf{X}_k), \dots, c_j(\mathbf{X}_k)]$  are used to build statistical surrogate models of  $f$  and  $c_j, j \in J$ , by modeling each blackbox output as a random variable. Statistical surrogate models provide information about the mean, variance, and probability density function of the modeled random variable. In this way, we can compute statistical measurements of the relevance of each candidate of the search step. To emphasize the global exploration of  $\mathcal{X}$ , we favor areas with uncertain blackbox outputs. Using the formulations presented in this paper (Sect. 3), we find a candidate  $x_k^{SP}$  (where  $SP$  denotes the surrogate problem) by solving a subproblem. This candidate is then projected onto the current mesh  $M_k$  to preserve the original convergence properties described in [4].

Although `dynaTree` predictions are piecewise linear, most of the relevance criteria used in the surrogate problem formulations are not. Moreover, the use of these formulations is not limited to `dynaTree`: any modeling method able to provide the necessary statistical information can be used. Surrogate models may have some of the typical properties of a blackbox. In particular, they may fail to return a value or derivatives may not be available. For these reasons, we solve the surrogate problem using a direct search algorithm as well. Specifically, we use MADS both for solving the original blackbox optimization problem and the surrogate subproblem of the search step in a nested optimization manner.

During the poll step, a set of candidates, called the *poll set*, is defined as  $P_k = \{x_k + d : d \in D_k\}$ , where  $D_k$  is a set of polling directions based on combinations of directions in  $D$ . The poll size parameter  $\Delta_k^p = \max_{d \in D_k} \|d\|$  defines the maximum norm of the directions of  $D_k$ . MADS controls  $\Delta_k^m$  and  $\Delta_k^p$  so that  $\Delta_k^m$  decreases faster than  $\Delta_k^p$ , which causes the set of poll directions to grow dense in the unit sphere, once normalized. This allows polling in all possible directions of  $\mathbb{R}^n$ .

The set of trial points  $T_k = x_k^{SP} \cup P_k$  is evaluated by the blackbox *opportunistically*, which means that if evaluating a point leads to an improvement, the evaluation of  $T_k$  is interrupted. Since this strategy makes the evaluation order-critical, the relevance criterion used in the search step is also used to sort the points in  $T_k$ . This is performed via a filter-like mechanism described in [19].

After the trial-point evaluations, MADS updates the poll and mesh size parameters depending on the success of the iteration. The incumbent solution and the cache  $\mathbf{X}_k$  are then updated, and a new iteration begins. The optimization terminates when the stopping criteria are satisfied, which means that the mesh size parameter is smaller than the machine precision or the evaluation budget is exhausted. Figure 1 illustrates the complete algorithm.

**Fig. 1** Optimization algorithm

```

[1] Initialization
  Set initial mesh and poll sizes  $\Delta_0^m, \Delta_0^p > 0$ 
  Set starting point  $x_0 \in \mathcal{X}$ 
   $\mathbf{X}_0 \leftarrow \emptyset$ 
   $k \leftarrow 0$ 
[2] Search
  Build or update dynaTree models
   $x_k^{SP} \leftarrow$  surrogate problem solution
  Project  $x_k^{SP}$  onto the mesh  $M_k$ 
[3] Poll
  Build poll directions  $D_k$  and trial points  $P_k$ 
[4] Evaluation and update
  Build trial set  $T_k = x_k^{SP} \cup P_k$ 
  Sort  $T_k$  according to the surrogate models
  Perform opportunistic evaluation of  $T_k$ 
   $k \leftarrow k + 1$ 
  Update mesh and poll size  $\Delta_k^m, \Delta_k^p$ 
  Update incumbent solution  $x_k$ 
  Update cache  $[\mathbf{X}_k, f(\mathbf{X}_k), c_1(\mathbf{X}_k), \dots, c_m(\mathbf{X}_k)]$ 
  goto [2] if no stopping condition is met
    
```

## 2.2 The *dynaTree* library

The *dynaTree* library [23, 39] is used in this work to build statistical surrogate models. It is based on a Bayesian framework for parameter-free regression on non-smooth data. From the data  $[\mathbf{X}, y(\mathbf{X})]$ , *dynaTree* provides statistical information on  $y(x)$ , namely the mean  $\hat{y}(x)$ , the standard deviation  $\hat{\sigma}_y(x)$ , and the cumulative density function  $\mathbb{P}[y(x) < y_0], \forall y_0 \in \mathbb{R}$ . Unlike Kriging, *dynaTree* does not consider  $y$  to be a continuous or stationary Gaussian process. Consequently, *dynaTree* is a non-interpolating method, which means that  $x \in \mathbf{X} \not\Rightarrow \hat{y}(x) = y(x)$ . Such methods are best suited for the approximation of nonsmooth data [23, 25]. Specifically, *dynaTree* implements a piecewise linear regression that allows global and robust regression in the presence of noncontinuous data or first-order discontinuities.

This regression relies on *trees*. As illustrated in the one-dimensional example (Fig. 2), a tree implements a partition of the design space  $\mathcal{X}$ . Each interior node implements a partitioning criterion, and each leaf represents a part of  $\mathcal{X}$ . In each leaf  $\eta$ , a linear regression model is built using the data  $[\mathbf{X}_\eta, y(\mathbf{X}_\eta)]$  where  $\mathbf{X}_\eta = \mathbf{X} \cap \eta$ . In the one-dimensional example of Fig. 2, the plot shows 24 data points, the partitioning of the design space, the piecewise linear prediction, and the standard deviation of  $y(x)$ . The diagram below the plot depicts the tree associated with the partition of the interval  $[0, 25]$ .

The Bayesian approach enables the computation of the *likelihood*  $\mathcal{L}(\eta)$  of each leaf  $\eta$ , which quantifies the ability of the model to fit the data in  $\eta$ . Then, a prior  $\pi(T)$  allows us to penalize overly complicated arborescences [11, 12].  $\pi(T)$  is defined by considering that each leaf  $\eta$  can be split with a probability  $p_{split}(T, \eta)$  that grows

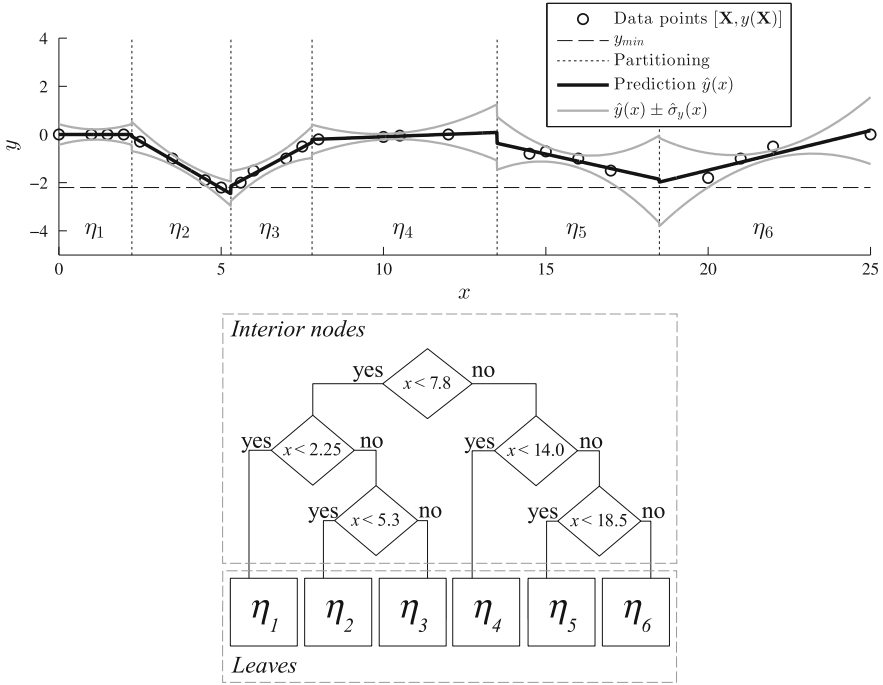


Fig. 2 dynaTree regression on 24 data points in  $\mathbb{R}$

with the depth of the leaf. The prior is the likelihood of the tree in relation to this splitting probability:

$$\pi(T) = \prod_{\eta \in \text{Leaves}(T)} p_{\text{split}}(T, \eta) \prod_{\eta \in \text{Interior}(T)} \overline{p_{\text{split}}(T, \eta)}. \tag{3}$$

Finally, the likelihood  $\mathcal{L}(T)$ , which quantifies the quality of  $T$ , is defined as:

$$\mathcal{L}(T) = \pi(T) \prod_{\eta \in \text{Leaves}(T)} \mathcal{L}(\eta). \tag{4}$$

Although one tree is sufficient to build a surrogate model, **dynaTree** generates a set of trees. This increases the likelihood of finding several efficient partitions and allows more robust regression. A particle learning sequential Monte Carlo algorithm [9, 10] adapts the set of trees to the observations, by reproducing and modifying the trees with the best likelihood. The modifications consist of three equally probable operations: splitting a leaf, merging two leaves, or making no change. Once the set is built, predictions are made by averaging all the trees. The interested reader can refer to [39] for more details.

This method offers several advantages: the Bayesian framework can handle noisy data and provides statistical information, while the space partitioning allows us to refine the model in areas of interest. Moreover, the selection of the number of trees allows us to find a balance between accuracy and computational cost. However, **dynaTree** is likely to be outperformed by Kriging and polynomial regression on smooth functions (see Appendix 2.2.1). In addition, the complexity of the space partitioning grows exponentially with the dimension of the design space. Thus, while **dynaTree** can theoretically be used for any problem size, its practical use is recommended for small to moderate size problems (not more than 10 to 20 variables depending on problem complexity). In high or very high dimensions, other surrogate methods such as Support Vector Machines [17] or Reduced Order Modeling [6, 31, 43] can be used.

### 2.2.1 Comparison of **dynaTree** to Kriging and Polynomial Regression

The **dynaTree** method is compared here to two other surrogate modeling methods: Kriging and polynomial regression (with 9 coefficients). The comparison is performed on 4 test functions:

$$y_1(x) = \cos(\pi x) \tag{5}$$

$$y_2(x) = \cos(\pi x) + 0.2g(x) \tag{6}$$

$$y_3(x) = \cos(\pi x)s(x) \tag{7}$$

$$y_4(x) = \cos(\pi x)s(x) + 0.2g(x) \tag{8}$$

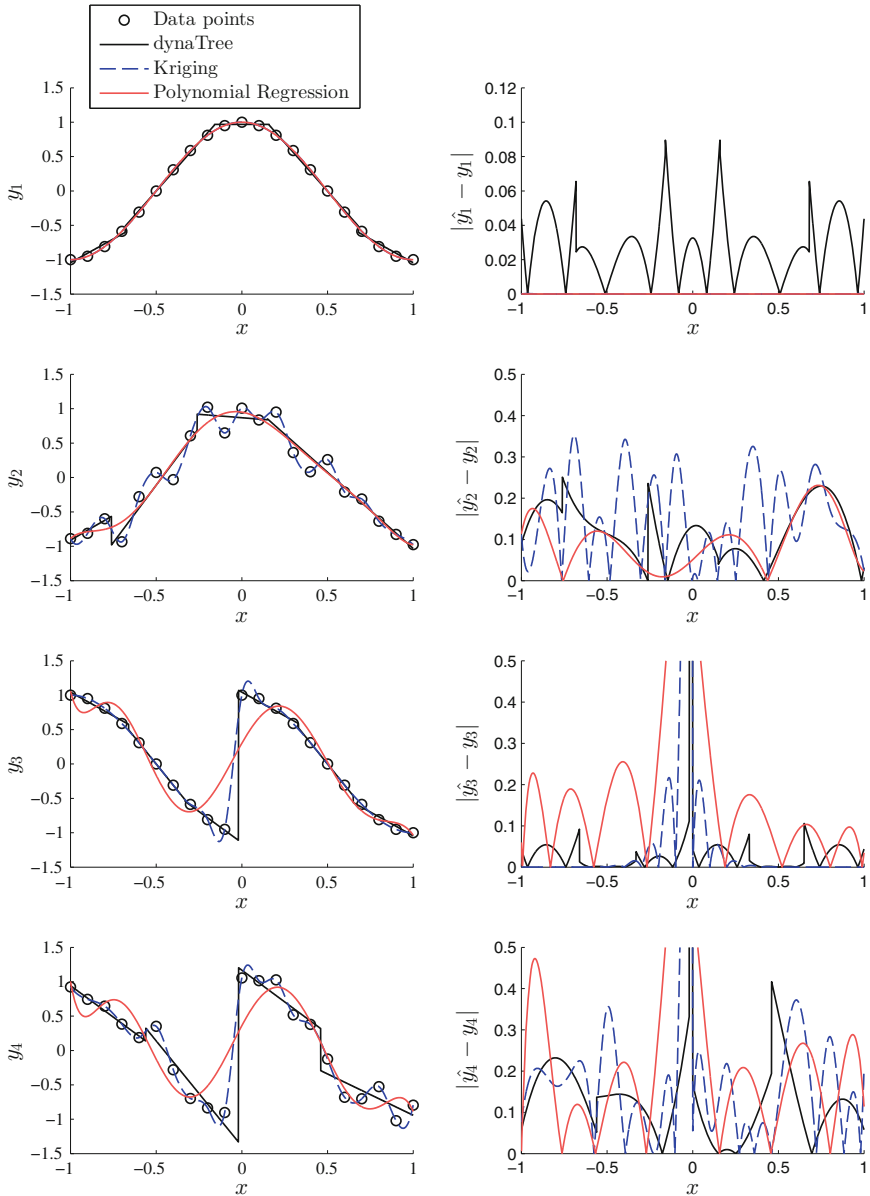
where  $g(x)$  is a normalized centered Gaussian noise and  $s$  is defined as

$$s(x) = \begin{cases} +1, & x \geq 0 \\ -1, & x < 0. \end{cases} \tag{9}$$

The models  $\hat{y}_i$  are built on 21 points regularly spaced in  $[-1, +1]$  (sampling period of 0.1). Then the mean square error between  $\hat{y}_i$  and  $y_i$  is computed on 2001 points regularly spaced in  $[-1, +1]$  (sampling period of 0.001). Figure 3 depicts the curves of  $\hat{y}_i$  for the three modeling methods. Table 1 shows the characteristics of each test function and the means square error (MSE) for each modeling method.

The polynomial regression performs very well on the smooth and clean test function  $y_1$ . It also returns the best MSE for  $y_2$ , but this error is not likely to decrease while more data points are available, because polynomial regression is not able to refine the model in well known area. The polynomial regression fails to fit the functions  $y_3$  and  $y_4$  due to the discontinuity in  $x = 0$ . Kriging also performs very well on  $y_1$  but is outperformed by **dynaTree** as soon as the function is nonsmooth and/or noisy ( $y_2$ ,  $y_3$  and  $y_4$ ). A comparison of surrogate modeling methods (including **dynaTree**) for five-dimensional problems can be found in [39].





**Fig. 3** Comparison of different surrogate modeling methods using 4 test functions (*left column: models, right column: absolute errors*)

**Table 1** Mean square errors on the 4 test functions

Function	Properties		Modeling method		
			dynaTree	Kriging	Polynomial
$y_1$	Smooth	Clean	0.0011	$3.10^{-9}$	$1.10^{-9}$
$y_2$	Smooth	Noisy	0.0183	0.0343	0.0125
$y_3$	Nonsmooth	Clean	0.0413	0.0709	0.0998
$y_4$	Nonsmooth	Noisy	0.0672	0.1047	0.1197

### 3 Formulations of the Surrogate Problem

The statistical information provided by the models built using `dynaTree` is used to compute other measures of candidate relevance. At most<sup>1</sup>  $m + 1$  surrogate models are built to evaluate the objective function and the  $m$  constraints. The mean and standard deviation of the surrogate objective and constraints are denoted  $\hat{f}$  and  $\hat{\sigma}_f$  and  $\hat{c}_j$  and  $\hat{\sigma}_j$ , respectively.

#### 3.1 Direct Surrogate of the Original Problem

The simplest surrogate formulation results from using surrogate models (in lieu of blackboxes) to evaluate the objective and constraints of the original problem (1):

$$\begin{aligned} & \min_{x \in \mathcal{X}} \hat{f}(x) \\ & \text{subject to } \hat{c}_j(x) \leq 0 \quad \forall j \in J. \end{aligned} \tag{10}$$

This formulation can be generalized to perform exploration of the design space. In [39], Taddy et al. propose solving unconstrained problems by sequentially solving the surrogate problem

$$\min_{x \in \mathcal{X}} -EI(x) - \lambda \hat{\sigma}_f(x), \tag{11}$$

where  $EI$  is some *expected improvement* function, and  $\lambda$  is an exploration parameter empirically chosen in  $[0, 1]$ . We use this concept to formulate the surrogate problem  $F\sigma$ :

$$(F\sigma) \left\{ \begin{array}{l} \min_{x \in \mathcal{X}} \hat{f}(x) - \lambda \hat{\sigma}_f(x) \\ \text{subject to } \hat{c}_j(x) - \lambda \hat{\sigma}_j(x) \leq 0 \quad \forall j \in J. \end{array} \right. \tag{12}$$

---

<sup>1</sup>There may be situations where the properties of the objective function or some of the constraints do not require the construction and use of surrogate models, e.g., if one of these functions is smooth and inexpensive and has an analytical expression.

This formulation is denoted  $F\sigma$ .  $F$  indicates that the objective of the surrogate problem is based on the surrogate model of the objective function, and  $\sigma$  indicates that the variance of the surrogate model is taken into account for the exploration.

Taddy et al. use the values  $\lambda \in \{1/100, 1/10, 1\}$  because the literature [26, 39] considers searching with  $\lambda = 0$  to be *myopic*. We will consider the values  $\lambda \in \{0, 1/100, 1/10, 1\}$ . Large values of  $\lambda$  imply that the search will favor candidates with an uncertain objective, which are in the ill-explored or nonsmooth areas of  $\mathcal{X}$ , generally outside the current attraction basin. In this formulation, for  $\lambda > 0$ , the feasible space is extended by the uncertainties in the constraints; the uncertainties in the objective are considered as potential improvements of the objective. Note that formulation  $F\sigma$  (Eq. (12)) is equivalent to the problem of Eq. (10) when  $\lambda = 0$ .

### 3.2 Probability of Feasibility

Another way to handle the constraints is to use the cumulative density function provided by the surrogate model to estimate the probability of a point being feasible. The value  $\mathbb{P}[c_j(x) \leq 0]$  is provided by the model and represents the probability that the constraint  $c_j$  is satisfied at  $x$ . An estimation of the probability that  $x$  is feasible is computed by

$$P(x) = \prod_{j \in J} \mathbb{P}[c_j(x) \leq 0] \approx \mathbb{P}[c_j(x) \leq 0, \forall j \in J]. \quad (13)$$

If the constraints are statistically independent, the above approximation is exact. Since we are considering blackbox output, it cannot be assumed that there is no correlation between the constraints; however,  $P(x)$  is the best approximation available. It is worth mentioning that the probability of feasibility of a point can also be estimated by building a model of an aggregate constraint  $h(x)$  and by computing  $P(x) = \mathbb{P}[h(x) \leq 0]$ . Several definitions are possible:

$$h(x) = \sum_{j \in J} \max \{c_j(x); 0\}^2, \quad (14)$$

$$h(x) = \max_{j \in J} \{c_j(x)\}, \text{ or} \quad (15)$$

$$h(x) = \begin{cases} 1 & \text{if } c_j(x) \leq 0 \quad \forall j \in J, \\ 0 & \text{otherwise.} \end{cases} \quad (16)$$

Aggregate constraints enable modeling feasibility by building just one surrogate model rather than  $m$ . This can reduce the computational time and avoid the question of the independence of the constraints. However, it also implies that fewer data contribute to building the model, which makes it less accurate than multiple-constraint surrogate models. Preliminary tests with `dynaTree` models have shown that building one model per constraint is more efficient. Thus, we use Eq. (13) to treat the constraints in this study.

This estimation of the probability of feasibility can be used as a chance constraint, meaning that a candidate must satisfy a constraint on  $P(x)$ , regardless of its objective. This leads to the formulation  $F\sigma P$  of the surrogate problem

$$(F\sigma P) \begin{cases} \min_{x \in \mathcal{X}} \hat{f}(x) - \lambda \hat{\sigma}_f(x) \\ \text{subject to } P(x) \geq p_c. \end{cases} \quad (17)$$

This formulation indicates that only candidates likely to be feasible will be evaluated. As a consequence, the candidates will remain distant from the boundary of the feasible domain and will approach it only when  $\sigma_j$  decreases. The choice of  $p_c$  can depend on the problem size and the number of constraints. If  $p_c$  is too high, the constraint on  $P$  can be impossible to satisfy, particularly at the beginning of the optimization when the constraints are uncertain. However, if  $p_c$  is too low, the candidate will rarely be in the feasible domain, leading to an inefficient search. In this study, we choose  $p_c = 0.5$ , which means that after the entire optimization run half of the points  $x_k^{SP}$  will be feasible, ensuring improvement of the models inside and outside the feasible domain.

### 3.3 Expected Improvement

*Improvement* is defined by

$$I(x) = \max\{f_{min} - f(x), 0\}, \quad (18)$$

where  $f_{min}$  is the objective function value of the currently best feasible point [36]. In the context of global optimization, evaluating a point that does not improve the objective is not considered counterproductive since this evaluation enhances the information about the problem [26, 27]. Evaluating a point that improves the objective is a step forward in the optimization and narrows the area where a global optimum can be found. Thus, the utility of evaluating a new point is always positive. This principle is manifested in the definition of  $EI$ , which is considered a major relevance criterion in global optimization [26, 27, 39]:

$$EI(x) = \mathbb{E}[I(x)] = \int_0^{+\infty} I \phi_f(f_{min} - I) dI, \quad (19)$$

where  $\phi_f$  is the probability density function of  $f$ , provided by the surrogate model. The formulation described in (11) can be adapted to handle constraints:

$$(EI\sigma) \begin{cases} \min_{x \in \mathcal{X}} -EI(x) - \lambda \hat{\sigma}_f(x) \\ \text{subject to } \hat{c}_j(x) - \lambda \hat{\sigma}_f(x) \leq 0. \end{cases} \quad (20)$$

The expected feasible improvement (*EFI*) considers in a single scalar criterion the objective and feasibility of a candidate [36]:

$$EFI(x) = EI(x) P(x). \tag{21}$$

The *EFI* represents a tangible measure of the relevance of a candidate in the context of constrained optimization. A promising candidate can be found by maximizing the *EFI*. This leads to an unconstrained formulation of the surrogate problem:

$$(EFI) \left\{ \begin{array}{l} \min_{x \in \mathcal{X}} \end{array} \right. - EFI(x). \tag{22}$$

Maximizing the *EFI* is an efficient method, but it would also be interesting to incorporate the exploration term proposed in (11) and used in the previous formulations:

$$(EFI\sigma) \left\{ \begin{array}{l} \min_{x \in \mathcal{X}} \end{array} \right. - EFI(x) - \lambda \hat{\sigma}_f(x). \tag{23}$$

The drawback of the formulation *EFI* $\sigma$  (Eq. (23)) is that the exploration term  $\hat{\sigma}_f(x)$  does not take into account uncertainties in the constraints. To address this,  $\hat{\sigma}_f(x)$  could be replaced by a norm on  $[\hat{\sigma}_f(x), \hat{\sigma}_1(x), \dots, \hat{\sigma}_m(x)]$ , but the uncertainty in the value of  $c_j$  is less significant than the uncertainty in the feasibility of the candidate. Given that the event “ $x$  is feasible” follows a Bernoulli law of probability  $P(x)$ , its variance is  $P(x)\overline{P(x)} \in [0, 1/4]$ . Thus, we propose a measure of the *uncertainty in the feasibility* ( $\mu$ ):

$$\mu(x) = 4 P(x) \overline{P(x)}. \tag{24}$$

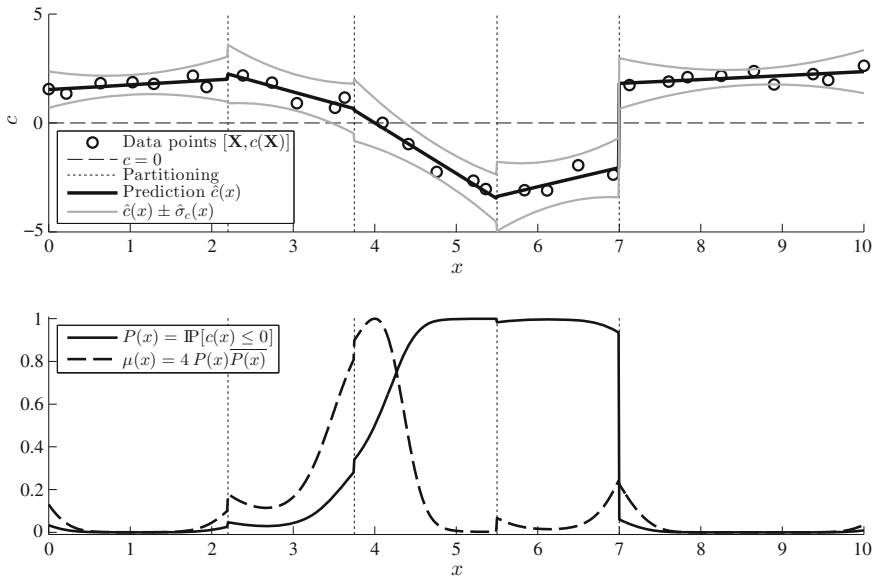
The multiplication by four is intended to normalize  $\mu$  in  $[0, 1]$ . The larger  $\mu$  is, the more uncertain is the feasibility of the candidate, which means that we cannot predict whether the candidate is feasible.  $\mu(x)$  is maximal for  $P(x) = 1/2$  and null for  $P(x) \in \{0, 1\}$ . Figure 4 illustrates this concept for a single constraint  $c(x)$ .

Using this measure, two formulations are derived. In the formulation *EFI* $\mu$ , the exploration term is multiplied by  $\mu(x)$ :

$$(EFI\mu) \left\{ \begin{array}{l} \min_{x \in \mathcal{X}} \end{array} \right. - EFI(x) - \lambda \hat{\sigma}_f(x) \mu(x). \tag{25}$$

This formulation encourages a search for candidates that are uncertain both in the objective and the feasibility. The drawback is that a promising candidate that is uncertain in only one of the two measures (objective or feasibility) will not be considered. To address this, the crossed formulation *EFIC* is introduced:

$$(EFIC) \left\{ \begin{array}{l} \min_{x \in \mathcal{X}} \end{array} \right. - EFI(x) - \lambda \left( EI(x) \mu(x) + P(x) \hat{\sigma}_f(x) \right). \tag{26}$$



**Fig. 4** Probability of feasibility and uncertainty in feasibility.  $\mu(x)$  is maximal for  $x = 4$ , where  $\hat{c}(x) = 0$ . In the neighborhood of  $x = 7$ , despite the sharp variation in  $c$ , the feasibility is predictable, so  $\mu(x)$  is small

The first part of the exploration term,  $EI(x)\mu(x)$ , favors candidates that have a promising objective and unpredictable feasibility. In contrast,  $P(x)\hat{\sigma}_f(x)$  favors candidates with an uncertain objective and good feasibility.

### 4 Multidisciplinary Design Optimization Examples

The proposed formulations are tested on 2 MDO applications related to aircraft design. These 2 problems are constrained; they may involve nonsmooth functions, may have several local optima, and may exhibit some numerical noise. To generate more results for the 2 MDO applications, we run a total of 100 optimizations by specifying 50 different feasible starting points for each application, using Latin hypercube sampling [32]. Thus, each formulation is tested on two sets of 50 optimization runs. For each optimization, we allow 1000  $n$  blackbox evaluations, but the optimization can stop earlier if the stopping criterion for the mesh size is satisfied. The two MDO applications have a computational time of 60 ms (Simplified Wing) and 5 ms (Aircraft Range) per evaluation on a standard desktop PC (Intel Core i7-2600, 16 Gb).

The numerical results were obtained using the MADS implementation of the NOMAD software package [1, 30] and the R dynaTree library [23] for building the statistical surrogate models. The different formulations are compared to MADS without a search step, and to MADS with the use of quadratic models inside the search

**Table 2** List of formulations

Name (Eq.) Section	Formulation	$\lambda$
MADS	MADS with no search [4]	N.A.
Quad	MADS with quadratic model [16]	
$F\sigma$ (12), Sect. 3.1	$\begin{cases} \min_{x \in \mathcal{X}} \hat{f}(x) - \lambda \hat{\sigma}_f(x) \\ st : \hat{c}_j(x) - \lambda \hat{\sigma}_j(x) \leq 0 \end{cases}$	$\lambda \in \{0, \frac{1}{100}, \frac{1}{10}, 1\}$
$F\sigma P$ (17), Sect. 3.2	$\begin{cases} \min_{x \in \mathcal{X}} \hat{f}(x) - \lambda \hat{\sigma}_f(x) \\ st : P(x) \geq p_c \end{cases}$	
$EI\sigma$ (20), Sect. 3.3	$\begin{cases} \min_{x \in \mathcal{X}} -EI(x) - \lambda \hat{\sigma}_f(x) \\ st : \hat{c}_j(x) - \lambda \hat{\sigma}_j(x) \leq 0 \end{cases}$	
$EFI$ (22), Sect. 3.3	$\begin{cases} \min_{x \in \mathcal{X}} -EFI(x) \end{cases}$	$\lambda = 0$
$EFI\sigma$ (23), Sect. 3.3	$\begin{cases} \min_{x \in \mathcal{X}} -EFI(x) - \lambda \hat{\sigma}_f(x) \end{cases}$	$\lambda \in \{\frac{1}{100}, \frac{1}{10}, 1\}$
$EFI\mu$ (25), Sect. 3.3	$\begin{cases} \min_{x \in \mathcal{X}} -EFI(x) - \lambda \hat{\sigma}_f(x) \mu(x) \end{cases}$	
$EFIC$ (26), Sect. 3.3	$\begin{cases} \min_{x \in \mathcal{X}} -EFI(x) - \lambda (EI(x)\mu(x) + P(x)\hat{\sigma}_f(x)) \end{cases}$	

step as defined in [16], which is denoted ‘‘Quad.’’ The formulations  $F\sigma$  (Eq. (12)),  $F\sigma P$  (Eq. (17)), and  $EI\sigma$  (Eq. (20)) are tested for  $\lambda \in \{0, 0.01, 0.1, 1\}$ . The formulations  $EFI\sigma$  (Eq. (23)),  $EFI\mu$  (Eq. (25)), and  $EFIC$  (Eq. (26)) are equivalent to formulation  $EFI$  (Eq. (22)) for  $\lambda = 0$ ; therefore, they are tested for  $\lambda \in \{0.01, 0.1, 1\}$ . A total of  $S = 25$  formulations are tested in this work, as summarized in Table 2.

### 4.1 Problem Description

The *Simplified Wing* problem [41] involves optimizing the geometry of a wing to minimize drag. The two disciplines involved are wing structures and aerodynamics. This problem is smooth but has many local minima. The best objective found in this study (for all formulations and initial guesses) is  $f^* = -16.60$ . The best objective value reported in [41] is  $f^* = -16.65$ . The problem can be summarized as

$$\begin{aligned} & \min \text{Wing drag} \\ & \text{subject to Shear stress} \leq 73,200 \text{ psi} \\ & \qquad \qquad \text{Tensile stress} \leq 47,900 \text{ psi} \\ & \qquad \qquad \text{Sum of the weights} \leq \text{total lift.} \end{aligned} \tag{27}$$

Two structural constraints guarantee wing integrity, and a constraint on the lift ensures sustentation. Table 3 lists the  $n = 7$  design optimization variables, their bounds, and the known optimal values. Five of the variables are related to the aerodynamics properties of the wing; the two other describe its structure.

**Table 3** Design optimization variables for the simplified wing MDO problem

Variables	Bounds		$x^*$
	Lower	Upper	
Wing span	30	45	44.132
Root chord	6	12	6.758
Taper ratio	0.28	0.50	0.282
Angle of attack at root	-1	3	3.0
Angle of attack at tip	-1	3	0.718
Tube external diameter	1.6	5.0	4.03
Tube thickness	0.3	0.79	0.3

The *Aircraft Range* problem [28] considers the design of a supersonic business jet by taking into account aerodynamics, structure, and propulsion. The problem is nonsmooth and has several local optima. The best objective value found in this work for all formulations and initial guesses is  $f^* = -3964.20$ . The best objective value reported in [41] is  $f^* = -3963.98$ . The range of the aircraft must be maximized while satisfying  $m = 10$  constraints related to structure, engine, and performance. The problem can be summarized as

$$\begin{aligned}
 & \max \text{ Aircraft range} \\
 & \text{subject to Normalized stress} \leq 1.09 \text{ (5 constraints)} \\
 & \quad \text{Pressure gradient} \leq 1.04 \text{ Pa m}^{-1} \\
 & \quad 0.5 \leq \text{Engine scale factor} \leq 1.5 \\
 & \quad \text{Normalized engine temperature} \leq 1.02 \\
 & \quad \text{Throttle setting} \leq \text{max throttle},
 \end{aligned} \tag{28}$$

where the max throttle is computed based on the altitude and Mach number using polynomial regression on Pratt & Whitney data [2]. The problem has  $n = 10$  design optimization variables, listed in Table 4 along with their bounds and known optimal values. Seven of them describe the wing aerodynamic properties. The others describe the flight conditions: engine command, altitude, and speed.

### 4.2 Comparison Metrics

Comparisons between the formulations are performed independently on the two sets of optimization runs using statistics of deviation from the best known solution, *data profiles* and *performance profiles* [33].

In each set, the optimization runs are denoted  $p \in \{1, \dots, P\}$ , with  $P = 50$  for each of the MDO problems. The dimension of the design space is denoted  $n$ .  $f_p^*$  is the best feasible objective value found among all formulations for optimization run



**Table 4** Design optimization variables for the aircraft range MDO problem

Variables	Bounds		$x^*$
	Lower	Upper	
Taper ratio	0.1	0.4	0.4
Wingbox cross-section	0.75	1.25	0.75
Skin friction coeff.	0.75	1.25	0.75
Throttle	0.1	1.0	0.156
Thickness/chord	0.01	0.09	0.06
Altitude	30,000	60,000	60,000
Mach number	1.4	1.8	1.4
Aspect ratio	2.5	8.5	2.5
Wing sweep	40	70	70
Wing surface area	50	1500	1500

$p$ . The progress of the optimization is represented by the number  $i \in \{1, \dots, i_{max}\}$  of groups of  $(n + 1)$  evaluations, which is equivalent to the number of *simplex gradient estimates* (SGEs) [33]. The formulations are denoted by  $s \in \{1, \dots, S\}$ , where  $S = 24$ .

#### 4.2.1 Statistics of Deviation from the Best Known Solution

For each formulation  $s$ , the best value of the objective for optimization run  $p$  after  $i$  SGEs is denoted  $f_{p,s,i}$ , which is infinite if no feasible point has been found. The relative deviation from the best known solution is defined as

$$d_{p,s,i} = \min \left\{ \frac{f_{p,s,i} - f_p^*}{|f_p^*|}, 1 \right\}, \tag{29}$$

where  $|f_p^*| > 0$ . To remove outliers, the deviation is bounded. This allows us to compute deviation statistics.

For formulation  $s$ , the average deviation from the best known solution of all optimization runs is defined as

$$d_s^{mean} = \frac{1}{P} \sum_{p=1}^P d_{p,s,i_{max}}. \tag{30}$$

The maximum deviation  $d_s^{max}$  and the standard deviation of the deviation  $d_s^{std}$  are defined accordingly. Table 5 reports deviation statistics (in %) for each formulation and for the two sets of optimization runs. In each column, the formulations that are better (worse) than MADS or Quad are followed by the sign (+) (preceded by the sign (-)). The best value of each column is highlighted in bold.

**Table 5** Relative deviation (%) for the two sets of optimization runs

Formulation	$\lambda$	MDO simp. wing.			MDO airc. range		
		Relative deviation (%)			Relative deviation (%)		
		$d_s^{max}$	$d_s^{mean}$	$d_s^{std}$	$d_s^{max}$	$d_s^{mean}$	$d_s^{std}$
MADS [4]	N.A.	11.9	2.17	1.87	31.7	0.827	4.52
Quad [16]	N.A.	8.79	2.09	1.58	37.2	0.999	5.52
$F\sigma$ Eq. (12), Sect. 3.1	0	<b>3.37(+)</b>	1.72(+)	0.916(+)	22.7(+)	0.598(+)	3.24(+)
	0.01	6.48(+)	1.82(+)	1.05(+)	(-) <b>53.7</b>	(-) <b>1.33</b>	(-) <b>7.67</b>
	0.1	3.59(+)	1.60(+)	0.968(+)	8.23(+)	0.237(+)	1.22(+)
	1.0	(-) <b>13.0</b>	1.90(+)	1.82	(-) <b>77.6</b>	(-) <b>1.73</b>	(-) <b>11.0</b>
$F\sigma P$ Eq. (17), Sect. 3.2	0	4.40(+)	1.59(+)	1.05(+)	(-) <b>86.0</b>	(-) <b>1.76</b>	(-) <b>12.2</b>
	0.01	11.0	1.83(+)	1.64	3.86(+)	0.0773(+)	0.546(+)
	0.1	4.89(+)	1.89(+)	1.12(+)	(-) <b>58.7</b>	(-) <b>1.71</b>	(-) <b>8.93</b>
	1.0	4.07(+)	1.67(+)	0.909(+)	(-) <b>84.8</b>	(-) <b>2.29</b>	(-) <b>12.3</b>
$EI\sigma$ Eq. (20), Sect. 3.3	0	3.83(+)	1.72(+)	1.04(+)	(-) <b>74.2</b>	(-) <b>1.74</b>	(-) <b>10.5</b>
	0.01	3.87(+)	<b>1.54(+)</b>	<b>0.823(+)</b>	(-) <b>91.5</b>	(-) <b>2.17</b>	(-) <b>12.9</b>
	0.1	9.30	1.86(+)	1.48(+)	2.59(+)	0.0658(+)	0.377(+)
	1.0	(-) <b>13.2</b>	(-) <b>2.18</b>	1.80	15.0(+)	0.832	2.61(+)
$EFI$ Eq. (22), Sect. 3.3	0	3.47(+)	1.74(+)	0.916(+)	5.18(+)	0.221(+)	0.873(+)

(continued)

**Table 5** (continued)

Formulation	$\lambda$	MDO simp. wing.		MDO airc. range	
		Relative deviation (%) $d_s^{max}$	$d_s^{mean}$	Relative deviation (%) $d_s^{max}$	$d_s^{mean}$
<i>EFF</i> $\sigma$ Eq. (23), Sect. 3.3	0.01	6.11(+)	1.67(+)	1.20(+)	8.84(+)
	0.1	9.89	1.78(+)	1.52(+)	<b>1.39(+)</b>
	1.0	5.80(+)	2.08(+)	1.02(+)	(-) <b>72.9</b>
<i>EFF</i> $\mu$ Eq. (25), Sect. 3.3	0.01	4.52(+)	1.85(+)	0.923(+)	1.72(+)
	0.1	4.44(+)	1.86(+)	1.19(+)	3.74(+)
	1.0	4.00(+)	1.80(+)	1.06(+)	(-) <b>60.6</b>
<i>EFF</i> $C$ Eq. (26), Sect. 3.3	0.01	4.60(+)	1.79(+)	1.02(+)	3.53(+)
	0.1	6.39(+)	1.82(+)	1.17(+)	(-) <b>50.3</b>
	1.0	9.92	1.89(+)	1.47(+)	(-) <b>90.1</b>

Formulations that are better (worse) than MADS or quad are followed by the sign (+) (preceded by the sign (-)). The best value in each column is highlighted in bold

### 4.2.2 Data and Performance Profiles

Data and performance profiles [33] allow the comparison of optimization methods on a set of runs for the same problem using different parameters (such as initial guess) and/or different problems. Instead of considering the mean (or other usual statistical metrics) of the objective, they consider the ratio of runs to meet a given precision  $\tau$ . As an example, in an engineering design situation, this precision may be required to consider the design as admissible or of practical use. Thus, data and performance profiles express the *ratio of solved problems*, regarding to a precision  $\tau$ .

This ratio, for formulation  $s$ , after  $i$  groups of  $(n + 1)$  evaluations and for a given precision  $\tau$  is defined as

$$r_{s,i}(\tau) = \frac{1}{P} \text{size} \{p \in \{1, \dots, P\}: d_{p,s,i} \leq \tau\}, \quad (31)$$

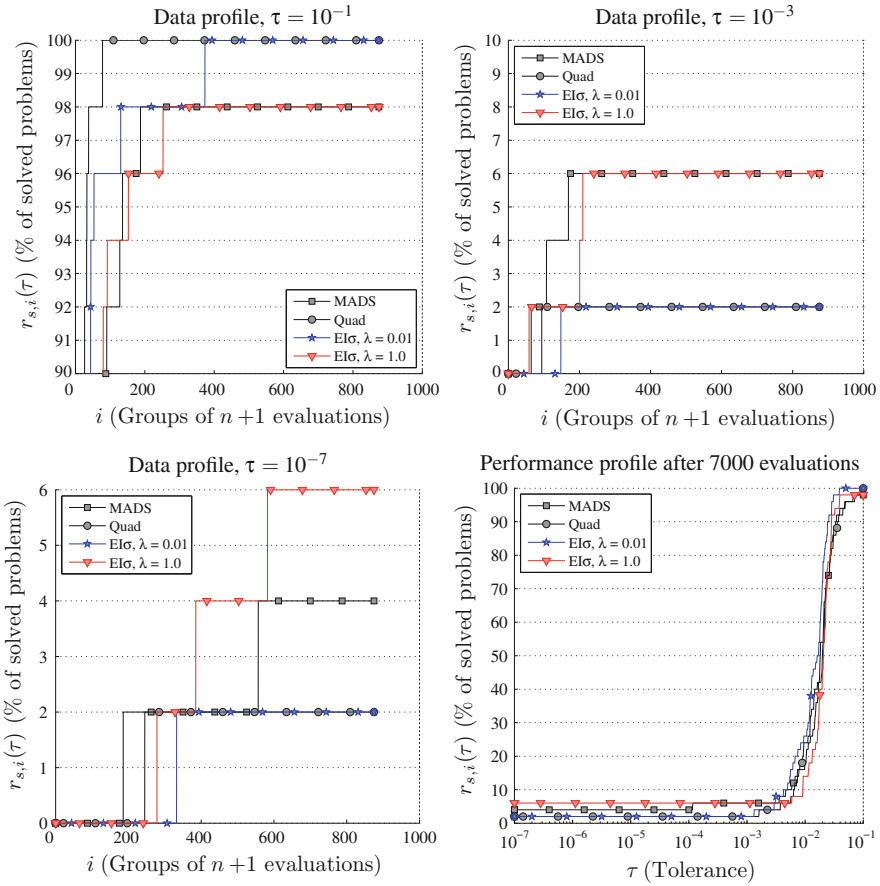
where  $\tau$  represents the tolerance on the deviation  $d_{p,s,i}$ . If the tolerance decreases, the number of optimization runs  $p$  satisfying the condition  $d_{p,s,i} \leq \tau$  will also decrease. For a given  $\tau$ , the proportion  $r_{s,i}(\tau)$  varies depending on the formulation  $s$  and on the number  $i$  of SGEs. As the optimization proceeds, the proportion is likely to increase since more evaluations are performed. For a given  $\tau$  and  $i$ ,  $r_{s_1,i}(\tau) > r_{s_2,i}(\tau)$  means that formulation  $s_1$  yields better results than  $s_2$ . In each profile, the proportion  $r_{s,i}(\tau)$  is plotted for several formulations  $s$  in order to compare them.

In the *data profiles*, the value of  $i$  varies in order to compare the formulations at various times of the optimization. The tolerance  $\tau$  is fixed and can take the values  $\{10^{-1}, 10^{-3}, 10^{-7}\}$ . Each curve in the profile represents the function  $i \rightarrow r_{s,i}(\tau)$  for a formulation  $s$ . The  $x$ -axis specifies the number  $i$  and the  $y$ -axis indicates the ratio  $r_{s,i}(\tau)$ . On a graph showing several data profiles, a higher curve indicates a more successful optimization method for a given number of blackbox evaluations. It is possible that one method leads in the beginning of the optimization and that another method (e.g., a method focusing more on exploration) becomes better as the optimization progresses.

In the *performance profiles*,  $\tau$  varies in order to compare the formulations for various tolerances. The progress of the optimization is fixed at  $i = i_{max}$ , which enables a comparison of the formulations in terms of performance. The tolerance  $\tau$  varies in  $[10^{-7}, 10^{-1}]$ . Each curve in the profile represents the function  $\tau \rightarrow r_{s,i_{max}}(\tau)$  for a formulation  $s$ . The  $x$ -axis represents the tolerance  $\tau$ . As in the data profiles, the  $y$ -axis indicates the proportion  $r_{s,i}(\tau)$ . A higher curve on a graph represents a more successful optimization method for a given tolerance. It is possible however, that a method has the best ratio for small tolerances but is surpassed by other methods for large tolerances.

These profiles are linked since  $r_{s,i_{max}}(\tau)$  appears both at the end of the data profile for the precision  $\tau$  and in the performance profile at abscissa  $\tau$ .

Given the number of formulations presented, the profiles can be displayed in a visually meaningful manner only for a small number of formulations. Therefore,

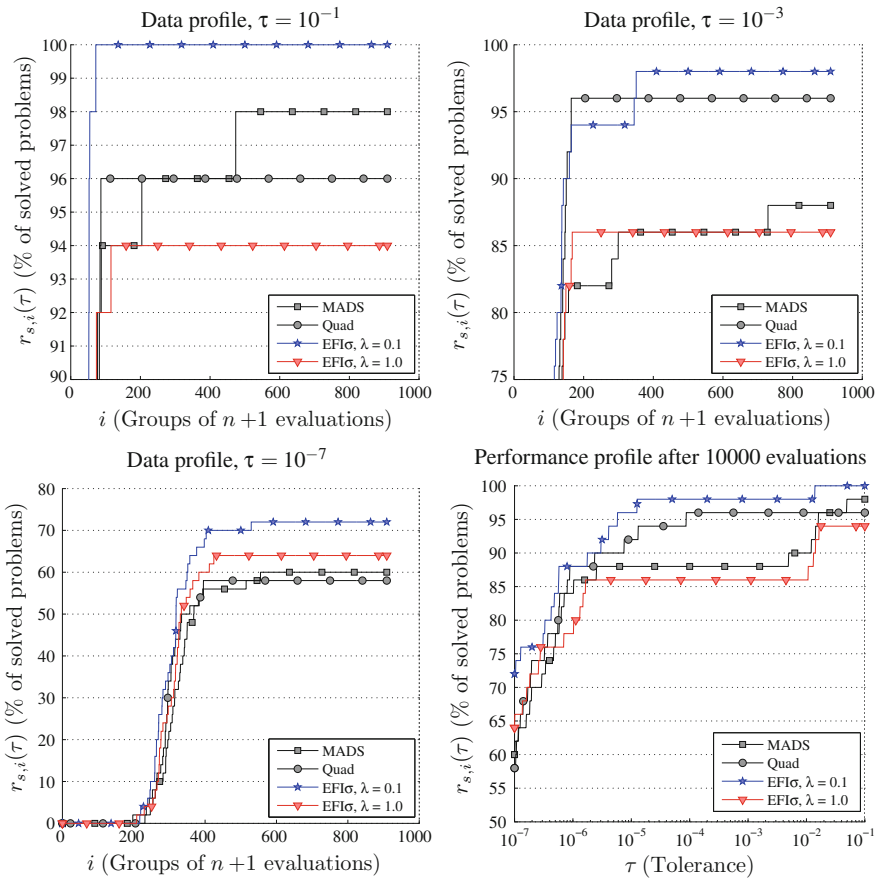


**Fig. 5** Data and performance profiles for the simplified wing MDO problem. MADS and Quad are used as a reference;  $EI\sigma, \lambda = 0.01$ , and  $EI\sigma, \lambda = 1.0$ , are the formulations with the best and worst mean deviation, respectively

for each set of optimization runs, the data and performance profiles are plotted for MADS, Quad, and two formulations: the best and the worst according to the mean deviation  $d_s^{mean}$  (Figs. 5 and 6).

### 4.3 Discussion

If the stopping criteria are met, the optimization algorithm stops before the budget of evaluations is consumed. On the two sets of optimization runs, the mean number of SGEs per run is 615, and 416, respectively. We observe a slight negative correlation ( $-4$ , and  $-6\%$  on the two sets) between the mean deviation of a formulation  $s$  and



**Fig. 6** Data and performance profiles for the aircraft-range MDO problem. MADS and Quad are used as a reference;  $EFl\sigma, \lambda = 0.1$ , and  $EFl\sigma, \lambda = 1.0$ , are the formulations with the best and worst mean deviation, respectively

the mean number of SGEs. This illustrates the need to explore the design space: if no mechanism enables a search for a better solution outside the current attraction basin, the algorithm may converge to a local optimum.

The statistical formulations exhibit a significant advantage for the simplified-wing MDO runs. All but one of the formulations yield a smaller deviation than that of MADS and Quad. Most formulations provide a reduction of more than 10% in the mean deviation. The data and performance profiles show that Quad outperforms MADS for large tolerances, but the opposite occurs for small tolerances. Similarly, if we consider the best and worst statistical formulations according to the mean deviation, formulation ( $EFl\sigma, \lambda = 0.01$ ) outperforms ( $EFl\sigma, \lambda = 1$ ) for large tolerances, and conversely for small tolerances.

The steep curves between  $\tau = 10^{-1}$  and  $\tau = 10^{-3}$  in the performance profile of Fig. 5 illustrate the existence of multiple local minima. Since the problem is not noisy, if the algorithm finds the proper attraction basin, it can quickly reach an accuracy of  $10^{-7}$ . Otherwise, it is unlikely to reach a deviation smaller than  $10^{-3}$ . There is no variation in the proportion of solved problems below  $\tau = 10^{-4}$ .

Finally, for the aircraft-range MDO runs, the mean and maximum deviations yielded by the statistical formulations can be up to three times higher than that of MADS or Quad. However, some formulations are very efficient: 10 formulations are better than MADS and Quad, 6 formulations reduce the deviation by five or more, and  $(EFI\sigma, \lambda = 0.1)$  divides the maximum deviation by 22 and the mean deviation by 29. This discrepancy is caused by a few underperforming runs which fail to find the attraction basin of  $f^*$  and impact negatively the max and mean deviation. The ratio of runs that find the proper attraction basin can better be described with data and performance profiles. For the data profiles in Fig. 6,  $(EFI\sigma, \lambda = 0.1)$  shows a significant advantage: for  $\tau = 0.1$ , a proportion of 100% is reached in less than 100 SGEs. The worst formulation,  $(EFI\sigma, \lambda = 1)$ , is outperformed by MADS and Quad for  $\tau = 10^{-1}$  and  $\tau = 10^{-3}$  but performs better for  $\tau = 10^{-7}$ . The data profile with  $\tau = 0.1$  shows that 94% of the runs of  $(EFI\sigma, \lambda = 1)$  found the attraction basin of  $f^*$ , which is not far from  $(EFI\sigma, \lambda = 0.1)$  (100%), MADS (98%) and Quad (96%). This illustrates the efficiency of a robust regression on noisy functions. In this problem, the statistical formulations show a significant advantage over MADS and Quad.

Based on the results of these numerical experimentations, it appears that mild values of  $\lambda$  (0.01) should be used with formulations  $EFI\sigma$  (Eq. (23)),  $EFI\mu$  (Eq. (25)) or  $EFIC$  (Eq. (26)) to balance global exploration and computational time.

## 5 Concluding Remarks

This work introduced seven novel problem formulations for using statistical surrogates and the MADS derivative-free optimization algorithm for blackbox engineering design. These formulations take advantage of the statistical features of the surrogate and emphasize the exploration of the design space. The presented surrogate management framework formulations can be used with any direct search method based on the search-and-poll paradigm. They have been implemented using the `dynaTree` library to build the statistical surrogate models, and were tested on 2 simulation-based MDO problems. They generally perform as good as or better than existing formulations but seem to exhibit significant advantages when used to solve nonsmooth, noisy and nonconvex problems.

The work emphasizes the appropriate use of statistical surrogate models during the search step of MADS. However, the presented formulations can be employed in any surrogate-based optimization method, for example search-and-poll methods [8] or EGO (Efficient Global Optimization) [26]. Similarly, the surrogate modeling method used in this work (`dynaTree`) can be replaced by any other modeling method that

provides the necessary statistical information: predictive mean, predictive variance and probabilistic distribution. Suitable methods include Kriging [29], Gaussian Processes [35] and Treed Gaussian Processes [22, 24]. Non-statistical surrogates cannot be used with these formulations.

In future work, the surrogate search could be improved by updating the exploration parameter  $\lambda$  depending on the result of the search and on the smoothness of the blackbox outputs. A similar strategy could be applied to the parameter  $p_c$  involved in the chance constraint of formulation  $F\sigma P$  (Eq. (17)). The correlation between the constraints could be analyzed to provide a more accurate estimation of the probability of feasibility. Further experimentations may allow to build a decision process to chose the most promising formulation depending on the characteristics and features of different problems and applications. Finally, we would like to note that the integration of the `dynaTree` statistical surrogate modeling tool with the MADS algorithm will be available in a future release of the free `NOMAD` software package [1, 30].

**Acknowledgments** This work was partially supported by NSERC Discovery Grants 418250-2012 and 436193-2013 and by a GERAD postdoctoral fellowship; such support does not constitute an endorsement by the sponsors of the opinions expressed in this chapter. The authors would like to thank Prof. Charles Audet of GERAD and École Polytechnique for his useful comments and Prof. Robert Gramacy of the University of Chicago for his help with implementing `dynaTree` within `NOMAD`.

## References

1. Abramson MA, Audet C, Couture G, Dennis JE, Jr, Le Digabel S Tribes C, The `NOMAD` project. <https://www.gerad.ca/nomad>
2. AIAA/UTC/Pratt & Whitney. Undergraduate individual aircraft design competition, 1995/1996
3. Audet C, Dennis JE Jr (2003) Analysis of generalized pattern searches. *SIAM J Optim* 13(3):889–903
4. Audet C, Dennis JE Jr (2006) Mesh adaptive direct search algorithms for constrained optimization. *SIAM J Optim* 17(1):188–217
5. Audet C, Bécharde V, Le Digabel S (2008) Nonsmooth optimization through mesh adaptive direct search and variable neighborhood search. *J Glob Optim* 41(2):299–318
6. Bai Z (2002) Krylov subspace techniques for reduced-order modeling of large-scale dynamical systems. *Appl Numer Math* 43(1–2):9–44
7. Bandler JW, Cheng QS, Dakroury SA, Mohamed AS, Bakr MH, Madsen K, Sondergaard J (2004) Space mapping: the state of the art. *IEEE Trans Microw Theory Tech* 52(1):337–361
8. Booker AJ, Dennis JE Jr, Frank PD, Serafini DB, Torczon V, Trosset MW (1999) A rigorous framework for optimization of expensive functions by surrogates. *Struct Multi Optim* 17(1): 1–13
9. Carvalho CM, Johannes M, Lopes HF, Polson NG (2010) Particle learning and smoothing. *Stat Sci* 25(1):88–106
10. Carvalho CM, Lopes HF, Polson NG, Taddy MA (2010) Particle learning for general mixtures. *Bayesian Anal* 5(4):709–740
11. Chipman HA, George EI, McCulloch RE (1998) Bayesian CART model search (with discussion). *J Am Stat Assoc* 93(443):935–960
12. Chipman HA, George EI, McCulloch RE (2002) Bayesian treed models. *Mach Learn* 48 (1–3):299–320



13. Clarke FH Optimization and nonsmooth analysis. Wiley, New York, 1983. Reissued in 1990 by SIAM Publications, Philadelphia, as, vol 5 in the series Classics in Applied Mathematics
14. Cohn DA (1996) Neural network exploration using optimal experimental design. *Adv Neural Inf Process Syst* 6(9):679–686
15. Conn AR, Scheinberg K, Vicente LN (2009) Introduction to derivative-free optimization. MOS/SIAM series on optimization. SIAM, Philadelphia
16. Conn AR, Le Digabel S (2013) Use of quadratic models with mesh-adaptive direct search for constrained black box optimization. *Optim Methods Softw* 28(1):139–158
17. Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20(3):273–297
18. Custódio AL, Rocha H, Vicente LN (2010) Incorporating minimum Frobenius norm models in direct search. *Comput Optim Appl* 46(2):265–278
19. Fletcher R, Leyffer S (2002) Nonlinear programming without a penalty function. *Math Program Ser A* 91:239–269
20. Forrester AIJ, Keane AJ (2009) Recent advances in surrogate-based optimization. *Prog Aerosp Sci* 45(1–3):50–79
21. Goldberg DE (1989) Genetic algorithms in search, optimization and machine learning. Wesley, Boston
22. Gramacy RB, Le Digabel S (2011) The mesh adaptive direct search algorithm with treed Gaussian process surrogates. Technical Report G-2011-37, Les cahiers du GERAD, 2011. To appear in the *Pac J Optim*
23. Gramacy RB, Taddy MA (2010) dynaTree: An R package implementing dynamic trees for learning and design. Software available at <http://CRAN.R-project.org/package=dynaTree>
24. Gramacy RB, Lee HKH (2008) Bayesian treed Gaussian process models with an application to computer modeling. *J Am Stat Assoc* 103(483):1119–1130
25. Gramacy RB, Taddy MA, Wild SM (2013) Variable selection and sensitivity analysis using dynamic trees, with an application to computer code performance tuning. *Ann Appl Stat* 7(1):51–80
26. Jones DR, Schonlau M, Welch WJ (1998) Efficient global optimization of expensive black box functions. *J Glob Optim* 13(4):455–492
27. Jones DR (2001) A taxonomy of global optimization methods based on response surfaces. *J Glob Optim* 21:345–383
28. Kodiyalam S (2001) Multidisciplinary aerospace systems optimization. Technical Report NASA/CR-2001-211053, Lockheed Martin Space Systems Company, Computational Aero-Sciences Project, Sunnyvale, CA
29. Krige DG (1951) A statistical approach to some mine valuations and allied problems at the Witwatersrand. Master's thesis, University of Witwatersrand
30. Le Digabel S (2011) Algorithm 909: NOMAD: Nonlinear optimization with the MADS algorithm. *ACM Trans Math Softw* 37(4):44:1–44:15
31. Liem RP (2007) Surrogate modeling for large-scale black-box systems. Master's thesis, School of Engineering, Computation for Design and Optimization Program
32. McKay MD, Beckman RJ, Conover WJ (1979) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 21(2):239–245
33. Moré JJ, Wild SM (2009) Benchmarking derivative-free optimization algorithms. *SIAM J Optim* 20(1):172–191
34. Queipo N, Haftka R, Shyy W, Goel T, Vaidyanathan R, Kevintucker P (2005) Surrogate-based analysis and optimization. *Prog Aerosp Sci* 41(1):1–28
35. Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. The MIT Press, Cambridge
36. Schonlau M, Jones DR, Welch WJ (1998) Global versus local search in constrained optimization of computer models. In: *New developments and applications in experimental design*, number 34 in IMS Lecture Notes–Monograph Series, pp 11–25. Institute of Mathematical Statistics
37. Serafini DB (1998) A framework for managing models in nonlinear optimization of computationally expensive functions. Ph.D. thesis, Department of Computational and Applied Mathematics, Rice University, Houston, Texas

38. Simpson TW, Korte JJ, Mauery TM, Mistree F (2001) Kriging models for global approximation in simulation-based multidisciplinary design optimization. *AIAA J* 39(12):2233–2241
39. Taddy MA, Gramacy RB, Polson NG (2011) Dynamic trees for learning and design. *J Am Stat Assoc* 106(493):109–123
40. Torczon V (1997) On the convergence of pattern search algorithms. *SIAM J Optim* 7(1):1–25
41. Tribes C, Dubé J-F, Trépanier J-Y (2005) Decomposition of multidisciplinary optimization problems: formulations and application to a simplified wing design. *Eng Optim* 37(8):775–796
42. Vaz AIF, Vicente LN (2007) A particle swarm pattern search method for bound constrained global optimization. *J Glob Optim* 39(2):197–219
43. Willcox K, Peraire J (2002) Balanced model reduction via the proper orthogonal decomposition. *AIAA J* 40(11):2323–2330
44. Williams BJ, Santner TJ, Notz WI (2000) Sequential design of computer experiments to minimize integrated response functions. *Stat Sin* 10(4):1133–1152

# Life Cycle Analysis and Optimization of a Steel Building

G.K. Bekas, D.N. Kaziolas and G.E. Stavroulakis

**Abstract** The present study seeks to couple the problem of the structural optimization of building frames, with that of the optimization of design options for their energy efficiency. The objective function is a cost function that takes into account both the structural cost and energy performance along the whole life of the building. Consequently, the following design parameters are involved: insulation thickness, wall and window insulation profile, window sizes, heating and air conditioning system sizing, sizing of steel cross-sections, as well as parameters related to the life cycle of the building. Modeling is based on acceptable from national and European regulations procedures. Optimization is solved using evolutionary algorithms. The optimization problem is implemented on a steel office building (10 × 15 m), in Chania, Crete, at the south part of Greece. This is a first attempt to combine Life Cycle Cost and Optimization with classical Structural Optimization for steel structures. Depending on the requirements from the users of the building further evaluation using building energy management system (BEMS) for the intelligent operation and management of heating, ventilation and air-conditioning (HVAC) may be performed.

**Keywords** Whole-life cost and optimization · Structural optimization · Energy performance and optimization

---

G.K. Bekas · G.E. Stavroulakis  
Department of Production Engineering and Management, Technical University of Crete,  
73100 Chania, Greece  
e-mail: gbekas@isc.tuc.gr

G.E. Stavroulakis  
e-mail: gestavroulakis@isc.tuc.gr; gestavr@dpem.tuc.gr

D.N. Kaziolas (✉)  
Eastern Macedonia and Thrace Institute of Technology, Kavala, Greece  
e-mail: dnkazio@yahoo.gr

## 1 Introduction

The total life cycle cost of a specific system depends on the most critical components of the system. The most important parameters that are usually examined in life cycle cost problems are the following [1, 2]:

- Construction costs
- Maintenance costs
- Operation costs
- Remaining cost at the end of the structure's expected life cycle.

The formula below is a generalized approach for a system's total life cycle cost (Eq. 1):

$$LCC = C + PV_{\text{RECURRING}} - PV_{\text{RESIDUAL-VALUE}} \quad (1)$$

Where:

LCC is the total life cycle cost.

C is the year 0 construction cost.

$PV_{\text{RECURRING}}$  is the present value of all recurring costs (utilities, maintenance costs, replacements, service costs etc.).

$PV_{\text{RESIDUAL-VALUE}}$  is the present value of the residual value at the end of the examined life cycle period, expressed at the reference year. The residual value is either considered to be equal to zero or it can be calculated through the following formula (Eq. 2):

$$PV_{\text{RESIDUAL-VALUE}} = \text{Subsystem's initial value}^*(\text{Current year})/(\text{Subsystem's total life cycle (in years)})^*\text{Factor accounting for the inflation rates} \quad (2)$$

Nevertheless, since in optimization problems the above formula generally has little practical importance, the present study will neglect the residual values in its optimization calculations.

In order for the life cycle cost of a specific building to be minimized, it is important to determine -during its design and construction stage- the subsystems that affect its life cycle cost with the view of taking optimal design decisions.

In general, the following subsystems have a considerable impact on the life cycle cost of a specific building [3]:

- Building Envelope (insulation profiles, shading systems, glazing, roofing etc.)
- Mechanical and Energy Systems (use of photovoltaic panels or alternative sources of energy, ventilation systems, water distribution systems)
- Structural Systems (selection of appropriate frame materials, sizing of the frame components)

- Siting (landscaping and irrigation-related design decisions).
- Electrical Systems (lighting sources and control, distribution)

For typical cases of buildings in Greece, practical experience as well as data derived from statutory sources in building construction cost analysis studies have shown that the most critical subsystems that affect its total whole life cost are those related to its structural and energy performance.

These subsystems also interact with one another as the building frame affects its energy performance and the insulation plays a role on the frame's structural design loads. Furthermore, the other subsystems such as the water distribution systems, landscaping options, electrical systems constitute an optimization problem that can be examined separately.

Apart from that, it is also necessary to consider the average life cycle of the above mentioned subsystems in order to predict any potential replacements that may occur during the examined life cycle period. According to various sources (Technical Chamber of Greece, Stanford university, CIBSE [4]), the average life cycle of the examined building components is as follows:

- Structural steel: 80 years (lifetime)
- Building Exteriors, Doors, and Windows: 80 years (lifetime)
- EPS insulation profiles: 100 years (lifetime)
- Mineral wool insulation profiles: 50 years
- HVAC systems: 15–20 years

Critical information about the building subsystems' service lives as well as their maintenance rates can also be found in the following software: ATHENA, BEES, Boustead, GaBi, SIMAPRO [5]. The purpose of this information is to reflect what would more likely happen in an average situation. It is meaningful to note that the rates are dependent on the geographic location of the building. In such software there is also provision for various scenarios of life cycle design decisions reflecting combinations of different scores of environmental friendliness and economic efficiency.

The methodology proposed in this paper can be extended to cover buildings made of timber or reinforced concrete material and combinations of them. The interaction between structural optimization and energy performance, in view of the whole life cycle analysis, have not been studied in other sources and seems to be an interesting and useful investigation.

## 2 Methodology

In terms of their contribution to total life cycle cost of typical building, the most important subsystems are the structural systems and the systems related to the energy design of a building. These ones can also be optimized from the early stages of the design of a building.

In order to test the capacity of software to optimize these subsystems together, it was decided to develop an algorithm unifying the structural and energy performance optimization of a building. This algorithm would also be one of the first published attempts to optimize the energy performance of buildings according to KENAK; the recent Greek code for the energy design of buildings, which is fully compatible with European codes [6, 7].

At first, a market research took place in an attempt to discover average, real-life cost figures of the subsystems that would be used in the algorithm. The market research took into consideration the costs of the following building components:

- Metallic wall or roof panels.
- EPS or mineral wool insulation of various thicknesses.
- A+++ or A energy class air-conditioning systems.
- Structural steel cost per kg.
- Double and triple-glazed aluminum windows (with regular or low-e values).

In order to save computational time and unify parameters that have an impact on each other and correlate the energy performance parameters with the resultant cost, curve-fitting and multiple linear regression has been used. The cost functions below are some of the ones that were used in the algorithms and they are demonstrated in order for the reader to be able to understand the logic behind that idea:

```

costwindowseast = (218.376 - 38.931*Uwineast +
47.888*gg1)*Awineast
costinsulationwallwest = (5.603* Uwwest^-1.21)*Awwest
(mineral wool)
costAC = -3461.45 + 172.5595*Ptherm + 190.222*SEER +
674.565*SCOP (A energy class air conditioning systems)

```

The correlation results revealed (basing any judgment on the computed R-squared values of the cost functions that were produced through multiple linear regression [Bekas Ph.D., in preparation]), a relatively high degree of correlation implying a logical relationship between cost and critical energy performance parameters. The use of other methods of data fitting such as neural networks could also be an effective alternative for the purpose of creating cost functions. Furthermore, the following scenarios are examined, for a life cycle period of 10 or 30 years:

**Scenario 1:**

-Mineral wool insulation profiles with A energy class A/C as HVAC system.

**Scenario 2:**

-EPS insulation profiles with A energy class A/C as HVAC system.

**Scenario 3:**

-EPS insulation profiles with A+++ energy class A/C as HVAC system.

### 3 Optimization Procedure

After a finite element analysis, the building frame components were optimized along with the following subsystems:

The steel frame cross-sections were modeled as discrete variables reflecting carefully selected predefined choices of cross-sections (In terms of the programming approach that was used in the algorithms, the characteristic dimensions  $b$ ,  $d$ ,  $t_w$ ,  $t_f$  of each cross-section, derive from a multiple-if algorithmic structure that associates each variable to the steel cross-section characteristic dimensions) [8].

- U-values of floor (the U-value measures the performance of a building element in terms of heat transfer; it is assumed that the building floor has a reinforced concrete slab (of 20 cm thickness) and below that a u-value results from the optimization procedure).
- U-values of walls (each orientation was examined separately).
- U-value of roof.
- Area of windows (south elevation).
- Area of windows (all other elevations; each orientation was examined separately).
- ggl value.
- Power of heating system.
- Power of cooling system.
- SCOP (Seasonal coefficient of performance of the thermal system).
- SEER (Seasonal coefficient of performance of the cooling system).

### 4 Constraints

The algorithm that was developed took into account the following constraints:

- Stress constraints were imposed on the steel frame cross-sections [8].
- The power of the heating system should be greater than the result of following formula that is used for the sizing of heating systems by the Greek specifications.

$$P_{\text{thermalsystem}} > 2.5 \times U_m \times A \times \Delta T$$

(Where:  $U_m$  is the average u-value of the exposed (to the atmospheric air) building envelope,  $A$  is the total area of the exposed building envelope,  $\Delta T$  is a temperature difference used for the sizing of the thermal system and is increased through the multiplication by a coefficient that co-estimates losses etc.)

- The same should apply for the air conditioning system, whose power (in kilowatts) must be sufficient for the most adverse day of the summer (21st of July) [9].
- All the components of the building envelope should have acceptable lower and upper limits of u-values. Therefore:

- **U-values of walls:**  
 $0.20 < U_{\text{walls}} < 0.60$
  - **U-value of the floor:**  
 $0.20 < U_{\text{floor}} < 1.20$
  - **U-value of the roof:**  
 $0.20 < U_{\text{roof}} < 0.50$
- The overall average u-value of the building should be lower than what is required by the relevant specification (KENAK) [6, 7].
  - The window u-values should be realistic and therefore they should not be lower than what can be encountered in the market.
  - The seasonal coefficients SCOP for the heating system and SEER for the air-conditioning system should represent the upper and lower limits that are encountered in the Greek market.
  - The total window area in the main elevation (therefore, the south oriented elevation with an acceptable deviation equal to plus or minus 30 degrees ( $\pm 30^{\text{circ}}$ )) of the building should be sufficiently big. Despite the fact that this consideration is generally a choice dependent on the architectural designer, for the current building it was decided that 45 % of the total window area should have south orientation.
  - The total area of the building windows should ensure sufficient natural illumination and ventilation. According to the Greek building codes, this area should represent at least 10 % of the total area of the building.
  - The ggl values (hence, g values multiplied by 0.75; therefore reduced due to the contribution of the window frame that was considered to approximately occupy 25 % of their total area) of windows should have a value between 0.29 and 0.55.

## 5 Model

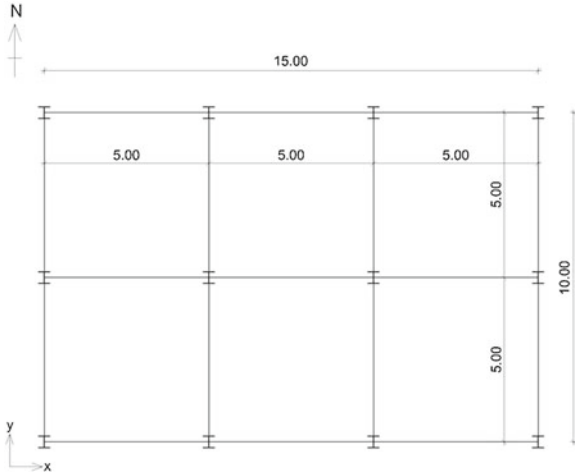
The building that was used in the simulation is a single-storey steel building located on Chania, Crete. A plan view of the building -which has a  $10 \times 15$  m rectangular shape, is shown below (Fig. 1).

At first it was assumed that the building will be used as an office building and this influenced the considerations that were used in the calculations (thermal or cooling loads generated by the theoretical population of building users, minimum required ventilation, characteristic electrical appliances expected to be used in the building).

Apart from that, the following data were used for the optimization of energy design of the building:

- The thermal bridges were calculated with the use of the approximate standardized values of the national standards [6, 7].
- The outer and inner walls are made of metallic panels and their color is a nuance of grey and the level of shading is considered to be known [6, 7].





**Fig. 1** Simplified plan view of the building

- The solar gains during the winter period (October to May) are not taken into account in the calculation of the total thermal load. The opposite however, applies for the summer period (May to October). The solar gains were calculated with the use of the approximate standardized values of the national standards for the specific geographic location [6, 7].
- Loads on the steel frame: 20.94 kN/m (middle span along x-x axis, mineral wool scenario).
- Loads on the steel frame: 10.24 kN/m (side spans along x-x axis, mineral wool scenario).
- Loads on the steel frame: 20.48 kN/m (middle spans along y-y axis, mineral wool scenario).
- Loads on the steel frame: 10.24 kN/m (side spans along y-y axis, mineral wool scenario).
- Loads on the steel frame: 18.76 kN/m (middle span along x-x axis, EPS scenarios).
- Loads on the steel frame: 9.38 kN/m (side spans along x-x axis, EPS scenarios).
- Loads on the steel frame: 18.76 kN/m (middle spans along y-y axis, EPS scenarios).
- Loads on the steel frame: 9.38 kN/m (side spans along y-y axis, EPS scenarios).
- Base temperature inside the building = 25 °C.
- Heating Degree days (Geographic location: Chania) = 2215.
- Cooling degree days (Geographic location: Chania) = 218.
- Uniform building height = 3 m.
- Examined life cycle period in years: 10 & 30 years
- Coefficient accounting for the electricity cost in Euros/kWh = 0.012269.
- Illumination load per square meter: 0.05 kWh/m<sup>2</sup>.

In accordance with the relevant specifications, it was also taken into account that the office building has an intermittent type of heating, a 5 day working week and 4 hr occupancy and this consideration resulted in the selection of an appropriate correction factor [4, 6, 7]. The heating and cooling costs derive from the energy balance of the building (losses minus gains) [6, 7, 10] and are multiplied by the previously mentioned coefficient that converts the energy needs (in kWh) into electricity costs. The maintenance rates for the building are considered to be equal to 1% of its initial value (therefore, unaffected by inflation rates) per year, with a start point five years after its construction. As regards the HVAC systems, the maintenance rate is considered to be equal to 2% of their initial value (unaffected by inflation rates) per year [11]. An inflation rate with a constant value equal to 3% per year is also taken into account in the calculation of the cost of their replacement at the end of their life cycle (20 years) [12].

As regards the heating and cooling costs, it is also possible to use the predicted UPV values of the electricity costs 30 years after the construction of the building, however only predicted values from countries such as the USA, can be found.

The objective function is the sum of the cost of the following subsystems:

```
total cost = cost of insulation + Heating cost*Number
of years + Cooling cost*Number of years + cost of frame
+ cost of A/C system + cost of windows + cost of roof +
cost of walls + HVAC maintenance + general building
maintenance + cost of the floor slab
```

The constraints incorporated in the objective function describing the total life cycle cost through the use of conditional penalty functions whose violation would result in very high cost values.

## 6 Results and Discussion

The optimization problem is possible to be solved with the use of simulated annealing and genetic algorithms and the first method seems to constantly produce better results.

The energy performance optimization results that were produced by running several scenarios for a life cycle period of 10 or 30 years are shown in the appendix (Tables 1, 2 and 3).

As regards the optimized cross-sections of the frame components (Scenario 1):

- Middle span beams: IPE 240.
- Side span beams: IPE 200.
- Corner columns: HEB 140.
- Middle columns of the west and east elevation: IPE 300
- Middle columns of the north and south elevation: IPE 300.
- Interior columns: IPE 360.

The optimized cross-sections of the frame components for the scenarios 2 & 3, are as follows:

- Middle span beams: IPE 240.
- Side span beams: IPE 200.
- Corner columns: IPE 100.
- Middle columns of the west and east elevation: IPE 100.
- Middle columns of the north and south elevation: IPE 100.
- Interior columns: IPE 120.

An interpretation of the results can lead to the following conclusions:

- It seems to be a cost-effective decision to use window panes with very low  $g$  values. Nevertheless, for the examined life cycle periods of the building the triple glazed window profiles with low  $g$  values, in no case constituted the optimal alternative. The area occupied by the windows is every time dependent on the optimization calculations.
- The floor generally seems to be the least important component to insulate and the roof the most important to insulate. Furthermore, the optimal insulation thickness of the walls slightly increases with the increase of the examined life cycle period.
- Subsystems with a high degree of homogeneity (e.g. A+++ or A energy class A/C systems and insulation profiles where the thickness of -merely one- specific material needs to be optimized) can be correlated with energy performance parameters through multiple linear regression, attaining very high R-squared values. This can save considerable computational time.
- The optimization program naturally selects larger -within reason- window areas on the south elevation. It seems that it may be a redundant constraint to place a lower bound on the window area of the south elevation.
- The heating and cooling requirements of the office building can be covered with a typical 12000 Btu, A/C system. The comparison of the market prices for the current building showed that an A energy class A/C system is by 67 % a cheaper alternative in comparison with an A+++ energy class A/C system. It should be born in mind that the algorithms also consider replacement of the HVAC system 20 years after the building construction [12]. It meaningful to note that the simulation logic that was used in the algorithm considered that merely one A/C unit would be used. Therefore, the cost function concerns only one unit of specific energy class. Evidently, in larger buildings there is potential for different simulation approaches and the number of the air conditioning system terminals could either be predefined or it could be a variable of the optimization problem.
- The figure below (Fig. 2) displays several well-known upper limits of building energy consumption levels. Level 1 is an approximate figure for current acceptable consumption levels for buildings in Germany, level 2 stands for the Minergie practice followed by Switzerland, levels 3 & 4 are regarded as low energy consumption levels and buildings whose energy consumption is below  $15 \text{ kWh/m}^2$

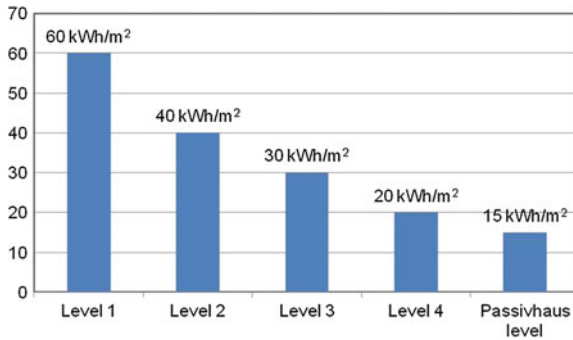


Fig. 2 Well-known building energy consumption levels

are classified as passivhaus. The results showed that 10 years after the construction of the building the optimal level is around  $32 \text{ kWh/m}^2$ , but 30 years after the construction of the building it escalates to slightly above  $30 \text{ kWh/m}^2$  [13–16].

All in all, a life cycle analysis of a steel building has been performed that takes into account energy considerations for both construction and material costs, as well as the energy consumption during the whole life of the structure. In this sense structural and energy optimization are combined and solved with practical global optimization algorithms. It must be emphasized that the complexity of the model restricts the applicability of classical, local numerical optimization algorithms.

By using the proposed model an optimal design of a new steel structure that takes into account its energy consumption during its whole life cycle can be attempted.

The cost functions proposed here can also be used for the evaluation of several alternative design scenarios, including the usage of different materials (for example, timber or reinforced concrete structures).

Since almost all involved quantities are contaminated with uncertainties, extension of the proposed method using fuzzy variables and fuzzy optimization seems to be reasonable. This extension remains open for further investigation.

**Acknowledgments** This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: ARCHIMEDES III. Investing in knowledge society through the European Social Fund.

## Appendix: Results of the Optimization Calculations 1

**Table 1** Scenario 1 results of the optimization calculations (Optimal Energy consumption level: Below level 2)

	U/floor	SCOP	ggl	Awin south	Awin north	Awin east	Awin west	Uroof	Uwall south	Uwall north	Uwall east	Uwall west	Uwin south	Uwin north	Uwin east	Uwinwest	Power of HVAC system	SEER	Time period
1	0.99	3.6	0.29	11.05	4.348	0.500	0.525	0.500	0.591	0.600	0.592	0.600	3.399	2.849	2.45	3.036	4.385	3.202	30 years
2	1.20	3.6	0.29	7.783	4.174	0.500	2.546	0.500	0.600	0.600	0.600	0.600	3.400	2.745	3.05	3.386	4.723	3.200	10 years

**Table 2** Scenario 2 results of the optimization calculations (Optimal Energy consumption level: Below level 2)

	U/floor	SCOP	ggl	Awin south	Awin north	Awin east	Awin west	Uroof	Uwall south	Uwall north	Uwall east	Uwall west	Uwin south	Uwin north	U win east	Uwinwest	Power of HVAC system	SEER	Time period
1	1.19	3.6	0.290	9.045	4.02	0.531	1.406	0.500	0.566	0.600	0.600	0.600	3.400	3.106	3.03	3.40	4.714	3.20	30 years
2	1.14	3.6	0.291	9.304	4.71	0.500	0.500	0.500	0.599	0.600	0.600	0.600	3.399	2.644	3.05	3.35	4.614	3.20	10 years

**Table 3** Scenario 3 results of the optimization calculations (Optimal Energy consumption level: Below level 2)

	U/floor	SCOP	ggl	Awin south	Awin north	Awin east	Awin west	Uroof	Uwall south	Uwall north	Uwall east	Uwall west	Uwin south	Uwin north	Uwin east	Uwinwest	Power of HVAC system	SEER	Time period
1	0.9	5.1	0.29	10.38	3.62	0.500	1.328	0.50	0.58	0.59	0.60	0.57	3.40	2.75	2.55	3.32	2.95	8.33	30 years
2	1.07	5.1	0.34	6.99	4.77	0.500	3.193	0.50	0.60	0.60	0.60	0.60	3.40	2.67	3.00	3.40	3.17	8.15	10 years

## References

1. Kaziolias DN, Zygomalas I, Stavroulakis GE, Emmanouloudis D, Baniotopoulos CC (2013) Evolution of environmental sustainability for timber and steel construction. In: Hakansson A, Häjer M, Howlett RJ, Jain LC (eds) Proceedings of the 4th international conference in sustainability in energy and buildings (SEB'12), 2013. Smart Innovation, Systems and Technologies, vol 22, pp 24–33
2. Kaziolias DN, Zygomalas I, Stavroulakis GE, Baniotopoulos CC (2013) Life cycle assessment of a steel-framed residential building. In: Topping BHV, Iványi P(eds) Proceedings of the fourteenth international conference on civil, structural and environmental engineering computing, 2013. Civil-Comp Press, Stirlingshire, UK, Paper 152. doi:[10.4203/ccp.102.152](https://doi.org/10.4203/ccp.102.152)
3. Stanford University: Land and Buildings (2005) Guidelines for life cycle cost analysis
4. Balaras CA (2011) Estimating energy consumption. National Observatory of Athens, Greece
5. BEES (Building for Environmental and Economic Sustainability) software homepage (2014) <http://ws680.nist.gov/Bees/>. Accessed May 2014
6. Technical Chamber of Greece (2010) T.O.T.E.E. 20701–1/2010 & T.O.T.E.E. 20701–2/2010. Athens, Greece
7. Technical Chamber of Greece (2012) T.O.T.E.E. 20701–3/2010. Athens, Greece
8. Cheng FY, Truman KZ (2010) Structural optimization: dynamic and seismic applications. Spon Press, USA
9. Bourkas PD (1998) Applications of building services in typical and industrial buildings. Dissertation, National Technical University of Athens, Greece, pp 214–233
10. Weber T et al (2004) The utilization factor for free heat in buildings: a statistical approach. KTH-The Royal Institute of Technology, Stockholm
11. Nielsen TR (2002) Optimization of buildings with respect to energy and indoor environment. Ph.D. Dissertation, Department of Civil Engineering, Technical University of Denmark, Denmark, pp 26–103
12. ISO 15686–5 (2008) Building and constructed assets—service life planning
13. Current German Energy Saving Regulations for Buildings (2014) EnEV 2014. Accessed March 2014
14. Minergie building energy design criteria (2014) [http://www.minergie.ch/standard\\_minergie.html](http://www.minergie.ch/standard_minergie.html). Accessed March 2014
15. Passive house certification criteria (2014) [http://www.passreg.eu/index.php?page\\_id=305](http://www.passreg.eu/index.php?page_id=305). Accessed March 2014
16. Thomsen KE, Wittchen KB (2008) European national strategies to move towards very low energy buildings. Danish Building Research Institute, Aalborg University, Denmark



# Optimization of Reinforced Concrete Columns Subjected to Uniaxial Loading

Gebrail Bekdaş and Sinan Melih Nigdeli

**Abstract** The distance from extreme compression fiber to neutral axis ( $c$ ) is depended to combinations of axial load and flexural moment capacities of reinforced concrete (RC) columns. Since  $c$  is depended to different internal forces, the value of  $c$  cannot be found without assuming the final design. Thus, it can be iteratively searched in order to find the flexural moment capacity of columns under an axial loading. By using the presented method, the solution with the minimum cost ensuring maximum flexural moment and axial load is found. A random search technique is explained in this chapter for optimum design of uniaxial RC columns with minimum cost. In optimization, design of RC columns is done by considering the design rules described in ACI 318- Building Code Requirements for Structural Concrete. The random search technique (RST) for optimization of RC uniaxial columns is effective on finding optimum cross-sections and reinforcement design with minimum cost.

**Keywords** Reinforced concrete · Columns · Random search technique · Optimization · ACI-318 · Cost optimization

## 1 Introduction

In design of reinforced concrete (RC) structures, structural members are defined according the architectural designs. The main goal of the design engineer is to find the solution by considering security measures given in design codes, esthetic and comfort requirements of people and economy in material. Although the architectural projects limit the independent design of engineer, the design is done by assuming the design variables between these limits. Then, the assumed design is modified according to design codes if the dimensions of the member are not suitable to carry

---

G. Bekdaş (✉) · S.M. Nigdeli  
Department of Civil Engineering, Istanbul University, 34320 Istanbul, Avcılar, Turkey  
e-mail: bekdas@istanbul.edu.tr

S.M. Nigdeli  
e-mail: melihnig@istanbul.edu.tr

out internal forces resulting from static and dynamic sources. The economy of the design is depended to the experience of design engineers. Although the cross-section dimensions are precisely assumed for RC members, the required reinforcement bar can never be provided as calculated since the bars in the market are constant in size. For these reasons, optimization is important for RC members. This chapter represents a numerical optimization technique for uniaxial RC columns. The design variable of RC columns such as cross-sectional dimensions and amount of steel bars (detailed design with diameter size and numbers) are randomly searched in the presented method for the minimum material cost ensuring the ACI-Building Code Requirements for Structural Concrete requirements. For several RC structural members, several optimization methodologies have been proposed. The reviews of several studies are presented in Sect. 2.

## 2 Literature Survey for Optimization of RC Members

The recent approaches contain optimization of RC structures (2D or 3D) or a detailed optimization of a member of a RC structure. Metaheuristic based methods are the leading ones for the last 15 years in search several design variables of the optimized RC member. The following contributions to the optimization science for RC application are given in this section.

Coello et al. employed genetic algorithm in development of an optimization approach for RC beams [1]. Genetic algorithm is also used in the approach of Rafiq and Southcombe for optimization biaxial RC columns [2]. Several RC member was optimized by genetic algorithm based approach of Koumousis and Arsenis [3]. The detailed reinforcement design of RC frame structure employing genetic algorithm was done by Rajeev and Krishnamoorthy [4]. By employing sequential quadratic programming technique, shape optimization of RC members was done and genetic algorithm was employed in cost optimization by Rath et al. [5]. By considering slenderness of the columns, RC frames was optimized by Camp et al. by employing genetic algorithm and the optimization process was carried out by grouping several members of RC structures [6]. Ferreira et al. optimally designed T-shaped RC beams according to different design codes [7].

Genetic algorithm was also combined with other metaheuristic method inspired from natural phenomena and these hybrid algorithms have been used in the optimization of RC members. A hybrid algorithm, which is combination of genetic algorithm and simulated annealing, was used by Leps and Sejnoha for optimum design of continuous beams [8]. By considering lateral equivalent static earthquake loads at the joints, Lee and Ahn optimized RC frame by using a genetic algorithm based method and a database including possible design of RC members [9].

Three dimensional RC frame structure under excitation of dead, live, snow and earthquake load was optimized by the method of Balling and Yao [10]. Ahmadkhanlou and Adeli proposed an optimization method with two stages for optimization of

RC slab. The neural dynamics model [11, 12] for the optimum solution of continuous variables and perturbation technique modify the values to practical ones were used in the optimization method [13]. Barros et al. developed expressions for the bending moment, steel area and ratio for singly or doubly RC beams for optimum design [14]. Optimum cost design of pre-stressed concrete bridges were done by Sirca Jr. and Adeli [15]. RC continuous beams were optimized by using a genetic algorithm based method and selecting design variables from a database in the study of Govindaraj and Ramasamy [16]. By combining genetic algorithm and discretized form of the Hook and Jeeves method, a hybrid algorithm was used in the optimization of RC flat slab buildings [17]. RC frames were optimized by the genetic algorithm based method of Govindaraj and Ramasamy [18]. Single-bay multi-story and multi-bay single story RC frames were optimized by Guerra and Kiousis [19]. A multi objective optimization approach for RC frames was developed by Paya et al. by employing a metaheuristic method called simulated annealing [20]. Two heuristic methods such as random walk and descent local search and two metaheuristic methods such as the threshold accepting and the simulated annealing based optimization was proposed for optimization of RC frames of bridges [21].

Generally RC member optimization studies consider the minimization of the cost. Several studies considered the value of embedded CO<sub>2</sub> emission. Two different approaches using simulated annealing algorithm [22] and big bang-big crunch optimization [23] were used for the optimization of RC frames in order to reduce cost and embedded CO<sub>2</sub> emission.

Gil-Martin et al. developed a reinforcement sizing diagram approach for RC beams and columns [24]. Barros et al. investigated the optimum depth and reinforcement of RC beam in rectangular shape [25]. According to Eurocode 2, Fedghouche and Tiliouine optimized singly reinforced T-shaped RC beams by employing genetic algorithm [26].

Several approaches employing metaheuristic algorithms such as simulated annealing [27, 28], harmony search [29], big bang-big crunch [30] and charged system search [31] have been used in the optimization of RC retaining walls. The music inspired metaheuristic algorithm called harmony search have been used in the optimization of several RC members such as continuous beams [32], T-shaped RC beams [33], columns [34] and frames [35]. Kaveh and Sabzi optimized RC frames by using several metaheuristic algorithms [29]. Optimum design of RC beams were done by Kaveh and Sabzi and big bang-big crunch was employed in their approach [36]. Rama Mohan Rao combined several algorithms such as simulated annealing and tabu search for optimization of hybrid fiber-reinforced composite plates [37]. A random search technique for the optimization of RC beams [38] and columns [39] was developed. In the following section, random search technique for optimization of RC columns is summarized.

### 3 Random Search Technique for Optimum Design of RC Columns

An RC column with a cross-sectional dimension and reinforcements can carry different combination of axial force and flexural moment. This reason is resulting from the change of stresses on the cross-section and location of steel reinforcements. Thus, the distance from the extreme compression fiber to the neutral axis ( $c$ ) changes according to loading conditions. In design of RC members, a reinforcement ratio is calculated for a constant cross-section. When the cross-section is assumed, the reinforcements for the axial force can be found and the flexural moment capacity can be calculated according to the value of  $c$ . For the ratio of reinforcements the moment capacity may be very different than the required one. In order to find the closest flexural moment value to the required one optimization techniques must be used.

The presented method; random search technique is numerical algorithm which iteratively search the best design of RC member according to design constraints, member loadings and objectives. The objective of the optimization is explained as material cost in this chapter. ACI-318 Building Code Requirements for Structural Concrete [40] rules were taken into consideration.

As mentioned in the introduction section of the chapter, design engineers are depended to architectural project. For that reason or esthetics of the building, the cross-section dimensions are limited with the ranges. These ranges may be also selected as practical dimensions for shortening the optimization process. Also, ranges for steel reinforcement must be used to shorten the optimization process and consider the supplying of the steel bars. The bar with big diameter sizes may not be found near to the construction. For that reason, the price of the steel may increase because of transportation costs.

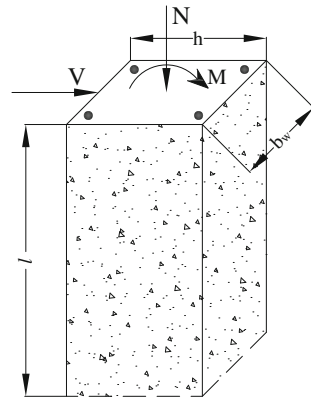
Generally, concrete is a cheap material comparing to steel but cost ratio of concrete to steel may change according to the region of the construction yard. Transportation and import costs play a great role in this factor. Also, if the travel time of the concrete form facility to construction yard is long, the use of admixture may increase the cost of the concrete. For that reason, numerical optimization of RC member must be done by considering specific conditions. Mathematic optimum result may not be optimum for all specific conditions.

Before the random search of design variables of RC columns, several design constants given in Table 1 are defined. These design constants are length of column ( $l$ ), clear cover ( $c_c$ ), maximum aggregate diameter ( $D_{max}$ ), elasticity modulus of steel ( $E_s$ ), specific gravity of steel ( $\gamma_s$ ), specific gravity of concrete ( $\gamma_c$ ), yield strength of steel ( $f_y$ ), compressive strength of concrete ( $f'_c$ ), cost of the concrete per  $m^3$  ( $C_c$ ), cost of the steel per ton ( $C_s$ ). Also, loading conditions such as axial force ( $N$ ), shear force ( $V$ ) and flexural moment ( $M$ ) are defined. In the Fig. 1, the loading of the column is shown.

**Table 1** Design constants of the optimization problem

Definition	Symbol
Length of column (l)	l
Clear cover	$c_c$
Range of reinforcement	$\phi$
Range of shear reinforcement	$\phi_v$
Max. aggregate diameter	$D_{max}$
Yield strength of steel	$f_y$
Comp. strength of concrete	$f'_c$
Elasticity modulus of steel,	$E_s$
Specific gravity of steel,	$\gamma_s$
Specific gravity of concrete	$\gamma_c$
Cost of the concrete per $m^3$	$C_c$
Cost of the steel per ton	$C_s$

**Fig. 1** Loadings of column



The ranges of design variables are also defined. The design variables are breadth of column ( $b_w$ ), height of the column ( $h$ ), number and diameter size of longitudinal reinforcement bars in two lines (including web reinforcements) and diameter size and distance of shear reinforcements. In Fig. 2, the design variables are shown. Symmetrical design is done for upper and lower section of the column.

After the design constants, loadings and ranges of design variables are defined, cross-section dimensions ( $b_w$  and  $h$ ) are randomly defined by considering the selected range. For productivity of the column in the construction yard, dimensions are assigned with productivity values which are multiple of a value. The ductile fracture conditions given in Eq. (1) and (2) are checked for randomly selected dimensions. The first condition is a shear force criterion with two inequalities while second condition is related with the axial capacity of columns.

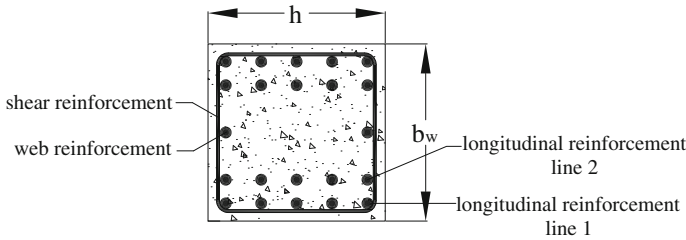


Fig. 2 Design variables

$$V < \begin{cases} 0.2f'_c A_c \\ 5.5A_c \end{cases} \tag{1}$$

$$N < 0.5f'_c A_c \tag{2}$$

In Eqs. (1) and (2),  $A_c$  represents the cross-sectional area ( $b_w h$ ) of column. If these conditions are not suitable for selected cross-section,  $b_w$  and  $h$  are iteratively randomized.

After cross-section dimension supporting ductility conditions, reinforcement design is started. Number and diameter size of the longitudinal reinforcement are randomly defined for upper and lower faces of column. In order to carry flexural moment in opposite directions, the same reinforcements were used for both faces of column. ACI-318 rules are checked for the orientation of reinforcement bars. If needed, the reinforcements are positioned in two lines. Placement condition defined in ACI-318 [40] for columns are shown in Eq. (3).  $\phi_{average}$  is the average of the diameter sizes in a line where the placement condition is checked.  $a_\phi$  is the clear distance between reinforcements. The reinforcements are iteratively randomized until placement condition is satisfied.

$$a_\phi > \begin{cases} 1.5 \phi_{average} \\ 40 \text{ mm} \\ \frac{4}{3} D_{max} \end{cases} \tag{3}$$

In the methodology, web reinforcements are also assigned with randomization. Also, minimum and maximum reinforcement conditions are also checked. Reinforcement ratio ( $\rho$ ), which is calculated by the ratio of all longitudinal reinforcements to cross-sectional area, must be between 0.01 and 0.06. If the limit conditions are not satisfied, iterative randomization of reinforcements continue.

After all design variable related with axial forces are randomly assigned with a practical value, the distance from extreme compression fiber to neutral axis ( $c$ ) is scanned for axial force capacity. Then, flexural moment capacity of random design is found. If the flexural moment capacity is lower than the required one or more than a defined percentage of the required value, the iterations are repeated. In the present

method used in the optimization of the numerical example, this percentage is taken as 100%. For every 500 iteration, it is iteratively increased with 1%.

After the random design of cross-section and longitudinal reinforcement, the design of shear reinforcements was done. Iteratively, diameter sizes are assigned with the values within the range and the required distance of shear reinforcement (stirrups) are found according to nominal shear strength of concrete ( $V_c$ ) and nominal shear strength of reinforcement ( $V_s$ ) given in Eqs. (4) and (5), respectively.

$$V_c = \frac{\sqrt{f'_c}}{6} b_w d \tag{4}$$

$$V_s = \frac{A_v f_y d}{s} \tag{5}$$

$A_v$  and  $s$  represents shear reinforcement area and distance between them.  $d$  is the effective depth of the concrete. Also, the  $V_s$  value must not exceed  $0.66\sqrt{f'_c} b_w d$ . In that situation, the objective function is penalized with a very big value. Also, the calculated results are compared with the minimum shear reinforcement ( $A_{v,min}$ ) value and maximum shear reinforcement distance ( $s_{max}$ ) defined in Eqs. (6) and (7), respectively. The result of shear reinforcements with the minimum cost is taken into consideration and the results modified according to Eqs. (6) and (7).

$$(A_v)_{min} = \frac{1}{3} \frac{b_w s}{f_y} \tag{6}$$

$$s_{max} \begin{cases} \leq \frac{d}{4} & \text{if } V_s \geq 0.33\sqrt{f'_c} b_w d \\ \leq \frac{d}{2} & \text{if not} \end{cases} \tag{7}$$

After a suitable design is found, the maximum material cost which is the objective function of the optimization is calculated. The objective function which is minimized is given in Eq. (8). The parameters used in Eq. (8) are listed in Table 2.

**Table 2** Parameters of objective function

Definition	Symbol
Material cost of the beam per unit meter	C
Gross area of cross-section	$A_g$
Area of nonprestressed longitudinal reinforcement	$A_{st}$
Area of shear reinforcement spacing $s$	$A_v$
Length of shear reinforcement spacing $s$	$u_{st}$
Material cost of the concrete per $m^3$	$C_c$
Material cost of the steel per ton	$C_s$
Specific gravity of steel	$\gamma_s$

$$\min C = (A_g - A_{st})C_c + (A_{st} + \frac{A_v}{s}u_{st})l\gamma_s C \quad (8)$$

The objective function is calculated by repeating the optimization process for several iteration numbers and design with the minimum cost is found. The flowchart of the optimization methodology is given Fig. 3.

Also, the strength of material such as  $f_y$  and  $f'_c$  may taken as a design variable, but in construction of a structure, using different material types may not be practical. The methodology is applied for different loading condition of axial force and flexural moment in Sect. 4.

## 4 Numerical Example

The optimum design of uniaxial columns was investigated for different flexural moment and axial force values. Design constant, shear force value ( $V$ ) and ranges of design variables used in the numerical examples are given in Table 3.

In the calculations, the compressive stress block to neutral axis depth was assumed as equivalent rectangular. The  $\beta_1$  value, which is a factor relating depth of equivalent rectangular stress block, was calculated as given in Eq. (9).

$$\begin{aligned} \beta_1 &= 0.85 & 17\text{MPa} < f'_c \leq 28\text{MPa} \\ \beta_1 &= 0.85 - 0.0071428(f'_c - 28) & f'_c > 28\text{MPa} \end{aligned} \quad (9)$$

If the value of  $\beta_1$  is lower than 0.65,  $\beta_1$  is taken as 0.65. The elasticity modulus of concrete was calculated by using Eq. (10).

$$E_c = 4700\sqrt{f'_c} \quad (10)$$

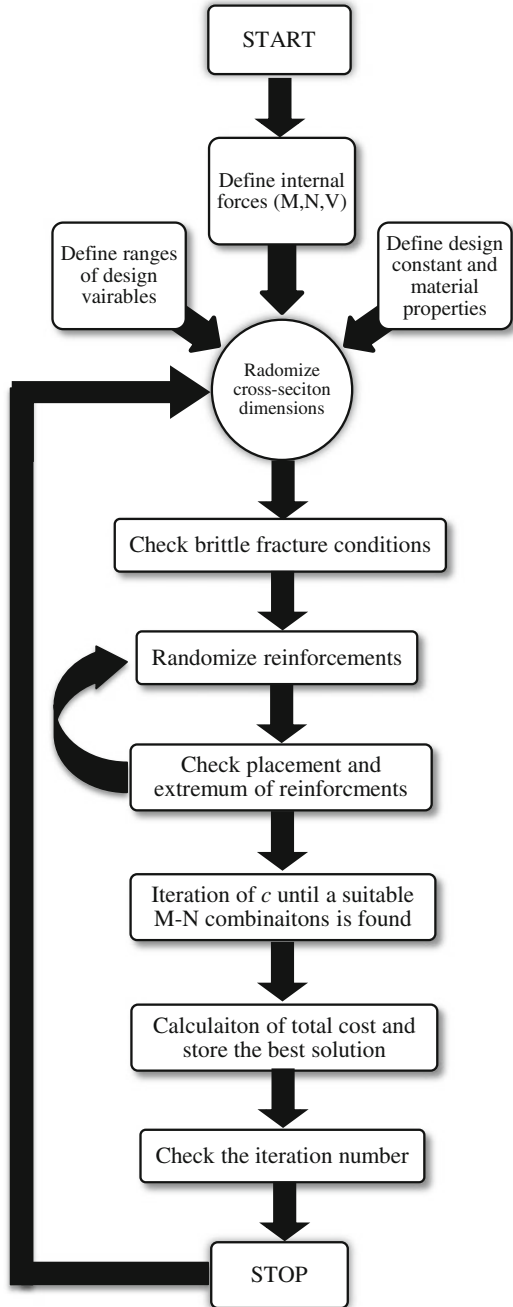
In searching of design variables, the values of  $b_w$  and  $h$  were chosen from (or rounded to) values which are divisible to 50 mm in order to produce a RC structure member practical in construction yard. Because of constant size of steel reinforcements, even integers are assigned for the diameter sizes. The optimum results of several M-N cases (Table 4) are given in Table 5.

The optimum cross-section dimensions of Case 1 such as  $b_w$  and  $h$  are 250 and 300 mm, respectively. In this case, reinforcement positioned in one line is suitable to carry the internal forces. Two reinforcements with the minimum diameter range (16 mm) is found for the optimum results. Also, cross-section dimensions are found as the range minimums. For that reason, the most shear reinforcement is needed for Case 1 since the nominal shear strength of concrete ( $V_c$ ) is low for the design with small cross-section dimensions. The total material cost of the Case 1 is 19.61\$ as seen in Table 5.

In Case 2, the optimum height of the column is 450 mm while  $b_w$  is 250 mm which is also minimum value as found as Case 1. In order to carry more flexural moment,



Fig. 3 Flowchart of the methodology



**Table 3** Design constant, shear force value and ranges of design variables

Description	Value
Length of column (l)	3 m
Clear cover, $c_c$	30 mm
Max. aggregate diameter, $D_{max}$	16 mm
Yield strength of steel, $f_y$	420 MPa
Comp. strength of concrete, $f'_c$	25 MPa
Elasticity modulus of steel, $E_s$	200,000 MPa
Specific gravity of steel, $\gamma_s$	7.86 t/m <sup>3</sup>
Specific gravity of steel, $\gamma_c$	2.5 t/m <sup>3</sup>
Cost of the concrete per m <sup>3</sup>	40\$
Cost of the steel per ton	400\$
Shear force, V	100 kN
Range of web width, $b_w$	250–400 mm
Range of height, h	300–600 mm
Range of reinforcement $\phi$	16–30 mm
Range of shear reinforcement $\phi_v$	8–14 mm

**Table 4** N-M cases for numerical example

Case	N (kN)	M (kNm)
1	500	100
2	1000	200
3	1500	300
4	2000	400
5	2500	500
6	3000	600

h value is increasing. Longitudinal reinforcements are also found as minimum values for allowed range. The total cost of the optimum design is 28.05\$ for Case 2.

In Case 3, increase of height of the column is also seen according to previous cases. Single line design of steel reinforcement is also possible for Case 3 in order to position the required optimum steel reinforcements. But in Case 3, the longitudinal reinforcements are not assigned with minimum range size.

In Case 4, the optimum height of the column is the maximum allowed value. In that case, the reinforcements or the breadth of the column must be increase. The breadth of the column has effective on placing more reinforcements in a line, but using the reinforcements in two lines for a section with 300 mm breadth is the optimum solution. By the increase of the cross-section dimensions, optimum shear reinforcement is getting lower and the distance between stirrups are increasing. The same shear reinforcement is optimum for the last three cases (Case 4–6).

**Table 5** The optimum results of design variable

	Case 1	Case 2	Case3	Case 4	Case 5	Case 6
$b_w$ (mm)	250	250	300	300	400	400
h (mm)	300	450	500	600	600	600
Bars in upper or lower section (line 1)	2 $\Phi$ 16	2 $\Phi$ 16	1 $\Phi$ 22 +1 $\Phi$ 16	1 $\Phi$ 20 +1 $\Phi$ 18	1 $\Phi$ 22 +1 $\Phi$ 20	3 $\Phi$ 20 +2 $\Phi$ 18
Bars in upper or lower section (line 2)	–	–	–	1 $\Phi$ 20	1 $\Phi$ 20	1 $\Phi$ 16
Web reinforcement in a face	–	1 $\Phi$ 16	1 $\Phi$ 16	1 $\Phi$ 18	1 $\Phi$ 16	1 $\Phi$ 16
Shear rein. diameter (mm)	$\Phi$ 8	$\Phi$ 8	$\Phi$ 8	$\Phi$ 8	$\Phi$ 8	$\Phi$ 8
Shear rein. distance (mm)	120	190	220	270	270	270
Optimum cost (\$)	19.61	28.05	36.11	46.19	55.12	67.11

For Cases 5 and 6, the cross-section is assigned with the range maximums. For these cases, the longitudinal moments were positioned in two lines. The cost of Case 5 and 6 are 55.12\$ and 67.11\$, respectively. In Case 6 comparing to Case 5, the longitudinal reinforcements are significantly increasing because of the limit of cross-sectional dimensions.

In the conclusion section, results for additional M-N combinations were given in several graphs in order to discuss the results of the proposed method. The optimum results were searched for flexural moment values between 100 kNm and 700 kNm by 100 kNm differences. In that cases, five different axial force value (500, 1000, 1500, 2000 and 2500 kN) were used.

## 5 Conclusion

The optimum cost of different M-N combinations are plotted in Fig. 4. As seen in the graph, the optimum costs are near to each other for 400 kNm flexural moment. This situation is also observed for flexural moment more than 400 kNm, but not for 500 kN axial force. For the flexural moments below 400 kNm, ACI-318 rules are critical constraints in design. Especially for 2000 and 2500 kN axial force, the optimum costs for flexural moment between 100 and 300 kNm are nearly equal to each other.

In Fig. 5, the optimum total reinforcement ratio of longitudinal reinforcements to cross-sectional area are plotted for M-N combinations. As seen in the Fig. 5, ranges of design variables are more critical than the minimum required reinforcement ratio defined as 0.01 in ACI-318 for 500 kN axial force and 100 kNm flexural moments. In the cases with 500 kN axial force, the optimum cost and total reinforcement ratios are very big compared to other axial force values. Since the compressive forces are low in the section, these forces are not so effective to reduce tensile forces resulting from flexural moments. To carry tensile stresses, steel reinforcement bars are needed. In most flexural moment cases of 2500 kN axial force, minimum reinforcements are optimums while big cross-sections are enough to carry compressive forces.

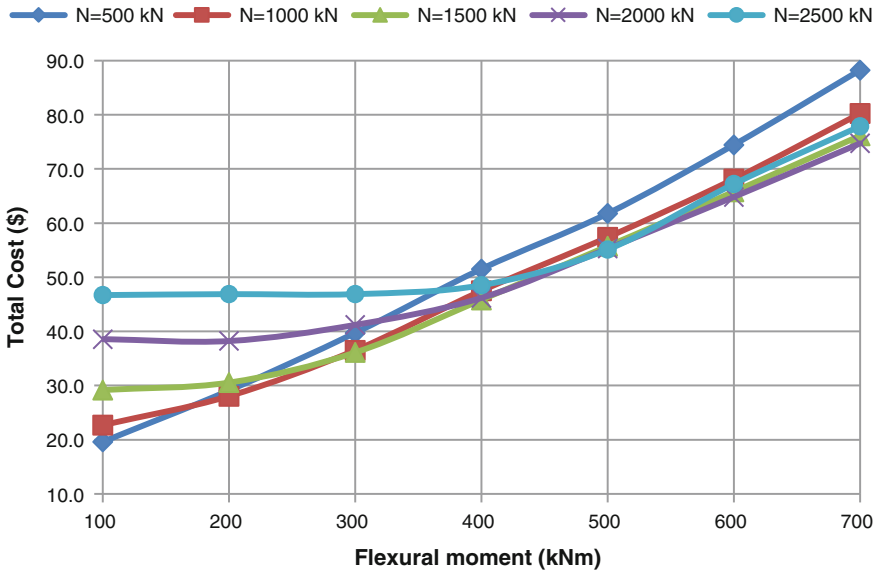


Fig. 4 Total cost values of the M-N cases

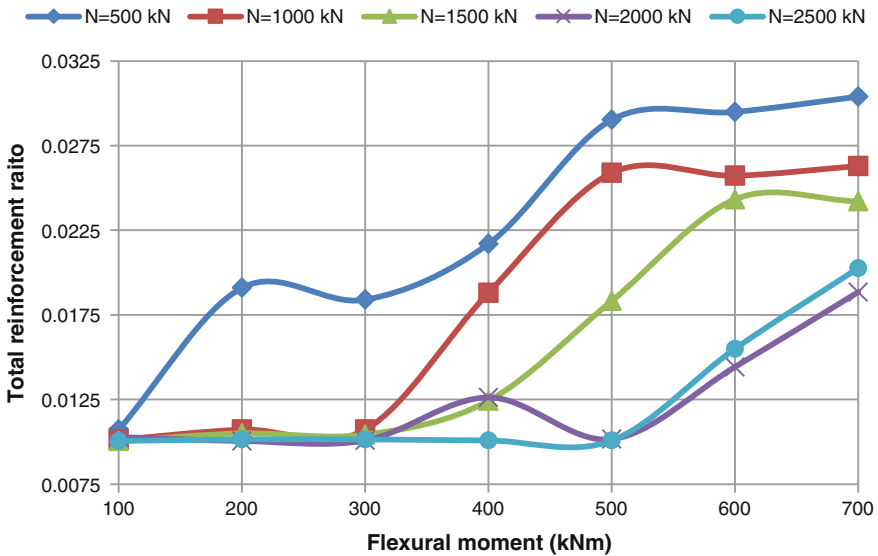


Fig. 5 Total reinforcement ratio values of the M-N cases

By using the presented approach, optimum solutions of RC uniaxial columns can be found for different M-N combinations. As seen in the optimum results, cross-sectional area of column was enlarged in order to carry more internal forces.

This situation is originated from the big cost difference of steel and concrete. The proposed method can assign reinforcements in two lines in order to ensure positioning rules about adherence between steel and concrete. Because of this ability, the optimum results are ready for production in construction yards without modification. In M-N combination with low internal forces, longitudinal reinforcements are positioned in single line. This results shows the effectiveness of the proposed method. As a conclusion, random search technique for the optimization of RC columns is a feasible approach.

## References

1. Coello CC, Hernandez FS, Farrera FA (1997) Optimal design of reinforced concrete beams using genetic algorithms. *Expert Syst Appl* 12:101–108
2. Rafiq MY, Southcombe C (1998) Genetic algorithms in optimal design and detailing of reinforced concrete biaxial columns supported by a declarative approach for capacity checking. *Comput Struct* 69:443–457
3. Koumousis VK, Arsenis SJ (1998) Genetic algorithms in optimal detailed design of reinforced concrete members. *Comput-Aided Civil Inf* 13:43–52
4. Rajeev S, Krishnamoorthy CS (1998) Genetic algorithm-based methodology for design optimization of reinforced concrete frames. *Comput-Aided Civ Inf* 13:63–74
5. Rath DP, Ahlawat AS, Ramaswamy A (1999) Shape optimization of RC flexural members. *J Struct Eng ASCE* 125(12):1439–1446
6. Camp CV, Pezeshk S, Hansson H (2003) Flexural design of reinforced concrete frames using a genetic algorithm. *J Struct Eng-ASCE* 129(1):105–111
7. Ferreira CC, Barros MHFM, Barros AFM (2003) Optimal design of reinforced concrete T-sections in bending. *Eng Struct* 25:951–964
8. Leps M, Sejnoha M (2003) New approach to optimization of reinforced concrete beams. *Comput Struct* 81:1957–1966
9. Lee C, Ahn J (2003) Flexural design of reinforced concrete frames by genetic algorithm. *J Struct Eng-ASCE* 129(6):762–774
10. Balling R, Yao X (1997) Optimization of reinforced concrete frames. *J Struct Eng-ASCE* 123(2):193–202
11. Adeli H, Park HS (1995) Optimization of space structures by neural dynamics. *Neural Netw* 8(5):769–781
12. Adeli H, Park HS (1998) *Neurocomput Des Autom*. CRC Press, Boca Raton, FL
13. Ahmadvkhanlou F, Adeli H (2005) Optimum cost design of reinforced concrete slabs using neural dynamics model. *Eng Appl Artif Intell* 18(1):65–72
14. Barros MHFM, Martins RAF, Barros AFM (2005) Cost optimization of singly and doubly reinforced concrete beams with EC2-2001. *Struct Multidiscip O.*, 30(3):236–242
15. Sirca G Jr, Adeli H (2005) Cost optimization of prestressed concrete bridges. *J Struct Eng, ASCE* 131(3):380–388
16. Govindaraj V, Ramasamy JV (2005) Optimum detailed design of reinforced concrete continuous beams using genetic algorithms. *Comput Struct* 84:34–48
17. Sahab MG, Ashour AF, Toropov VV (2005) Cost optimisation of reinforced concrete flat slab buildings. *Eng Struct* 27:313–322
18. Govindaraj V, Ramasamy JV (2007) Optimum detailed design of reinforced concrete frames using genetic algorithms. *Eng Optim* 39(4):471–494
19. Guerra A, Kioussis PD (2006) Design optimization of reinforced concrete structures. *Comput Concr* 3:313–334

20. Paya I, Yepes V, Gonzalez-Vidosa F, Hospitaler A (2008) Multiobjective optimization of concrete frames by simulated annealing. *Comput-Aided Civ Inf* 23(8):596–610
21. Perea C, Alcalá J, Yepes V, Gonzalez-Vidosa F, Hospitaler A (2008) Design of reinforced concrete bridge frames by heuristic optimization. *Adv Eng Softw* 39:676–688
22. Paya-Zaforteza I, Yepes V, Hospitaler A, Gonzalez-Vidosa F (2009) CO<sub>2</sub>-optimization of reinforced concrete frames by simulated annealing. *Eng Struct* 31:1501–1508
23. Camp CV, Huq F (2013) CO<sub>2</sub> and cost optimization of reinforced concrete frames using a big bang-big crunch algorithm. *Eng Struct* 48:363–372
24. Gil-Martin LM, Hernandez-Montes E, Aschheim M (2010) Optimal reinforcement of RC columns for biaxial bending. *Mater Struct* 43:1245–1256
25. Barros AFM, Barros MHFM, Ferreira CC (2012) Optimal design of rectangular RC sections for ultimate bending strength. *Struct Multidiscip Optim* 45(6):845–860
26. Fedghouche F, Tiliouine B (2012) Minimum cost design of reinforced concrete T-beams at ultimate loads using Eurocode2. *Eng Struct* 42:43–50
27. Ceranic B, Freyer C, Baines RW (2001) An application of simulated annealing to the optimum design reinforced concrete retaining structure. *Comput Struct* 79:1569–1581
28. Yepes V, Alcalá J, Perea C, Gonzalez-Vidosa F (2008) A parametric study of optimum earth-retaining walls by simulated annealing. *Eng Struct* 30:821–830
29. Kaveh A, Abadi ASM (2011) Harmony search based algorithms for the optimum cost design of reinforced concrete cantilever retaining walls. *Int J Civil Eng* 9(1):1–8
30. Camp CV, Akin A (2012) Design of retaining walls using big bang-big crunch optimization. *J Struct Eng-ASCE* 138(3):438–448
31. Talatahari S, Sheikholeslami R, Shadfaran M, Pourbaba M (2012) Optimum design of gravity retaining walls using charged system search algorithm. *Math Probl Eng* 2012:1–10
32. Akin A, Saka MP (2010) Optimum detailed design of reinforced concrete continuous beams using the harmony search algorithm. In: Topping BHV, Adam JM, Pallarés FJ, Bru R, Romero ML (eds) *Proceedings of the tenth international conference on computational structures technology*, Civil-Comp Press, Stirlingshire, UK, Paper 131, doi:[10.4203/ccp.93.131](https://doi.org/10.4203/ccp.93.131)
33. Bekdaş G, Nigdeli SM (2013) Optimization of T-shaped RC flexural members for different compressive strengths of concrete. *Int J Mech* 7:109–119
34. Bekdaş G, Nigdeli SM (2014) Optimization of slender reinforced concrete columns. 85th annual meeting of the international association of applied mathematics and mechanics, Erlangen, Germany, 10–14 Mar
35. Bekdaş G, Nigdeli SM (2014) Optimization of RC frame structures subjected to static loading. In: 11th world congress on computational mechanics, Barcelona, Spain, 20–25 July
36. Kaveh A, Sabzi O (2012) Optimal design of reinforced concrete frames using big bang-big crunch algorithm. *Int J Civil Eng* 10(3):189–200
37. Rao Rama Mohan AR, Shyju PP (2010) A meta-heuristic algorithm for multi-objective optimal design of hybrid laminate composite structures. *Comput-Aided Civil Infrastruct Eng* 25(3):149–170
38. Nigdeli SM, Bekdas G (2013) Optimization of RC beams for various cost ratios of steel/concrete. In: 4th European conference of civil engineering ECCIE'13, Antalya, Turkey, 8–10 Oct
39. Bekdaş G, Nigdeli SM (2014) Optimum design of uniaxial RC columns. In: An international conference on engineering and applied sciences optimization, Kos Island, Greece, 4–6 June
40. ACI 318M–05 (2005) Building code requirements for structural concrete and commentary. American Concrete Institute, Farmington Hills, MI

# The Effect of Stakeholder Interactions on Design Decisions

Garrett Waycaster, Christian Bes, Volodymyr Bilotkach, Christian Gogu, Raphael Haftka and Nam-Ho Kim

**Abstract** The success of an engineering project typically involves multiple stakeholders beyond the designer alone, such as customers, regulators, or design competitors. Each of these stakeholders is a dynamic decision maker, optimizing their decisions in order to maximize their own profits. However, traditional design optimization often does not account for these interactions, or relies on approximations of stakeholder preferences. Utilizing game theory, we propose a framework for understanding the types of interactions that may take place and their effect on the design optimization formulation. These effects can be considered as an economic uncertainty that arises due to limited information about interactions between stakeholders. This framework is demonstrated for a simple example of interactions between an aircraft designer and an airline. It is found that even in the case of very simple interactions, changes in market conditions can have a significant impact on stakeholder behaviors and therefore on the optimal design. This suggests that these interactions should be given consideration during design optimization.

---

G. Waycaster · R. Haftka · N.H. Kim  
University of Florida, Gainesville, FL, USA  
e-mail: gcwaycaster@ufl.edu

R. Haftka  
e-mail: haftka@ufl.edu

N.H. Kim  
e-mail: nkim@ufl.edu

C. Bes · C. Gogu (✉)  
Université Toulouse III, Toulouse, France  
e-mail: christian.bes@univ-tlse3.fr

C. Gogu  
e-mail: christian.gogu@gmail.com

V. Bilotkach  
Newcastle University, Newcastle upon Tyne, UK  
e-mail: volodymyr.bilotkach@newcastle.ac.uk

## 1 Introduction

Many modern engineered systems involve multiple stakeholders, each providing some inputs and receiving some outputs with respect to the system. In the simplest case, this might be a designer who determines system characteristics and a customer who determines how to utilize the system. In more complex systems, we might also have system operators, regulators, or suppliers. We may additionally have multiple stakeholders within each of these groups competing with one another, for example multiple designers each providing similar products to their customers. Each of these stakeholders acts as a dynamic decision maker, acting and reacting based on the decisions made by other stakeholders. These types of interactions can have a dramatic effect on the success or failure of a design.

There are several methods designers currently use to attempt to understand these interactions, mostly by attempting to uncover the preferences of other stakeholders. Most frequently, designers use legacy information based on the types of designs they and their competitors have produced before and the success of those designs. A designer may also use direct communication with other stakeholders, such as via a market study, to attempt to determine the relative importance of different performance metrics. However, these methods are not exact, and the resulting understanding of stakeholder preferences will have some error. This may be due to sampling bias of legacy designs, extrapolation into a new design space, or in cases of direct communication, miscommunication of preferences, either through a stakeholder's ignorance of their own preferences or a deliberate attempt to sway the designers' decisions. We can consider these errors in understanding stakeholder preferences as an economic uncertainty, directly changing a designer's true objective function and therefore affecting the design optimization process.

In order to understand the effects of these stakeholder interactions, we can utilize game theory [1]. Game theory has been developed in economics as a way to model strategic decision making between rational stakeholders, or players. Depending on the way players interact and the information shared between them, we can arrive at different outcomes for the same basic design problem. From the perspective of our optimization problem, game theory allows us to adaptively update our objective function, relating the performance characteristics of our design to designer profits, based on our location in the design space, changes in the market, and actions of other stakeholders. We will introduce this idea in more detail with some simple examples in the next section.

Previous works such as Vincent [2], Rao [3], Badhrinath and Rao [4], and Lewis and Mistree [5] have demonstrated the use of game theory for solving multidisciplinary design problems, but have not addressed the application of game theory to economic uncertainty and interactions. Li and Azarm study the design of a product [6] or product family [7] in the presence of competitive products in the market and uncertain customer preferences, but do not model customers or competitors as dynamic decision makers. Subrahmanyam [8] also considers the idea of market uncertainties as affecting design optimality, but these uncertainties are taken as given



values and are not affected by design decisions. Morrison [9] applies game theory to a case study of fuel efficiency innovation among competing airlines, but does not consider additional stakeholders or applications to design optimization. The present work also draws from the ideas of decision based design [10] and value driven design [11] as tools for explaining design value as a function of performance attributes. The objective of this work is to reformulate a multidisciplinary design optimization problem to account for dynamic interactions between multiple stakeholders and market changes using a game theory model with both simultaneous and sequential interactions considered. We will additionally demonstrate, using an example from the aerospace industry, why considering these interactions during design optimization is important, and how it provides a designer with more information about design trade-offs.

The remaining part of the work is organized as follows. In Sect. 2 we provide our method of reformulating an optimization problem to account for different types of stakeholder interactions. In Sect. 3, we apply this method to a simple example problem of interactions between aircraft designers and regulators. Section 4 summarizes our conclusions, some limitations of the proposed framework, and plans for future work.

## 2 Problem Formulation

For the purpose of this work, we will focus on how we can reformulate an optimization problem when considering the effects of the interactions between  $l$  stakeholders. Readers interested in the principles of game theory can find more information from introductory game theory text books such as Fudenberg and Tirole [1]. First, let us consider a basic multidisciplinary design optimization problem formulation:

$$\begin{aligned} & \text{maximize } \sum_{i=1}^n w_i f_i(\mathbf{X}) \\ & \text{s.t. } g_j(\mathbf{X}) \geq 0 \text{ for } j = 1, \dots, m \end{aligned} \tag{1}$$

where  $\mathbf{X}$  is our vector of design variables,  $f_i$  describes the  $i$ th performance metric of the design,  $w_i$  is the weight of the  $i$ th performance metric in the optimization, and  $g_j$  describes the  $j$ th of  $m$  many design constraints

By varying the vector  $w$  in this optimization, we can calculate a set of Pareto optimal designs for different performance values. Now consider that for each design and set of performance values (that is, each weight vector  $\mathbf{w}$ ) we can define some profit function for our designer,

$$\Pi_1(\mathbf{w}, \mathbf{Y}, \mathbf{E}) \tag{2}$$

where  $\mathbf{Y}$  describes the decision vector of the other stakeholders in the design and  $E$  describes a set of exogenous variables not directly controlled by any stakeholders. This function is used to transform our design performance and other stakeholder decisions directly into the profit for the designer. Note that our designer is labeled as the first stakeholder ( $\mathbf{Y}_1 = \mathbf{w}$ ) and there are  $l - 1$  other stakeholders.

The decision vector  $\mathbf{Y}$  will be determined by the other stakeholders attempting to maximize their own expected profits, such that

$$\mathbf{Y}_k = \mathit{argmax}(\Pi_k(\mathbf{Y}_k, \mathbf{w}, \mathbf{Y}_{\sim k}, \mathbf{E})) \text{ for } k = 2, \dots, l \quad (3)$$

where  $\Pi_k$  describes the profit of the  $k$ th stakeholder,  $\mathbf{Y}_k$  is the decision vector of the  $k$ th out of  $l$  many stakeholders,  $\mathbf{Y}_{\sim k}$  and is the decision vector of the other  $l - 2$  stakeholders.

We now have  $l$  profit functions and  $l$  decision sets. This can be thought of as  $l$  different optimization problems, each dependent on the same decision vector for all players, forming an overdetermined set of equations. In order to determine a solution, we must apply a set of rules; in our case this is based on a certain game structure that describes the amount of information shared between stakeholders and the order in which decisions are made. Information shared between stakeholders refers to how well each stakeholder is able to approximate the profit functions of the others. For example, a designer may not explicitly know the profit function of their customer, but may make an approximation based on prior designs. We will also show that there may arise situations where one stakeholder may have an incentive to deliberately mislead another stakeholder in order to create a more favorable situation for themselves. This type of behavior need not be detrimental for the stakeholder being misled, and can in some cases be advantageous for both parties.

The order of decisions may be either simultaneous, sequential, or partially both. Sequential decision making means one stakeholder chooses their decision vector first and passes that decision on to the next stakeholder in the sequence. Stakeholders moving first will approximate the reaction of each subsequent stakeholder based on their available information about those stakeholders' profit functions. These approximated reactions are known as a best reply function [1]; that is, given that stakeholder one chooses  $Y_1$ , stakeholder 2 will maximize their expected profit by playing  $Y_2$ , or simply

$$\mathbf{Y}_i = \varphi_{ij}(\mathbf{Y}_j, \hat{\mathbf{Y}}) \quad (4)$$

where  $\varphi_{ij}$  is the best reply function that relates the given  $Y_j$  to the best reply  $Y_i$  and  $\hat{\mathbf{Y}}$  is the vector of decisions of all the other stakeholders, some of which may be known based on the sequence of the game, and others which require their own best reply function to determine. Each of these can be solved recursively to determine a best reply function for each subsequent decision maker.

We can therefore formulate our profit maximization problem for the designer by combining Eqs. (1), (2), and (4), where the decisions of stakeholder acting in

sequence before the designer are given as inputs, and the best reply function for stakeholders acting after the designer act as constraints. This problem will be subject to uncertainty in the exogenous inputs,  $\mathbf{E}$ , as well as uncertainty due to approximations made in determining the best reply function,  $\varphi$ .

$$\begin{aligned}
 & \text{maximize } \Pi_1(\mathbf{w}, \mathbf{Y}, \mathbf{E}) \\
 (\mathbf{X}) &= \text{argmax} \sum_{i=1}^n w_i f_i(\mathbf{X}) \\
 & \text{s.t. } g_j(\mathbf{X}) \geq \mathbf{0} \text{ for } j = 1, \dots, m \\
 & \mathbf{Y}_k = \varphi_{k1}(\mathbf{w}, \mathbf{Y}_{\sim k}) \text{ for } k = 2, \dots, l
 \end{aligned} \tag{5}$$

In the case of simultaneous decisions, we must use the concept of a Nash equilibrium [1] to determine a solution. A Nash equilibrium is a point in the decision space where no stakeholder can improve their own profit function by changing their decision vector. This means that a Nash equilibrium acts as a self-enforcing agreement between the players. That is to say,  $(X, Y)$  is a Nash equilibrium if and only if

$$\begin{aligned}
 & \Pi_1(\mathbf{w}, \mathbf{Y}, \mathbf{E}) > \Pi_1(\mathbf{w}^*, \mathbf{Y}, \mathbf{E}) \text{ for all } \mathbf{w}^* \neq \mathbf{w}, \text{ and} \\
 & \Pi_k(\mathbf{Y}_k, \mathbf{w}, \mathbf{Y}_{\sim k}, \mathbf{E}) > \Pi_k(\mathbf{Y}_k^*, \mathbf{w}, \mathbf{Y}_{\sim k}, \mathbf{E}) \text{ for all } \mathbf{Y}_k^* \neq \mathbf{Y}_k, k = 2, \dots, l
 \end{aligned} \tag{6}$$

We can find any pure strategy Nash equilibria by formulating a best reply function for each stakeholder and solving that system of equations to determine where all the best replies intersect. A pure strategy Nash equilibrium means a stakeholder plays a single deterministic decision vector, while a mixed strategy means a stakeholder randomly selects from multiple pure strategies with some predetermined probability of each. It should be noted that there is no guarantee of a single unique Nash equilibrium, and equilibria can exist in both pure and mixed strategies. To solve our problem using simultaneous decision making, we are no longer performing an optimization. Instead, we are looking for the intersection of the surfaces defined by the best reply functions for each of our stakeholders. These intersections represent pure strategy equilibria, of which there may be multiple or none. In cases of multiple Nash equilibria, we can sometimes eliminate some equilibria through so called refinements. For the purposes of this work, we will present all Nash equilibria as possible outcomes, and we will only deal with simultaneous decision making in the discrete decision context for simplicity.

### 3 Example Problem

Having defined how we may formulate an optimization problem considering interactions with other stakeholders, let us consider a simple example. We have two stakeholders, an aircraft designer and builder and their customer the airline. Both are

monopolists, meaning they face no competition. We assume that the designer leases aircraft to the airline at a per flight cost that is fixed, regardless of the aircraft design or the number of flights.

The designer's only decision variable is the level of technology to invest in the aircraft,  $T$ . This can be thought of as the design effort and material and labor cost associated with producing the aircraft. For our problem, we will consider to be bounded between 0 and 1.  $T$  acts as the only weighting variable  $w$  as described in Eq. (1), where a value of 0 is the optimal manufacturing cost, and a value of 1 is the optimal customer value.

The airline's decision variable is the number of flights that they will offer,  $Q$ , which will determine the price they charge per ticket based on a fixed linear demand for air travel. The airline has some fixed cost of operation per flight, some cost that is proportional to the price of jet fuel,  $c_F$ , and some benefit based on the level of technology invested in the aircraft. We can then formulate the profit functions for both stakeholders as follows

$$\Pi_d(T, Q) = Q(L - c_T T) \quad (7)$$

$$\Pi_a(T, Q, c_F) = Q(P(Q)N_p - c_F F - c_L L + v_T T) \quad (8)$$

where  $c_T$  is the cost to implement new technology for the designer,  $F$  is the fuel consumption per flight,  $L$  is the lease cost per flight,  $c_L$  is some factor greater than 1 describing the total fixed costs for the airline including lease cost,  $v_T$  is the value of technology to the airline,  $N_p$  is the number of passengers per flight, and  $P(Q)$  is the price per ticket based on the linear demand function, given by

$$P(Q) = a - bQN_p \quad (9)$$

To create a meaningful example, we first find some reasonable estimates for some of the unknown coefficients in our problem. We select a Boeing 737-700 as the baseline aircraft for our analysis. Considering the standard configuration capacity of 128 passengers [12] and an average load factor of roughly 0.8 [13], we take the number of passengers per flight,  $N_p$ , as 100. Given an average flight length of 1000 miles [13], we calculate the fuel consumption per flight,  $F$ , as roughly 1500 gallons [14]. Average recent jet fuel prices are around \$3.00 per gallon [15], and we consider a range up to \$5.00 to account for possible future changes. Based on the 737-700 list price of \$76M [16] and a useful life of 60,000 flights [17] we find a per flight cost of \$1,300. Considering additional storage and maintenance costs as roughly doubling this expense, we select the per flight lease cost of the aircraft,  $L$ , as \$3000. Based on available airfare cost breakdown data [18], we consider that ranges  $c_L$  from 10 to 12, meaning that the capital cost of the aircraft ranges from 8 to 10% of the total cost per flight, depending on the airline. In order to determine characteristic numbers for the cost and value of new technology, we consider a new aircraft design project. We consider that this new design will cost an additional \$850 per flight, roughly a 25% increase from the initial design, and provides a benefit

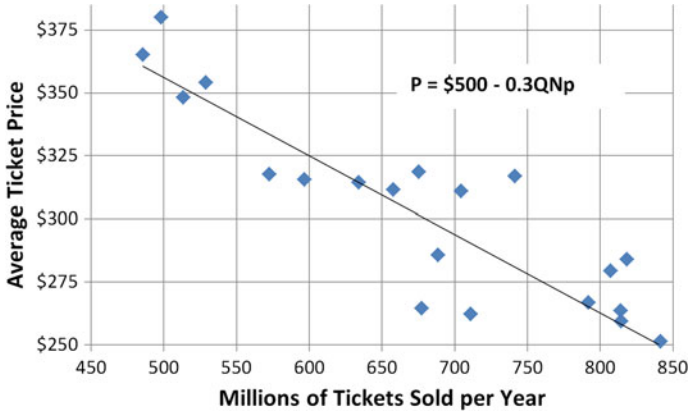


Fig. 1 Historical ticket price versus quantity sold [13, 19]

of \$4200 per flight through increased capacity, efficiency, and passenger comfort. Finally, by collecting data on tickets sold and average ticket price over the past 20 years, we fit the linear relationship between quantity and price as shown in Fig. 1. This approximation assumes that the airline uses this single aircraft design to service all of their routes.

Now let us consider the simplest case of interaction, where the designer first decides on the level of technology investment with full information about the airline profit function, and the airline then determines the quantity of flights in a sequential game. Note that both profit functions, Eqs. (7) and (8), are concave functions. We can therefore calculate a best reply function for the airline by setting to zero the first derivative of the airline profit function with respect to  $Q$  and solving for  $Q$ , such that

$$\frac{d\Pi_a}{dQ} = v_T T - c_L L - c_F F + N_p(a - bN_p Q) - N_p^2 Q b \tag{10}$$

$$Q^* = \varphi_{da}(T) = \frac{aN_p + v_T T - c_F F - c_L L}{2bN_p^2} \tag{11}$$

We can substitute this best reply function into the designer’s profit function to replace and solve for the designer’s optimal value of  $T$  by setting to zero the derivative of the designer’s profit function with respect to  $T$  and solving for  $T$ ,

$$\frac{d\Pi_d}{dT} = \frac{v_T(L - c_T T) + c_T(c_F F - c_L L - aN_p - v_T T)}{2N_p^2 b} \tag{12}$$

$$T^* = \frac{v_T L + c_F c_T F + c_L c_T L - a c_T N_p}{2c_T v_T} \tag{13}$$

Using our values for our various coefficients, we can calculate the decision of the designer and airline and the profit for each. Since we have ranges of values for

**Table 1** Solution values for sequential game with no uncertainty

$c_F$	$c_L$	$Q^*$	$T^*$	$\Pi_d$	$\Pi_a$
\$3.00	10	2.58M	0 (-0.08)	\$7.75B	\$20.0B
\$3.00	12	2.02M	0.63	\$4.99B	\$12.32B
\$5.00	10	2.28M	0.27	\$6.30B	\$15.55B
\$5.00	12	1.78M	0.99	\$3.83B	\$9.57B

fuel price and the airline cost factor, we perform this analysis at the 4 extreme cases of these coefficients as shown in Table 1. Because our problem is linear in these values, we can interpolate between these 4 points to find the decisions and profits at any combination. Note that in the first case, the designer would choose an optimal value of slightly negative technology investment, however we restrict this value to be between 0 and 1. It can be seen that the optimal decisions and resulting profits for both the designer and airline vary greatly with these possible changes in parameters  $c_F$  and  $c_L$ .

In a realistic design problem, we will likely consider that a designer must make design decisions without knowledge of future fuel prices. These prices will be unknown to the airline as well. A designer will then maximize expected profits based on the possible distribution of future fuel prices. Due to the simple linear nature of our example problem, this will be the same as designing based on the mean value of future fuel prices.

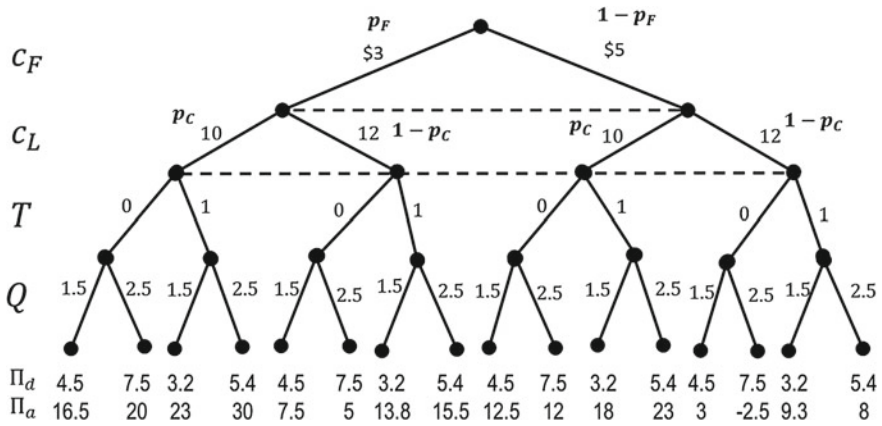
A designer may face additional uncertainty in their understanding of the airlines’ profit function, for example in the value of  $c_L$ . However, the airline will be able to know this value exactly. This is known in game theory as a game of “incomplete information” [1]. This means the designer will face some error in their prediction of the best reply function of the designer, specifically

$$Q^* = \varphi_{da}^-(T) = \frac{aN_p + v_T T - c_F F - (c_L + \varepsilon) L}{2bN_p^2} \tag{14}$$

where  $\varepsilon$  describes the error in the designers understanding of airline costs.

We can see from our previous example that the designer will invest more in technology if they believe the airlines fixed cost,  $c_L$ , is higher. This is because higher fixed costs mean the effect of technology on airline marginal profits is more significant, and therefore more technology investment will have a greater effect on the quantity of flights. This relationship implies that airlines will have an incentive to mislead designers into believing that their costs are higher than in reality, shifting profits away from designers and toward airlines. Without considering the effects of these interactions, designers will be unable to understand the effects of these potential uncertainties.

To explore these interactions in more detail, let us switch from a continuous game to a discrete one. In this case, the designer must either decide to invest in new technology ( $T = 1$ ) or not ( $T = 0$ ). The airline will decide whether to expand their market by offering a higher number of flights ( $Q = 2.5M$ ), or to maintain their



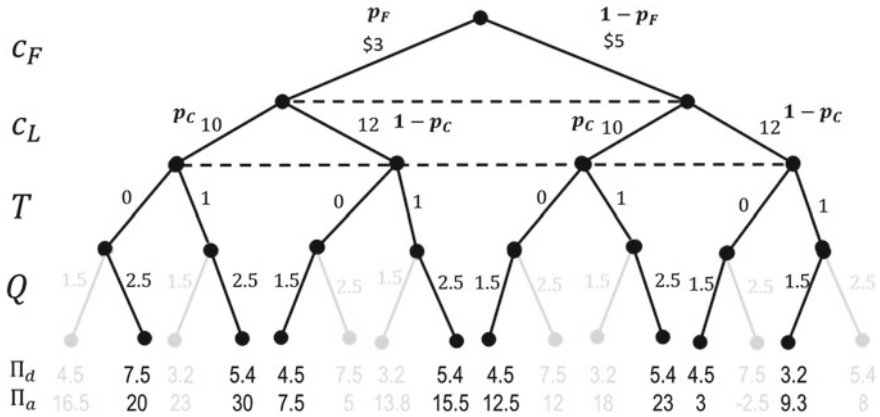
**Fig. 2** Extensive form game with uncertainty in fuel prices  $p_F$  and in fixed cost  $p_C$ , where designers choose technology  $T$  and airlines choose quantity of flights  $Q$  with payoffs for the designer and the airline, respectively

current levels ( $Q = 1.5M$ ). We consider that fuel prices will either be \$3 per gallon with probability  $p_F$  or \$5 per gallon with probability  $1 - p_F$ . Finally, the designer assumes the airline is a low cost carrier ( $c_L = 10$ ) with probability  $p_C$  or a high cost carrier ( $c_L = 12$ ) with probability  $1 - p_C$ . We can express this problem using a decision tree (cf. Fig. 2), known in game theory as an extensive form game [1].

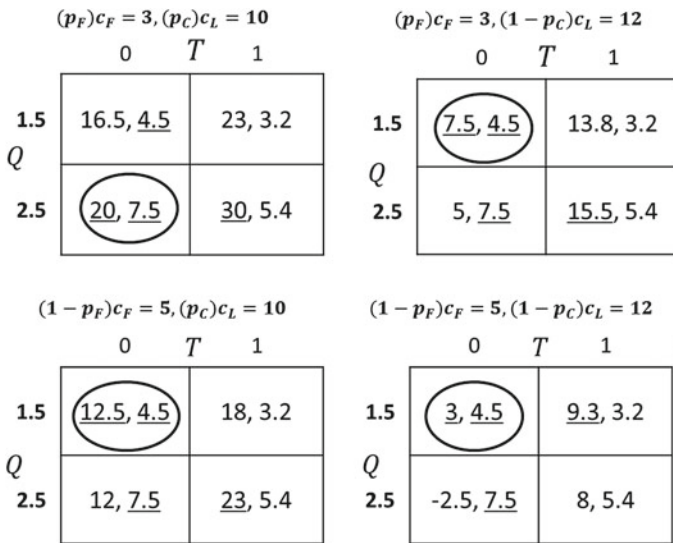
In Fig. 2, each node represents a decision, and dashed lines between nodes indicate an information set, where the decision maker must act without knowing for certain which node in the information set they are currently in. The solution will therefore depend on the decision maker’s beliefs about the values of  $p_F$  and  $p_C$ . The payoffs for each resulting set of decisions are given at the end of each path, where the top number is the designer’s profit, and the bottom number is the airline’s profit, both in billions of dollars. We can simplify this game by eliminating dominated strategies for the airline, since we know at the last branch of the decision tree the airline will choose the value that maximizes their own profits; this is known as backwards induction. Fig. 3 shows these dominated strategies in gray.

We see that, based on this discrete example, the designer can only influence the airline to utilize more flights by increasing technology investment if fuel prices are low and airline costs are high, or fuel prices are high and costs are low. In the remaining two cases, the designer will strictly prefer not to invest in new technology, since they will lease the same number of flights regardless and will have a higher profit margin for each. Airlines will always prefer the case where designers invest in technology, as they always gain higher profits.

From this simple example, we would conclude that if fuel prices are high, airlines will attempt to convince designers that they have low costs, as designers will believe they can then influence flight quantity by investing in technology. If fuel prices are low, airlines will attempt to convince designers that their costs are high, again in an effort to encourage designers to invest in technology.



**Fig. 3** Backwards induction indicating strictly dominated choices (*gray*) for the airline when choosing quantity  $Q$



**Fig. 4** Simultaneous game solution

We may also be interested to know if the possible solutions of this game change if we consider that designers and airline make decision simultaneously. For example, airlines submit orders for new aircraft without knowing future fuel prices or precise aircraft specifications. We can represent this sort of game using strategic form, with 4 payoff matrices representing the 4 possible combinations of fuel price and airline costs as shown in Fig. 4.

The numbers in each box represent the payoffs for the airline and the designer, respectively. Numbers that are underlined indicate a best reply for that stakeholder.



When both numbers are underlined in the same box, meaning the best replies intersect, we have a Nash equilibrium for that individual game, represented by circling that square. We can see that for the simple game we have constructed, it is never advantageous for the designer to invest in technology. This happens because since decisions are made at the same time, the designer's choice cannot influence the quantity selected by the airline. We can also see that when airline costs are high ( $c_L = 12$ ), meaning we are on the two matrices on the right side, the equilibrium solution for this game will be  $(T = 0)$ ,  $(Q = 1.5M)$ . When airline costs are low, the equilibrium will depend on the probability of low fuel prices,  $p_F$ , as the airline will attempt to maximize their expected profits. If the airline believes  $p_F$  is less than 0.11, they will always choose the low quantity ( $Q = 1.5M$ ), and if they believe  $p_F$  is greater than 0.11 the airline will choose the high quantity, ( $Q = 2.5M$ ). When  $p_F$  is equal to 0.11, the airline is indifferent between these two strategies and may play either one, or play a mixed strategy where they randomly select between both options. It should be noted that the designer would strictly prefer the airline select the higher quantity, but based on this game structure, they have no way to influence that decision.

It should be noted that the solutions we have found for each of these different types of games need not be Pareto optimal in terms of profits for both stakeholders. For example, in Fig. 4, we can see that both the designer and a high cost airline ( $c_L = 12$ ) would be strictly better off playing the strategy  $(T = 1)$ ,  $(Q = 2.5M)$  as compared to the equilibrium strategy  $(T = 0)$ ,  $(Q = 1.5M)$ , regardless of the values of fuel price and airline costs. However, that strategy is not an equilibrium because one or both of the stakeholders can improve their profits by modifying their decision. For example, in the case of  $[c_F = 5, c_L = 12]$  starting at  $(T = 1)$ ,  $(Q = 2.5M)$ , we see that the designer would strictly prefer to select  $(T = 0)$  when the airline plays  $(Q = 2.5M)$ , and similarly the airline prefers  $(Q = 1.5M)$  against  $(T = 1)$ . Because the strategies and payoffs are known, each player will realize the other will try to change their own strategy, and will respond accordingly, resulting in selecting  $(T = 0)$ ,  $(Q = 1.5M)$ . This is a variation on the classical game theory example known as the prisoner's dilemma [1].

## 4 Conclusions

We have presented a framework for how game theory can be utilized in design optimization to better model and understand interactions between multiple stakeholders. We demonstrated how, based on the order in which interactions take place and the information shared between stakeholders, the optimal decision for the designer can change significantly. By incorporating these interactions into the design problem, we can directly anticipate these changes and can quantify the uncertainty in the profit expected for our final design based on approximations of other stakeholders. Additionally, this framework is able to directly provide information for the designer regarding trade-offs between multiple disciplines during design, since we are able to

adaptively update the designer's objective function based on changes in stakeholder preferences due to changes in performance.

Using our simple example problem, we demonstrate that for the sequential game between the designer and airline, small changes in the value of certain profit function coefficients can have a large effect on optimal design choices and profits for both stakeholders. We observe that for the values we have selected in our sequential game, the airline may have an incentive to obscure their true costs from designers in order to encourage investment in new technology. Looking at the same problem but using a simultaneous structure, the designer will never elect to invest in technology, based on the cases considered. From these two examples we have shown that understanding the structure of the game can greatly change the outcome, and that, within that structure, approximations by one stakeholder in the preferences of another can have a large impact on design decisions and profits.

We do note that, depending on the game structure utilized, a stakeholder may need to approximate the decisions of the designer in their profit maximization, requiring them to solve the design optimization problem within their own profit optimization. For expensive design problems, this creates computational limitations, and future work is needed to address this issue. It can also be difficult in a practical problem to quantify the type of interactions between multiple stakeholders. The authors have previously proposed a method to understand these interactions by using causal models [20]. Future work in this area will focus on applying the methods described to a realistic design problem and understanding the relative importance of uncertainty in stakeholder preferences as compared to traditional design uncertainties like variations in material properties and operating conditions.

## References

1. Fudenberg D, Tirole J (1991) *Game theory*. MIT Press, Cambridge
2. Vincent TL (1983) Game theory as a design tool. *J Mech Trans Autom Des* 105(2):165–170
3. Rao SS (1987) Game theory approach for multiobjective structural optimization. *Comput Struct* 25(1):119–127
4. Badhrinath K, Jagannatha Rao JR (1996) Modeling for concurrent design using game theory formulations. *Concurr Eng* 4(4):389–399
5. Lewis K, Mistree F (2001) Modeling subsystem interactions: a game theoretic approach. *J Des Manuf Autom* 4(1):17–35
6. Li H, Azarm S (2000) Product design selection under uncertainty and with competitive advantage. *J Mech Des* 122(4):411–418
7. Li H, Azarm S (2002) An approach for product line design selection under uncertainty and competition. *J Mech Des* 124(3):385–392
8. Subrahmanyam S, Pekny JF, Reklaitis GV (1994) Design of batch chemical plants under market uncertainty. *Indus Eng Chem Res* 33(11):2688–2701
9. Morrison D, James K, Hansman RJ, Sgouridis S (2012) Game theory analysis of the impact of single-aisle aircraft competition on emissions. *J Aircr* 49(2):483–494
10. Hazelrigg GA (1998) A framework for decision-based engineering design. *J Mech Des* 120(4):653–658
11. Collopy PD, Hollingsworth PM (2011) Value-driven design. *J Aircr* 48(3):749–759

12. Boeing (2014) Boeing 737–700 technical characteristics. [http://www.boeing.com/boeing/commercial/737family/pf/pf\\_700tech.page](http://www.boeing.com/boeing/commercial/737family/pf/pf_700tech.page). Accessed 7 April 2014
13. Bureau of Transportation Statistics (2014) TranStats database: air carrier summary: T2: U.S. air carrier TRAFFIC and capacity statistics by aircraft type. [http://www.transtats.bts.gov/Fields.asp?Table\\_ID=254](http://www.transtats.bts.gov/Fields.asp?Table_ID=254). Accessed 7 April 2014
14. Puget Sound Business Journal (2014) Mindful of rivals, Boeing keeps tinkering with its 737. <http://www.bizjournals.com/seattle/stories/2008/08/11/story13.html?page=all>. Accessed 7 April 2014
15. Airlines for America (2014) Annual crude oil and jet fuel prices. <http://www.airlines.org/Pages/Annual-Crude-Oil-and-Jet-Fuel-Prices.aspx>. Accessed 7 April 2014
16. Boeing (2014) Boeing: jet prices. <http://www.boeing.com/boeing/commercial/prices/>. Accessed 7 April 2014
17. Jiang H (2013) Key findings on airline economic life. Boeing Technical Report
18. McCartney S (2014) How airlines spend your airfare. Wall Street J. <http://online.wsj.com/news/articles/SB10001424052702303296604577450581396602106>. Accessed 7 April 2014
19. Airlines for America (2014) Average round-trip fares and fees. <http://www.airlines.org/Pages/Annual-Round-TripFares-and-Fees-Domestic.aspx>. Accessed 7 April 2014
20. Waycaster G, Haftka RT, Kim NH, Bilotkach V, Gogu C, Bes C (2014) Relationship Between fleet characteristics and welfare of stakeholders in commercial air transportation: manufacturers, airlines, and the public. In: Proceedings of the 52<sup>nd</sup> aerospace sciences meeting, Washington, DC, 13–17 Jan 2014

# A Fixed Point Approach to Bi-level Multi-objective Problems

Carla Antoni and Franco Giannessi

**Abstract** The present note aims at introducing a new approach for handling bi-level multi-objective problems. The advantage consists in the fact that, for solving the upper level, it does not require to know explicitly the lower level. Here, the linear case is fully treated. Hints are given on how to extend it to the important class of the cono-functions, which contains that of the convex functions, and is one of the few extensions of convex functions which are numerically viable. The final section gives suggestions for further research in the field.

**Keywords** Bi-level vector optimization · Multi-objective optimization · Scalarization · Cone-functions

**AMS Mathematics Subject Classification (2000):** 65K · 90C

## 1 Introduction

Many real world problems can be formulated mathematically as extremum problems, where there are several objective functions. Rarely, such functions achieve the extremum at a same point. This has led, in the last decades, to a rapid mathematical development of this field, whose origin goes back to more than one century ago. Almost independently of this, in some fields of engineering dealing with the design, the need has gradually emerged of taking into account the competition of some variables, which were previously condensed in just one. Roughly speaking, the researches, carried on in the mathematical optimization area, can be split into those which aim at detecting properties of the set of solutions, and those which aim

---

C. Antoni (✉)  
Naval Academy, 72 Viale Italia, Livorno, Italy  
e-mail: carla.antoni5@gmail.com

F. Giannessi  
Department of Mathematics, University of Pisa, Pisa, Italy  
e-mail: gianness@dm.unipi.it

at providing us with methods for finding such a set by using, in general, scalarizing techniques; the former are extremely important as a base for any other research; the latter should take into consideration the fact that, in most of the applications, the designer has a bi-level problem: an extremum problem (upper level), whose feasible region is the set of the extremum points of a multi-objective constrained extremum problem (lower level); consequently, the methods of solution of the bi-level problem should require to run on such a feasible set, namely the set of vector extremum points, as less as possible (unlike what some existing methods try to do). Here, based on previous results [12], a method for solving the bi-level problem is described. Our main scope consists in outlining the method and let it be easily understood to a wide audience more than deliver a detailed, rigorous exposition of the method; this will be done in a forthcoming paper [2]. Consequently, to make the text plain, we take some assumptions, which are somewhat strong, and which can be easily weakened; moreover, again for the sake of simplicity, we take for granted the existence of the extrema we meet. In Sect. 2, after having proved some properties of such a class, we define a new type of scalarization for a multiobjective problem (lower level), and then we outline an approach to the bi-level problem. While this will be the subject of a forthcoming paper, in this note (Sect. 4) we will develop the case where the multiobjective problem (lower level) is linear and the upper level is convex, and we consider strong solutions with Pareto-cone. It will be shown that, in this case, we improve the existing literature, in as much as the lower level requires to handle a linear problem (while the existing literature is faced with a nonlinear one and, in general, for weak solutions). Section 5 contains some numerical examples. In the final section, we discuss shortly some further developments.

In Appendix, we consider a class of nonconvex functions, which enjoy the nice property to have convex level sets, and for which constructive sufficient condition can be established in order to state whether or not a given function belongs to such a class.

Many real world problems lead to the minimization (or maximization) of a scalar function over the set of minimum points of a multi-objective problem. Hence, we are faced with a bi-level problem.

Let  $l, m$  and  $n$  be positive integer,  $X \subseteq \mathbb{R}^n, C \subseteq \mathbb{R}^l$  be a convex, closed and pointed cone with apex at the origin; the functions  $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}, f : \mathbb{R}^n \rightarrow \mathbb{R}^l, g : \mathbb{R}^n \rightarrow \mathbb{R}^m$  are given. In the sequel,  $int S$  and  $ri S$  will denote the topological interior and relative interior of the set  $S$ , respectively.

Consider the problem (lower level):

$$\min_{C_0} f(x) \text{ s.t. } x \in K := \{x \in \mathbb{R}^n : g(x) \geq 0\}, \tag{1}$$

where  $\min_{C_0}$  marks vector minimum with respect to the cone  $C_0 := C \setminus \{0\}$ ;  $y$  is a (global) multi-objective minimum point (in short, MMP) of (1) if and only if

$$f(y) \not\prec_{C_0} f(x), \quad \forall x \in K, \tag{2}$$

where the inequality in (2) means  $f(y) - f(x) \notin C_0$ . At  $C = \mathbb{R}_+^l$ , (1) becomes the classic Pareto problem.

Finally, consider the problem (upper level):

$$\min \Phi(x) \text{ s.t. } x \in K^0, \tag{3}$$

where  $K^0$  is the set of VMPs of (1).

## 2 Problems in Engineering

In the design of an aircraft, there are many quantities to be taken into account [17]. The designers must define the performance of an aircraft. In other words, they must carefully specify the optimization problem and the trade-off between all these quantities; some can be considered as objectives, some as constraints; some of them can be seen as a set of concurrent objectives, are in conflict each other. Hence, the designers are faced with a very hard multi-objective optimization. In a much simplified version, the objectives are minimum induced drag problem of a wing system, maximum lifting and minimum cost. The designer could, of course, consider these 3 objectives as concurrent without any other distinction, and formulate a 3-objective optimization problem. However, it is evident that this would be a coarse approach. In fact, the comparison of the cost with, e.g., the lift would lead to compare heterogeneous magnitudes. The “natural” comparison is between induced drag and lift. Then, among all the aircrafts, for which none “dominates” the others, in the sense that an improvement of lift (drag) implies a worsening of drag (lift), it is meaningful to search for one, which minimizes the cost. Hence the bi-level approach is that which makes sense.

In the general case, when designing a machine one has to take into account design variable, functional, and criterion constraints. The design variable constraints have the form

$$\underline{x}_j \leq x_j \leq \overline{x}_j, \quad j = 1, \dots, n. \tag{4}$$

In the case of mechanical systems the  $x_j$  represent the stiffness coefficients, the moments of inertia, masses, damping factors, geometric dimensions, etc. The functional constraints may be written as follows

$$\underline{C}_i \leq g_i(x) \leq \overline{C}_i, \quad i = 1, \dots, m, \tag{5}$$

where the functional dependences (relationships)  $g_i(x)$  may be either functional depending on the integral curves of the differential equations or explicit functions of  $x$ ;  $\underline{C}_i$  and  $\overline{C}_i$  are the lower and the upper admissible values of the quantity  $g_i(x)$ . The functional constraints can specify the range of the allowable stresses in structural elements, the track gauge, etc. Also, there exists particular performance criteria such

as productivity, material consumption, and efficiency. It is desired that, with other things being equal, these criteria, denoted by  $f_j$ ,  $j = 1, \dots, l$  would have extremal values. For simplicity we assume that  $f_i$  are to be minimized. In order to avoid situations in which the designer regards the values of some criteria as unacceptable, we introduce criterion constraints

$$f_i(x) \leq \bar{f}_i, \quad i = 1, \dots, l, \tag{6}$$

where  $\bar{f}_i$  is the worst value of criterion  $f_i(x)$  to which the designer may agree. Criterion constraint  $\bar{f}_i$  cannot be chosen before solving the problem. Constraints (4)–(6) define the feasible solution set  $K$ :

$$K = \{x \in \mathbb{R}^n : \underline{x} \leq x \leq \bar{x}, \bar{g} \leq g \leq \bar{g}, f \leq \bar{f}\}. \tag{7}$$

If functions  $g$  and  $f$  are continuous, then the set  $K$  is closed. One of the basic problems of multicriteria optimization is the following: find a set  $K^0 \subseteq K$  for which  $y \in K^0$  is a solution of

$$\min_C f(x). \tag{8}$$

$K^0$  is the Pareto optimal set. It plays an important role in vector optimization problems because it can be analyzed more easily than the feasible solution set and because the optimal vector always belongs to the Pareto optimal set irrespective of the system of preferences used by the designer for comparing vectors belonging to the feasible solution set. The importance of this set is determined to a great extent by the following well-known theorem.

**Theorem 1** *If the feasible set  $K$  is closed and  $f_i$ ,  $i = 1, \dots, l$  are continuous, then the Pareto optimal set is nonempty.*

### 3 Scalarization of the Lower Level

Now, let us consider the scalarization of (1) by exploiting the method, which was introduced in [11]; see also Sect. 6 of [12].

For each  $y \in X$ , and  $p \in C^*$ , consider the sets:

$$S(y) := \{x \in X : f(x) \in f(y) - C\}$$

$$S_p(y) := \{x \in X : \langle p, f(x) \rangle \leq \langle p, f(y) \rangle\}$$

$S(y)$  is evidently a level set of  $f$  with respect to  $C$ . Indeed, when  $X = \mathbb{R}^n$  and  $C = \mathbb{R}_+^l$ , then it is precisely the lower set of  $f$ ; in the affine case,  $S(y)$  is a cone with apex at  $y$ , and  $S_p(y)$  becomes a supporting half-space at its apex.

Note that, apart from  $S(y)$ ,  $S_p(y)$  and Proposition 3.1 (where  $p$  is a parameter),  $p$  will be considered fixed, in particular in the case of the algorithm. Instead,  $y$  will now play the role of a parameter and, later, that of the unknowns; this will be reported. The concept of level set, in strict or extended sense, plays a fundamental role in the present scalarization; in order to describe it, we need to establish some properties.

**Proposition 3.1**

(i) If  $f$  is a convex function on  $X$ , then,  $\forall y \in X$ ,  $S(y)$  is convex.

(ii) If  $p \in C^*$ , then,  $\forall y \in X$ ,

$$S(y) \subseteq S_p(y), \quad y \in S(y) \cap S_p(y).$$

*Proof* (i) For  $x^i \in S(y)$  there exist  $c^i$  such that  $f(x^i) = f(y) - c^i$ ,  $i = 1, 2$ . Then,  $\forall \alpha \in [0, 1]$ ,

$$(1 - \alpha)f(x^1) + \alpha f(x^2) = f(y) - ((1 - \alpha)c^1 + \alpha c^2).$$

Since  $C$  is convex  $(1 - \alpha)c^1 + \alpha c^2 \in C$ . Moreover, if  $f$  is a  $C$  function, then  $\forall \alpha \in [0, 1]$  there exists  $c'(\alpha)$  such that

$$f(x(\alpha)) = (1 - \alpha)f(x^1) + \alpha f(x^2) - c'(\alpha).$$

It follows that

$$f(x(\alpha)) = f(y) - (1 - \alpha)c^1 - \alpha c^2 - c'(\alpha)$$

and then  $x(\alpha) \in S(y)$ ,  $\forall \alpha \in [0, 1]$ ,  $\forall y \in X$ .

(ii)  $0 \in C \iff y \in S(y)$ ;  $y \in S_p(y)$  is trivial. The thesis follows. □

Now, let  $p \in C^*$ ; as announced, *unlike what in general happens in the field of scalarization,  $p$  will remain fixed* in the rest of this section. Let us introduce the following problem (in the unknown  $x$ , depending on the parameter  $y$ ):

$$\min \langle p, f(x) \rangle, \quad x \in K(y) := K \cap S(y). \tag{9}$$

Borrowing the terminology of Variational Inequalities, we call (9) *Quasi-Minimum Problem*. Its feasible region depends (parametrically) on  $y$ ; we will see that, for our scalarization method, it will be important to consider the case  $y = x$ , where the feasible region depends on the unknown; i.e. the feasible points are “fixed points” of the point-to-set map  $y \mapsto K(y)$ .



*Remark 3.1* Under suitable assumptions the first order necessary condition of (9) is:

$$\langle p^T \nabla f(x), y - x \rangle \geq 0, \quad x \in K(y), \tag{10}$$

which is a particular case of a Quasi-Variational Inequality.

In general, problem (10) looks difficult. The following proposition identifies a class of (9), which can be handled easily.

**Proposition 3.2** *Let  $X$  be convex,  $f$  be a convex,  $g$  be concave on  $X$  and  $p \in C^*$ . Then (9) is convex.*

*Proof* We have to show that the restriction of  $\langle p, f(\cdot) \rangle$  to  $X$  and  $K(y)$  are convex. Since  $p \in C^*$  and  $f$  is a convex,  $\forall x^1, x^2 \in X$  it holds

$$\langle p, (1 - \alpha)f(x^1) + \alpha f(x^2) - f(x(\alpha)) \rangle \geq 0, \quad \forall \alpha \in [0, 1],$$

and, equivalently,

$$(1 - \alpha)\langle p, f(x^1) \rangle + \alpha\langle p, f(x^2) \rangle - \langle p, f(x(\alpha)) \rangle \geq 0, \quad \forall \alpha \in [0, 1],$$

that is the convexity of  $x \mapsto \langle p, f(x) \rangle$ . The convexity of  $X$  and the concavity of  $g$  implies the convexity of  $K$ ; then, the convexity of  $S(y)$  (Proposition 3.1), implies that of  $K(y)$ . □

As announced, we want to run on the set of VMPs of (1) in such a way to make the resolution of (1) as easy as possible. In other words, by exploiting the properties of (9), it will be possible to define a method which solves (1) without having obliged to find in advance  $K^0$ .

**Proposition 3.3**  $y \in X, x \in S(y) \implies S(x) \subseteq S(y)$ .

*Proof*  $x \in S(y)$  if and only if there is  $c \in C$  such that  $f(x) = f(y) - c$ , and  $\hat{x} \in S(x)$  if and only if there is  $\hat{c} \in C$  such that  $f(\hat{x}) = f(x) - \hat{c}$ . It follows that  $f(\hat{x}) = f(y) - (c + \hat{c})$ , that is  $\hat{x}$  belongs to  $S(y)$ . Then, the inclusion  $S(x) \subseteq S(y)$  is proved. □

The above proposition shows that, if the “apex” of the level set  $S(y)$  is shifted to a point belonging to it, then the translated level set is contained in it. This property will allow us to find a VMP of (1).

**Proposition 3.4** *If  $x^0$  is a (global) minimum point of (9) at  $y = y^0$ , then  $x^0$  is a (global) minimum point of (9) at  $y = x^0$ .*

*Proof* Proposition 3.1 guarantees that  $x^0 \in S(x^0)$ . Ab absurdo, suppose that  $x^0$  be not a (global) minimum point of (9) at  $y = x^0$ . Then,

$$\exists \hat{x} \in K \cap S(x^0) : \langle p, f(\hat{x}) \rangle < \langle p, f(x^0) \rangle. \tag{11}$$

Proposition 3.3 implies  $\hat{x} \in S(y_0)$ , and the conditions

$$\hat{x} \in K \cap S(y^0) : \langle p, f(\hat{x}) \rangle < \langle p, f(x^0) \rangle. \tag{12}$$

contradict the assumptions. Necessarily  $x^0$  is a global minimum point of (9) at  $y = x^0$ .  $\square$

*Remark 3.2* Taking into account Propositions 3.3 and 3.4, problem (9) can be formulated as:

$$\text{find } x^0 \in K \text{ s.t. } \min_{x \in K(x^0)} \langle p, f(x) \rangle = \langle p, f(x^0) \rangle, \tag{13}$$

which justifies, once more, the terminology Quasi-minimum Problem.

The following proposition connects the optimality of (1), its image, and the optimality of (13). It is trivial to note that (2) is satisfied if and only if the system (in the unknown  $x$ ):

$$f(y) - f(x) \geq_{C_0} 0, \quad g(x) \geq 0, \quad x \in X \tag{14}$$

is impossible.

**Proposition 3.5** *Let  $p \in \text{int } C^*$ .*

(i)  *$y$  is a VMP of (1) if and only if the system (in the unknown  $x$ ):*

$$\langle p, f(y) - f(x) \rangle > 0, \quad f(y) - f(x) \in C, \quad g(x) \geq 0, \quad x \in X, \tag{15}$$

*is impossible.*

(ii) *The impossibility of (15) is a necessary and sufficient condition for  $y$  to be a (scalar) minimum point of (9) or (13).*

*Proof* (i) If  $\hat{x}$  satisfies (15), then  $f(y) - f(\hat{x}) \neq 0$ ; consequently  $\hat{x}$  satisfies (14). Viceversa, suppose there is  $\hat{x} \in X$  such that  $f(y) - f(\hat{x}) \geq_{C_0} 0, g(\hat{x}) \geq 0$ . This implies  $\langle p, f(y) - f(\hat{x}) \rangle > 0$ . In fact, since  $p \in \text{int } C^*$ , there is  $r > 0$  such that  $p + N_r \subseteq C^*$ , where  $N_r = \{\delta \in \mathbb{R}^l : \|\delta\| < r\}$ , then

$$\langle p + \delta, f(y) - f(\hat{x}) \rangle \geq 0, \quad \forall \delta \in N_r. \tag{16}$$

Ab absurdo, suppose

$$\langle p, f(y) - f(\hat{x}) \rangle = 0. \tag{17}$$

Since there exists  $\varepsilon > 0$  such that  $\varepsilon \| f(y) - f(\hat{x}) \| < r$ , and  $f(y) - f(\hat{x}) \neq 0$ , then, from (16) and (17) it follows that

$$0 \leq \langle p - \varepsilon(f(y) - f(\hat{x})), f(y) - f(\hat{x}) \rangle = -\varepsilon \| f(y) - f(\hat{x}) \| < 0. \tag{18}$$

This is absurd.

(ii) follows from the definition of scalar minimum point. □

*Remark 3.3* System (15) allows one to associate (1) with its Image Space (IS) and perform an useful analysis. To this end pose:

$$u = f(y) - f(x), \quad v = g(x), \quad x \in X; \tag{19}$$

the image of  $X$  through the function  $x \mapsto (f(y) - f(x), g(x))$  is the IS associated with (1). For details, see [9, 12].

We are now able to define the steps of an approach for finding (all) the VMPs of (1).

- (A) Choose any  $p \in \text{int } C^*$  ( $p$  will remain fixed in the sequel).
- (B) Choose any  $y^0 \in K$  and solve the (scalar) problem (9) at  $y = y^0$ ; let  $x^0$  be a solution; according to Proposition 3.4,  $x^0$  is a VMP of (1).
- (C) Consider (9) as a parametric problem in the parameter  $y$ : start at  $y = x^0$  and find its solutions. According to Propositions 3.4 and 3.5 all the solutions of (1) will be found. This approach, which will be developed in [2], seems promising independently of the bi-level problem. If we apply it to the bi-level problem, then it becomes the following set of steps.

As said in Sect. 1, in general, in the real applications, the problem to solve is just (3), and not that of finding all the solutions of (1); hence, it is desired to meet, among the solutions of (1), only those, which allow one to solve (3). The approach described above serves to satisfy such a need. To this end, the method described at the end of the previous section can be integrated this way:

- (A) Choose any  $p \in \text{int } C^*$  ( $p$  will remain fixed in the sequel).
- (B) Choose any  $y^1 \in K$  and solve the (scalar) problem (9) at  $y = y^1$ ; let  $x^1$  be a minimum point; call  $K_1^0$  the set of solutions of (9) obtained by varying  $y$  from  $y = y^1$ ; of course  $x^1 \in K_1^0$ ; according to Propositions 3.4 and 3.5, all the elements of  $K_1^0$  are VMPs of (1).
- (C) Solve the problem:

$$\min \Phi(y) \text{ s.t. } y \in K_1^0; \tag{20}$$

if we can conclude that the solutions of this problem are such also on  $K^0$ , then (3) is solved; otherwise, we must continue.

(D) Jump to a subset of  $K^0$ , adjacent to  $K_1^0$ ; let it be  $K_2^0$ ; repeat (C) on it; and so on.

As is easily seen, thanks to the method of the above section, in solving (3) we do not meet all the solutions of (1). The above method is a general scheme, which requires to be implemented; the implementation takes advantage, if it is done within a certain class of functions; an instance of this is shown below.

## 4 Reduction of the Scalar Problem

### 4.1 The Upper Level: The Linear Case

The symbols of this section are independent of those of the previous sections. Let us now consider problem (3), where  $\Phi$  is convex; even if it is not necessary, for the sake of simplicity,  $\Phi$  will be assumed to be differentiable.

Now, suppose that the lower level be linear, and consider the case where  $X = \mathbb{R}_+^n$  and  $C = \mathbb{R}_+^l$ , which, although a particular one, is among the most important formats in the applications. Thus, without any loss of generality, we can set:

$$f_i(x) = \langle d^i, x \rangle, \quad d^i \in \mathbb{R}^n, \quad i = 1, \dots, l, \quad D = \begin{pmatrix} d^1 \\ \vdots \\ d^l \end{pmatrix}, \quad f(x) = Dx$$

$$\langle p, f(x) \rangle = \langle pD, x \rangle$$

where  $d^i$  and  $p$  are considered as row-vectors. The set  $K(y)$  can be cast in the form:

$$K(y) = \{x \in \mathbb{R}_+^n : -Dx \geq -Dy, \Gamma x \geq \gamma\},$$

with  $\Gamma \in \mathbb{R}^{m \times n}$ ,  $\gamma \in \mathbb{R}^m$ . By setting

$$A = \begin{pmatrix} -D \\ \Gamma \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ \gamma \end{pmatrix}, \quad E = \begin{pmatrix} -D \\ 0 \end{pmatrix}, \quad c = pD$$

problem (9) takes the form:

$$\min \langle c, x \rangle, \quad Ax \geq b + Ey, \quad x \geq 0, \tag{21}$$

where, without any loss of generality, we assume that the rank of  $A$  be  $n$ , and that:

$$A = \begin{pmatrix} B \\ N \end{pmatrix}, \quad b = \begin{pmatrix} b_B \\ b_N \end{pmatrix}, \quad E = \begin{pmatrix} E_B \\ E_N \end{pmatrix},$$

where  $B$  is a feasible basis, that is

$$B^{-1}(b_B + E_B y) \geq 0, \quad N(B^{-1}(b_B + E_B y)) \geq b_N + E_N y. \quad (22)$$

The vector

$$x = B^{-1}(b_B + E_B y) \quad (23)$$

is a solution of (21) if and only if

$$cB^{-1} \geq 0, \quad (24)$$

and remains optimal until  $y$  fulfils (22). Note that the performance of step (B) of the method leads at a point, namely  $x^0$ , where the inequalities, which defines  $S(y)$ , are verified as equalities; consequently, in general,  $x^0$  will be overdetermined. Therefore, we assume that an anti-cycling ordering is adopted. Moreover, it is not restrictive to suppose that  $B$  contains at least one row of  $D$ ; this will understood in the sequel.

Now, with a small abuse of notation, problem (20) becomes:

$$\min \Phi(y), \quad \text{s.t. } y \in K_{B_1}^0, \quad (25)$$

where

$$K_{B_1}^0 = \{y \in \mathbb{R}^n : (I_n - B_1^{-1}E_{B_1})y = B_1^{-1}b_{B_1}, \quad (N_1 B_1^{-1}E_{B_1} - E_{N_1})y \geq b_{N_1} - N_1 B_1^{-1}b_{B_1}\},$$

and  $B_1$  is a base which identifies  $K_0^1$ . We can now specify the method (A)–(D) to the present case; it finds a local minimum point of (3).

- (a) Choose any  $p$  such that  $p \in \text{int } C^*$ ;  $p$  will remain fixed in the sequel.
- (b) Choose any  $y^1 \in K$ . We have to solve (21), which is assumed to have minimum. By a standard use of Simplex Method, we find an optimal basis, say  $B_1$  and the minimum point given by:

$$x^1 = B_1^{-1}(b_{B_1} + E_{B_1}y^1).$$

Now, replace  $y^1$  with the parameter  $y$ , but keep  $B_1$  as basis. According to the Propositions 3.4 and 3.5, all the VMPs of (1), corresponding to  $B_1$ , are obtained as those solutions of (21) which equal the very parameter  $y$ ; this is equivalent to say that  $y$  must be such that  $B_1$  is both primal and dual feasible and  $y$  must be a fixed point of the map:  $y \mapsto B_1^{-1}(b_{B_1} + E_{B_1}y)$ . Since  $B_1$  is dual feasible if and only if  $cB_1^{-1} \geq 0$  (which is a by-product of the construction of  $x^1$  and does not depend on  $y$ ), in conclusion, such VMPs are the solutions of the system:

$$\begin{cases} B_1(N_1 B_1^{-1} E_{B_1} - E_{N_1})y \geq b_{N_1} - N_1 B_1^{-1} b_{B_1} & \text{primal feasibility} \\ y = B_1^{-1}(b_{B_1} + E_{B_1}y) & \text{fixed point} \end{cases} \quad (26)$$

Call  $K_{B_1}^0$  the set of solutions of (26).

(c) Consider the problem

$$\min \Phi(y), \text{ s.t. } y \in K_{B_1}^0, \quad (27)$$

and let  $y^2$  be a minimum point of it. If  $y^2 \in ri K_{B_1}^0$ , then stop; otherwise, perform next step.

(d) If  $\Phi$  does not decrease, when we try, through a pivot, to exchange  $B_1$  with one of its adjacent bases (such an exchange will be performed under an anti-cycling rule), then again with  $y^2$  we have reached a solution of (27) and we stop. Otherwise, we replace  $B_1$  with an adjacent basis, which allows  $\Phi$  to decrease, and repeat the step (c).

*Remark 4.1* It is worthy to stress the fact that the previous method may reduce the bi-level problem to a finite sequence of scalar extremum problems. For instance, if both (1) and (3) are linear, then, performing the steps (a)–(d) of this section amounts to execute a finite steps of Simplex Method; if  $\Phi$  is convex and (1) is linear (as assumed in this section), then performing (a)–(d) in this section amounts to execute a finite number of steps of the Gradient Method.

Now, we will give a justification of the above method. Let  $K$  be a polyhedron of  $\mathbb{R}^n$  and  $Q$  a convex cone having, as apex, the origin, which does not belong to it. Given a vector  $x$ ,  $Q_x$  denotes the translation of  $Q$ , which has  $x$  as apex or

$$Q_x = \{y \in \mathbb{R}^n : y = x + q, q \in Q\}.$$

Moreover, in the sequel,  $H^0$  denotes any hyper-plane of  $\mathbb{R}^n$ , defined by

$$\{x \in \mathbb{R}^n : \langle a, x \rangle = b\}$$

and  $H^-$  and  $H^+$  denote, respectively, the half-spaces

$$\{x \in \mathbb{R}^n : \langle a, x \rangle \leq b\}, \quad \{x \in \mathbb{R}^n : \langle a, x \rangle \geq b\}.$$

**Definition 4.1** Let  $\mathcal{F}$  be a set of proper faces of a polyhedron  $K$ .  $\mathcal{F}$  is said to be connected, if and only if, for each pair of element of  $\mathcal{F}$ , say  $F'$  and  $F''$ , there exists a set of proper faces of  $K$ , say  $F_0, F_1, \dots, F_r$ , contained in  $\mathcal{F}$ , such that  $F_0 = F'$ ,  $F_r = F''$ ,  $F_{i-1} \cap F_i \neq \emptyset$  and  $F_i$  is not a subspace of  $F_{i-1}$ , for  $i = 1, \dots, r$ .

**Lemma 4.1** *Let  $F$  be a face of  $K$  and  $x^0 \in \text{int } F$ . If*

$$Q_{x^0} \cap K = \emptyset, \tag{28}$$

*then, for all  $x \in F$ ,*

$$Q_x \cap K = \emptyset.$$

*Proof* Let the polyhedron  $K$  be the set

$$\{x \in \mathbb{R}^n : \Gamma x \geq \gamma\},$$

where  $\Gamma \in \mathbb{R}^{m \times n}$ ,  $\gamma \in \mathbb{R}^m$ . Without any loss of generality, as face  $F$  we can consider

$$F = \{x \in \mathbb{R}^n : \Gamma_1 x = \gamma_1, \Gamma_2 x \geq \gamma_2\}$$

where

$$\Gamma = \begin{pmatrix} \Gamma_1 \\ \Gamma_2 \end{pmatrix}, \quad \gamma = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}.$$

The hypothesis  $x^0 \in \text{ri } F$  means

$$\Gamma_1 x^0 = \gamma_1, \quad \Gamma_2 x^0 > \gamma_2. \tag{29}$$

From (28) we draw that, not only one of the inequalities which define  $K$  is violated, but, account taken of (29), such inequality corresponds to  $\Gamma_1$ : with obvious notation, let us denote it by

$$\langle (\Gamma_1)_i, x^0 + q \rangle < (\gamma_1)_i. \tag{30}$$

(nel prodotto scalare servono vettori colonna?)

In fact, if ab absurdo such a violated inequality corresponded to  $\Gamma_2$ , then, by letting  $q \rightarrow 0$  we would obtain

$$\langle (\Gamma_2)_i, x^0 \rangle \leq (\gamma_2)_i$$

which contradicts (29). Now, let  $x \in F$ . The equalities

$$\langle (\Gamma_1)_i, x + q = \langle (\Gamma_1)_i, x^0 + q \rangle,$$

and (30) lead to  $Q_x \cap K = \emptyset$ . □

**Lemma 4.2**  *$K^0$  is connected.*

*Proof* First, observe that  $K^0 = \{x \in K : Q_x \cap K = \emptyset\}$ . If  $F_1$  and  $F_2$  are proper faces of  $K$ , which belong also to  $K^0$ , there exist two hyper-planes, which are not parallel (contradicting this leads to contradict that one of the two faces does not belong to  $K^0$ ):

$$H_1^0 = \{x : \langle a_1, x \rangle = b_1\}, \quad H_2^0 = \{x : \langle a_2, x \rangle = b_2\}$$

which support  $K$  and, respectively, contain  $F_1$  and  $F_2$ , and such that

$$Q_x \subseteq H_i^-, \quad \forall x \in H_i^0, \quad i = 1, 2.$$

Put

$$V^- := H_1^- \cap H_2^-, \quad V^+ := H_1^+ \cap H_2^+,$$

and set,  $\forall t \in [0, 1]$ ,

$$H_t^0 := \{x \in \mathbb{R}^n : \langle (1-t)a_1 + ta_2, x \rangle = (1-t)b_1 + tb_2\}.$$

Since,  $\forall t \in [0, 1]$ ,

$$K \subseteq V^+ \subseteq H_t^+, \quad Q_x \subseteq V^-, \quad \forall x \in H_1^0 \cap H_2^0,$$

then  $\forall t \in [0, 1]$ , there is a hyper-plane, say  $\mathcal{H}_t^0$ , parallel to  $H_t^0$ , supporting  $K$  and such that

$$K \subseteq \mathcal{H}_t^+, \quad Q_x \subseteq \mathcal{H}_t^-, \quad \forall x \in \mathcal{H}_t^0.$$

This means that the face  $\mathcal{H}_t^0 \cap K$  is a subset of  $K^0$ . Finally, the interval  $[0, 1]$  is partitioned into a finite number of sub-intervals, and, this way, two consecutive intervals correspond to adjacent faces of  $K^0$ .  $\square$

**Proposition 4.1** *Suppose that the function  $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}$  be convex and differentiable, and suppose that its infimum (minimum) occurs on  $\mathbb{R}^n \setminus K$ . Assume that  $f$  and  $g$  be as above. Then, the algorithm (a)–(d) finds a local minimum point of (3) in a finite number of steps.*

*Proof* First of all observe that  $K^0$  is a connected set of faces of  $K$  (Lemma 4.2). Observe also that, in going from basis  $B$  to an adjacent one, maintaining a solution of (21) (note that  $B$  contains at least one row of  $D$ ; such an assumption is not restrictive because of Proposition 3.4 and allows us to parametrize the faces of  $K^0$ ), we pass from a face of  $K^0$  to an adjacent face of  $K$  still included in  $K^0$ . Then, by adopting any (but fixed) ordering of the combinations of class  $n$  extracted from  $\{1, 2, \dots, l+m\}$ , and an anti-cycling order, the algorithm (a)–(d) can visit all the faces of  $K$  and then of  $K^0$  if it is necessary for the minimization of (3).



The stationary point at which the algorithm stops is a local minimum point of (3). In fact, it holds that

$$\nabla\Phi(x^0) \in \text{conv}\{a_i, i \in \mathcal{I}(F^0)\}, \tag{31}$$

where  $\mathcal{I}(F^0)$  is the set of indexes of the constraints of (21), which are also of  $K$  and are binding at  $x^0$ . From (31) we draw that zero belongs to a convex combination, at  $x^0$ , of the gradients of  $\Phi$  and of the constraints of (21), which are binding at  $x^0$ , identified by  $\mathcal{I}(F^0)$ . Due to the convexity of all the implicated functions, such a condition is sufficient besides necessary.  $\square$

## 5 Reduction of the Upper Level

### 5.1 Examples

*Example 1* In (1) and (3) set  $n = m = l = 2$ ,  $X = C = \mathbb{R}_+^2$  and

$$f(x) = \begin{pmatrix} 2x_1 - x_2 \\ -x_1 + 2x_2 \end{pmatrix}, \quad g(x) = \begin{pmatrix} 2x_1 + x_2 - 1 \\ -x_1 + 2x_2 - 1 \end{pmatrix}, \quad \Phi(x) = x_1^2 + (x_2 - 1/2)^2.$$

Let us perform (a)–(d) of Sect. 4.

(a) Since  $C^* = C$ , we can choose  $p = (2, 3)$ ; it will remain fixed.

(b) Now we have:

$$D = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \quad \Gamma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad \gamma = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad c = (1 \ 4),$$

so that (25) becomes

$$\begin{aligned} &\min(x_1 + 4x_2) \\ &\text{s.t.} \end{aligned} \tag{32}$$

$$\begin{pmatrix} -2 & 1 \\ 1 & -2 \\ 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} -2 & 1 \\ 1 & -2 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \quad x \in X$$

We choose  $y^1 = (2/3 \ 4/9)$ . By means of a straightforward use of Simplex Method, we find that the basis of  $A$ , formed with the 1st and 4th rows, or:

$$B_1 = \begin{pmatrix} -2 & 1 \\ 1 & 2 \end{pmatrix}$$

gives the unique solution of (32) (with  $y = y^1$ ) i.e.  $x^1 = (5/9 \ 2/9)$ . The set  $K_{B_1}^0$ , given by the system (26), becomes

$$K_{B_1}^0 = \{y \in \mathbb{R}_+^2 : y_1 + 2y_2 = 1, \ 3y_1 \geq 1\}.$$

(c) Problem (27) becomes

$$\min \left( y_1^2 + (y_2 - 1/2)^2 \right), \text{ s.t. } y \in K_{B_1}^0.$$

Now replace  $B_1$  with

$$B_2 = \begin{pmatrix} -2 & 1 \\ 2 & 1 \end{pmatrix}.$$

By means of the Gradient Method, we easily find its unique minimum point and minimum:

$$y^2 = (1/3 \ 1/3), \quad \Phi(y^2) = 5/36.$$

Since  $y^2 \in \partial K_{B_1}^0$ , we perform (d).

(d) From the equalities

$$\nabla \Phi(y_2) = (-2/3 \ -1/3)$$

we draw that  $y^2$  is not either a global or a local minimum point of  $\Phi$  on  $K^0$  and that (the 1st constraint of  $K$  being binding and the 2nd being redundant), by replacing  $B_1$  with  $B_2$ ,  $\Phi$  decreases with respect to  $5/36$ . Then perform again the step (b) with  $B_2$  and  $y^2$  in place of  $B_1$  and  $y^1$ , respectively. (b)' When  $B_2^B$  System (26) gives the set

$$K_{B_2}^0 = \{y \in \mathbb{R}_+^2 : 2y_1 + y_2 = 1, \ 3y_1 \leq 1\}.$$

(c)' Problem (27) becomes

$$\min \left( y_1^2 + (y_2 - 1/2)^2 \right), \text{ s.t. } y \in K_{B_2}^0.$$

By means of the Gradient Method, we easily find its unique minimum point and minimum:

$$y^3 = (1/5 \ 3/5), \quad \Phi(y^3) = 1/20 < \Phi(y^2).$$

Since  $y^3 \in \partial K_{B_2}^0$ ,  $y^3$  is a global besides local minimum point of (3).

*Example 2* Consider the previous example, replacing  $\Phi$  with the following one:

$$\Phi(x) = (x_1 - 2)^2 + (x_2 - 5/6)^2.$$

Perform the steps (a), (b) and (c), but with the present  $\Phi$ . Solving

$$\min \left( (x_1 - 2)^2 + (x_2 - 5/6)^2 \right), \quad s.t. y \in K_{B_1}^0, \tag{33}$$

we find

$$y^2 = (2/5 \ 3/10) \in ri K_{B_1}^0, \quad \Phi(y^2) = 16/45.$$

Despite of this, if we consider in (33)  $K_{B_2}^0$  instead of  $K_{B_1}^0$  we find

$$\tilde{y} = (1/5 \ 3/5), \quad \Phi(\tilde{y}) = 49/180 < 16/45,$$

which shows that, notwithstanding the fact that  $y^2$  be a global minimum point of  $\Phi$  on  $K_{B_1}^0$ , it is not a global minimum point on  $K^0$ .

*Example 3* Let us now briefly discuss a classic scalarization method, namely that introduced in [8]; see also [7, 15]. It aims a finding the weak VMPs of (1), and thus the comparison with the method described in the previous sections is not perfectly fitting; however, we disregard this aspect since one might think of extending it; hence, we want to see what would happen if it were extended to the case of a cone  $C$  and not  $int C$ . Consider again Example 1. To find a (weak) VMP, we must consider the problem:

$$\forall (u_1, u_2) \in f(K), \quad find F(u) = \min_{x \in K} \max \left( f_1(x) - u_1, f_2(x) - u_2 \right); \tag{34}$$

the result is a (weak) VMP. Note that the minimization in (34) is a nonsmooth problem. For instance, when  $f$  is the function of Example 1, and  $u = (2 \ 1)$  such a minimization becomes:

$$\min_{x \in K} \max \left( 2x_1 - x_2 - 2, -x_1 + 2x_2 - 1 \right), \tag{35}$$

which leads to  $x^1 = (5/9 \ 2/9)$  of Example 1. The practically impossible problem is to express  $\Phi$  as function of a vector running on the set of solutions of (34), even if only those of a subset of  $K^0$  like  $K_{B_1}^0$ .

*Example 4* Let us set  $n = 1, l = m = 2, X = \mathbb{R}, C = \mathbb{R}_+^2$ ,

$$f_1(x) = x, f_2(x) = x^2, g_1(x) = 0x + 1, g_2(x) = -x.$$

Obviously,  $K = [-1, 0]$ , and all the elements of  $K$  are VMPs of (1). Set  $y = 0$ . For  $p_1, p_2 > 0$ , consider the classic scalarized problem:

$$\min(p_1x + p_2x), x \in K.$$

Note that  $x = 0$  is not a (global) minimum point of the classic scalarized problem whatever  $p_1, p_2 > 0$  may be.

*Example 5* Let us set  $n = 1, l = m = 2, X = \mathbb{R}, C = \mathbb{R}^2$ , and

$$f = (2x - x^2 \ 1 - x^2), g = (x \ 1 - x).$$

we find  $S(y) = \{y\}, \forall y \in [0, 1]$ . Hence, the unique solution of (3) is  $y$  itself. By varying  $y$ , (3) gives, with its solutions, the interval  $\sigma = [0, 1]$ , which is the set of VMP of (1), as it is trivial to check. Now, let us use the classic scalarization [14] outside the classic assumption of convexity, i.e. the scalar parametric problem which, here, becomes:

$$\min(c_1f_1(x) + c_2f_2(x)), x \in \sigma$$

that is

$$\min - (c_1 + c_2)x^2 + 2c_1x + c_2, x \in \sigma \tag{36}$$

where  $(c_1 \ c_2) \in \text{int } C^* = \text{int } \mathbb{R}^2$  is a pair of parameters. Every minimum point of (36) is a VMP of (1). In the present example it is easy to see that the only solutions of (36) are  $x = 0$ , or  $x = 0$  and  $x = 1$ , or  $x = 1$  according to respectively  $c_2 < c_1$  or  $c_2 = c_1$  or  $c_2 > c_1$ . Hence, the scalarized problem (36) does not detect all the solutions of (1) (the same happens obviously to (3), if  $S(y)$  is deleted).

## 6 Further Developments

The development carried out in the previous sections is, deliberately, much simplified. In fact, the scope of this paper is to stress the importance, for the applications, of addressing some research efforts to the study of the bi-level vector problem. Some possible extensions are outlined below.

(i) A first effort will be devoted to let the previous method be able to find global minima. Some of the assumptions, made to simplifying the exposition, are too restrictive; it should be useful to remove them.

(ii) In order to stress the importance of the bi-level approach, let us bring an example. An extremely important application of vector optimization is to aerospace design. In this field, the first fundamental quantities are lift, drug and cost (of course, in reality, besides them, we have many other quantities or their splitting). To formulate (1) with such 3 objectives ( $l = 3$ ) should be meaningless; a competition between the cost and the lift or the drug should be a nonsense. A meaningful approach is to formulate (1) with 2 objectives, the lift and the drug ( $l = 2$ ), and (3) with  $\Phi$  to represent the cost.

(3i) In Sect. 4, the general method of Sect. 3 has been applied to a particular (even if particularly important) class of problems, and it has been shown how the bi-level problem can be reduced to a (finite) sequence of single-problems. It should be interesting to obtain a similar result for other classes of problems; for instance, exploiting Sect. 2, the class of C-functions. Extensions to infinite dimensional spaces are also of great importance. We note that the method of the previous sections may reduce the bi-level problem to a finite sequence of scalar extremum problems. For instance, if both (1) and (3) are linear, then, performing (a)–(d) of Sect. 4, amounts to execute a finite steps of Simplex Method; if  $\Phi$  is convex and (1) is linear (as assumed in Sect. 4), then performing (a)–(d) of Sect. 4, amounts to execute a finite steps of the Gradient Method. It should be interesting to identify other classes of bi-level problems for which such a reduction holds.

(4i) As it is well known, not always an equilibrium can be expressed as the extremum of any functional; this led to formulate the theory of Variational Inequalities (VI). Furthermore, some equilibria are characterized by more than one operator and a blending of the involved operator may be not sufficient; this led to formulate the theory of Vector Variational Inequalities (VVI). As shown for VOP in the previous sections, also in this case a bi-level approach is suitable for the applications. Consequently, it should be useful to extend the method of Sect. 4 to the case of VVI. In other words, (1) must be replaced by a VVI: let  $F : \mathbb{R}^n \rightarrow \mathbb{R}^{l \times n}$  be a matrix-valued function and consider the VVI, which consists in finding  $y \in K$  such that:

$$F(y)(x - y) \not\prec_{C_0} 0, \quad \forall x \in K, \tag{37}$$

where  $C_0$  and  $K$  are as in Sect. 1. Denote by  $K^0$  the set of solutions to (37). Now, consider the (scalar) VI, which consists in finding  $y \in K^0$  such that:

$$\langle \Psi(y), x - y \rangle \geq 0, \quad \forall x \in K^0, \tag{38}$$

where  $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}^n$ . When both (37) and (38) admit the primitives (see the so-called Symmetry Principle), then they can be cast in the formats (1) and (3) respectively. The scalarization method for (37) described in Sect. Appendix of [12] should allow one to define, for (37) and (38), a method like that of Sect. 4, avoiding

to be obliged to find necessarily all the solutions of (37), namely  $K^0$ . The above VVI and VI are of Stampacchia type; same question about scalarization can be posed for the Minty type VVI and VI.

(5i) Another development may deal with the perturbation function of (3). There exists a wide literature as it concerns with scalar optimization and a few with (37), but they are independent each other. In as much as the important problem is (3), the study of the perturbation function of (1) should be *auxiliary* to (3) and not autonomous. Let the constraints of (1) be  $g \geq \xi$ , where  $\xi$  plays the role of a parameter. Then  $K$  and  $K^0$  depend on  $\xi$ ; denote them by  $K(\xi)$  and  $K^0(\xi)$ , respectively. Hence, the minimum in (3) will depend on  $\xi$ , say  $\Phi^\downarrow(\xi)$ . The study of the properties of  $K^0(\xi)$  is extremely important, while to find it is, in general, very difficult, but also useless, if (3) is the main scope.

(6i) Another subject, strictly connected with the previous one, is that of duality. The literature on duality for (1) is wide. Here too, in as much as the important problem is (3), the study of duality of (1) should be dependent on that of (3). Let us restrict to the Lagrangian duality, whose study is naturally located in the Image Space associated with the given problem. In fact the dual space is that of the functionals, whose zero level sets are considered to separate two suitable sets of the IS [9]. Hence, we have an IS associated with (1) and an IS associated with (3). In general, (3) has a positive duality gap. Sensitivity is a further topic, which is fundamental for the applications.

(7i) An extension of the present approach to set-valued, in particular interval-valued, extremum problems is conceivable. The infinite dimensional vector extremum problems, especially those of isoperimetric type, and the stochastic version of the previously mentioned problems, are surely interesting fields of research.

## Appendix

The definitions of  $A$  and  $D$  in this section are independent of those of the other sections.

**Definition A.1** The (positive) polar of the cone  $C$  is given by:

$$C^* := \{x \in \mathbb{R}^n : \langle y, x \rangle \geq 0, \forall y \in C\}. \tag{39}$$

The following set of functions will be the base of the present approach. More precisely, in the present paper, we establish the theory, based on  $C$ -functions, which is the background of the approach to the bi-level problems we want to carry on. In the present paper, we begin with the class of problems, say convex-linear problems, which have  $\Phi$  convex,  $f$  linear and  $K$  polyhedral; other classes will be studied in furthercoming papers.

**Definition A.2** Let  $X$  be convex;  $f$  is a convex [9] if and only if  $\forall x^1, x^2 \in X, \forall \alpha \in [0, 1]$ :

$$(1 - \alpha)f(x^1) + \alpha f(x^2) - f((1 - \alpha)x^1 + \alpha x^2) \in C. \tag{40}$$

When  $C \subseteq \mathbb{R}^l$  or  $C \supseteq \mathbb{R}^l$ , then  $f$  is called  $C$ -convex. At  $l = 1$  and  $C = \mathbb{R}_+$ ,  $f$  is the classic convex function. In most of the literature, regardless of the occurrence of such inclusions, a convex is often called  $C$ -convex; this is not suitable. For instance, the  $\mathbb{R}_-$ -function, which turns out to be a concave function, should be called  $\mathbb{R}_-$ -convex; this, even if formally correct, is unnecessarily far from the

intuitive sense and the common language. The definition is a cornerstone of mathematics; consequently, it should be handled very cautiously, without distorting the already possessed concepts. The following property of convexs will be fundamental in the sequel.

**Proposition A.1** *If  $f$  is a convex on  $X$  and  $c^* \in C^*$ , then  $\langle c^*, f \rangle$  is convex on  $X$ .*

*Proof* Since  $f$  is a convex,  $\forall c^* \in C^*, \forall x^1, x^2 \in X, \forall \alpha \in [0, 1]$ ,

$$\langle c^*, (1 - \alpha)f(x^1) + \alpha f(x^2) - f((1 - \alpha)x^1 + \alpha x^2) \rangle \geq 0$$

or, equivalently,

$$(1 - \alpha)\langle c^*, f(x^1) \rangle + \alpha\langle c^*, f(x^2) \rangle - \langle c^*, f((1 - \alpha)x^1 + \alpha x^2) \rangle \geq 0.$$

Hence, the convexity of  $\langle c^*, f \rangle$  follows. □

As it is well known, the drawback of most of the extensions of convex functions is the lack of conditions which allow one to detect, through viable numerical calculus, whether or not a given function fulfils the definition of such an extension. The  $C$ -functions (see the Appendix) are among the few extensions for which some viable conditions can be established. Suppose that the cone  $C$  be polyhedral, so that there exists a matrix  $A \in \mathbb{R}^{r \times l}$ , whose generic entry is denoted by  $a_{ij}$ , such that:

$$C = \{u \in \mathbb{R}^l : Au \geq 0\}. \tag{41}$$

In this case,  $f$  is a  $C$ -function if and only if,  $\forall x^1, x^2 \in X, \forall \alpha \in [0, 1]$ ,

$$A[(1 - \alpha)f(x^1) + \alpha f(x^2) - f((1 - \alpha)x^1 + \alpha x^2)] \geq 0 \tag{42}$$

or

$$(1 - \alpha)\phi_i(x^1) + \alpha\phi_i(x^2) - \phi_i((1 - \alpha)x^1 + \alpha x^2) \geq 0, \quad i = 1, \dots, r,$$

where

$$\phi_i(x) := \sum_{j=1}^l a_{ij} f_j(x), \quad i = 1, \dots, r. \tag{43}$$

Thus, the following result holds.

**Proposition A.2**  *$f$  is a  $C$ -function on  $X$  with respect to the polyhedral cone (41) if and only if the functions  $\phi_i, i = 1, \dots, r$  of (43) are convex on  $X$ .*

Observe that the functions  $\phi_1, \dots, \phi_r$  can be convex, even if some (all) the functions  $f_1, \dots, f_l$  are not, as the following example shows.

*Example A.1* Let  $X = \mathbb{R}^2, C = \{u \in \mathbb{R}^2 : 2u_1 + u_2 \geq 0, u_1 + 2u_2 \geq 0\}$ . Let  $f_1(x) = -x_1^2/2 + 3x_2^2, f_2(x) = 3x_1^2 - x_2^2/2$ .  $f_1$  and  $f_2$  are not convex but  $\phi_1 = 2f_1 + f_2$  and  $\phi_2 = f_1 + 2f_2$  are convex and then  $f = (f_1, f_2)$  is a  $C$ -function.

The preceding example suggests a condition for  $f$  to be a  $C$ -function when  $C$  is like in (41). Set:

$$f = (f_1, \dots, f_l), \quad \text{where } f_i(x) = \langle x, D_i x \rangle, \quad D_i \in \mathbb{R}^{l \times l}, \quad i = 1, \dots, l, \tag{44}$$

and put  $x(\alpha) := (1 - \alpha)x^1 + \alpha x^2$ ,  $\alpha \in [0, 1]$ . Condition (42) is fulfilled if and only if,  $\forall i = 1, \dots, l$ ,  $\forall x^1, x^2, \forall \alpha \in [0, 1]$ ,

$$(1 - \alpha)\langle x^1, \sum_{j=1}^l a_{ij} D_j x^1 \rangle + \alpha \langle x^2, \sum_{j=1}^l a_{ij} D_j x^2 \rangle - \langle x(\alpha), \sum_{j=1}^l a_{ij} D_j x(\alpha) \rangle \geq 0.$$

Thus, the following result holds.

**Proposition A.3** *The (vector) quadratic function (44) is a C-function on X with respect to the cone (41), if and only if each of the matrices*

$$Q_i = \sum_{j=1}^l a_{ij} D_j, \quad i = 1, \dots, l,$$

has non-negative eigenvalues.

*Remark A.1* The cone of Example 2.1 contains  $\mathbb{R}_+^2$ , but it does not differ much from  $\mathbb{R}_+^2$ . In several applications, like e.g. the design of aircrafts, the cone is the Pareto one; however, the designers may desire to explore what happens, if such a cone is relaxed a little bit.

Now, let us consider the case, where C is not necessarily polyhedral; let it be defined by its supporting half-spaces, or

$$C := \bigcap_{t \in T} \{x \in \mathbb{R}^l : \langle a_t, x \rangle \geq 0\}, \tag{45}$$

where  $T$  is an interval of  $\mathbb{R}$ , and,  $\forall t \in T$ ,  $a_t \in \mathbb{R}^l$ .

When the cone  $C$  is given by (45), a function  $f$  is a  $C$ -function if and only if,  $\forall x^1, x^2, \forall \alpha \in [0, 1]$ ,

$$\langle a_t, (1 - \alpha)f(x^1) + \alpha f(x^2) - f(x(\alpha)) \rangle \geq 0, \quad \forall t \in T,$$

that is

$$(1 - \alpha)\varphi_t(x^1) + \alpha\varphi_t(x^2) - \varphi_t(x(\alpha)) \geq 0, \quad \forall t \in T,$$

where

$$\varphi_t = \langle a_t, f \rangle. \tag{46}$$

We have thus obtained:

**Proposition A.4** *The function f is a C-function on X with respect to the cone C defined in (45) if and only if,  $\forall t \in T$ , the function  $\varphi_t$  defined in (46) is convex on X.*

The functions  $\varphi_t, t \in T$ , can be convex on  $X$  even if some (all) the functions  $f_j, j = 1, \dots, l$  are not, as the following example shows.

*Example A.2* Let  $X = \mathbb{R}^2$  and  $C = \{u \in \mathbb{R}^3 : u_3 \geq \sqrt{u_1^2 + u_2^2}\}$ . The family of all the supporting halfspaces of  $C$ , namely (45), is easily found to be:

$$\bigcap_{t \in [-\sqrt{2}, \sqrt{2}]} \{u \in \mathbb{R}^3 : -tu_1 \pm \sqrt{2 - t^2}u_2 + \sqrt{2}u_3 \geq 0\}.$$



Consider the vector function  $f = (f_1, f_2, f_3)$  with:

$$f_i(x) = \langle x, D_i x \rangle,$$

being

$$D_1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad D_2 = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad D_3 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

We have now:

$$a(t) = (-t, \pm\sqrt{2-t^2}, \sqrt{2}), \quad t \in [-\sqrt{2}, \sqrt{2}],$$

so that

$$\phi_t(x) = \left\langle x, \begin{pmatrix} -t \pm \sqrt{2-t^2} + \sqrt{2} & 0 \\ 0 & -t \pm \sqrt{2-t^2} + \sqrt{2} \end{pmatrix} x \right\rangle.$$

It is easy to see that, for each  $t \in [-\sqrt{2}, \sqrt{2}]$ ,  $\phi_t$  is convex on  $X$ , while  $f_1$  and  $f_2$  are not convex.

*Remark A.2* Example A.2 suggests a condition for  $f$  to be a  $C$ -function on  $X$  with respect to a not necessarily polyhedral cone  $C$ .

## References

1. Antoni C, Giannessi F, Some remarks on bi-level vector extremum problems. Constructive nonsmooth analysis and related topics. In: Demianov, Pardalos e Batsyn (eds.) Springer Optimizations and Applications
2. Antoni C, Al-Shahrani M, On bi-level vector optimization, in preparation
3. Bonnel H, Morgan J (2006) Semivectorial bilevel optimization problem: penalty approach. *J Optim Theory Appl* 131:365–382
4. Chen GY, Huang X, Yang X (2005) Vector optimization, set-valued and variational analysis. Springer, Berlin
5. Dempe S, Dinh N, Dutta J (2010) Optimality conditions for a simple convex bilevel programming problem, variational analysis and generalized differentiation in optimization and control. In: Burachik RS, Yao JC (eds) Springer optimization and Its applications, vol 47. Springer, Berlin, pp 149–162
6. Dempe S, Gadhi N, Zemkoho AB, New optimality conditions for the semivectorial bilevel optimization problem. *J Optim Theory Appl*, to appear
7. Eichfelder G (2008) Adaptive scalarization methods in multiobjective optimization. Springer, Heidelberg
8. Gerstewitz (Tammer) C (1983) Nichtkonvexe Dualität in der Vektoroptimierung. *Wissensch. Zeitschr. TH Leuna-Merseburg* 25:357–364
9. Giannessi F (2005) Constrained optimization and image space analysis. Separation of sets and optimality conditions, vol I. Springer, New York
10. Giannessi F (2007) On the theory of lagrangian duality. *Optimization letters* 1 9–20, Springer, New York
11. Giannessi F, Pellegrini L (2001) Image space analysis for vector optimization and variational inequalities. Scalarization. In: Pardalos PM, Migdalas A, Burkard RE (eds) Combinatorial and global optimization, Kluwer, Dordrecht, pp 97–110
12. Giannessi F, Mastroeni G, Pellegrini L (2000) On the theory of vector optimization and variational inequalities. Image space analysis and separation. Vector variational inequalities and vector equilibria., *Mathematical theories, Series nonconvex optimization and its application-* Kluwer, Dordrecht, pp 141–215

13. Jahn J (2011) In: vector Optimization—Theory, applications, and extensions, 2nd edn. Springer, Heidelberg
14. Luc DT (1989) Theory of vector optimization. Springer, Berlin
15. Mastroeni G (2012) Optimality conditions in image space analysis for vector optimization problems. In: Ansari QH, Yao J-C (eds) Recent developments in vector optimization, series vector optimization, Springer, Berlin, pp 169–220
16. Pascoletti A, Serafini P (1984) Scalarizing vector optimization problems. *J Optim Theory Appl* 42:499–524
17. Roman B, Statnikov, Multicriteria Design — Optimization and Identification, Kluwer Academic Publishers
18. Solodov MV (2007) An explicit descent method for bilevel convex optimization. *J Convex Anal* 14:227237

# Reliability-Based Shape Design Optimization of Structures Subjected to Fatigue

Manolis Georgioudakis, Nikos D. Lagaros and Manolis Papadrakakis

**Abstract** Fatigue has been played a key role into the design process of structures, since many failures of them are attributed to repeated loading and unloading conditions. Crack growth due to fatigue, represents a critical issue for the integrity and resistance of structures and several numerical methods mainly based on fracture mechanics have been proposed in order to address this issue. Apart from loading, the shape of the structures is directly attributed to their service life. In this study, the extended finite element is integrated into a shape design optimization framework aiming to improve the service life of structural components subject to fatigue. The relation between the geometry of the structural component with the service life is also examined. This investigation is extended into a probabilistic design framework considering both material properties and crack tip initialization as random variables. The applicability and potential of the formulations presented are demonstrated with a characteristic numerical example. It is shown that with proper shape changes, the service life of structural component can be enhanced significantly. Comparisons with optimized shapes found for targeted service life are also addressed, while the choice of initial imperfection position and orientation was found to have a significant effect on the optimal shapes.

## 1 Introduction

The failure process of structural systems is considered among the most challenging phenomena in solid and structural mechanics. Despite of the advances achieved over the past decades in developing numerical simulation methods for modeling

---

M. Georgioudakis (✉) · N.D. Lagaros · M. Papadrakakis  
Institute of Structural Analysis & Antiseismic Research, School of Civil Engineering,  
National Technical University of Athens, Zografou Campus, 15780 Athens, Greece  
e-mail: geoem@mail.ntua.gr

N.D. Lagaros  
e-mail: nlagaros@central.ntua.gr

M. Papadrakakis  
e-mail: mpapadra@central.ntua.gr

© Springer International Publishing Switzerland 2015  
N.D. Lagaros and M. Papadrakakis (eds.), *Engineering and Applied Sciences Optimization*, Computational Methods in Applied Sciences 38,  
DOI 10.1007/978-3-319-18320-6\_24

such phenomena, there are issues still open to be addressed for accurately describing failure mechanisms at the macro as well as at the micro level. Reliability and accuracy of the numerical description of failure process, plays an important role for the design of new materials as well as for understanding their durability and resistance to various loading conditions.

In case of structural components subjected to cycling loading, fatigue plays an important role to their residual service life. When loading exceeds certain threshold value, microscopic cracks begin to form that propagate and possibly lead to fracture. Additionally, the shape of these components is a key parameter that significantly affects their residual service life and designers can control. Square holes or sharp corners lead to increased local stresses where fatigue cracks can initiate whereas round holes and smooth transitions or fillets will increase fatigue strength of the component. Hence, it is required not only a reliable simulation tool for crack growth analysis able to predict the crack paths and accurately describe the stiffness degradation due to damage, but also there is a need for an optimization procedure capable to identify improved designs of the structural components with regard to a targeted service life.

Limitations of the analytical methods in handling arbitrary complex geometries and crack propagation phenomena led to the development of numerical techniques for solving fracture mechanics problems. In recent years, finite elements with enrichments have gained increasing interest in modeling material failure, with the extended finite element method (XFEM) being the most popular of them. XFEM [50] is capable of modeling discontinuities within the standard finite element framework and its efficiency increases when coupled with the level set method (LSM) [52]. In this framework, Edke and Chang [13] presented a shape sensitivity analysis method for calculating gradients of crack growth rate and crack growth direction for 2D structural components under mixed-mode loading, by overcoming the issues of calculating accurate derivatives of both crack growth rate and direction. This work was further extended [14] to a shape optimization framework to support design of 2D structural components again under mixed-mode fracture for maximizing the service life and minimizing their weight. Furthermore, Li et al. [46] proposed elegant XFEM schemes for LSM based structural optimization, aiming to improve the computational accuracy and efficiency of XFEM, while Wang et al. [64] considered a reanalysis algorithm based on incremental Cholesky factorization which is implemented into an optimization algorithm to predict the angle of crack initiation from a hole in a plate with inclusion.

Many numerical methods have been developed over the last four decades in order to meet the demands of design optimization. These methods can be classified in two categories, gradient-based and derivative-free ones. Mathematical programming methods are the most popular methods of the first category, which make use of local curvature information, derived from linearization of objective and constraint functions and by using their derivatives with respect to the design variables at points obtained in the process of optimization. Heuristic and metaheuristic algorithms are nature-inspired or bio-inspired procedures that belong to the derivative-free category of methods. Metaheuristic algorithms for engineering optimization

include genetic algorithms (GA) [34], simulated annealing (SA) [38], particle swarm optimization (PSO) [37], ant colony algorithm (ACO) [12], artificial bee colony algorithm (ABC) [27], harmony search (HS) [23], cuckoo search algorithm [67], firefly algorithm (FA) [66], bat algorithm [68], krill herd [20], and many others. Evolutionary algorithms (EA) are among the most widely used class of meta-heuristic algorithms and in particular evolutionary programming (EP) [19], genetic algorithms [26, 34], evolution strategies (ES) [55, 59] and genetic programming (GP) [39].

The advancements in reliability theory of the past 30 years and the development of more accurate quantification of uncertainties associated with system loads, material properties and resistances have stimulated the interest in probabilistic treatment of systems [58]. The reliability of a system or its probability of failure constitute important factors to be considered during the design procedure, since they characterize the system's capability to successfully accomplish its design requirements. First and second order reliability methods, however, that have been developed to assess reliability, they require prior knowledge of the means and variances of component random variables and the definition of a differentiable limit-state function. On the other hand, simulation based methods are not restricted by form and knowledge of the limit-state function but many of them are characterized by high computational cost.

In this study, XFEM and LSM are integrated into a shape design optimization framework, aiming to investigate the relation between geometry and fatigue life in the design of 2D structural components. Specifically, shape design optimization problems are formulated within the context of XFEM, where the volume of the structural component is to be minimized subjected to constraint functions related to targeted service life (minimum number of fatigue cycles allowed) when material properties and crack tip initialization are considered as random variables. XFEM is adopted to solve the crack propagation problem as originally proposed by Moës et al. [50] and Stolarska et al. [61], with the introduction of adaptive enrichment technique and the consideration of asymptotic crack tip fields and Heaviside functions. XFEM formulation is particularly suitable for this type of problem since mesh difficulties encountered into a CAD-FEM shape optimization problem are avoided by working with a fixed mesh approach. In association to XFEM, the level set description is used to describe the geometry providing also the ability to modify the CAD model topology during the optimization process. Nature inspired optimization techniques have been proven to be computationally appealing, since they have been found to be robust and efficient even for complex problems and for this purpose are applied in this study. An illustrative example of a structural component is presented, and the results show that, with proper shape changes, the service life of structural systems subjected to fatigue loads can be enhanced. Comparisons between optimized shapes obtained for various targeted fatigue life values are also addressed, while the location of the initial imperfection along with its orientation were found to have a significant effect on the optimal shapes for the components examined [24].

## 2 Handling Fatigue Using XFEM

Fatigue growth occurs because of inelastic behavior at the crack tip. The present study is focused on 2D mixed-mode linear elastic fracture mechanics (LEFM) formulation, where the size of plastic zone is sufficiently small and it is embedded within an elastic singularity zone around the crack tip.

### 2.1 Fatigue Crack Growth Analysis at Mixed-Mode Loading

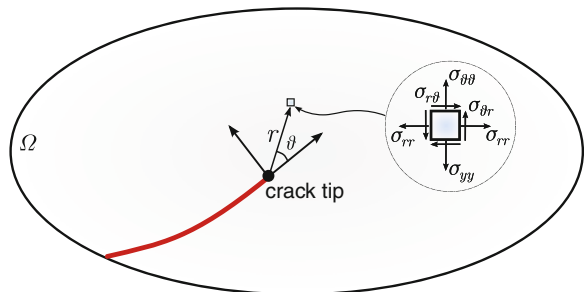
In order to quantify crack growth around the crack tip in the presence of constant amplitude cyclic stress intensity, the basic assumptions of LEFM are employed. The conditions at the crack tip are uniquely defined by the current value of the stress intensity factors (SIFs)  $K$ . For extracting mixed-mode SIFs, the domain form of the interaction energy integral [70] is used, based on the path independent J-integral [56], providing mesh independency and easy integration within the finite element code. When both stress intensity factors ( $K_I, K_{II}$ ) are known, the critical direction of crack growth  $\theta_c$  as well as the number of fatigue cycles  $N$  can easily be computed.

#### 2.1.1 Computation of the Crack Growth Direction

The accuracy and reliability of the analysis of a cracked body depends primarily on continuity and accurate determination of the crack path. It is therefore important to select the crack growth criteria very carefully. Among the existing criteria, the maximum hoop stress criterion [17], is used in this study. The crack growth criterion states that (i) crack initiation will occur when the maximum hoop stress reaches a critical value and (ii) crack will grow along direction  $\theta_{cr}$  in which circumferential stress  $\sigma_{\theta\theta}$  is maximum.

Then the circumferential stress  $\sigma_{r\theta}$  (see Fig. 1) along the direction of crack propagation is a principal stress, hence the crack propagation direction  $\theta_{cr}$  is determined by setting the shear stress equal to zero, i.e.:

**Fig. 1** Polar coordinates in the crack tip coordinate system



$$\sigma_{r\theta} = \frac{1}{2\pi r} \cos \frac{\theta}{2} \left( \frac{1}{2} K_I \sin \theta + \frac{1}{2} K_{II} (3 \cos \theta - 1) \right) = 0 \quad (1)$$

This leads to the expression for defining the critical crack propagation direction  $\theta_{cr}$  in terms of local crack tip coordinate system as:

$$\theta_{cr} = 2 \operatorname{atan} \frac{1}{4} \left( \frac{K_I}{K_{II}} \pm \sqrt{\frac{K_I^2}{K_{II}^2} + 8} \right) \quad (2)$$

It is worth mentioning that according to this criteria, the maximum propagation angle  $\theta_{cr}$  is limited to  $70.5^\circ$  for pure Mode II crack propagation problems.

### 2.1.2 Computation of Fatigue Cycles

Fatigue crack growth is estimated using Paris law [53], which is originally proposed for single mode deformation cases, relating the crack propagation rate under fatigue loading to SIFs. For the case of mixed-mode loading, a modified Paris law can be expressed using the effective stress intensity factor range  $\Delta K_{\text{eff}} = K_{\text{max}} - K_{\text{min}}$ . For a certain fatigue loading level, where the crack grows by length  $\Delta a$  in  $\Delta N$  cycles, Paris law reads:

$$\frac{\Delta a}{\Delta N} \approx \frac{da}{dN} = C (\Delta K_{\text{eff}})^m \quad (3)$$

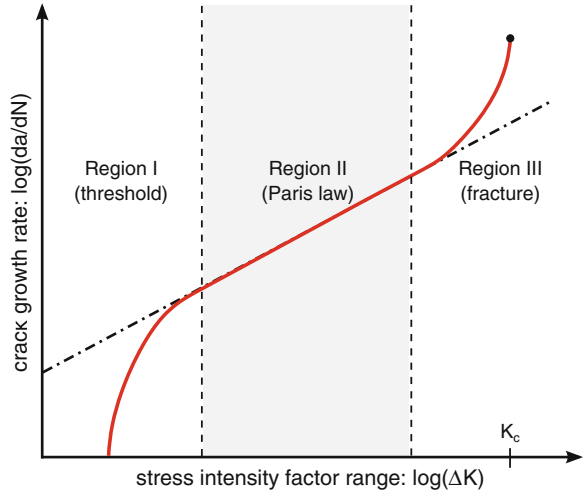
where  $C$  and  $m$  are empirical material constants.  $m$  is often called as the *Paris exponent* and is typically defined in the range of 3–4 for common steel and aluminium alloys. Equation 3 represents a linear relationship between  $\log(\Delta K_{\text{eff}})$  and  $\log(\frac{da}{dN})$  which is used to describe the fatigue crack propagation behavior in region II (see Fig. 2). For calculating the effective mixed-mode stress intensity factor  $\Delta K_{\text{eff}}$ , various criteria have been proposed in the literature. In this study, the energy release rate model has been adopted, leading to:

$$\Delta K_{\text{eff}} = \sqrt{\Delta K_I^2 + \Delta K_{II}^2} \quad (4)$$

and consequently, the number of the corresponding cycles is computed according to [2]:

$$\Delta N = \frac{\Delta a}{C (\Delta K_{\text{eff}})^m} \quad (5)$$

**Fig. 2** Logarithmic crack growth rate and effective region of Paris law



### 2.1.3 Fracture Toughness

Similar to the strength of materials theory where the computed stress is compared with an allowable stress defining the material strength, LEFM assumes that unstable fracture occurs when SIF  $K$  reaches a critical value  $K_c$ , called *fracture toughness*, which represents the potential ability of a material to withstand a given stress field at the crack tip and to resist progressive tensile crack extension. In other words,  $K_c$  is a material constant and is used as a threshold value for SIFs in each pure fracture.

In XFEM, special functions are added to the finite element approximation based on the partition of unity (PU) [3]. Finite element mesh is generated and then additional degrees of freedom are introduced to selected nodes of the finite element model near to the discontinuities in order to provide a higher level of accuracy. Hence quasi-static crack propagation simulations can be carried out without remeshing, by modeling the domain with standard finite elements without explicitly meshing the crack surfaces.

## 2.2 Modeling the Crack Using XFEM

For crack modeling in XFEM, two types of enrichment functions are used: (i) The Heaviside (step) function and (ii) the asymptotic crack-tip enrichment functions taken from LEFM [2]. The displacement field can be expressed as a superposition of the standard  $u^{std}$ , crack-split  $u^H$  and crack-tip  $u^{tip}$  fields as:

$$\mathbf{u}(\mathbf{x}) = \mathbf{u}^{std} + \mathbf{u}^{enr} = \mathbf{u}^{std} + \mathbf{u}^H + \mathbf{u}^{tip} \tag{6}$$



or more explicitly:

$$\mathbf{u}(\mathbf{x}) = \sum_{j=1}^n N_j(\mathbf{x})\mathbf{u}_j + \sum_{h=1}^{n_h} N_h(\mathbf{x})H(\mathbf{x})\mathbf{a}_h + \sum_{k=1}^{n_t} N_k(\mathbf{x}) \left( \sum_{l=1}^{n_f} F_l(\mathbf{x})\mathbf{b}_k^l \right) \quad (7)$$

where  $n$  is the number of nodes in each finite element with standard degrees of freedom  $\mathbf{u}_j$  and shape functions  $N_j(\mathbf{x})$ ,  $n_h$  is the number of nodes in the elements containing the crack face (but not crack tip),  $\mathbf{a}_h$  is the vector of additional degrees of freedom for modeling crack faces by the Heaviside function  $H(\mathbf{x})$ ,  $n_t$  is the number of nodes associated with the crack tip in its influence domain,  $\mathbf{b}_k^l$  is the vector of additional degrees of freedom for modeling crack tips. Finally,  $F_l(\mathbf{x})$  are the crack-tip enrichment functions, given by:

$$\{F_l(r, \theta)\}_{l=1}^4 = \left\{ \sqrt{r} \sin\left(\frac{\theta}{2}\right); \sqrt{r} \cos\left(\frac{\theta}{2}\right); \sqrt{r} \sin\left(\frac{\theta}{2}\right) \sin(\theta); \sqrt{r} \cos\left(\frac{\theta}{2}\right) \sin(\theta) \right\} \quad (8)$$

The elements which are completely cut by the crack, are enriched with the Heaviside (step) function  $H$ . The Heaviside function is a discontinuous function across the crack surface and is constant on each side of the crack. Splitting the domain by the crack causes a displacement jump and Heaviside function gives the desired behavior to approximate the true displacement field.

The first contributing part ( $u^{std}$ ) on the right-hand side of Eq. (7) corresponds to the classical finite element approximation to determine the displacement field, while the second part ( $u^{enr}$ ) refers to the enrichment approximation which takes into account the existence of any discontinuities. This second contributing part utilizes additional degrees of freedom to facilitate modeling of the discontinuous field, such as cracks, without modeling it explicitly.

### 3 The Structural Optimization Problem

Structural optimization problems are characterized by objective and constraint functions that are generally non-linear functions of the design variables. These functions are usually implicit, discontinuous and non-convex. In general there are three classes of structural optimization problems: sizing, shape and topology problems. Structural optimization was focused at the beginning on sizing optimization, such as optimizing cross sectional areas of truss and frame structures, or the thickness of plates and shells and subsequently later, the problem of finding optimum boundaries of a structure and optimizing its shape was also considered. In the former case the structural domain is fixed, while in the latter case it is not fixed but it has a predefined topology.

The mathematical formulation of structural optimization problems can be expressed in standard mathematical terms as a non-linear programming problem, which in general form can be stated as follows:

$$\begin{aligned}
& \text{opt: } F(\mathbf{s}) \\
& \text{subject to: } g_j(\mathbf{s}) \leq 0, \quad j = 1, \dots, k \\
& \quad s_i^{low} \leq s_i \leq s_i^{up}, \quad i = 1, \dots, n
\end{aligned} \tag{9}$$

where  $\mathbf{s}$  is the vector of design variables,  $F(\mathbf{s})$  is the objective function to be optimized (minimized or maximized),  $g_j(\mathbf{s})$  are the behavioral constraint functions, while  $s_i^{low}$  and  $s_i^{up}$  are the lower and upper bounds of the  $i$ th design variable. Due to fabrication limitations the design variables are not always continuous but discrete since cross-sections or dimensions belong to a certain design set. A discrete structural optimization problem can be formulated in the form of Eq. (9) where  $s_i \in \mathfrak{N}_d$ ,  $i = 1, 2, \dots, n$  where  $\mathfrak{N}_d^n$  is a given set of discrete values representing for example the available structural member cross-sections or dimensions and design variables  $\mathbf{s}$  can take values only from this set.

### 3.1 Shape Optimization

In structural shape optimization problems the aim is to improve the performance of the structural component by modifying its boundaries [4, 6, 28, 60]. All functions are related to the design variables, which are coordinates of key points in the boundary of the structure. The shape optimization methodology proceeds with the following steps:

- (i) At the outset of the optimization, the geometry of the structure under investigation has to be defined. The boundaries of the structure are modeled using cubic B-splines that, are defined by a set of key points. Some of the coordinates of these key points will be considered as design variables.
- (ii) An automatic mesh generator is used to create the finite element model. A finite element analysis is carried out and displacements, stresses are calculated.
- (iii) The optimization problem is solved; the design variables are improved and the new shape of the structure is defined. If the convergence criteria for the search algorithm are satisfied, then the optimized solution has been found and the process is terminated, else a new geometry is defined and the whole process is repeated from step (ii).

### 3.2 XFEM Shape Optimization Considering Uncertainties

In this study, two problem formulations are considered, a deterministic and a probabilistic one. According to the deterministic formulation, the goal is to minimize the material volume expressed by optimized geometry of the structural component subject to constraints related to the minimum service life allowed (calculated using fatigue cycles as described in Sect. 2.1.2).

### 3.2.1 Deterministic Formulation (DET)

The design problem for the deterministic formulation (DET) is defined as:

$$\begin{aligned}
 & \min: V(\mathbf{s}) \\
 & \text{subject to: } N(\mathbf{s}) \geq N_{min} \\
 & \quad s_i^{low} \leq s_i \leq s_i^{up}, \quad i = 1, 2, \dots, n
 \end{aligned} \tag{10}$$

$V$  is the volume of the structural component,  $s_i$  are the shape design variables with lower and upper limits  $s_i^{low}$  and  $s_i^{up}$ , respectively, and  $N$  is the service life in terms of number of fatigue cycles with the lower limit of  $N_{min}$ .

### 3.2.2 Probabilistic Formulation (PROB)

The probabilistic design problem (PROB) is defined as:

$$\begin{aligned}
 & \min: V(\mathbf{s}) \\
 & \text{subject to: } \bar{N}(\mathbf{s}, \mathbf{x}) \geq N_{min} \\
 & \quad s_i^{low} \leq s_i \leq s_i^{up}, \quad i = 1, 2, \dots, n \\
 & \quad x_j \sim f_x(\mu_x, \sigma_x^2) \quad j = 1, 2, \dots, nr
 \end{aligned} \tag{11}$$

where  $\mathbf{s}$  and  $\mathbf{x}$  are the vectors of the design and random variables, respectively,  $\bar{N}$  is the mean number of fatigue cycles.

The probabilistic quantity  $\bar{N}$  of Eq. (11) is calculated by means of the Latin hypercube sampling (LHS) method. LHS was introduced by McKay et al. [49] in an effort to reduce the required computational cost of purely random sampling methodologies. LHS can generate variable number of samples well distributed over the entire range of interest. A Latin hypercube sample is constructed by dividing the range of each of the  $nr$  uncertain variables into  $M$  non-overlapping segments of equal marginal probability. Thus, the whole parameter space, consisted of  $M$  parameters, is partitioned into  $M^{nr}$  cells. A single value is selected randomly from each interval, producing  $M$  sample values for each input variable. The values are randomly matched to create  $M$  sets from the  $M^{nr}$  space with respect to the density of each interval for the  $M$  simulations.

## 4 Metaheuristic Search Algorithms

Heuristic algorithms are based on trial-and-error, learning and adaptation procedures in order to solve problems. Metaheuristic algorithms achieve efficient performance for a wide range of combinatorial optimization problems. Four metaheuristic

algorithms that are based on the evolution process are used in the framework of this study. In particular evolution strategies (ES), covariance matrix adaptation (CMA), elitist covariance matrix adaptation (ECMA) and differential evolution (DE) are employed. Details on ES, CMA and ECMA can be found in the work by Lagaros [42], while the version of DE implemented in this study is briefly outlined below.

#### 4.1 Evolution Strategies (ES)

Evolutionary strategies are population-based probabilistic direct search optimization algorithm gleaned from principles of Darwinian evolution. Starting with an initial population of  $\mu$  candidate designs, an offspring population of  $\lambda$  designs is created from the parents using variation operators. Depending on the manner in which the variation and selection operators are designed and the spaces in which they act, different classes of ES have been proposed. In ES algorithm employed in this study [55, 59], each member of the population is equipped with a set of parameters:

$$\begin{aligned} \mathbf{a} &= [(s_d, \boldsymbol{\gamma}), (s_c, \boldsymbol{\sigma}, \boldsymbol{\alpha})] \in (I_d, I_c) \\ I_d &= D^{n_d} \times R_+^{n_\gamma} \\ I_c &= D^{n_c} \times R_+^{n_\sigma} \times [-\pi, \pi]^{n_\alpha} \end{aligned} \quad (12)$$

where  $s_d$  and  $s_c$ , are the vectors of discrete and continuous design variables defined in the discrete and continuous design sets  $D^{n_d}$  and  $R^{n_c}$ , respectively. Vectors  $\boldsymbol{\gamma}$ ,  $\boldsymbol{\sigma}$  and  $\boldsymbol{\alpha}$ , are the distribution parameter vectors taking values in  $R_+^{n_\gamma}$ ,  $R_+^{n_\sigma}$  and  $[-\pi, \pi]^{n_\alpha}$ , respectively. Vector  $\boldsymbol{\gamma}$  corresponds to the variances of the Poisson distribution. Vector  $\boldsymbol{\sigma} \in R_+^{n_\sigma}$  corresponds to the standard deviations ( $1 \leq n_\sigma \leq n_c$ ) of the normal distribution. Vector  $\boldsymbol{\alpha} \in [-\pi, \pi]^{n_\alpha}$  is related to the inclination angles ( $n_\alpha = (n_c - n_\sigma/2)(n_\sigma - 1)$ ) defining linearly correlated mutations of the continuous design variables  $s_d$ , where  $n = n_d + n_c$  is the total number of design variables.

Let  $P(t) = \{a_1, \dots, a_\mu\}$  denotes a population of individuals at the  $t$ -th generation. The genetic operators used in the ES method are denoted by the following mappings:

$$\begin{aligned} rec: (I_d, I_c)^\mu &\longrightarrow (I_d, I_c)^\lambda \quad (\text{recombination}) \\ mut: (I_d, I_c)^\lambda &\longrightarrow (I_d, I_c)^\lambda \quad (\text{mutation}) \\ sel_\mu^k: (I_d, I_c)^k &\longrightarrow (I_d, I_c)^\mu \quad (\text{selection, } k \in \{\lambda, \mu + \lambda\}) \end{aligned} \quad (13)$$

A single iteration of the ES, which is a step from the population  $P_p^t$  to the next parent population  $P_p^{t+1}$  is modeled by the mapping:

$$opt_{EA}: (I_d, I_c)_t^\mu \longrightarrow (I_d, I_c)_{t+1}^\mu \quad (14)$$

**Algorithm 1** ES algorithm

---

```

1:  $t = 0$ 
2: initialize( $P(t = 0)$ )
3: evaluate( $P(t = 0)$ )
4: repeat:
5:    $P_p(t) = \text{selectBest}(\mu, P(t))$ 
6:    $P_c(t) = \text{reproduce}(\lambda, P_p)$ 
7:   mutate( $P_c(t)$ )
8:   evaluate( $P_c(t)$ )
9:   if UsePlusStrategy then
10:     $P(t + 1) = P_c(t) \cup P(t)$ 
11:   else
12:     $P(t + 1) = P_c(t)$ 
13:   end if
14:    $t = t + 1$ 
15: until isNotTerminated()

```

---

**4.2 Covariance Matrix Adaptation (CMA)**

The covariance matrix adaptation, proposed by Hansen and Ostermeier [30] is a completely de-randomized self-adaptation scheme. First, the covariance matrix of the mutation distribution is changed in order to increase the probability of producing the selected mutation step again. Second, the rate of change is adjusted according to the number of strategy parameters to be adapted. Third, under random selection the expectation of the covariance matrix is stationary. Further, the adaptation mechanism is inherently independent of the given coordinate system. The transition from generation  $g$  to  $g + 1$ , given in the following steps, completely defines the Algorithm 2.

*Generation of offsprings.* Creation of  $\lambda$  new offsprings as follows:

$$s_k^{g+1} \sim N(\mathbf{m}^{(g)}, \sigma^{(g)^2} \mathbf{C}^{(g)}) \sim \mathbf{m}^{(g)} + \sigma^{(g)} N(\mathbf{0}, \mathbf{C}^{(g)}) \quad (15)$$

where  $s_k^{g+1} \in \mathfrak{R}^n$  is the design vector of the  $k$ th offspring in generation  $g + 1$ , ( $k = 1, \dots, \lambda$ ),  $N(\mathbf{m}^{(g)}, \mathbf{C}^{(g)})$  are normally distributed random numbers where  $\mathbf{m}^{(g)} \in \mathfrak{R}^n$  is the mean value vector and  $\mathbf{C}^{(g)}$  is the covariance matrix while  $\sigma^{(g)} \in \mathfrak{R}_+$  is the global step size. To define a generation step, the new mean value vector  $\mathbf{m}^{(g+1)}$ , global step size  $\sigma^{(g+1)}$ , and covariance matrix  $\mathbf{C}^{(g+1)}$  have to be defined.

*New mean value vector.* After selection scheme  $(\mu, \lambda)$  operates over the  $\lambda$  offsprings, the new mean value vector  $\mathbf{m}^{(g+1)}$  is calculated according to the following expression:

$$\mathbf{m}^{(g+1)} = \sum_{i=1}^{\mu} w_i \mathbf{s}_{i:\lambda}^{(g+1)} \quad (16)$$

**Algorithm 2** CMA algorithm

- 
- 1: initialize  $\lambda, \mu, w_{i=1, \dots, \mu}, \mu_{\text{eff}}, c_\sigma, d_\sigma, c_c, \mu_{\text{cov}}, c_{\text{cov}}$
  - 2: initialize  $\mathbf{C}(t) \in \mathbb{R}^n, \mathbf{m}(t) = \text{ones}(n \times 1), \mathbf{p}(t) = \text{zeros}(n \times 1)$
  - 3: **repeat**:
  - 4:  $\mathbf{x}_i(t) \sim N(\mathbf{m}(t), \sigma^2(t)\mathbf{C}(t))$  for  $i = 1, \dots, \lambda$
  - 5:  $\mathbf{m}(t+1) = \sum_{i=1}^{\mu} w_i \mathbf{x}_i(t)$
  - 6:  $\mathbf{p}_c(t) = (1 - c_c)\mathbf{p}_c(t-1) + \sqrt{c_c(2 - c_c)\mu_{\text{eff}}} \left( \frac{\mathbf{m}(t+1) - \mathbf{m}(t)}{\sigma(t)} \right)$
  - 7:  $\mathbf{C}(t+1) = (1 - c_{\text{cov}})\mathbf{C}(t) + c_{\text{cov}} \left( 1 - \frac{1}{\mu_{\text{cov}}} \sum_{i=1}^{\mu} w_i \text{OP} \left( \frac{\mathbf{x}_i(t) - \mathbf{m}(t)}{\sigma(t)} \right) + \frac{c_{\text{cov}}}{\mu_{\text{cov}}} \text{OP}(\mathbf{p}_c(t)) \right)$
  - 8:  $\mathbf{p}_\sigma(t) = (1 - c_\sigma)\mathbf{p}_\sigma(t-1) + \sqrt{c_\sigma(2 - c_\sigma)\mu_{\text{eff}}}\mathbf{C}(t)^{-\frac{1}{2}} \frac{\mathbf{m}(t+1) - \mathbf{m}(t)}{\sigma(t)}$
  - 9:  $\sigma(t+1) = \sigma(t) \exp \left( \frac{c_\sigma}{d_\sigma} \left( \frac{\|\mathbf{p}_\sigma(t)\|}{E\|N(0, 1)\|} - 1 \right) \right)$
  - 10: **until** stopping criterion is met
- 

where  $\mathbf{s}_{i:\lambda}^{(g+1)}$  is the  $i$ th best offspring and  $w_i$  are the weight coefficients.

*Global step size.* The new global step size is calculated according to the following expression:

$$\sigma^{(g+1)} = \sigma^{(g)} \exp \left( \frac{c_\sigma}{d_\sigma} \left( \frac{\|\mathbf{p}_\sigma^{(g+1)}\|}{E\|N(0, 1)\|} - 1 \right) \right) \quad (17)$$

while the matrix  $\mathbf{C}^{(g)-\frac{1}{2}}$  is given by:

$$\mathbf{C}^{(g)-\frac{1}{2}} = \mathbf{B}^{(g)} \mathbf{D}^{(g)-1} \mathbf{B}^{(g)T} \quad (18)$$

where the columns of  $\mathbf{B}^{(g)}$  are an orthogonal basis of the eigenvectors of  $\mathbf{C}^{(g)}$  and the diagonal elements of  $\mathbf{D}^{(g)}$  are the square roots of the corresponding positive eigenvalues.

*Covariance matrix update.* The new covariance matrix  $\mathbf{C}^{(g+1)}$  is calculated from the following equation:

$$\begin{aligned} \mathbf{C}^{(g+1)} &= (1 - c_{\text{cov}})\mathbf{C}^{(g)} + \frac{c_{\text{cov}}}{\mu_{\text{cov}}} \mathbf{p}_c^{(g+1)} \mathbf{p}_c^{(g+1)T} \\ &+ c_{\text{cov}} \left( 1 - \frac{1}{\mu_{\text{cov}}} \right) \sum_{i=1}^{\mu} w_i \text{OP} \left( \frac{\mathbf{s}_{i:\lambda}^{(g+1)} - \mathbf{m}^{(g)}}{\sigma^{(g)}} \right) \end{aligned} \quad (19)$$

*OP* denotes the outer product of a vector with itself and  $\mathbf{p}_c^{(g)} \in \mathbb{R}^n$  is the evolution path ( $\mathbf{p}_c^{(0)} = \mathbf{0}$ ).

### 4.3 Elitist Covariance Matrix Adaptation (ECMA)

Elitist CMA evolution strategies algorithm is a combination of the well-known  $(1 + \lambda)$ -selection scheme of evolution strategies [55], with covariance matrix adaptation [35]. The original update rule of the covariance matrix is applied to the  $(1 + \lambda)$ -selection while the cumulative step size adaptation (path length control) of the CMA( $\mu/\mu, \lambda$ ) is replaced by a success rule based step size control. Every individual  $a$  of the ECMA algorithm is comprised of five components:

$$a = \{s, \bar{p}_{succ}, \sigma, \mathbf{p}_c, \mathbf{C}\} \tag{20}$$

where  $s$  is the design vector,  $\bar{p}_{succ}$  is a parameter that controls the success rate during the evolution process,  $\sigma$  is the step size,  $\mathbf{p}_c$  is the evolution path and  $\mathbf{C}$  is the covariance matrix. Contrary to CMA, each individual has its own step size  $\sigma$ , evolution path  $\mathbf{p}_c$  and covariance matrix  $\mathbf{C}$ . A pseudo code of the ECMA algorithm

---

**Algorithm 3**  $(1 + \lambda)$ -ECMA

---

- 1:  $\mathbf{g} = \mathbf{0}$ , initialize  $a_{parent}^{(g)}$
  - 2: **repeat**
  - 3:  $a_{parent}^{(g+1)} \leftarrow a_{parent}^{(g+1)}$
  - 4: **for**  $k = 1, \dots, \lambda$  **do**
  - 5:  $s_k^{(g+1)} \sim N(\mathbf{s}_{parent}^{(g)}, \sigma^{(g)^2} \mathbf{C}^{(g)})$
  - 6: **end for**
  - 7:  $UpdateStepSize\left(a_{parent}^{(g+1)}, \frac{\lambda_{succ}^{(g+1)}}{\lambda}\right)$
  - 8: **if**  $f(s_{1:\lambda}^{(g+1)}) < f(s_{parent}^{(g)})$  **then**
  - 9:  $\mathbf{x}_{parent}^{(g+1)} \leftarrow \mathbf{x}_{1:\lambda}^{(g+1)}$
  - 10:  $UpdateCovariance\left(a_{parent}^{(g+1)}, \frac{s_{parent}^{(g+1)} - s_{parent}^{(g)}}{\sigma_{parent}^{(g)}}\right)$
  - 11: **end if**
  - 12: **until** stopping criterion is met
- 

is shown in Algorithm 3. In line #1 a new parent  $a_{parent}^{(g)}$  is generated. In lines #4–6,  $\lambda$  new offsprings are generated from the parent vector  $a_{parent}^{(g)}$ . The new offsprings are sampled according to Eq. (8), with variable  $\mathbf{m}^{(g)}$  being replaced by the design vector  $s_{parent}^{(g)}$  of the parent individual. After the  $\lambda$  new offsprings are sampled, the parent’s step size is updated by means of *UpdateStepSize* subroutine (see Procedure 4). The arguments of the subroutine are the parent  $a_{parent}^{(g)}$  and the success rate  $\lambda_{succ}^{(g+1)} / \lambda$ , where  $\lambda_{succ}^{(g+1)}$  is the number of offsprings having better fitness function than the parent.

The step size update is based upon the 1/5 success rule, thus when the ratio  $\lambda_{succ}^{(g+1)}/\lambda$  is larger than 1/5 step size increases, otherwise step size decreases. If the best offspring has a better fitness value than the parent, it becomes the parent of the next generation (see lines #8–9), and the covariance matrix of the new parent is updated by means of *UpdateCovariance* subroutine (see Procedure 5). The arguments of the subroutine are the current parent and the step change:

$$\frac{s_{parent}^{(g+1)} - s_{parent}^{(g)}}{\sigma_{parent}^{(g)}} \quad (21)$$

The update of the evolution path and the covariance matrix depends on the success rate:

$$\bar{p}_{succ} = \frac{\lambda_{succ}}{\lambda} \quad (22)$$

If the success rate is below a given threshold value  $p_{thresh}$  then the step size is taken into account and the evolution path and the covariance matrix is updated (see lines #2–3 of Procedure 5). If the success rate is above the given threshold  $p_{thresh}$  the step change is not taken into account and evolution path and covariance matrix happens are updated (see lines #5–6).

---

**Procedure 4** *UpdateSizeState* ( $a = \{s, \bar{p}_{succ}, \sigma, \mathbf{p}_c, \mathbf{C}\}, p_{succ}$ )

---

- 1:  $\bar{p}_{succ} \leftarrow (1 - c_p)\bar{p}_{succ} + c_p p_{succ}$
  - 2:  $\sigma \leftarrow \sigma \exp\left(\frac{1}{d} \left(\bar{p}_{succ} - \frac{p_{succ}^{target}}{1 - p_{succ}^{target}}(1 - \bar{p}_{succ})\right)\right)$
- 

---

**Procedure 5** *UpdateCovariance* ( $a = \{s, \bar{p}_{succ}, \sigma, \mathbf{p}_c, \mathbf{C}\}, s_{step} \in \mathbb{R}^n$ )

---

- 1: **if**  $\bar{p}_{succ} < p_{thresh}$  **then**
  - 2:  $\mathbf{p}_c \leftarrow (1 - c_c)\mathbf{p}_c + \sqrt{c_c(2 - c_c)}\mathbf{x}_{step}$
  - 3:  $\mathbf{C} \leftarrow (1 - c_{cov})\mathbf{C} + c_{cov}\mathbf{p}_c\mathbf{p}_c^T$
  - 4: **else**
  - 5:  $\mathbf{p}_c \leftarrow (1 - c_c)\mathbf{p}_c$
  - 6:  $\mathbf{C} \leftarrow (1 - c_{cov})\mathbf{C} + c_{cov}(\mathbf{p}_c\mathbf{p}_c^T + c_c(2 - c_c)\mathbf{C})$
  - 7: **end if**
-



### 4.4 Differential Evolution (DE)

Storn and Price [62] proposed a floating point evolutionary algorithm for global optimization and named it differential evolution (DE), by implementing a special kind operator in order to create offsprings from parent vectors. Several variants of DE have been proposed so far [9]. According to the variant implemented in the current study, a donor vector  $\mathbf{v}_{i,g+1}$  is generated first:

$$\mathbf{v}_{i,g+1} = \mathbf{s}_{r_1,g} + F \cdot (\mathbf{s}_{r_2,g} - \mathbf{s}_{r_3,g}) \tag{23}$$

Integers  $r_1, r_2$  and  $r_3$  are chosen randomly from the interval  $[1, NP]$  while  $i \neq r_1, r_2$  and  $r_3$ .  $NP$  is the population size and  $F$  is a real constant value, called the *mutation factor*. In the next step the crossover operator is applied by generating the trial vector  $\mathbf{u}_{i,g+1}$  which is defined from the elements of  $\mathbf{s}_{i,g}$  or  $\mathbf{v}_{i,g+1}$  with probability  $CR$ :

$$u_{j,i,g+1} = \begin{cases} v_{j,i,g+1}, & \text{if } \text{rand}_{i,j} \leq CR \text{ or } j = I_{\text{rand}} \\ s_{j,i,g}, & \text{if } \text{rand}_{i,j} \geq CR \text{ or } j \neq I_{\text{rand}} \end{cases} \tag{24}$$

$i = 1, 2, \dots, NP$  and  $j = 1, 2, \dots, n$

where  $\text{rand}_{j,i} \sim U[0, 1]$ ,  $I_{\text{rand}}$  is a random integer from  $[1, 2, \dots, n]$  which ensures that  $\mathbf{v}_{i,g+1} \neq \mathbf{s}_{i,g}$ . The last step of the generation procedure is the implementation of the selection operator where the vector  $\mathbf{s}_{i,g}$  is compared to the trial vector  $\mathbf{u}_{i,g+1}$ :

$$\mathbf{s}_{i,g+1} = \begin{cases} \mathbf{u}_{i,g+1}, & \text{if } F(\mathbf{u}_{i,g+1}) \leq F(\mathbf{s}_{i,g}) \\ \mathbf{s}_{i,g}, & \text{otherwise} \end{cases} \tag{25}$$

$i = 1, 2, \dots, NP$

where  $F(\mathbf{s})$  is the objective function to be optimized, while without loss of generality the implementation described in Eq. (25) corresponds to minimization.

---

**Algorithm 6** Classical DE

---

- 1: initialize  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{NP}\} \in \mathfrak{N}^n$
  - 2:  $f_i = f(\mathbf{x}_i)$  for  $i = \{1, \dots, NP\}$
  - 3: **repeat:**
  - 4:    $\mathbf{v}_i = \text{CreateDonor}(\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{NP}\})$ , for  $i = \{1, \dots, NP\}$
  - 5:    $\mathbf{v}_i = \text{Crossover}(\mathbf{x}_i, \mathbf{v}_i)$  for  $i = \{1, \dots, NP\}$
  - 6:    $f_i^{\text{offs}} = f(\mathbf{u}_i)$  for  $i = \{1, \dots, NP\}$
  - 7:    $\{\mathbf{x}_i, f_i\} = \text{Selection}(\mathbf{x}_i, \mathbf{u}_i, f_i, f_i^{\text{offs}})$  for  $i = \{1, \dots, NP\}$
  - 8: **until** stopping criterion is met
-

## 5 Towards the Selection of the Optimization Algorithm

In the past a number of studies have been published where structural optimization with single and multiple objectives are solved implementing metaheuristics. A sensitivity analysis is performed for four metaheuristic algorithms in benchmark multimodal constrained functions highlighting the proper search algorithm for solving the structural optimization problem.

### 5.1 Literature Survey on Metaheuristic Based Structural Optimization

Perez and Behdinan [54] presented the background and implementation of a particle swarm optimization algorithm suitable for constraint structural optimization problems, while improvements that effect of the setting parameters and functionality of the algorithm were shown. Hasançebi [31] investigated the computational performance of adaptive evolution strategies in large-scale structural optimization. Bureerat and Limtragool [5] presented the application of simulated annealing for solving structural topology optimization, while a numerical technique termed as multiresolution design variables was proposed as a numerical tool to enhance the searching performance. Hansen et al. [29] introduced an optimization approach based on an evolution strategy that incorporates multiple criteria by using nonlinear finite-element analyses for stability and a set of linear analyses for damage-tolerance evaluation, the applicability of the approach was presented for the window area of a generic aircraft fuselage. Kaveh and Shahrouzi [36] proposed a hybrid strategy combining indirect information share in ant systems with direct constructive genetic search, for this purpose some proper coding techniques were employed to enable testing the method with various sets of control parameters. Farhat et al. [18] proposed a systematic methodology for determining the optimal cross-sectional areas of buckling restrained braces used for the seismic upgrading of structures against severe earthquakes, for this purpose single-objective and multi-objective optimization problems were formulated. Chen and Chen [7] proposed modified evolution strategies for solving mixed-discrete optimization problems, in particular three approaches were proposed for handling discrete variables.

Gholizadeh and Salajegheh [25] proposed a new metamodeling framework that reduces the computational burden of the structural optimization against the time history loading, for this purpose a metamodel consisting of adaptive neuro-fuzzy inference system, subtractive algorithm, self-organizing map and a set of radial basis function networks were used to accurately predict the time history responses of structures. Wang et al. [65] studied an optimal cost base isolation design or retrofit design method for bridges subject to transient earthquake loads. Hasançebi et al. [32] utilized metaheuristic techniques like genetic algorithms, simulated annealing, evolution strategies, particle swarm optimizer, tabu search, ant colony optimization and

harmony search in order to develop seven optimum design algorithms for real size rigidly connected steel frames. Manan et al. [47] employed four different biologically inspired optimization algorithms (binary genetic algorithm, continuous genetic algorithm, particle swarm optimization, and ant colony optimization) and a simple meta-modeling approach on the same problem set. Gandomi and Yang [21] provide an overview of structural optimization problems of both truss and non-truss cases. Martínez et al. [48] described a methodology for the analysis and design of reinforced concrete tall bridge piers with hollow rectangular sections, which are typically used in deep valley bridge viaducts. Kripakaran et al. [40] presented computational approaches that can be implemented in a decision support system for the design of moment-resisting steel frames, while trade-off studies were performed using genetic algorithms to evaluate the savings due to the inclusion of the cost of connections in the optimization model. Gandomi et al. [22] used the cuckoo search (CS) method for solving structural optimization problems, furthermore, for the validation against structural engineering optimization problems the CS method was applied to 13 design problems taken from the literature.

Kunakote and Bureerat [41] dealt with the comparative performance of some established multi-objective evolutionary algorithms for structural topology optimization, four multi-objective problems, having design objectives like structural compliance, natural frequency and mass, and subjected to constraints on stress, were used for performance testing. Su et al. [63] used genetic algorithm to handle topology and sizing optimization of truss structures, in which a sparse node matrix encoding approach is used and individual identification technique is employed to avoid duplicate structural analysis to save computation time. Gandomi and Yang [21] used firefly algorithm for solving mixed continuous/discrete structural optimization problems, the results of a trade study carried out on six classical structural optimization problems taken from literature confirm the validity of the proposed algorithm. Degertekin [11] proposed two improved harmony search algorithms for sizing optimization of truss structures, while four truss structure weight minimization problems were presented to demonstrate the robustness of the proposed algorithms. The main part of the work by Muc and Muc-Wierzgoń [51] was devoted to the definition of design variables and the forms of objective functions for multi-layered plated and shell structures, while the evolution strategy method was used as the optimization algorithm. Comparative studies of metaheuristics on engineering problems can be found in two recent studies by the authors Lagaros and Karlaftis [43], Lagaros and Papadrakakis [44] and in the edited book by Yang and Koziel [69].

## ***5.2 Sensitivity Analysis of Metaheuristic Algorithms***

Choosing the proper search algorithm for solving an optimization problem is not a straightforward procedure. In this section a sensitivity analysis of four search algorithms is performed for five constrained multimodal benchmark test functions in order to identify the best algorithm and to be used for solving the structural shape

optimization problem studied in the next section. This sensitivity analysis is carried out to examine the efficiency of the four metaheuristic algorithms and thus proving their robustness. In particular, for the solution of the five problems ES, CMA, ECMA and DE methods are implemented, since they were found robust and efficient in previous numerical tests [43, 44]. This should not be considered as an implication related to the efficiency of other algorithms, since any algorithm available can be considered for the solution of the optimization problem based on user's experience.

The control parameters for DE are the population size ( $NP$ ), probability ( $CR$ ) and constant ( $F$ ), while for ES, CMA and ECMA the control parameters are the number of parents ( $\mu$ ) and offsprings ( $\lambda$ ). The characteristic parameters adopted for the implementation are as follows: (i) for DE method, population size  $NP = 15$ , probability  $CR = 0.90$  and constant  $F = 0.60$ , while (ii) for all three ES, CMA and ECMA methods, number of parents  $\mu = 1$  and offsprings  $\lambda = 14$  for the case of ES and ECMA and number of parents  $\mu = 5$  and offsprings  $\lambda = 15$  for the case of CMA.

For all four algorithms the initial population is generated randomly using LHS in the range of design space for each test example examined, while for the implementation of all algorithms, the real valued representation of the design vectors is adopted. For the purposes of the sensitivity analysis 50 independent optimization runs were performed, for the combination of the algorithmic parameters given above. The 50 independent optimization runs, represents a necessary step since non deterministic optimization algorithms do not yield the same results when restarted with the same parameters [57]. Using the optimum objective function values achieved for the 50 independent optimization runs, mean and coefficient of variation of the optimum objective function value are calculated.

For comparative reasons the method adopted for handling the constraints and the termination criterion is the same for all metaheuristic optimization algorithms. In particular, the simple yet effective, multiple linear segment penalty function [44] is used in this study for handling the constraints. According to this technique if no violation is detected, then no penalty is imposed on the objective function. If any of the constraints is violated, a penalty, relative to the maximum degree of constraints' violation, is applied to the objective function, otherwise the optimization procedure is terminated after 10,000 function evaluations. For the results found in the literature and used for our comparative study different constraint handling techniques and termination criteria were implemented.

### 5.2.1 Test Case S-6ACT

The first test case considered in this sensitivity analysis study is the so called S-6ACT [33] problem that is defined as follows:

$$\begin{aligned} \min: \quad F(x) = & x_1^2 + x_2^2 + x_1x_2 - 14x_1 - 16x_2 + (x_3 - 10)^2 + 4(x_4 - 5)^2 \\ & + (x_5 - 3)^2 + 2(x_6 - 1)^2 + 5x_7^2 + 7(x_8 - 11)^2 + 2(x_9 - 10)^2 \\ & + (x_{10} - 7)^2 + 45 \end{aligned}$$

$$\begin{aligned}
 \text{subject to: } & g_1(\mathbf{x}) = 105 - 4x_1 - 5x_2 + 3x_7 - 9x_8 \geq 0 \\
 & g_2(\mathbf{x}) = -10x_1 + 8x_2 + 17x_7 - 2x_8 \geq 0 \\
 & g_3(\mathbf{x}) = 8x_1 - 2x_2 + 17x_7 - 2x_8 \geq 0 \\
 & g_4(\mathbf{x}) = -3(x_1 - 2)^2 - 4(x_2 - 3)^2 - 2x_3^2 + 7x_4 + 120 \geq 0 \\
 & g_5(\mathbf{x}) = -5x_2 - 8x_2 - (x_3 - 6)^2 + 2x_4 + 40 \geq 0 \\
 & g_6(\mathbf{x}) = -x_1^2 - 2(x_2 - 2)^2 + 2x_1x_2 - 14x_5 + 6x_6 \geq 0 \\
 & g_7(\mathbf{x}) = -0.5(x_1 - 8)^2 - 2(x_2 - 4)^2 + 3x_5^2 + x_6 + 30 \geq 0 \\
 & g_8(\mathbf{x}) = 3x_1 - 6x_2 - 12(x_9 - 8)^2 \geq 0 \\
 & -10 \leq x_i \leq 10, \quad i = 1, \dots, 10
 \end{aligned}$$

It is a 10 design variables problem with 8 inequality constraints. As it can be observed in Table 1 the better COV value is achieved by CMA and the worst one by ES algorithm, while the best mean value is obtained by DE algorithm and the worst by ES.

The best optimized designs achieved by the four metaheuristics among the 50 independent optimization runs is given in Table 2. Although, the best optimized design is achieved by CMA and DE algorithm, DE algorithm had slightly better performance with reference to the statistical data of Table 1. It should be noted also that for all 50 independent optimization runs performed for each algorithm, feasible optimized designs were obtained.

### 5.2.2 Test Case S-CRES

This test case problem was proposed by Deb [10] and is formulated with 2 design variables and 2 inequality constraints:

$$\begin{aligned}
 \text{min: } & F(x) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2 \\
 \text{subject to: } & g_1(\mathbf{x}) = 4.84 - (x_1 - 0.05)^2 - (x_2 - 2.5)^2 \geq 0 \\
 & g_2(\mathbf{x}) = x_1^2 + (x_2 - 2.5)^2 - 4.84 \geq 0 \\
 & 0 \leq x_1 \leq 6 \\
 & 0 \leq x_2 \leq 6
 \end{aligned}$$

**Table 1** Results comparison for test case S-6ACT

Algorithm	$\mu$	$\lambda$	Selection	Obj. function		
				Best	Mean	COV (%)
ES	1	14	+	14.962	49.0379	1.56E+02
CMA	5	10	,	14.257	15.1669	8.57E-02
ECMA	1	14	+	14.436	14.2681	5.17E+00
DE				14.257	14.2608	3.42E-01

**Table 2** Results comparison for test case S-6ACT

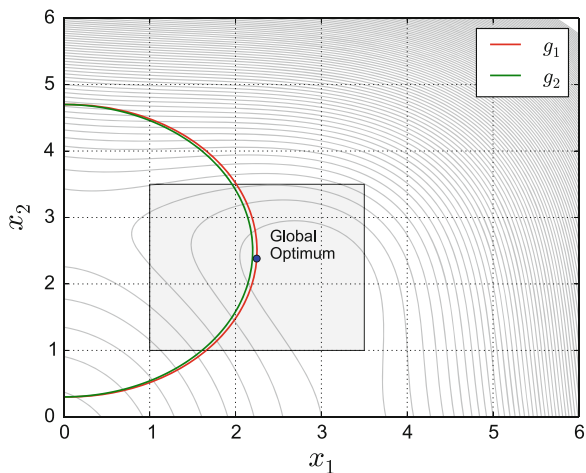
$x$	Deb [10]	ES	CMA	ECMA	DE
$x_1$	2.171996	1.5859	1.576076	1.6996902	1.5760762
$x_2$	2.363683	2.8712	2.731987	2.6947086	2.7319869
$x_3$	8.773926	8.7952	8.791763	8.7832448	8.7917633
$x_4$	5.095984	5.0471	5.059531	4.9932193	5.0595309
$x_5$	0.990655	1.1745	0.976753	1.0675614	0.9767532
$x_6$	1.430574	1.9129	1.436430	1.6072484	1.4364296
$x_7$	1.321644	0.7489	0.783778	0.8738167	0.7837782
$x_8$	9.828726	9.6163	9.709677	9.7054379	9.7096767
$x_9$	8.280092	9.7648	9.774489	9.7654962	9.7744885
$x_{10}$	8.375927	7.1255	7.064255	6.9290318	7.0642553
$F$	24.30621	14.962	14.257	14.436	14.257

In Fig. 3 the feasible and infeasible domain of the problem is shown. The feasible domain is approximately 0.7% of the total search space. The two constraint functions  $g_1, g_2$  create a crescent shape for the feasible domain, as it is shown in Fig. 4 with the zoomed area around the optimal point.

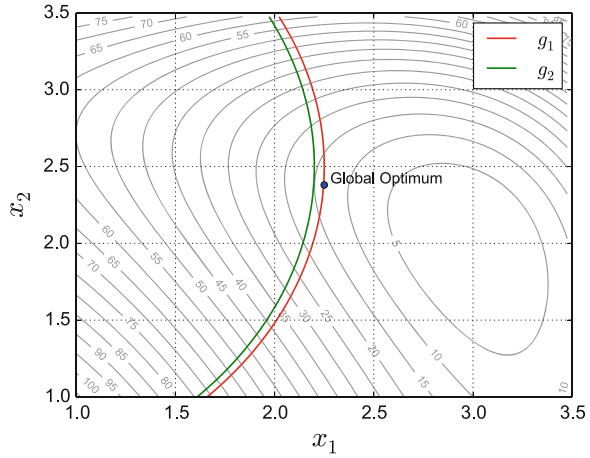
Similar to the previous test case, statistical results (mean value and COV) are given in Table 3. Furthermore, in Table 4, the results are compared with the best result found in literature [10]. It should be noted also that for all 50 independent optimization runs performed for each algorithm, feasible optimized designs were obtained.

The CMA and DE algorithms had better performance, since COV values of the optimized objective function value obtained at the end of the evolution process was orders of magnitude smaller than the one obtained by the other two algorithms.

**Fig. 3** Feasible and infeasible domain for S-CRES problem



**Fig. 4** Enlarged space around the optimal point



**Table 3** Results comparison for test case S-CRES

Algorithm	$\mu$	$\lambda$	Selection	Obj. function		
				Best	Mean	COV (%)
ES	1	14	+	13.59085	13.6897	2.53E+00
CMA	5	15	,	13.59084	13.5909	1.34E-03
ECMA	1	14	+	13.59087	13.6096	3.49E-01
DE				13.59084	13.5957	1.79E-02

**Table 4** Results comparison for test case S-CRES

$x$	Deb [10]	ES	CMA	ECMA	DE
$x_1$	2.246826	2.246841	2.246826	2.246811	2.246826
$x_2$	2.381865	2.382141	2.381865	2.381597	2.381865
$F$	13.59085	13.59085	13.59084	13.59087	13.59084

**5.2.3 Test Case S-0.5F**

The optimization problem S-0.5F [8] is formulated with 7 design variables and 4 inequality constraints:

$$\begin{aligned}
 \text{min: } & F(x) = (x_1 - 10)^2 + 5(x_2 - 12)^2 + x_3^4 + 3(x_4 - 11)^2 \\
 & \quad + 10x_5^6 + 7x_6^2 + x_7^4 - 4x_6x_7 \\
 \text{subject to: } & g_1(\mathbf{x}) = 127 - 2x_1^2 - 3x_2^2 - x_3 - 4x_4^2 - 5x_5^2 \geq 0 \\
 & g_2(\mathbf{x}) = 282 - 7x_1 - 3x_2 - 10x_3^2 - x_4 - x_5 \geq 0 \\
 & g_3(\mathbf{x}) = 196 - 23x_1 - x_2^2 - 6x_6^2 + 8x_7 \geq 0 \\
 & g_4(\mathbf{x}) = -4x_1^2 - x_2^2 + 3x_1x_2 - 2x_3^2 - 5x_6 + 11x_7 \geq 0 \\
 & -10 \leq x_i \leq 10, \quad i = 1, \dots, 7
 \end{aligned}$$

**Table 5** Results for test case S-0.5F

Algorithm	$\mu$	$\lambda$	Selection	Obj. function		
				Best	Mean	COV (%)
ES	1	14	+	680.7721	705.7945	1.24E+01
CMA	5	15	,	680.6301	680.6301	1.20E-07
ECMA	1	14	+	680.6848	681.5228	6.33E-02
DE				680.6301	680.6551	1.67E-01

**Table 6** Results comparison for test case S-0.5F

$x$	Deb [10]	ES	CMA	ECMA	DE
$x_1$	2.330499	2.320378	2.330501	2.299430	2.330501
$x_2$	1.951372	1.967625	1.951373	1.947076	1.951373
$x_3$	-0.477541	-0.281803	-0.477539	-0.468747	-0.477539
$x_4$	4.365723	4.319129	4.365723	4.382807	4.365723
$x_5$	-0.624487	-0.615799	-0.624484	-0.611883	-0.624484
$x_6$	1.038131	1.057470	1.038125	1.001823	1.038125
$x_7$	1.594227	1.560759	1.594225	1.541608	1.594225
$F$	680.63	680.77	680.63	680.69	680.63

In this problem, only 0.5% of the space is feasible. Similar to the previous test functions for all 50 independent optimization runs performed for each algorithm, feasible optimized designs were obtained. Statistical results (mean value and COV) are given in Table 5. Table 6 shows that even though all algorithms managed to locate the optimal design domain, only CMA and DE algorithms found the global optimum design. CMA algorithm had the best performance, since COV value of the optimized objective function is almost zero.

### 5.2.4 Test Case S-HIM

The optimization problem S-HIM [8] is formulated with 5 design variables and 6 inequality constraints:

$$\begin{aligned}
 \text{min: } & F(\mathbf{x}) = 5.3578547x_3^2 + 0.8356891x_1x_5 + 37.293239x_1 - 40792.141 \\
 \text{subject to: } & g_1(\mathbf{x}) = 85.334407 + 0.0056858x_2x_5 + 0.0006262x_1x_4 \\
 & \quad - 0.0022053x_3x_5 \geq 0 \\
 & g_2(\mathbf{x}) = 92 - g_1(\mathbf{x}) \geq 0 \\
 & g_3(\mathbf{x}) = 80.51249 + 0.0071317x_2x_5 + 0.0029955x_1x_2 \\
 & \quad + 0.0021813x_3^2 - 90 \geq 0 \\
 & g_4(\mathbf{x}) = 20 - g_3(\mathbf{x}) \geq 0
 \end{aligned}$$



**Table 7** Results for test case S-HIM

Algorithm	$\mu$	$\lambda$	Selection	Obj. function		
				Best	Mean	COV (%)
ES	1	14	+	-30665.5	-30190.0	1.81E+00
CMA	5	15	,	-25273.7	-24258.6	2.07E+01
ECMA	1	14	+	-30665.5	-30477.6	8.45E-01
DE				-30665.5	-30700.5	2.68E-01

**Table 8** Results comparison for test case S-HIM

$x$	ES	CMA	ECMA	DE
$x_1$	78.000	78.000	78.000	78.000
$x_2$	33.000	33.000	33.000	33.000
$x_3$	29.996	45.000	29.995	29.995
$x_4$	45.000	45.000	45.000	45.000
$x_5$	36.776	27.000	36.776	36.776
$F$	-30665.5	-25272.7	-30665.5	-30665.5

$$g_5(\mathbf{x}) = 9.300961 + 0.0047026x_3x_5 + 0.0012547x_1x_3 + 0.0019085x_3x_4 - 20 \geq 0$$

$$g_6(\mathbf{x}) = 5 - g_4(\mathbf{x}) \geq 0$$

$$78 \leq x_1 \leq 102$$

$$33 \leq x_2 \leq 45$$

$$-27 \leq x_i \leq 45, \quad i = 3, 4, 5$$

Statistical results (mean value and COV) are given in Table 7. The optimal value of the objective function value is equal to -31005.7966 [1], which was achieved after 350,000 function evaluations. From Table 8 is shown that for all 50 independent optimization runs performed only for ES, CMA and DE algorithms, feasible optimized designs were obtained. In contrast to the previous test functions, CMA algorithm failed to identify the area of the optimal solution.

### 5.2.5 Test Case S-G08

The optimization problem S-G08 [1] is formulated with 2 design variables and 2 inequality constraints:

$$\min: F(x) = \frac{\sin(2\pi x_1)^3 \sin(2\pi x_2)}{x_1^3(x_1 + x_2)}$$

$$\text{subject to: } g_1(\mathbf{x}) = x_1^2 - x_2 + 1 \geq 0$$

$$g_2(\mathbf{x}) = 1 - x_1 + (x_2 - 4)^2 \geq 0$$

$$= 1 \leq x_1 \leq 3$$

$$= 1 \leq x_2 \leq 5$$

**Table 9** Results for test case S-G08

Algorithm	$\mu$	$\lambda$	Selection	Obj. function		
				Best	Mean	COV (%)
ES	1	14	+	-0.10546	-0.08413	4.14E+01
CMA	5	15	,	-0.10546	-0.06765	1.49E+02
ECMA	1	14	+	-0.10546	-0.10398	4.11E+01
DE				-0.10566	-0.10546	7.02E-01

**Fig. 5** Design variables domain for test case S-G08

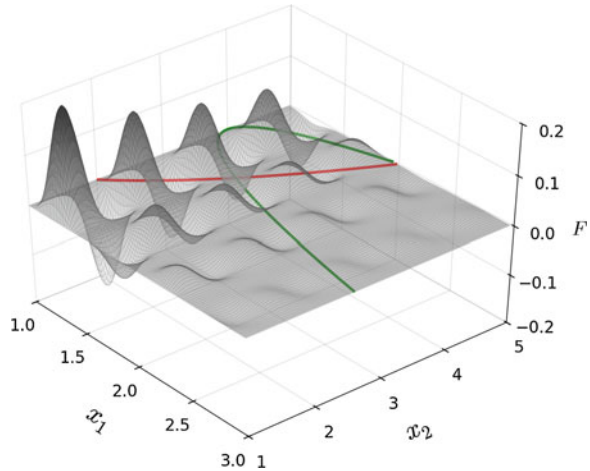
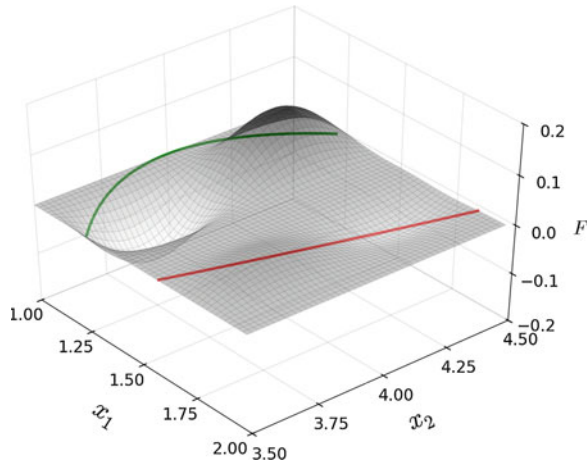


Figure 5 depicts the search space, while Fig. 6 depicts the area around the optimal solution found in the literature. Similar to the previous test case, statistical results (mean value and standard deviation) are given in Table 9. The DE algorithm had better performance, since COV value of the optimized objective function value obtained at the end of the evolution process was orders of magnitude smaller than that obtained for the other three algorithms. The optimal value of the objective function found in the literature is equal to  $-0.09582$  [1], achieved after 350,000 function evaluations. Similar to the previous test functions, in Table 10 is shown that for all algorithms feasible optimized designs were obtained.

### 5.3 Selection of the Appropriate Search Algorithm

The sensitivity of the four algorithms with respect to different optimization runs characterized by the mean and coefficient of variation of the optimized objective function values for each metaheuristic algorithm was identified in the corresponding tables of Sect. 5.2. The lower mean and COV values are, the better the algorithm is.

**Fig. 6** Domain around global minimum for test case S-G08



**Table 10** Results comparison for test case S-G08

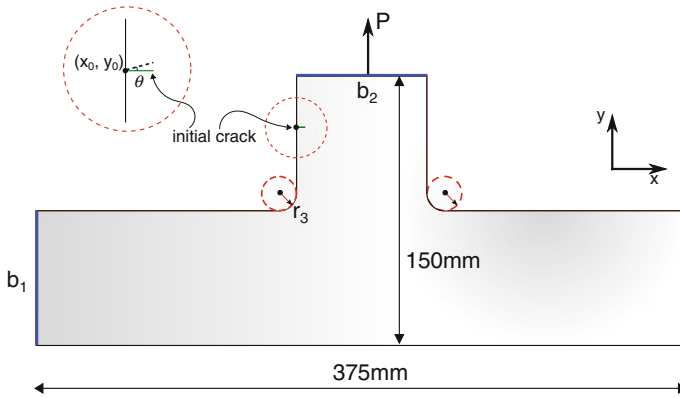
$x$	Aguirre et al. [1]	ES	CMA	ECMA	DE
$x_1$	1.227971	1.227818	1.227818	1.227818	1.227817
$x_2$	4.245373	3.744911	3.744911	3.744911	3.744911
$F$	-0.09582	-0.10546	-0.10546	-0.10546	-0.10546

This is due to the fact that low COV values mean that the algorithm is not influenced by the independent runs. Overall, the algorithm resulting to the lower mean value (in case of minimization problem) and COV is used for performing the optimization run with the specific algorithm, i.e. the DE algorithm.

## 6 Numerical Examples

A fillet from a steel structural member [61] is analyzed in this section to illustrate the capabilities of the proposed methodology described in the previous sections of this study. The geometry, loading conditions, and design variables of the structural component are shown in Fig. 7. Four-node linear quadrilateral elements under plane stress conditions with constant thickness equal to 5 mm and isotropic material properties are assumed. For the purposes of this study two boundary conditions are considered; in the first one, designated as *fillet rigid*, all nodes of the bottom edge are fixed while in the second one, denoted as *fillet flexible*, only the two end nodes of the bottom edge of the component are fixed.

For both test examples deterministic and probabilistic shape optimization problems are solved. The objective function to be minimized, corresponds to the material volume while two sets of constraints are enforced, i.e. deterministic and probabilistic constraints on the fatigue cycles. Furthermore, due to manufacturing limitations



**Fig. 7** Fillet geometry, loading and design variables of the problem

**Table 11** Upper and lower bounds of design variables and corresponding steps (in mm)

Design variables	$l_{up}$	$l_{low}$	Step
$b_1$	100.0	50.0	1.0
$b_2$	100.0	50.0	1.0
$r_3$	30.0	10.0	1.0

the design variables are treated as discrete in the same way as in a single objective design optimization problems with the discrete version of Evolution Strategies [45]. The design variables correspond to the dimensions of the structural component taken from Table 11. The design load  $P$  (see Fig. 7), is applied as a concentrated tensile load at the midpoint of the top edge and is equal to 20 KN.

It is common in probabilistic analysis to distinguish between uncertainty that reflects the variability of the outcome of a repeatable experiment and uncertainty due to ignorance. The last one is sometimes referred as “randomness”, commonly known as “aleatoric uncertainty”, which cannot be reduced. However, both deterministic and probabilistic approaches rely on various model assumptions and model parameters that are based on the current state of knowledge on the behavior of structural systems under given conditions. There is uncertainty associated with these conditions, which depends upon the state of knowledge that is referred as “epistemic uncertainty”.

In this study various sources of uncertainty are considered: on crack tip initialization (aleatoric randomness) which influences the shape of the crack propagation path and on modeling (epistemic uncertainty) which affects the structural capacity. The structural stiffness is directly connected to the Young modulus  $E$ , of structural steel, while the number of fatigue cycles is influenced by the material properties  $C$  and  $m$ . The crack length increment  $\Delta a$  and the poisson ratio are taken equal to 5.0 mm and 0.3, respectively, both implemented as deterministic. Thus, for the structural component five random variables are used, i.e. the ordinate  $y_0$  of the crack tip initialization and the corresponding angle  $\theta$  along with the Young modulus  $E$  and parameters  $C$ ,  $m$ . The material properties for the structural steel of the component are implemented

**Table 12** Random variables with the type of distribution and each statistical parameters: mean value ( $\mu$ ) and standard deviation ( $\sigma$ )

Random variables	$\mu$	$\sigma$	cov (%)	Distribution type
$y_0$ (in mm)	$(150 - b_1)/(2 + b_1)$	–	5	Normal
$\theta$ (in °)	0.0	0.50	–	Normal
$E$ (in GPa)	207.0	35.19	17	Lognormal
$C$	2.45e-11	4.16e-12	17	Lognormal
$m$	2.37	0.40	17	Lognormal

as independent random variables whose characteristics were selected according to Ellingwood et al. [16], Ellingwood and Galambos [15] and are given in Table 12.

The numerical study that follows comprises of two parts: in the first part a parametric investigation is performed in order to find the number of simulations required for computational efficiency and robustness regarding the calculation of the statistical quantities required and the identification of the most appropriate one that can be used in order to characterize the influence of randomness on the fatigue cycles. In the second part, the performance of structural components under fatigue is investigated within a probabilistic shape design optimization framework.

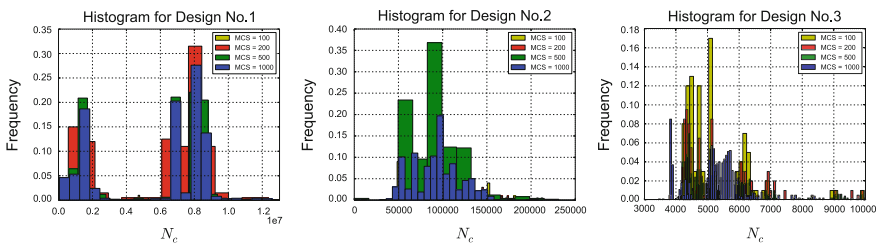
## 6.1 Parametric Investigation

For the purpose of this parametric investigation the fillet rigid case is examined and three designs, corresponding to the upper (Design 1), lower (Design 3) bounds of the designs variables and an intermediate one (Design 2) are chosen. The scope of this investigation is to find the lower number of simulations for a reliable calculation of certain statistical quantities that are related to the number of fatigue cycles. To this end, Monte Carlo (MC) simulations based on LHS are performed for the three designs described above and the mean, median and standard deviation of the number of fatigue cycles are calculated (see Table 13).

The performance of the different number of MC simulations is depicted in the histograms of Fig. 8. For the needs of this investigation, the three designs are subjected to the ensemble of different number of simulations (100 + 200 + 500 + 1000). Thus, 5400 XFEM analyses have been postprocessed for the three designs in order to create a response databank with the quantities of interest. The propagation of uncertainties is performed by means of the MC simulation method in connection to the LHS technique which has been incorporated into the XFEM framework as described above. According to LHS a given design is run repeatedly, for each MC simulation using different values for the uncertain parameters, drawn from their probability distributions as provided in Table 12. It is worth mentioning that the characteristic mesh size generated for the nested XFEM analysis in both probabilistic analysis and optimization cases, is kept constant in the region of the crack path.

**Table 13** Statistical quantities of the parametric investigation for the three designs of the rigid fillet case

Design	MCS	Mean	Median	Std. dev.
Design 1	100	5382911.9	6895308.5	3176784.9
Design 1	200	6848983.6	7024024.0	11271233.5
Design 1	500	6568327.3	7026526.0	13717577.0
Design 1	1000	653674.8	7026013.0	19043699.3
Design 2	100	90222.3	91794.2	27363.2
Design 2	200	94371.9	86020.0	28170.8
Design 2	500	96950.6	93982.8	109963.6
Design 2	1000	95214.8	94768.0	60920.3
Design 3	100	5260.5	4858.5	1141.3
Design 3	200	5371.9	4992.9	1308.9
Design 3	500	5369.6	5005.9	1278.3
Design 3	1000	5328.8	5360.0	994.2



**Fig. 8** Histograms of each design

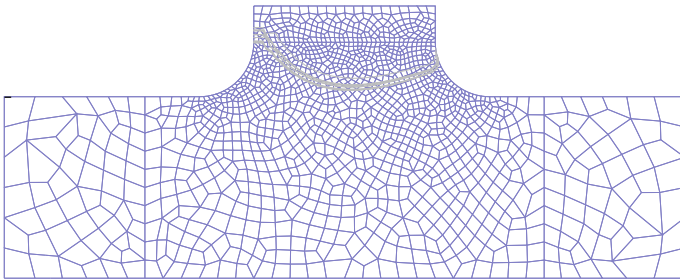
In the group of histograms of Fig. 8 the variability of the number of fatigue cycles with respect to the number of simulations is depicted. These histograms show the probabilistic distribution of the fatigue cycles value for different number of simulations implemented into XFEM and for the three designs, respectively. The frequency on the occurrence of the number of fatigue cycles is defined as the ratio of the number of simulations, corresponding to limit state values in a specific range, over the total number of simulations ( $N_{tot}$ ).  $N_{tot}$  is equal to 100, 200, 500 or 1000 depending on the number of simulations used.

Comparing the histograms of Fig. 8, it can be noticed that the width of the confidence bounds corresponding to the intermediate design is narrower compared to the other two, while for the case corresponding to the upper bounds of the design variables there are two zones of concentration for the frequency values. Furthermore, comparing the mean versus median values of the number of fatigue cycles, the median value is considered more reliable since it is not influenced by the extreme lower and upper values obtained. Specifically, in the framework of an optimization problem, search procedure might lead to designs where such extreme lower and upper values

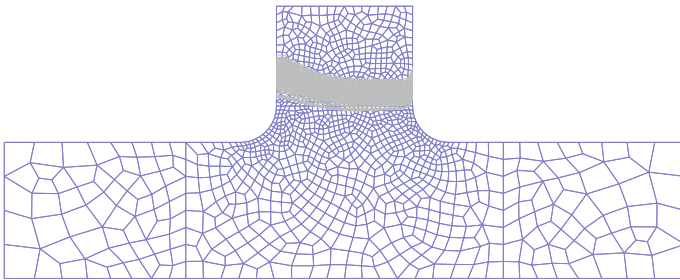
might be often encountered. In addition, 200 LH simulations were considered as an acceptable compromise between computational efficiency and robustness. To this extend an equal number of simulations are applied for the solution of the probabilistic formulation of the shape optimization problem which is investigated in the second part of this study.

The influence of the uncertain variables on the shape of the crack propagation paths is presented in Figs. 9, 10 and 11, where the cloud of the typical crack paths obtained for 200 simulations is depicted. A crack path is defined as typical, if its shape is similar to deterministic one. Especially, for Design 1, due to its geometric characteristics, many not typical crack paths were obtained, however only the typical ones are shown in Fig. 9. This is an additional reason for choosing the median versus mean value as the statistical quantity to be incorporated into the probabilistic formulations of the problems studied in the second part of this work.

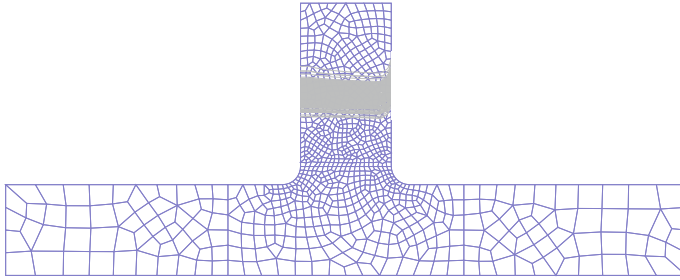
From the results obtained, it can be concluded that the crack paths obtained by means of XFEM is highly influenced by the random parameters considered in this study, thus the importance of incorporating them into the design procedure is examined in the following second part.



**Fig. 9** Design 1



**Fig. 10** Design 2



**Fig. 11** Design 3

## 6.2 Optimization Results

In the second part of this study four optimization problems are solved with the differential evolution (DE) metaheuristic optimization algorithm. The abbreviations DET\*K and PROB\*K correspond to the optimum designs obtained through a deterministic (DET) and probabilistic (PROB) formulation where the lower number of fatigue cycles allowed is equal to \* thousands.

### 6.2.1 Design Optimization Process

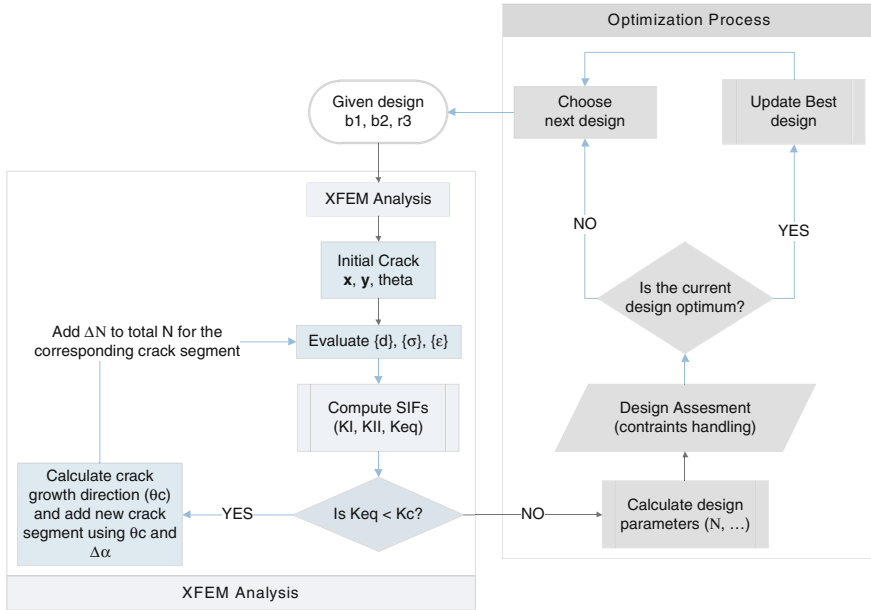
The optimization process that is based on the integration of XFEM into a deterministic and a probabilistic formulation of structural shape optimization is shown in Fig. 12. Within each design iteration of the search process there is a nested crack growth analysis loop performed for each candidate optimum design. Thus, a complete crack growth analysis is conducted until the failure criterion is met, i.e.  $K_{eq} < K_c$  and the corresponding service life is evaluated in order to assess the candidate optimum design.

The parameters used for the DE algorithm are as follows: population size  $NP = 30$ , the probability  $CR = 0.90$  and the mutation factor  $F = 0.60$ . For comparative reasons the method adopted for handling the constraints and the termination criterion is the same for all test cases. On the other hand, the optimization procedure is terminated when the best value of the objective function in the last 30 generations remains unchanged.

### 6.2.2 Fillet Rigid Test Case

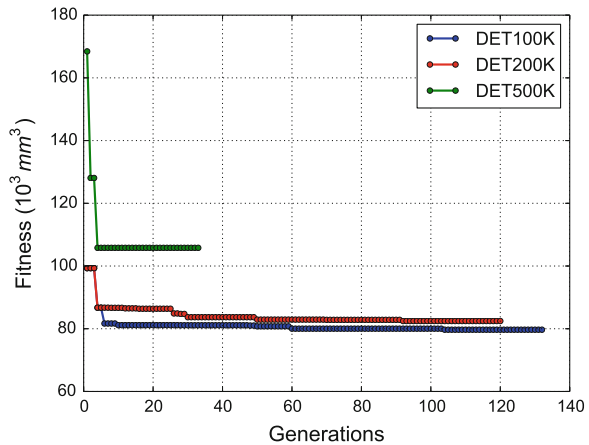
The fillet rigid structural component examined in the previous section is the test example of this study. For this case two groups of formulations were considered, deterministic and probabilistic ones (defined in Eqs. (10)–(11), respectively), where  $N_{min}$  was taken equal to 100, 200 and 500 thousands of fatigue cycles. The objective





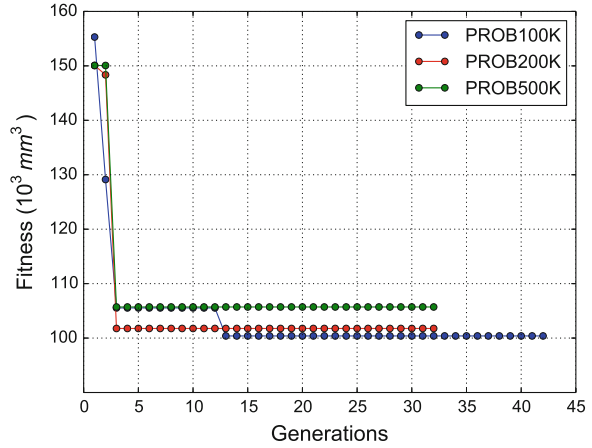
**Fig. 12** XFEM shape optimization process for deterministic and probabilistic formulation

**Fig. 13** Objective function versus generation for DET case (rigid fillet)



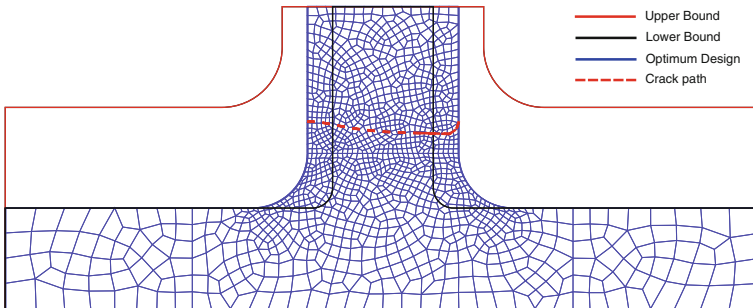
function to be minimized in this problem formulation, is the material volume. DE managed to reach optimum designs as shown in Figs. 13 and 14 together with the optimization history for the deterministic and probabilistic formulation respectively. The optimized designs achieved are presented in Table 14 along with the material volume, while the shapes of deterministic optimized designs are shown in Figs. 15, 16 and 17.

**Fig. 14** Objective function versus generation for PROB case (rigid fillet)



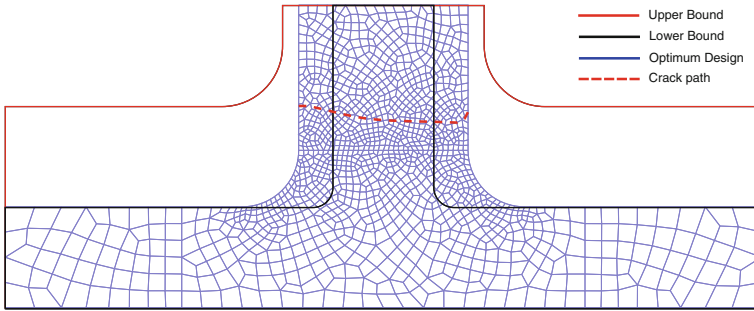
**Table 14** Optimum design for each problem formulation and corresponding statistical parameters for fillet rigid (MCS = 200)

Design	$b_1$	$b_2$	$r_3$	$V$	$N_c^{(det)}$	$\bar{N}_c$	$N_c^{med}$	COV (%)
DET100K	50.0	75.0	27.0	79,690	136,024	99,055	106,573	30.43
DET200K	50.0	84.0	28.0	82,461	201,728	261,604	198,703	36.84
DET500K	73.0	100.0	21.0	105,793	553,038	506,188	505,190	33.02
PROB100K	66.0	100.0	27.0	100,390	118,124	107,584	100,894	45.14
PROB200K	88.0	100.0	19.0	118,065	83,143	353,434	200,288	23.55
PROB500K	100.0	93.0	18.0	169,157	498,856	269,721	554,890	47.19

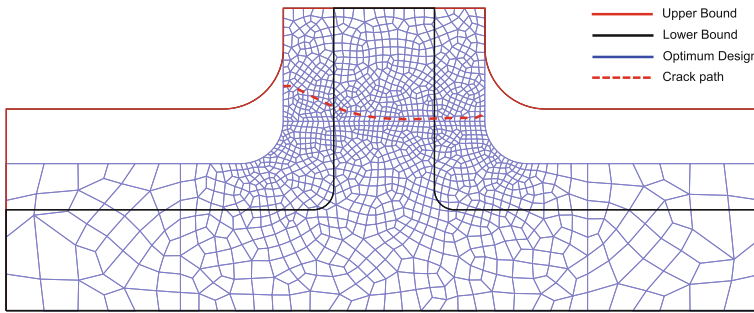


**Fig. 15** Optimum design for deterministic formulation DET100 for rigid fillet

From Table 14, comparing the three designs achieved by means of the deterministic formulation it can be said that the material volume of DET500K is increased by 33 and 28 % compared to DET100K and DET200K respectively, while that of DET200K is increased by almost 3.5 % compared to DET100K. Furthermore, it can be seen



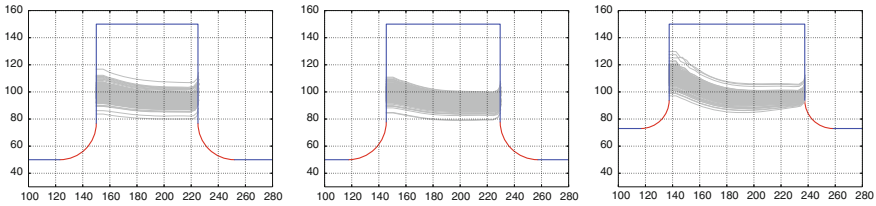
**Fig. 16** Optimum design for deterministic formulation DET200 for rigid fillet



**Fig. 17** Optimum design for deterministic formulation DET500 for rigid fillet

that there are differences to almost all design variables considered to formulate the optimization problem. The results obtained for the probabilistic formulation revealed that the material volume of PROB500K is increased by 68 and 43 % compared to PROB100K and PROB200K respectively, while that of PROB200K is increased by almost 17.5 % compared to PROB100K. In addition, it can be seen that the material volume of designs PROB100K, PROB200K and PROB500K is increased by 26, 43 and 60 % compared to DET100K, DET200K and DET500K, respectively.

In order to justify the formulation of the shape optimization problem considering uncertainties, probabilistic analyses are performed for all six optimized designs obtained through the corresponding problem formulations and the statistical quantities related to the number of fatigue cycles are calculated. These quantities are provided in Table 14 and as it can be seen there are cases where deterministic formulation overestimates the number of fatigue cycles compared to the median value when considering uncertainty. Furthermore, it can be seen that the mean value of the fatigue cycles is not a reliable statistical quantity since it is highly influenced by the crack paths due to high COV values (see Table 14 and Fig. 18). The high COV values which found from the reliability analysis proposed in emerges the necessity of a robust design formulation for the optimization problem, by minimizing these COV values and find the “real” optimum.



**Fig. 18** Crack patterns for DET100, DET200, DET500 case respectively (rigid fillet)

## 7 Conclusions

In this study structural shape optimization problems are formulated for designing structural components under fatigue. For this reason the extended finite element and level set methods are integrated into a shape design optimization framework, solving the nested crack propagation problem and avoiding the mesh difficulties encountered into a CAD-FEM shape optimization problem by working with a fixed mesh approach.

Based on observations of the numerical test presented the deterministic optimized design is not always a “safe” design with reference to the design guidelines, since there are many random factors that affect the design. In order to find a realistic optimized design the designer has to take into account all important random parameters. In the present work a reliability analysis combined with a structural shape design optimization formulation is proposed where probabilistic constraints are incorporated into the formulation of the design optimization problem. In particular, structural shape optimized designs are obtained, considering the influence of various sources of uncertainty. Randomness on the crack initialization along with the uncertainty on the material properties are considered. Shape design optimization problems were formulated for a benchmark structure, where the volume of the structural component is minimized subjected to constraint functions related to targeted service life (minimum number of fatigue cycles allowed) when material properties and crack tip initialization are considered as random variables.

A sensitivity analysis of four optimization algorithms based on evolution process was conducted in order to identify the best algorithm for the particular problem at hand to be used for solving the structural shape optimization problem. This sensitivity analysis is carried out in order to examine the efficiency and robustness of four metaheuristic algorithms. Comparing the four algorithms it can be said that evolutionary based algorithms can be considered as efficient tools for single-objective multi-modal constrained optimization problems. In all test cases examined, a large number of solutions need to be found and evaluated in search of the optimum one. The metaheuristics employed in this study have been found efficient in finding an optimized solution.

The aim of this work in addressing a structural optimization problem considering uncertainties was twofold. First the influence of the uncertain parameters and the number of Latin hypercube samples was examined and in particular those related to the statistical quantities and consequently to the number of fatigue cycles. In the second part of this study the two formulations of the optimization problem were considered feasible for realistic structures. The analysis of the benchmark structure has shown that with proper shape changes, the service life of structural systems subjected to fatigue loads can be enhanced significantly. Comparisons with optimized shapes found for targeted fatigue life are also performed, while the choice of the position and orientation of initial imperfection was found to have a significant effect on the optimal shapes for the structural components examined.

## References

1. Aguirre AH, Rionda SB, Coello Coello CA, Lizárraga GL, Montes EM (2004) Handling constraints using multiobjective optimization concepts. *Int J Numer Methods Eng* 59(15): 1989–2017
2. Anderson TL (2004) *Fracture mechanics: fundamentals and applications*, 3rd edn. CRC Press, Boca Raton
3. Babuška I, Melenk JM (1997) The partition of unity method. *Int J Numer Methods Eng* 40(4):727–758
4. Bletzinger KU, Ramm E (2001) Structural optimization and form finding of light weight structures. *Comput Struct* 79(22–25):2053–2062
5. Bureerat S, Limtragool J (2008) Structural topology optimisation using simulated annealing with multiresolution design variables. *Finite Elem Anal Des* 44(12–13):738–747
6. Chen S, Tortorelli DA (1997) Three-dimensional shape optimization with variational geometry. *Struct Optim* 13(2–3):81–94
7. Chen TY, Chen HC (2009) Mixed-discrete structural optimization using a rank-niche evolution strategy. *Eng Optim* 41(1):39–58
8. Coelho RF (2004) *Multicriteria optimization with expert rules for mechanical design*. Ph.D. Thesis, Universite Libre de Bruxelles, Faculte des Sciences Appliquees, Belgium
9. Das S, Suganthan P (2011) Differential evolution: a survey of the state-of-the-art. *IEEE Trans Evolut Comput* 15(1):4–31
10. Deb K (2000) An efficient constraint handling method for genetic algorithms. *Comput Methods Appl Mech Eng* 186(2–4):311–338
11. Degertekin SO (2012) Improved harmony search algorithms for sizing optimization of truss structures. *Comput Struct* 92–93:229–241
12. Dorigo M, Stützle T (2004) *Ant colony optimization*. MIT Press, Cambridge
13. Edke MS, Chang KH (2010) Shape sensitivity analysis for 2D mixed mode fractures using extended FEM (XFEM) and level set method (LSM). *Mech Based Des Struct Mach* 38(3): 328–347
14. Edke MS, Chang KH (2011) Shape optimization for 2-d mixed-mode fracture using extended FEM (XFEM) and level set method (LSM). *Struct Multidiscip Optim* 44(2):165–181
15. Ellingwood B, Galambos TV (1982) Probability-based criteria for structural design. *Struct Saf* 1(1):15–26
16. Ellingwood B, Galambos T, MacGregor J, Cornell C (1980) Development of a probability based load criterion for American National Standard A58: building code requirements for minimum design loads in buildings and other structures. U.S, Department of Commerce, National Bureau of Standards, Washington, DC

17. Erdogan F, Sih GC (1963) On the crack extension in plates under plane loading and transverse shear. *J Fluids Eng* 85(4):519–525
18. Farhat F, Nakamura S, Takahashi K (2009) Application of genetic algorithm to optimization of buckling restrained braces for seismic upgrading of existing structures. *Comput Struct* 87 (1–2):110–119
19. Fogel D (1992) Evolving artificial intelligence. Ph.D. Thesis, University of California, San Diego
20. Gandomi AH, Alavi AH (2012) Krill herd: a new bio-inspired optimization algorithm. *Commun Nonlinear Sci Numer Simul* 17(12):4831–4845
21. Gandomi AH, Yang XS (2011) Benchmark problems in structural optimization. In: Koziel S, Yang XS (eds) *Computational optimization, methods and algorithms*, no. 356 in studies in computational intelligence. Springer, Berlin Heidelberg, pp 259–281
22. Gandomi AH, Yang XS, Alavi AH (2013) Cuckoo search algorithm: a metaheuristic approach to solve structural optimization problems. *Eng Comput* 29(1):17–35
23. Geem ZW, Kim JH, Loganathan GV (2001) A new heuristic optimization algorithm: harmony search. *Simulation* 76(2):60–68
24. Georgioudakis M (2014) Stochastic analysis and optimum design of structures subjected to fracture. Ph.D. Thesis, School of Civil Engineering, National Technical University of Athens (NTUA)
25. Gholizadeh S, Salajegheh E (2009) Optimal design of structures subjected to time history loading by swarm intelligence and an advanced metamodel. *Comput Methods Appl Mech Eng* 198(37–40):2936–2949
26. Goldberg DE (1989) *Genetic algorithms in search, optimization and machine learning*, 1st edn. Addison-Wesley Longman Publishing Co., Inc., Boston
27. Haddad OB, Afshar A, Mariño MA (2006) Honey-bees mating optimization (HBMO) algorithm: a new heuristic approach for water resources optimization. *Water Resour Manage* 20(5):661–680
28. Haftka RT, Grandhi RV (1986) Structural shape optimization—a survey. *Comput Methods Appl Mech Eng* 57(1):91–106
29. Hansen LU, Häusler SM, Horst P (2008) Evolutionary multicriteria design optimization of integrally stiffened airframe structures. *J Aircr* 45(6):1881–1889
30. Hansen N, Ostermeier A (2001) Completely derandomized self-adaptation in evolution strategies. *Evolut Comput* 9(2):159–195
31. Hasançebi O (2008) Adaptive evolution strategies in structural optimization: enhancing their computational performance with applications to large-scale structures. *Comput Struct* 86 (1–2):119–132
32. Hasançebi O, Çarbaş S, Doğan E, Erdal F, Saka MP (2010) Comparison of non-deterministic search techniques in the optimum design of real size steel frames. *Comput Struct* 88 (17–18):1033–1048
33. Hock W, Schittkowski K (1980) Test examples for nonlinear programming codes. *J Optim Theory Appl* 30(1):127–129
34. Holland JH (1975) *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. University of Michigan Press, Holland
35. Igel C, Hansen N, Roth S (2007) Covariance matrix adaptation for multi-objective optimization. *Evolut Comput* 15(1):1–28
36. Kaveh A, Shahrouzi M (2008) Dynamic selective pressure using hybrid evolutionary and ant system strategies for structural optimization. *Int J Numer Methods Eng* 73(4):544–563
37. Kennedy J, Eberhart R (1995) Particle swarm optimization. *IEEE Int Conf Neural Netw* 4: 1942–1948
38. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220(4598):671–680
39. Koza JR (1992) *Genetic programming: on the programming of computers by means of natural selection*. MIT Press, Cambridge

40. Kripakaran P, Hall B, Gupta A (2011) A genetic algorithm for design of moment-resisting steel frames. *Struct Multidiscip Optim* 44(4):559–574
41. Kunakote T, Bureerat S (2011) Multi-objective topology optimization using evolutionary algorithms. *Eng Optim* 43(5):541–557
42. Lagaros ND (2014) A general purpose real-world structural design optimization computing platform. *Struct Multidiscip Optim* 49(6):1047–1066
43. Lagaros ND, Karlaftis MG (2011) A critical assessment of metaheuristics for scheduling emergency infrastructure inspections. *Swarm Evolut Comput* 1(3):147–163
44. Lagaros ND, Papadrakakis M (2012) Applied soft computing for optimum design of structures. *Struct Multidiscip Optim* 45(6):787–799
45. Lagaros ND, Fragiadakis M, Papadrakakis M (2004) Optimum design of shell structures with stiffening beams. *AIAA J* 42(1):175–184
46. Li L, Wang MY, Wei P (2012) XFEM schemes for level set based structural optimization. *Front Mech Eng* 7(4):335–356
47. Manan A, Vio GA, Harmin MY, Cooper JE (2010) Optimization of aeroelastic composite structures using evolutionary algorithms. *Eng Optim* 42(2):171–184
48. Martínez FJ, González-Vidosa F, Hospitaler A, Alcalá J (2011) Design of tall bridge piers by ant colony optimization. *Eng Struct* 33(8):2320–2329
49. McKay MD, Beckman RJ, Conover WJ (2000) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 42(1):55–61
50. Moës N, Dolbow J, Belytschko T (1999) A finite element method for crack growth without remeshing. *Int J Numer Methods Eng* 46(1):131–150
51. Muc A, Muc-Wierzoń M (2012) An evolution strategy in structural optimization problems for plates and shells. *Compos Struct* 94(4):1461–1470
52. Osher S, Sethian JA (1988) Fronts propagating with curvature dependent speed: algorithms based on hamilton-jacobi formulations. *J Comput Phys* 79(1):12–49
53. Paris P, Gomez M, Anderson W (1961) A rational analytic theory of fatigue. *Trend Eng* 13:9–14
54. Perez RE, Behdinan K (2007) Particle swarm approach for structural design optimization. *Comput Struct* 85(19–20):1579–1588
55. Rechenberg I (1973) *Evolutionstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Frommann-Holzboog, Stuttgart-Bad Cannstatt
56. Rice JR (1968) A path independent integral and the approximate analysis of strain concentrations by notches and cracks. *J Appl Mech* 35:379–386
57. Riche RL, Haftka RT (2012) On global optimization articles in SMO. *Struct Multidiscip Optim* 46(5):627–629
58. Schuëller GI (2006) Developments in stochastic structural mechanics. *Arch Appl Mech* 75(10–12):755–773
59. Schwefel HP (1981) *Numerical optimization of computer models*. Wiley, Chichester, New York
60. Sienz J, Hinton E (1997) Reliable structural optimization with error estimation, adaptivity and robust sensitivity analysis. *Comput Struct* 64(1–4):31–63
61. Stolarska M, Chopp DL, Moës N, Belytschko T (2001) Modelling crack growth by level sets in the extended finite element method. *Int J Numer Methods Eng* 51(8):943–960
62. Storm R, Price K (1997) Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *J Glob Optim* 11(4):341–359
63. Su R, Wang X, Gui L, Fan Z (2011) Multi-objective topology and sizing optimization of truss structures based on adaptive multi-island search strategy. *Struct Multidiscip Optim* 43(2):275–286
64. Su Y, Wang SN, Du YE (2013) Optimization algorithm of crack initial angle using the extended finite element method. *Appl Mech Mater* 444–445:77–84
65. Wang Q, Fang H, Zou XK (2010) Application of micro-GA for optimal cost base isolation design of bridges subject to transient earthquake loads. *Struct Multidiscip Optim* 41(5):765–777

66. Yang XS (2010) Nature-inspired metaheuristic algorithms, 2nd edn. Luniver Press, Bristol
67. Yang XS, Deb S (2010) Engineering optimisation by cuckoo search. *Int J Math Modell Numer Optim* 1(4):330–343
68. Yang XS, Gandomi AH (2012) Bat algorithm: a novel approach for global engineering optimization. *Eng Comput* 29(5):464–483
69. Yang XS, Koziel S (2011) Computational optimization and applications in engineering and industry. Springer, New York
70. Yau JF, Wang SS, Corten HT (1980) A mixed-mode crack analysis of isotropic solids using conservation laws of elasticity. *J Appl Mech* 47(2):335–341



# A Stress-Test of Alternative Formulations and Algorithmic Configurations for the Binary Combinatorial Optimization of Bridges Rehabilitation Selection

Dimos C. Charmpis and Loukas Dimitriou

**Abstract** Optimal surface transport asset management is a major concern with multiple economic and operational implications developed in various infrastructure areas. Although relevant ‘mature’ analytical frameworks have been proposed and developed, the problem setup and the algorithmic choices are still issues requiring thorough and detailed investigation. In this chapter, an optimal budget allocation framework is developed and stress-tested for the optimal scheduling of a bridges upgrading program. A suitable test case is developed for performing in-depth analysis that takes into consideration the most important features involved in such scheduling problems, while alternative formulations are also presented and discussed. The proposed frameworks are applied on a real large-scale dataset from the highway system of US, able to provide an adequate test-bed for investigating the optimal upgrade problem. The paper aims in the investigation of the effects that alterations of the problem setup, but also the effects that algorithmic configurations are introducing, when addressing real-world applications. The binary/selection problem is handled with a suitably coded Branch-and-Bound (BaB) algorithm, which is regarded as a robust and fast heuristic for such optimization problems. BaB is tested in alternative standard and extreme configurations, offering insights on its performance. Interestingly enough, although the continuous relaxation introduced by the BaB enables fast convergence, the NP-hard problem’s nature should be cautiously taken into consideration. The results are discussed in order to provide insights of applying the proposed framework in realistic infrastructure upgrading schemes.

**Keywords** Road bridges upgrade · Optimal scheduling · Sequential combinatorial optimization · Branch-and-bound · Pareto front

---

D.C. Charmpis (✉) · L. Dimitriou  
Department of Civil and Environmental Engineering,  
University of Cyprus, 75 Kallipoleos Str., P.O. Box 20537, 1678 Nicosia, Cyprus  
e-mail: charmpis@ucy.ac.cy

L. Dimitriou  
e-mail: lucdimit@ucy.ac.cy

## 1 Introduction

It is widely recognized that the development of civil infrastructure corresponds to a twofold process: infrastructure additions and the maintenance of the existing. Both require significant efforts (monetary and other) and it is essential to handle them in such manner that the resulting infrastructure meets the multiple requirements and covers the multiple objectives contemporary societies depend on. Alternative practices and large research efforts have been proposed and used worldwide for cases of optimal infrastructure handling, belonging to the broad multidisciplinary areas of asset management, optimal budget allocation problems, optimal planning, maintenance programming and life-cycle engineering analysis, elements that highlight the importance and complexity of the optimal (in multiple ways) infrastructure management case.

In order to be able to utilize the methodological and technical advances that have emerged in the recent decades for the optimal asset management of civil infrastructure, many important issues need to be considered, related to the data availability, reliability and consistency that adds to the burden involved in each case.

This chapter aims to offer results from an in-depth investigation that incremental programming modification as well as algorithmic configurations have on realistic implementations and in the overall model performance for such computationally complex cases. In particular, alternative problem setups are analyzed and discussed. Additionally, a well-documented optimization algorithm is stress-tested for addressing the above configurations, providing valuable results on the problem characteristics. The algorithm selected is the Branch-and-Bound (BaB), which is based on a relaxation of the binary problem to a series of linear programming approximations.

This comprehensive formulation that is offered and discussed, accompanied with computational results, are suitably selected and presented in order to contribute to the understanding of the nature of the bridge upgrade scheduling problem and the expected outcomes using optimization frameworks. The real data used in this work is obtained from the database of the Federal Highway Administration (FHWA) of the United States (US) (National Bridge Inventory-NBI).

Regarding the organization of the chapter, it starts with a brief but targeted literature review of optimal asset management of road infrastructure and budget allocation for bridges' maintenance purposes. Then, the case and threads of the optimal bridges' upgrade programming is discussed based on the realistic information extracted from the NBI database. The test-bed used here corresponds to the bridges stock of the State of New York, which yields a data sample of about 15,000 bridges and is thoroughly and reliably appraised in operational and economic terms. Results from a comprehensive case setup are analytically presented and discussed next. Also, the performance of the proposed optimization framework is presented in detail, giving information about its performance and some outlook. From the discussion section, useful insights both for the methodological part as well for the state-of-the-practice are provided. The last section concludes this work.

## 2 Literature Review

Optimal investments planning for infrastructure maintenance and generally asset management of the civil infrastructure has always been recognized as an important issue and as so substantial research effort is invested over the years [1]. A complete literature review of this important issue will not be presented in the current paper. Hence, focusing on road infrastructure, the importance for maintaining adequate operational conditions have been highlighted mainly for three reasons:

- i the importance of maintaining the necessary connectivity is closely related with economic activity,
- ii the necessary efforts for maintenance correspond to large amounts of money that should be invested in the most ‘prosperous’ manner, and
- iii in cases of emergency (natural disasters, accidents or deliberate malevolent actions) road networks are providing vital lifelines.

Following the above very broad categorization of the road asset management problems, the ‘rational’ way of treating (funding) asset maintenance is based on the general area of optimal budget allocation type of problems, while allocation involves both selecting the assets/elements that should be maintained as well as the time that maintenance efforts should be scheduled. For these reasons, many optimization frameworks have been tested and used [2, 3] focusing on the road asset management on a life-cycle basis and taking into consideration reliability issues [4–7]. The specific case of the road asset management and the corresponding fund allocation problem give rise to multiple objectives and concerns, therefore many approaches have been based on multi-objective optimization problem setups [8, 9], incorporating elements of stochasticity and uncertainty in the important assets of bridges [10]. Moreover, optimal allocation and scheduling problems have been also considered [11], while some –at least methodological– issues, emerging in cases of optimal programming for multidistrict agencies, are also reported [12]. Moving to the asset management of the road infrastructure, viewed as important lifelines in cases of emergency, optimal recovery planning has been treated again using optimal planning procedures [13, 14]. Seemingly to the recovery planning, infrastructure security planning may be treated as optimal allocation problem [15].

As can be observed by the above brief review of the recent research efforts in the optimal asset management and resource allocation type of problems, the approach typically used lies within the optimal selection, organization, classification, scheduling or hierarchy formation type of problems, which are treated using optimization routines and frameworks. In realistic cases, the above types of mathematical programming cases correspond to optimization paradigms of high computational complexity and as so several issues emerge, either due to the high dimensionality of the problems at hand, or by the difficulties to adequately tackle them. This is why the optimization routines used in the vast majority of the demonstrative cases presented in the literature employs (meta-)heuristics [16], the final solutions of which cannot be guaranteed to be absolute optimal solutions.

The current paper aims to offer an additional application of optimal maintenance programming for the important road infrastructure of road bridges. In particular, a real large-scale database is used for this case, which is presented in the following section.

### 3 Database Information

The data used here comes from the NBI program, which provides an extensive database that contains information on several hundreds of thousands bridges, culverts and tunnels in the US [17]. NBI was initiated in 1972 and is now yearly updated by the FHWA through inspections conducted from specially qualified personnel and it can be considered as a suitable test-bed for relevant research and ‘tools’ development [18]. The information in this inventory is stored in coded form and concerns location, structural condition, age, materials, traffic etc. for each construction.

#### 3.1 Database Items

The NBI-items exploited herein are synopsisized in Table 1 and described below.

##### *Bypass, Detour Length (DLEN)*

When a bridge is closed and cannot be used for whatever reason (failure, repair, maintenance, etc.), vehicles have to use a bypass to move around the closed bridge. The present item indicates the detour length (km), which corresponds to the total *additional* travel for a vehicle that results from the closure of a bridge. The longest detour length allowed to be coded is 199 km.

##### *Average Daily Traffic (ADT)*

This item reports a recent count for the annual average daily traffic volume of each bridge. This count includes all types of traffic (light vehicles, trucks, etc.). Even if a bridge is closed, an *ADT*-value is given and represents the vehicle count from before the bridge closure.

**Table 1** NBI items used in the present work

No.	Symbol	Description
19	<i>DLEN</i>	Bypass, Detour length
29	<i>ADT</i>	Average Daily traffic
96	<i>CIMPR</i>	Total improvement cost

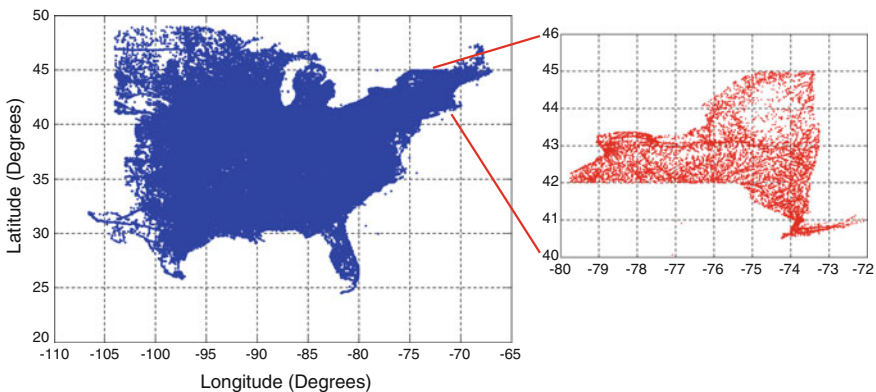
*Total Improvement Cost (CIMPR)*

This item offers an estimation of the total project cost (US \$) for improving each bridge. It includes all costs that can be associated with the particular bridge improvement project, i.e. the costs for structural upgrade, roadway construction and other incidental costs. The provided total cost estimation is current. When difficulties are encountered in making a reasonable cost estimation, setting *CIMPR* equal to 150 % of the bridge cost is recommended in [17].

**3.2 Description of Bridge Stock Data**

The bridge stock considered in this work for optimally allocating a budget for bridge improvement is the one of the State of New York (NY) with NBI state code 362. Figure 1 illustrates the respective bridge locations. The total number of NBI records for NY is 17,442. This number is reduced by 1948 records to exclude culverts. Moreover, as we are herein interested in steel and reinforced/prestressed concrete bridges only, another 615 records are eliminated to exclude bridges made of other materials (wood, masonry, aluminum, etc.). Finally, another 84 records are deleted to exclude a few bridges, for which essential data are missing, e.g. no bridge improvement cost (*CIMPR*) or traffic (*ADT*) is given in the database. Thus, the total number of NY bridges processed in this work is 14,795.

The NY bridge data extracted for the items of Table 1 from the NBI database are organized into vectors **DLEN**, **ADT** and **CIMPR**. The data are graphically depicted in Fig. 2, while a zoom-in view of traffic versus improvement cost data is provided in Fig. 3.



**Fig. 1** Locations of bridges in the NBI database: east US (*left*) and zoom-in view of NY State (*right*)

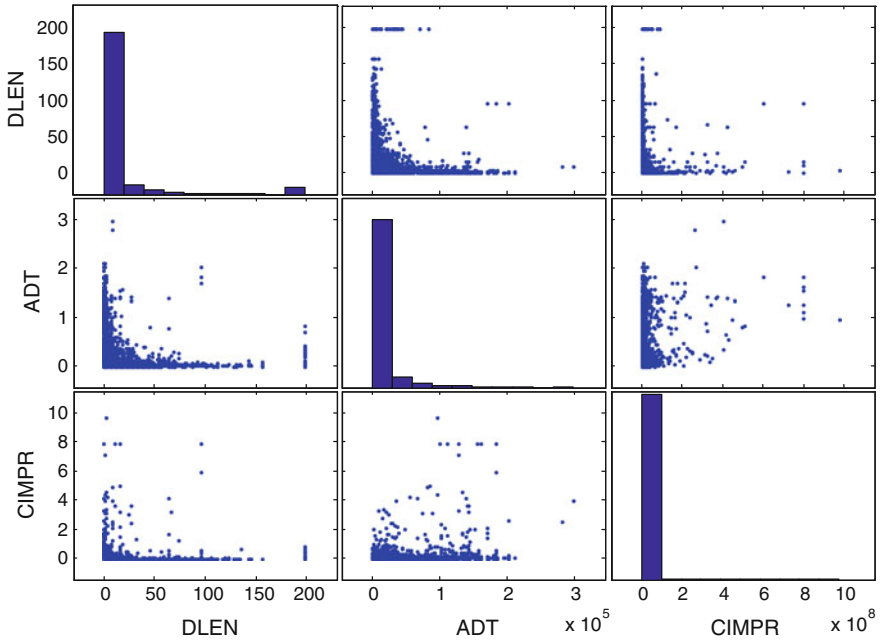


Fig. 2 Scatter plot matrix of NY bridge stock data

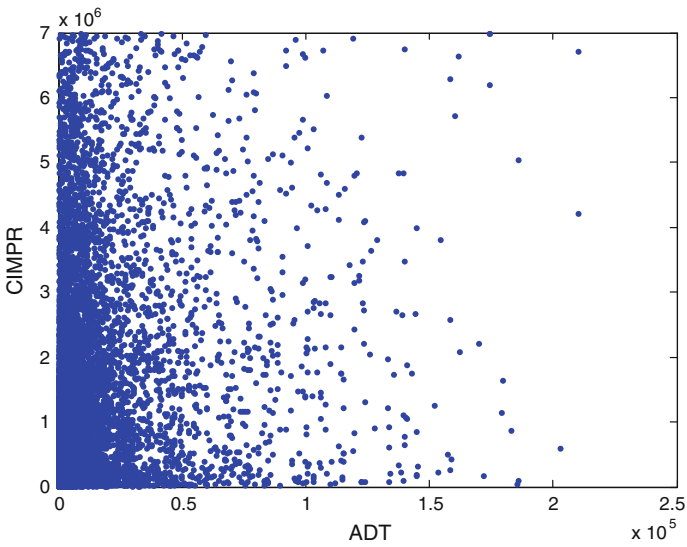


Fig. 3 Zoom-in view of scatter plot of traffic versus improvement cost bridge stock data

## 4 Optimization of Budget Allocation for Bridge Stock Improvement

An optimization problem is formulated in the present work to allocate the available budget for the improvement of a bridge stock. In order to facilitate the reference to various options to allocate the budget, a binary integer vector  $\mathbf{x}$  is used to indicate which bridges of the stock are improved. In particular, the terms of vector  $\mathbf{x}$  can only take the values 0 or 1 as follows:

$$x_i = \begin{cases} 0 \Rightarrow \text{bridge } i \text{ is not improved} \\ 1 \Rightarrow \text{bridge } i \text{ is improved} \end{cases} \quad (1)$$

Vector  $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_{n_b}]^t$  has  $n_b$  terms, where  $n_b$  is the number of bridges in the processed stock (for the NY stock used herein,  $n_b = 14,795$ ). Thus, for a particular budget allocation, vector  $\mathbf{x}$  provides the decision regarding the improvement or not of each bridge in the stock. The number of bridges improved is obtained simply by adding together all terms of vector  $\mathbf{x}$ :

$$n_b^{\text{impr}}(\mathbf{x}) = \sum_{i=1}^{n_b} x_i. \quad (2)$$

In the two formulations considered in the present work, the aim of the optimization procedure is to decide on the improvement programme of a bridge stock in a way that the social benefit is maximized. In the first formulation, the aim is to maximize the total traffic serviced by improved bridges using a pre-specified budget. This way, the highest possible traffic will take advantage of the budget spent. In mathematical programming terms, the optimization problem is expressed as:

$$\begin{aligned} &\text{find} && \mathbf{x} \\ &\text{that maximizes} && F(\mathbf{x}) = \mathbf{ADT}'\mathbf{x} \\ &\text{subject to} && \mathbf{CIMPR}'\mathbf{x} \leq \mathit{BUDGET} \\ &&& \mathbf{x} \text{ binary} \end{aligned} \quad (3)$$

In the above formulation,  $\mathbf{x}$  is the vector of decision variables  $x_i$  ( $i = 1, \dots, n_b$ ), which numerically control the improvement decision for each bridge. The user-specified parameter  $\mathit{BUDGET}$  is the maximum allowable total improvement cost (in US \$) of the bridges. Thus, a solution  $\mathbf{x}$  is feasible only when it satisfies the budget constraint; otherwise, it is infeasible.

The second formulation considered is obtained by appending one more constraint to (3). Specifically, the indirect service costs due to bridge closure for improvement works are additionally taken into account. For this purpose, a new vector  $\mathbf{CCLOS}$  is formed to provide a measure of the costs due to user inconvenience, delay, increased

**Table 2** Basic statistics of NY bridge stock data

	Minimum	Maximum	Average	Sum
<b>ADT</b> (vehicles)	1	297,700	11,400	168.7e6
<b>CIMPR</b> (\$)	2.0e3	975.3e6	4.0e6	58.7e9
<b>CCLOS</b> (vehicles × km)	0	19.5e6	74.0e3	1.1e9

Minimum, maximum and average values refer to one bridge for a particular data vector. The ‘sum’ refers to the total value obtained by adding together all bridge values for a particular data vector

fuel consumption, etc. associated with the temporary closure of each bridge under improvement. Each term  $CCLOS_i$  ( $i = 1, \dots, n_b$ ) of **CCLOS** is calculated as:

$$CCLOS_i = ADT_i DLEN_i \tag{4}$$

and its units are (vehicles × km). Thus, the second optimization formulation is expressed as:

$$\begin{aligned}
 &\text{find} && \mathbf{x} \\
 &\text{that maximizes} && F(\mathbf{x}) = \mathbf{ADT}'\mathbf{x} \\
 &\text{subject to} && \mathbf{CIMPR}'\mathbf{x} \leq \mathit{BUDGET} \\
 & && \mathbf{CCLOS}'\mathbf{x} \leq CCLOS_{\max} \\
 & && \mathbf{x} \text{ binary}
 \end{aligned} \tag{5}$$

The user-specified parameter  $CCLOS_{\max}$  is the maximum allowable total indirect service cost (in vehicles × km) of the bridges. Now, a solution  $\mathbf{x}$  is feasible only when it satisfies both budget and indirect service cost constraints; otherwise, it is infeasible.

Table 2 provides basic statistical properties for the bridge data in the three vectors **ADT**, **CIMPR** and **CCLOS** involved in the presented optimization formulations.

## 5 Binary Optimization Algorithm

The optimization formulations (3) and (5) define binary integer programming problems with a linear objective function  $F(\mathbf{x})$ , linear constraints and a binary solution vector  $\mathbf{x}$ . There is a number of optimization routines used for addressing such binary programming problems, all of them belonging to the heuristics class of algorithms. In this work, the Linear Programming (LP)-based Branch-and-Bound (BaB) algorithm is used [19, 20], in order to address the particular binary/combinatorial programming problem. This algorithm searches for an optimal solution to the binary programming problem by solving a series of *LP-relaxation* problems, in which the binary requirement on the decision variables is replaced by the ‘weaker’ constraint  $0 \leq x_i \leq 1$ . The algorithmic steps roughly are:



- search for a binary feasible solution by solving the problem as a continuous LP problem and by rounding to the appropriate/nearest integer in  $\{0,1\}$  (the search starts from a current solution);
- update the best binary feasible point found so far as the search ‘tree’ grows (updating);
- verify that no better binary feasible solution is possible by solving a series of LP problems (convergence).

The algorithm performs an incomplete, deterministic exploration of the decision space and avoids the computationally extremely demanding exhaustive search, which involves all possible  $2^{nb}$  solution vectors. By using a LP-relaxation feature within the BaB algorithm, convergence time is significantly improved while maintaining a level of stratification in the deterministic search process.

In this particular binary optimization algorithm, two features describe the computational burden involved in each case processed: (i) the *total number of iterations* needed by the LP method used (here the Simplex method) subject to particular stopping criteria for estimating upper and lower solution bounds and (ii) the *total number of nodes* constructed for exploring the search space. Few nodes and LP iterations correspond to fast BaB convergence, while large numbers of nodes and LP iterations signify that the BaB algorithm encounters difficulties in converging to a final (optimal) solution. The particular problem size and the alternative configurations solved herein, form a suitable test-bed for a crash-test of the BaB algorithm. Results interpretation and discussion are provided in subsequent sections.

## 6 Computational Experiments

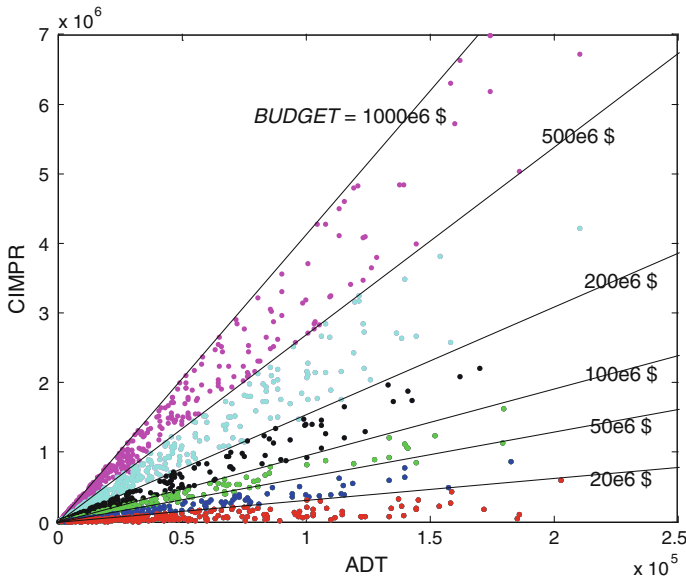
### 6.1 Maximization of Total Traffic Serviced by Improved Bridges for Pre-specified Budget

The budget allocation problem for the NY State bridge stock is first addressed with the BaB algorithm using formulation (3) for various available budgets ranging from 20 to 1000 million \$. The results attained from this parametric study are summarized in Table 3. The percentage values in this table compare optimization data/results with respective total values. Thus, with a budget of just 20 million \$, which is much less than 1 % of the total necessary improvement cost for all bridges (see sum of **CIMPR** in Table 2), 2.0 % of NY bridges can be improved servicing more than 10 % of the state’s overall bridge traffic (see sum of **ADT** in Table 2). With a high budget of 1000 million \$ (1.7 % of the total improvement cost for all bridges), more than 14 % of NY bridges can be improved servicing 37.5 % of the total bridge traffic. These results demonstrate the effectiveness of the developed budget allocation procedure, which optimally exploits the available amount of money for bridge improvements.

To gain insight into the optimal solutions attained using formulation (3), scatter plots for the *ADT* and *CIMPR* values of the improved bridges in each opti-

**Table 3** Optimization results obtained using formulation (3)

Optimization run	1	2	3	4	5	6
<i>BUDGET</i> (\$)	20e6	50e6	100e6	200e6	500e6	1000e6
	(0.3 ‰)	(0.9 ‰)	(1.7 ‰)	(3.4 ‰)	(8.5 ‰)	(1.7 ‰)
Unused budget (\$)	1000	0	0	18,000	2000	1,034,000
<i>F</i> (vehicles)	12.8e6	19.6e6	26.2e6	34.2e6	48.6e6	63.3e6
	(7.6 ‰)	(11.6 ‰)	(15.5 ‰)	(20.2 ‰)	(28.8 ‰)	(37.5 ‰)
Improved bridges $n_b^{impr}$	298	540	765	1069	1611	2179
	(2.0 ‰)	(3.6 ‰)	(5.2 ‰)	(7.2 ‰)	(10.9 ‰)	(14.7 ‰)



**Fig. 4** Scatter plot of *ADT* versus *CIMPR* data for improved bridges according to the optimal solutions and available *BUDGET*

mization case are given in Fig. 4. The optimizer clearly prefers to improve bridges, which have a relatively high *ADT*-value combined with a relatively low improvement cost (*CIMPR*) value. In fact, the optimizer seems to select for improvement all bridges, which have points in the *ADT*-*CIMPR* scatter plot below a straight line passing through the origin, relating to the deterministic search strategy the BaB algorithm is based on. The value of the *BUDGET*-parameter actually specifies the slope (*CIMPR*/*ADT*) of this ‘decision’-line, with a higher *BUDGET*-value inducing a higher slope.

Table 4 presents the slopes of the ‘decision’-line for the 6 optimization runs, as well as for the entire NY bridge sample. The slope for a bridge sample is actually equal to the maximum ratio *CIMPR*/*ADT* among all bridges of the sample. The results

**Table 4** Slopes of the ‘decision’-line

Bridge sample	<i>BUDGET</i> (\$)	Max ratio <i>CIMPR</i> / <i>ADT</i>
Optimal solution	20e6	3.1
	50e6	6.5
	100e6	9.5
	200e6	15.4
	500e6	26.9
	1000e6	41.3
Entire sample	$\infty$	6,108,000.0

of the table verify the positive correlation among available budget and slope of the ‘decision’-line. The extremely high slope-value for the entire sample is explained by the presence of many bridges in the sample, which have very low *ADT*-values combined with high *CIMPR*-values (see Fig. 3). This behavior is a consequence of the dependence mentioned earlier among the available budget and the slope of the ‘decision’-line in the *ADT-CIMPR* scatter plot. Figure 4 graphically illustrates this behavior.

Another interesting property of the results obtained is that the bridges improved according to the optimal solution for a lower *BUDGET*-value are a subset of the bridges improved according to the optimal solution for a higher *BUDGET*-value. In other words, when the *BUDGET*-value is increased, the optimizer adds improved bridges to the improved bridges selected for a lower *BUDGET*-value (see Fig. 4).

It is finally noted that, in order to use up the available *BUDGET* in each optimization case, the final BaB solution may need to be post-processed using another optimization routine. Thus, the ‘optimal’ BaB solution can be used as the starting point for a sequential meta-optimization round of stochastic search performed e.g. with a Genetic Algorithm [21]. The additional stochastic search can eliminate the unused budgets of Table 3 by selecting additional bridges to improve.

### 6.2 Maximization of Total Traffic Serviced by Improved Bridges for Pre-specified Budget and Indirect Service Cost

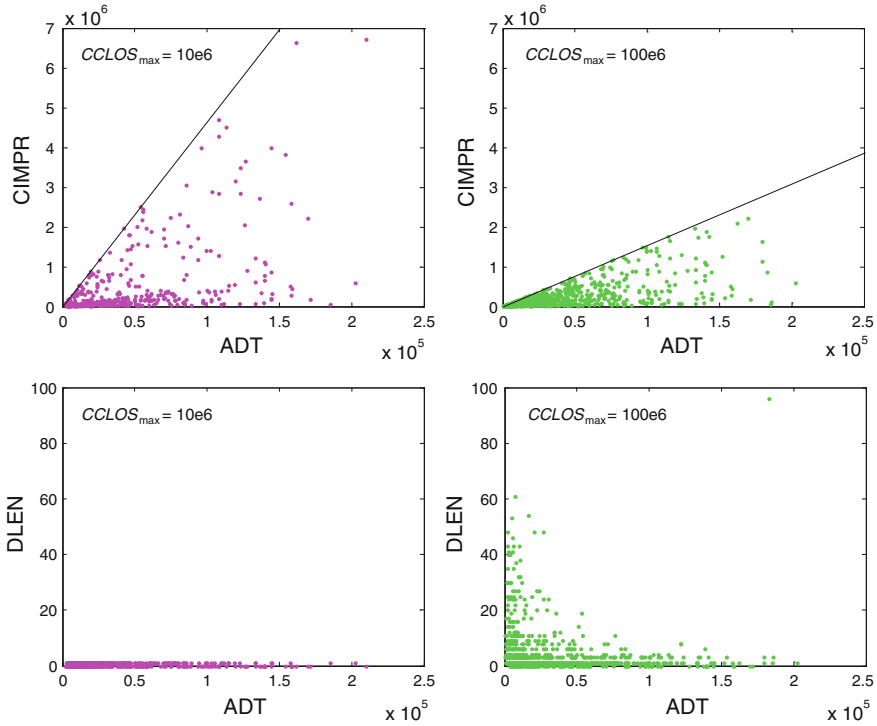
The budget allocation problem for the NY State bridge stock is also addressed with the BaB algorithm using formulation (5) for available budgets ranging from 50 to 500 million\$. The results attained from this parametric study are summarized in Table 5. Lines in bold correspond to the results of Table 3, i.e. with the indirect service cost actually deactivated. Thus, for values higher than the first  $CLOS_{max}$ -value given for each *BUDGET*, the indirect service cost constraint is satisfied anyway in the optimal solution, whether this constraint is taken into account or not (i.e. only the budget constraint is critical). The results of Table 5 demonstrate the effect of the

**Table 5** Optimization results obtained using formulation (5)

Constraints		Optimal solution	
<i>BUDGET</i> (\$)	<i>CLOS</i> <sub>max</sub> (vehicles × km)	<i>F</i> (vehicles)	Improved bridges <i>n</i> <sub>b</sub> <sup>impr</sup>
<b>50e6 (0.9 ‰)</b>	<b>84.2e6</b>	<b>19.6e6</b>	<b>540</b>
	50.0e6	19.5e6	544
	30.0e6	19.0e6	479
	20.0e6	17.7e6	434
	10.0e6	14.8e6	372
	5.0e6	11.1e6	227
<b>100e6 (1.7 ‰)</b>	<b>106.7e6</b>	<b>26.2e6</b>	<b>765</b>
	50.0e6	25.7e6	710
	30.0e6	24.1e6	642
	20.0e6	21.8e6	560
	10.0e6	17.3e6	421
	5.0e6	13.1e6	262
<b>200e6 (3.4 ‰)</b>	<b>145.6e6</b>	<b>34.2e6</b>	<b>1069</b>
	100.0e6	34.0e6	1036
	50.0e6	32.7e6	911
	30.0e6	29.5e6	803
	20.0e6	26.5e6	732
	10.0e6	20.2e6	482
	5.0e6	15.6e6	322
	2.0e6	11.1e6	227
<b>500e6 (8.5 ‰)</b>	<b>207.1e6</b>	<b>48.6e6</b>	<b>1611</b>
	100.0e6	47.9e6	1495
	50.0e6	43.7e6	1243
	30.0e6	38.3e6	1073
	20.0e6	33.3e6	892
	10.0e6	24.9e6	577
	5.0e6	20.1e6	417
	2.0e6	17.1e6	336

indirect service cost constraint on the optimal solution attained. It is worth noting that lower *CLOS*<sub>max</sub>-values consistently yield lower total *ADT*-values; these are generally combined with lower *n*<sub>b</sub><sup>impr</sup>-values. In other words, imposing a more strict indirect service cost constraint leads to an optimal improvement selection, which involves fewer bridges servicing a smaller percentage of the total traffic.

Insight into the optimal solutions attained using formulation (5) is gained through the scatter plots for the *ADT*, *CIMPR* and *DLEN* values of the improved bridges in two optimization cases depicted in Fig. 5. For the stricter *CLOS*<sub>max</sub>-value, the optimizer is obliged to seek bridges with lower detour lengths, even if this means that corresponding improvement costs are increased. This leads to a ‘compensation’



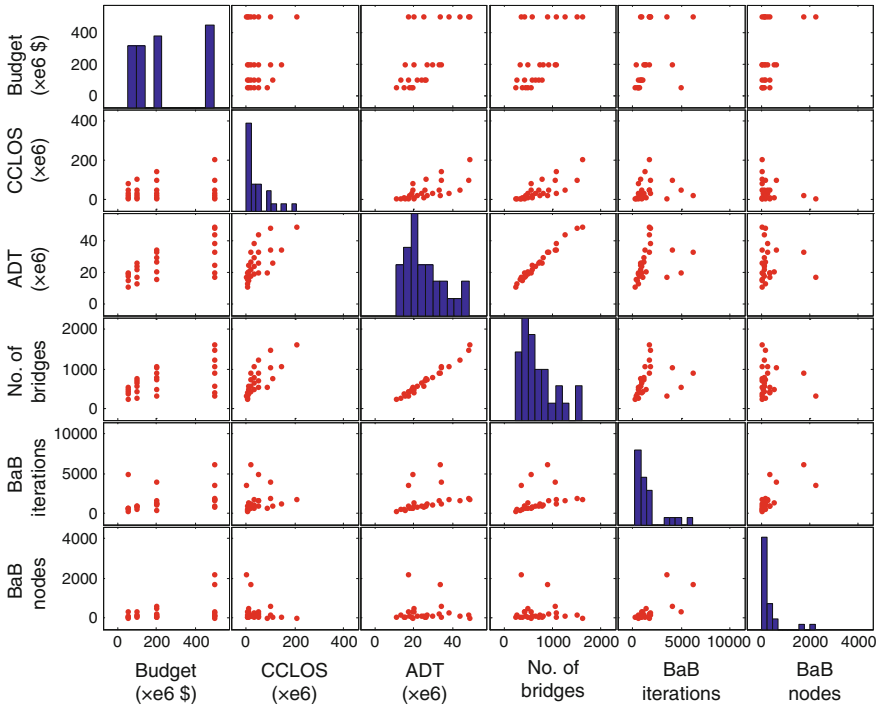
**Fig. 5** Scatter plot of  $ADT$  versus  $CIMPR$  and  $DLEN$  data for improved bridges according to the optimal solutions for two different  $CCLOS_{max}$ -values ( $BUDGET = 200$  million\$)

among the two constraints imposed and to a ‘decision’-line in the  $ADT$ - $CIMPR$  plot with much higher slope. Thus, when using optimization formulation (5), the slope of the ‘decision’-line is directly affected by both values controlling the problem’s constraints ( $BUDGET$  and  $CCLOS_{max}$ ).

By imposing the indirect service cost constraint, the property that the bridges improved according to the optimal solution for a lower  $BUDGET$ -value are a subset of the bridges improved according to the optimal solution for a higher  $BUDGET$ -value is lost. This is clearly evidenced in Fig. 5. With formulation (5), every optimization problem is actually a case of its own.

### 6.3 Discussion on the Performance of the Optimization Algorithm

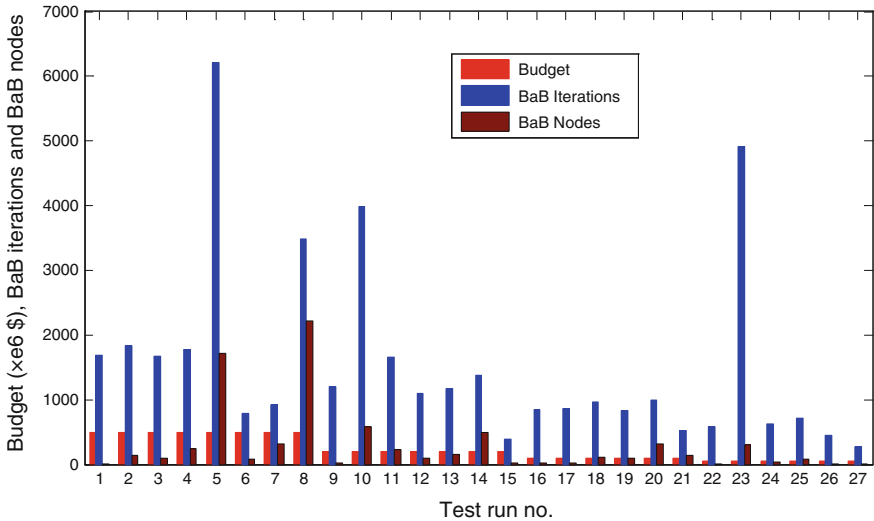
As shown from the above analysis, the BaB algorithm is able to treat alternative problem setups adequately. Though, some interesting features related to the BaB performance that cannot be observed in the results presented earlier are provided



**Fig. 6** Matrix of scatter and frequency plots for all problem’s parameters

below. In particular, in Fig. 6, a matrix of pairwise scatter diagrams and frequency distributions of the problem’s variables and algorithm’s operational characteristics is presented for 27 optimization cases conducted using formulation (5) in the framework of the earlier demonstrated stress-tests. Throughout the diagram, the stochastic nature of the relations between each pair of variables is evident. The only strict correlation emerges between the optimal total *ADT*-values and the corresponding numbers of selected bridges for upgrade. This positive correlation is expected, since all experiments aim in the maximization of *ADT* by selecting a number of bridges subject to constraints.

The number of nodes used and the corresponding algorithm’s LP iterations can be used as valid indices to assess the problem complexity and the BaB performance. Interestingly enough, the available budget, the total *ADT* achieved or the service cost *CCLOS* cannot ‘explain’ the number of necessary nodes or iterations needed for addressing each problem configuration, bearing also in mind that the same feasible solution is always used as the algorithm’s initial/starting conditions ( $X_i^0$ ). This issue is related to the problem’s complexity; although the BaB procedure transforms a NP-hard problem to a series of polynomial problems (efficiently solved by a LP relaxation), this alteration is not crucial for the algorithm’s required computational effort (typically associated with combinatorial problems). A clearer view is offered



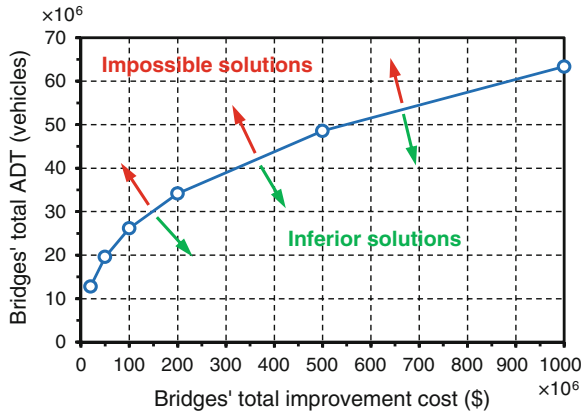
**Fig. 7** Graph of computational burden (number of nodes and algorithm’s LP iterations) as related to alternative budget availability for the 27 test runs

in Fig. 7, where the budget and the numbers of nodes and LP iterations are presented for the 27 test runs. Although the search space is the same, the computational burden (and the associated computational time), as depicted in the number of search nodes and the iterations, are non-systematically developed.

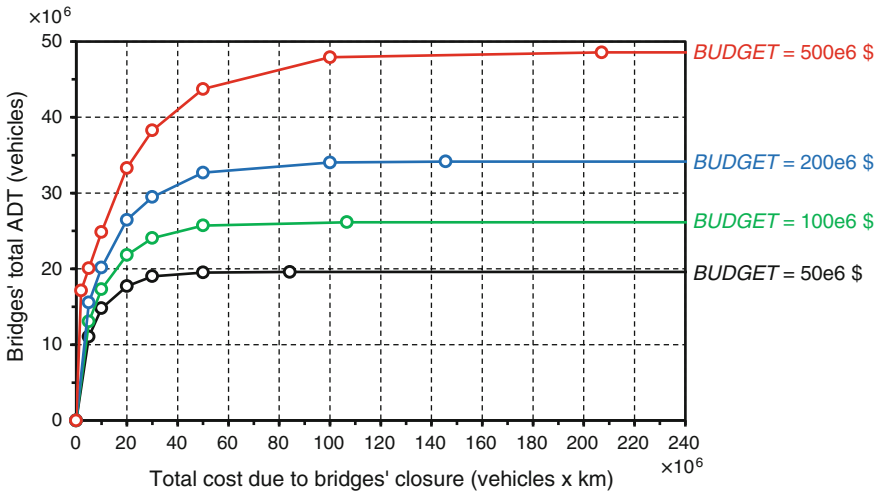
The above brief analysis can be regarded as a BaB stress-test for the particular combinatorial/selection type of problems. It is noted that marginal alterations of the problem setup (either in the constraints set formation or their thresholds) can have significant—or at least non-expected—effect on the computational burden involved, especially in such large-scale datasets. This element should be taken into consideration in the results appraisal meta-analysis, especially in cases where such algorithms are used under operational circumstances (e.g. when decisions need to be made in very short time) and not under strategic ones, like in the present work.

## 7 Policy Implications

A final point worth mentioning is that the relation among total improvement cost for the bridges and total traffic serviced by improved bridges is nonlinear. This conclusion is reached based on the results of Tables 3 and 5, while it is clearer illustrated in the graphs of Figs. 8 and 9. More specifically, the 6 points in the graph of Fig. 8, which correspond to the optimization runs conducted for 6 different *BUDGET*-values using formulation (3), actually provide a coarse view of the Pareto Front (PF) for a multi-objective optimization problem, which aims in maximizing *F* and minimizing the



**Fig. 8** Nonlinear relation ('Pareto Front') among total improvement cost and traffic serviced by improved bridges

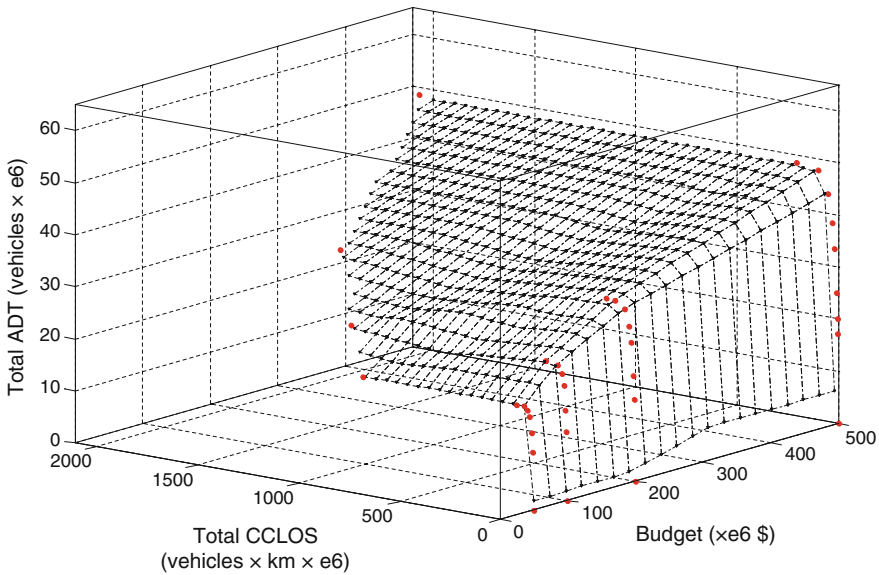


**Fig. 9** Pareto Fronts set for alternative budget availability relating total cost for bridges closure and ADT serviced by improved bridges

total improvement cost. The PF distinguishes the area of feasible solutions (which includes the optimal solutions on the front and inferior solutions below the front) from the area of impossible solutions (which correspond to points above the front with non-attainable total ADT-values for the respective total improvement costs). Figure 9 provides the similar PF information for various values controlling the indirect service cost constraint using formulation (5).

The nonlinear PF suggests that increasing an already high budget has a—as expected—positive effect on the total traffic serviced by improved bridges, but the





**Fig. 10** A nonlinear Pareto Front surface approximation for the complete problem setup

PF exhibits an asymptotically decreasing slope (Fig. 8). As regards the relationship among the total cost due to bridges closure and *ADT*, the corresponding PF is again nonlinear and asymptotically convergent to a particular rate, which in turn is subject to the available budget (Fig. 9).

The PFs provided here expose the optimal trade-of among all available variables and correspond to a useful decision tool in cases of financial fluctuations and investments programming. Such (non-obvious) relationships among various variables are detectable using an optimization analysis like the one proposed herein. The results of the above analysis can be used for estimating the complete PF for all variables. Hence, in Fig. 10, an approximation of the PF by means of interpolation is estimated, offering a valuable for practical purposes PF surface.

Such analysis and results as the above presented and especially the PF surface approximation are valuable for ‘quick-response’ analysis for supporting strategic policy formation. It is noted that the above PF surface is valid for the problem’s configuration, despite the fact that the bridges’ selection sets for each point of the PF can be quite different.

## 8 Conclusions

This paper aims to offer some insights on the application of optimization frameworks in the maintenance programming of road bridges under budgetary constraints and social considerations. In particular, a comprehensive formulation is offered and

discussed, while computational results are also provided, suitably selected and presented in order to contribute to the understanding of the nature of this problem and the expected outcomes of such optimization problems. The real data used herein are extracted from the database of FHWA (National Bridge Inventory-NBI), after suitable preprocessing to adjust these to the scope of the present research effort.

The computational experiments conducted were focused on the most comprehensive, though not trivial, case of maximizing the total bridges traffic (which reflects the social element of the bridges users' benefit) subject to budgetary constraints under a parametric setup for alternative budget availability. An alternative optimization formulation was also considered additionally accounting for indirect costs due to bridge closure for improvement works. The results of these problem setups were investigated based on the general information that can be extracted by these runs. In brief, it can be concluded that the Branch-and-Bound algorithm was able to provide an adequate search of the domain in reasonable time, coming up with solutions to large and demanding problem setups while exploiting the available budget in all cases (which is a qualitative indication for its algorithmic performance). Additionally, the optimization setups formed here allowed 'wise' and consistent selections of the most 'prosperous' sets of bridges that should be upgraded. Interestingly enough, the results exposed relationships among the variables used, in particular between the ratio of the improvement cost and the total traffic, which are useful for practical purposes on one hand, but could be an issue that could worth further investigation on the other.

Finally, as a point of outlook, the problem of optimal maintenance programming realistically involves several other, more elaborative problem formulations. For example, multiple objectives and constraints should be introduced into the problem setup of a useful optimal budget allocation program, mainly reflecting on qualitative requirements about the existing bridges conditions, the type of bridges and several others. Nevertheless, the results offered in the current test runs can be regarded as the necessary starting point for directing further research effort.

**Acknowledgments** The authors would like to thank Mr. F. Alogdianakis, MSc, for extracting from the NBI database the raw data used in this paper.

## References

1. Organization for Economic Co-operation and Development—OECD (2001) Asset management for the roads sector. Paris, 52
2. Lagaros N, Kepaptsoglou K, Karlaftis M (2013) Fund allocation for civil infrastructure security upgrade. *J Manage Eng* 29(2):172–182
3. Lagaros N, Karlaftis M (2011) A critical assessment of metaheuristics for scheduling emergency infrastructure inspections. *Swarm Evol Comput* 1(3):147–163
4. Hu X, Madanat S (2014) Determination of optimal MR&R policies for retaining life-cycle connectivity of bridge networks. *J Infrastruct Syst*. [http://dx.doi.org/10.1061/\(ASCE\)IS.1943-555X.0000226](http://dx.doi.org/10.1061/(ASCE)IS.1943-555X.0000226)

5. Liu M, Frangopol DM (2006) Optimizing bridge network maintenance management under uncertainty with conflicting criteria: life-cycle maintenance, failure, and user costs. *J Struct Eng* 132(11):1835–1845
6. Robelin CA, Madanat SM (2008) Reliability-based system-level optimization of bridge maintenance and replacement decisions. *Transp Sci* 42(4):508–513
7. Kong JS, Frangopol DM (2003) Life-cycle reliability-based maintenance cost optimization of deteriorating structures with emphasis on bridges. *J Struct Eng* 129(6):818–828
8. Karlaftis M, Kepaptsoglou K, Lambropoulos S (2007) Fund allocation for transportation network recovery following natural disasters. *J Urban Plan Dev* 133(1):1–8
9. Kepaptsoglou K, Sinha KC (2001) Optimal programming under uncertainty in the indiana bridge management system. In: *Proceedings of the 80th transportation research board annual meeting*, Washington, DC, USA
10. Patidar V, Labi S, Sinha KC, Thompson P (2007) Multi-objective optimization for bridge management systems. NCHRP Report 590. Transportation Research Board, Washington, DC
11. Chan WT, Fwa T, Tan J (2003) Optimal fund-allocation for multidistrict highway agencies. *ASCE J Infrastruct Syst* 9(4):167–175 (American Society of Civil Engineers)
12. Wu Z, Flintsch G, Ferreira A, Picado-Santos L (2012) Framework for multiobjective optimization of physical highway assets investments. *J Transp Eng* 138(12):1411–1421
13. Fwa T, Farhan J (2012) Optimal multiasset maintenance budget allocation in highway asset management. *J Transp Eng* 138(10):1179–1187
14. Chang L, Peng F, Ouyang Y, Elnashai AS, Spencer BF Jr (2012) Bridge seismic retrofit program planning to maximize post-earthquake transportation network capacity. *J Infrastruct Syst* 18(2):75–88
15. Augeri M, Colombrita R, Greco S, Lo Certo A, Matarazzo B, Slowinski R (2011) Dominance-based rough set approach to budget allocation in highway maintenance activities. *J Infrastruct Syst* 17(2):75–85
16. Kuhn K (2010) Network-level infrastructure management using approximate dynamic programming. *J Infrastruct Syst* 16(2):103–111
17. Federal Highway Administration (FHWA) (1995) Recording and coding guide for the structure inventory and appraisal of the nation's bridges. Report No. FHWA-PD-96-001, US Department of Transportation, Washington, DC
18. Golabi K, Shepard R (1997) Pontis: a system for maintenance optimization and improvement of US bridge networks. *Interfaces* 27(1):71–88
19. Hillier FS, Lieberman GJ (2001) *Introduction to operations research*. McGraw-Hill, New York
20. Nemhauser GL, Wolsey LA (1988) *Integer and combinatorial optimization*. Wiley, New York
21. Goldberg DE (1989) *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley, Reading, MA