

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Asadi Bagloee, Saeed

Title:

Road space optimisation for multiclass and multimodal traffic networks

Date:

2017

Persistent Link:

<https://hdl.handle.net/11343/191654>

Terms and Conditions:

Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.

ROAD SPACE OPTIMISATION FOR MULTICLASS AND MULTIMODAL TRAFFIC NETWORKS



THE UNIVERSITY OF

MELBOURNE

Saeed Asadi Bagloee
Melbourne School of Engineering

Department of Infrastructure Engineering
The University of Melbourne

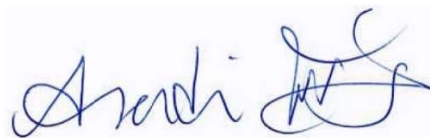
This dissertation is submitted for the degree of Doctor of Philosophy
August 2017

DECLARATION

This dissertation is the result of my own work and includes nothing, which is the outcome of work done in collaboration except where specifically indicated in the text. It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

In accordance with The University of Melbourne guidelines, this thesis does not exceed 100,000 words.

Signed: _____

A handwritten signature in blue ink, appearing to read 'Saeed Asadi Bagloee', is written over a light blue rectangular background.

Date: _____ 31/01/2017 _____

Saeed Asadi Bagloee

SUMMARY

Traffic congestion has become a serious concern and hindrance to the prosperity of many societies. Among a variety of solutions two approaches are of significant importance: constructing new roads and bridges to ease traffic congestion and promoting public transport. For the latter, the aim is to provide more space in the heart of cities for public transport (buses, trams, etc) aiming to get more commuters to their destinations. Therefore, two central questions have been addressed in this research; (i) investment in the road construction: given a number of candidate projects associated with construction expenses and a limited budget, what is the best choice of projects. This is known as the road network design problem (NDP), and (ii) transit priority lanes: given a road network, which roads should be selected to provide a lane to be exclusively used by public transport modes such that the overall performance of the transport system is not adversely affected. This problem is called the, “transit priority lane design problem” (TPLDP). For the former, (NDP) a hybridized method consisting of the branch and bound algorithm and Benders decomposition method has been developed. For the latter (TPLDP), the concept of Braess paradox was employed to seek for “mis-utilized” space in congested networks to be utilized by public transport. To this end, a merit index aiming to spot potentially some Braess-tainted roads is introduced first. Then a branch and bound algorithm was developed to find the best subset of the Braess tainted roads that have no adverse impact on the overall performance of the network.

This study advances the state of knowledge in the above mentioned problems in five areas:

- (i) the authenticity of the traffic model is enhanced by subjecting all the analysis to multimodal and multiclass traffic circulation,
- (ii) the methodologies developed in this study are tailored to real world applications as illustrated with numerical analysis,
- (iii) a RAM-efficient branch and bound algorithm (BB) has been developed such that the expansion of the BB’s tree structure becomes memoryless,
- (iv) inclusion of the Braess paradox in the pursuit of the transit priority lane would nullify possible adverse effects on the private modes, and

(v) a new method for the capacitated traffic assignment has been developed which is called inflated travel time.

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Majid Sarvi for his continuous support of my Ph.D study. I would also like to thank the rest of my advisory committee: Prof. Abbas Rajabifard, A/Prof. Russell G. Thompson, for their insightful comments and encouragement.

I am also indebted to two anonymous external reviewers for their meticulous and constructive comments on my thesis.

Last but not the least, I would like to thank my family for supporting me throughout this journey.

To my late father

... who gave his all to his kids, but didn't stay longer for the harvest season

Peer reviewed journal papers:

1. Bagloee, S.A., Sarvi, M., Patriksson, M. (in press) A hybrid branch-and-bound and Benders decomposition algorithm for the network design problem. *Computer-Aided Civil and Infrastructure Engineering*. **IF 5.28**
2. Bagloee, S., Sarvi, M., Patriksson, M., Rajabifard, A. (in press) A mixed user-equilibrium and system-optimal traffic assignment for connected vehicles stated as a complementarity problem. *Computer-Aided Civil and Infrastructure Engineering*. **IF 5.28**
3. Bagloee, S.A., Sarvi, M., Wolshon, B., Dixit, V. (in press) Identifying critical disruption scenarios and a global robustness index in road transport networks. *Transportation Research Part E: Logistics and Transportation Review*. **IF 2.80**
4. Bagloee, S.A., Sarvi, M., Wallace, M. (2016) Bicycle lane priority: Promoting bicycle as a green mode even in congested urban area. *Transportation Research Part A: Policy and Practice* 87, 102-121. **IF 2.39**
5. Bagloee, S., Sarvi, M., Rajabifard, A., Thompson, R., (in press) Identifying Achilles-heel roads in real sized networks, *Journal of Modern Transportation*, Springer. Submission code: **IF 0.73**.
6. Bagloee, S.A., Sarvi, M. (2015) Heuristic Approach to Capacitated Traffic Assignment Problem for Large-Scale Transport Networks. *Transportation Research Record: Journal of the Transportation Research Board*, 1-11. **IF 0.44**.

Book Chapter:

7. Sarvi, M., Bagloee, S.A., Bliemer, M., 2016. Network design for road transit priority in: Bliemer, M.C.J., Corinne, M., Claudine, M. (Eds.), *Handbook on Transport and Urban Planning in the Developed World*. Edward Elgar Publishing Ltd, UK.

Papers under review

8. Bagloee, S., Sarvi, M., Ceder, A., Transit priority lanes in the congested road networks, *Public Transport – Springer*, Submission code: PUTR-D-16-00018R2
9. Bagloee, S., Sarvi, M., Rajabifard, A., Thompson, R.G., System optimal relaxation and Benders decomposition algorithm for the large sized road network design problem, *International Journal of Logistics Systems and Management (IJLSM)*, Submission code: IJLSM-161310.
10. Bagloee, S., Sarvi, M., A modern congestion pricing policy for urban traffic: subsidy plus toll, *Journal of Modern Transportation*, Springer, Submission code: JMTR-D-16-00121R2

Peer reviewed Conference:

11. Bagloee, S., Sarvi, M. (2016) Shannon entropy to measure road network redundancy and reliability *Proceedings of Traffic Flow Theory (TFT2016)*, Sydney, Australia.
12. Bagloee, S., Sarvi, M. (2016) Capacitated traffic assignment problem subject to variable demand, a nonlinear formulation cum solution code in GAMS. *Proceedings of Australasian Transport Research Forum (ATRF 2016)*, Melbourne, Australia.

13. Bagloee, S., Sarvi, M., Rajabifard, A., Thompson, R.G., Saberi, M. (2016) A solution to the road network design problem for multimodal flow Proceedings of IEEE 19th International Conference on Intelligent Transportation Systems (ITSC 2016), Rio de Janeiro, Brazil.
14. Bagloee, S.A., Sarvi, M. (2015) A Heuristic Approach to Capacitated Traffic Assignment Problem Tailored to Large Scale Networks. Proceedings of Transportation Research Board, Washington D.C., United States.
15. Bagloee, S.A., Sarvi, M. (2016) Autonomous Vehicles: past, present and future implications. Intelligent Transportation Systems (ITS) World Congress, Melbourne, Australia.
16. Sarvi, M., Bagloee, S., Rajabifard, A., Thompson, R.G. (2016) Urban transport system: Large scale multiclass modelling; challenges, opportunities and future trend. Proceedings of Australasian Transport Research Forum (ATRF 2016), Australia, Melbourne.

Report:

17. Bagloee, SA, Florian, M, Sarvi, M., Centre Interuniversitaire de Recherche sur les Réseaux d'Entreprise (2016) A New Policy in Congestion Pricing: Why only Toll? Why not Subsidy? la Logistique et le Transport (CIRRELT), Montreal, Canada.

CONTENTS

1 INTRODUCTION	15
1.1 RESEARCH BACKGROUND	15
1.2 RESEARCH QUESTIONS	16
1.2.1 Road construction as a hard solution for traffic congestion.....	16
1.2.2 Priority as an efficient soft approach for reducing traffic congestion.....	17
1.3 KNOWLEDGE GAP	19
1.4 STRUCTURE OF THE PROBLEMS	19
1.4.1 Bi-level programing.....	20
1.4.2 Discrete or integer variables.....	20
1.4.3 Multiclass and multimodal traffic assignment	21
1.5 RESEARCH BARRIERS	21
1.6 STRATEGIC NATURE OF GENERALISED NETWORK DESIGN PROBLEMS	22
1.7 RESEARCH ROAD MAP	23
1.8 CONTRIBUTIONS OF THE RESEARCH	24
1.9 COMMUNICATION OF THE RESEARCH RESULTS.....	24
2 LITERATURE REVIEW	28
2.1 LITERATURE ON ROAD NETWORK DESIGN	28
2.1.1 Literature on the discrete road network design problem	29
2.1.2 A summary of the review of the road network design	32
2.2 LITERATURE ON THE TRANSIT PRIORITY LANE DESIGN	32
2.2.1 Literature on the transit priority lane design problem.....	33
2.2.2 A summary of literature on the TPLDP	35
2.3 SUMMARY AND CONCLUSION	35
3 MATHEMATICAL METHODS	37
3.1 SOLVING MIXED INTEGER BI-LEVEL PROGRAMING	37
3.2 SOLUTION ALGORITHMS FOR THE MINLP	39
3.2.1 Generalized Benders decomposition (GBD)	39
3.2.2 Outer approximation (OA)	40
3.2.3 Branch and bound (BB).....	41
3.3 A NUMERICAL EXAMPLE FOR GBD, OA AND BB.....	42
3.4 CONCLUSION	47
4 CAPACITATED TRAFFIC ASSIGNMENT PROBLEM.....	50
4.1 INTRODUCTION	50

4.2 LITERATURE REVIEW	52
4.3 MATHEMATICAL FEATURES	53
4.3.1 Formulation for capacitated traffic assignment problem (CTAP)	53
4.3.2 Mathematical features	55
4.3.3 Heuristic method to update the (initial) beta	56
4.3.4 Termination conditions.....	58
4.3.5 Capacity feasibility.....	58
4.4 NUMERICAL RESULTS	59
4.4.1 Hearn's benchmark case study.....	59
4.4.2 Large sized Winnipeg case study.....	61
4.5 CONCLUSION	66
5 THE ROAD NETWORK DESIGN PROBLEM	69
5.1 INTRODUCTION	69
5.2 FORMULATION OF THE DISCRETE EQUILIBRIUM NETWORK DESIGN PROBLEM.....	71
5.2.1 Treatment of multiclass and multimodal traffic	74
5.3 BRANCH-AND-BOUND ALGORITHM	75
5.3.1 Discreteness of the DNDP over the BB.....	75
5.3.2 Navigation in the BB's tree	76
5.3.3 Branching rule based on merit index:	77
5.3.4 Node selection rule	78
5.4 BENDERS DECOMPOSITION	78
5.4.1 Formulation of the lower bound.....	80
5.4.2 Benders decomposition method for a MNLP.....	81
5.4.3 Benders decomposition for a system optimal DNDP	82
5.4.4 Evaluation of the merit index in the performance of the BB algorithm	85
5.4.5 Some remarks on the methodology.....	87
5.5 NUMERICAL EVALUATIONS.....	91
5.5.1 Example 1: Gao's network	94
5.5.2 Example 2, Sioux-Falls network.....	96
5.5.3 Example 3: Winnipeg large-scale network.....	97
5.6 CONCLUSIONS.....	103
6 TRANSIT PRIORITY LANES	105
6.1 INTRODUCTION	105
6.2 THE TRANSIT PRIORITY LANES DESIGN PROBLEM (TPLDP)	107
6.3 METHODOLOGY	111

6.3.1 <i>Branch and bound in the context of optimisation methods</i>	111
6.3.2 <i>Merit index to find candidate roads</i>	111
6.3.3 <i>A tight lower bound</i>	114
6.4 NUMERICAL DEMONSTRATION	116
6.5 CONCLUSIONS.....	120
7 CONCLUSION	124
7.1 SUMMARY OF THE RESEARCH	124
7.2 SOLUTION METHODOLOGIES	126
7.3 CONTRIBUTIONS OF THE RESEARCH	127
7.4 SUGGESTIONS FOR FURTHER RESEARCH.....	127
8 REFERENCES	130

LIST OF TABLES

TABLE 3.1 RESULTS OF GENERALIZED BENDERS DECOMPOSITION METHOD	45
TABLE 3.2 RESULTS OF OUTER APPROXIMATION METHOD	46
TABLE 3.3 RESULTS OF BRANCH AND BOUND METHOD.....	47
TABLE 4.1 HEARN NETWORK: COMPARISONS RESULTS.....	60
TABLE 5.1 EXAMPLE 1, GAO’S NETWORK: GBD (GAO ET AL., 2005) VERSUS PROPOSED BB-B	96
TABLE 5.2, EXAMPLE 2;SIOUX-FALLS: BB-OA (FARVARESH AND SEPEHRI, 2013) VS PROPOSED BB-B.....	97
TABLE 5.3 WINNIPEG EXAMPLE, TWO-WAY (CANDIDATE) ROAD PROJECTS SORTED BASED ON THE MERIT INDEX	99
TABLE 5.4 WINNIPEG CASE STUDY: RESULTS OF THE PROPOSED BB-B.....	102
TABLE 6.1 CANDIDATE LINKS TO BE CONSIDERED AS TRANSIT PRIORITY LANES, DATA FOR WINNIPEG, CANADA.....	118
TABLE 6.2 NUMERICAL RESULT FOR DATA FOR WINNIPEG,	119

LIST OF FIGURES

FIGURE 1.1 THE IDEA OF TRANSIT PRIORITY LANE	17
FIGURE 1.2 PUBLIC TRANSPORT AS A MEANS TO ALLEVIATE TRAFFIC CONGESTION.....	18
FIGURE 1.3 HOW HARD IS NP-HARD IN COMPUTATIONAL COMPLEXITY THEORY	22
FIGURE 4.1 CONCEPTUAL REPRESENTATION OF THE PROPOSED METHODOLOGY ON THE ROAD DELAY FUNCTIONS	57
FIGURE 4.2 HEARN NETWORK: (A) CONVERGENCE, (B) FLUCTUATION OF FLOWS ON THE SATURATED LINKS.	63
FIGURE 4.3 WINNIPEG NETWORK RESULTS: (A) RESIDUAL FLOWS, (B) PACE VALUES, (C) CONVERGENCE.....	64
FIGURE 4.4 WINNIPEG NETWORK RESULTS, 4 TOP SATURATED LINKS: (A) FLUCTUATION OF FLOW, (B) INITIAL-BETA.....	65
FIGURE 4.5 WINNIPEG NETWORK RESULTS, SENSITIVITY ANALYSIS BETWEEN WITH/WITHOUT EXCESSIVE DEMAND SCENARIOS: (A) TOTAL FLOW ON DUMMY LINKS, (B) VARIATIONS OF BECKMANN VALUES (EXCLUDING DUMMY LINKS) OVER ITERATIONS	68
FIGURE 5.1 PROPOSED NODE SELECTION AND BRANCHING IN THE BRANCH-AND-BOUND ALGORITHM	79
FIGURE 5.2 BB'S PERFORMANCE WITH/WITHOUT THE MERIT INDEX (NOTE: v^j IS VALUE OF THE OBJECTIVE FUNCTION AT ITERATION i , AND ub^* IS THE BEST UPPER BOUND OR THE INCUMBENT VALUE)	86
FIGURE 5.3 ILLUSTRATION OF HOW THE LOWER BOUND VALUES ARE INHERITED THROUGH THE TREE	88
FIGURE 5.4 BENDERS ALGORITHM: THE FLOWCHART GRAPHICALLY REPRESENTS THE STEPS	92
FIGURE 5.5, A ONE-WAY SWITCH LINK (DASHED LINE) REPRESENTING A TWO-WAY ROAD (RED AND GREEN LINES)	94
FIGURE 5.6 GAO'S TEST NETWORK	95
FIGURE 5.7 WINNIPEG'S CENTRAL BUSINESS DISTRICT,	98
FIGURE 5.8 WINNIPEG EXAMPLE, WITH 20 ROAD PROJECTS.	100
FIGURE 6.1 IMPACT OF THE ALPHAS ON THE COMPUTATIONAL TIME AND THE OBJECTIVE FUNCTION	120
FIGURE 6.2 WINNIPEG TRANSPORT NETWORK AND SELECTED TRANSIT PRIORITY LANES	121

LIST OF ABBREVIATIONS AND ACRONYMS

ALM	augmented Lagrangian method
ATRF	Australasian Transport Research Forum
BB	branch-and-bound
B-MINLP	Bi-level mixed integer nonlinear programming problem
BP	Braess paradox
BPR	bureau of public roads (BPR) delay function
CBD	central business district
CNLP	continuous nonlinear programming
CTAP	capacitated traffic assignment problems
DNDP	Discrete network design problem
DPF	dynamic penalty function
DTA	dynamic traffic assignment
FW	Frank-Wolfe
GA	genetic algorithm
GBD	generalized Benders decomposition
HOV	high occupancy vehicles
IPF	inner penalty function
ITT	inflated travel time
KKT	Karush–Kuhn–Tucker
LOS	level of service
MGBD	master generalized Benders decomposition
MINLP	mixed integer nonlinear programming
MILP	mixed integer linear programming
MIP	mixed integer problems
MMMC-UE-TAP	multimodal, multiclass user equilibrium traffic assignment problem

NDP	network design problem
NLP	nonlinear Programming
NP hard	NP hard (non-deterministic polynomial-time hard)
OA	outer approximation
OD	origin-destination
OR	operational research
PCE	passenger car equivalency
PCU	passenger car unit
PSP	primal sub-problem
RMP	relaxed master problem
SO	system optimal
SO-DNDP	system-optimal discrete network design problem
TAP	traffic assignment problem
TDM	travel demand management
TPLDP	transit priority lane design problem
TRB	transportation research board
UE	user-equilibrium
UE-TAP	UE traffic assignment problem
UE-TAP	user equilibrium traffic assignment problem
VI	variational inequality
vphpl	vehicle per hour per lanes

1 INTRODUCTION

In this chapter a snapshot of the research, background, research questions, aims and scope as well as contributions are provided. An outline of the research plan and a roadmap as well as the overall structure of the thesis is also presented.

1.1 Research background

Traffic congestion is emerging as a major constraint to the achievement of national economic goals in many cities around the world. Studies have shown that the total amount of travel undertaken by residents of Australian cities has grown ten-fold in the last 60 years, and the cost of traffic congestion to the economy totalled \$9.4 billion in 2005. These costs are amongst the highest in the world when compared with Australia's gross domestic product (GDP) (Sarvi et al., 2016). As urban populations continue to grow, traffic congestion is expected to increase in developed and developing countries. Even if there is still a need to build new freeway links and roads, particularly in outer city areas, the congestion problem cannot for, economic and social and environmental reasons, be solved simply by building more and more roads. Although, in some cases, new roads, bridges, highways, tunnel etc. will be inevitable, making the best use of the existing infrastructure is equally, if not more important. The former is laborious, capital intensive and time consuming (called a hard approach) whereas the latter as a soft approach is much less onerous.

In this study, in order to address traffic congestion the both approaches are investigated. The hard approach is clear, it is all about investment in new road infrastructure which must be as efficient and wise as possible. The viability of such investments can be elaborated as follows. There exists a limited budget and a number of road construction projects associated with expenses pertaining to the design, planning, land acquisition, construction, etc. The question then becomes what would be the best choice of projects.

The soft approach seeks to secure the most efficient use of existing road space. Its mandate is to maximise the efficiency of the traffic circulation which is emerging as one of the key issues for urban transport planners, local government officials, and representatives of national governments. A traffic system comprises of a variety of

distinct modes such as public transport, private vehicles, freight, high occupancy vehicles, etc. In the context of road space optimisation, the main question that arises is: what is the best split of the road space between different modes. To this end the key point is to first realise that the purpose of mobility is to move people not vehicles, therefore, persons should be given the highest priority. This is the fundamental tenet of road space optimisation which assigns mass transit as the highest priority. In the following section the research questions are defined and the soft and hard approaches are elaborated on.

1.2 Research questions

The aim of this research is to alleviate traffic congestion for which two solutions were investigated. Therefore, two questions are defined as follows.

1.2.1 Road construction as a hard solution for traffic congestion

Given the constant increase in travel demand, road construction sometimes is the only solution. This includes new roads, bridges, grade-separated interchanges, tunnels as well as lane widening. Though the pace of such investment in developed countries has slowed down remarkably, it is on a sharp upward trend in some developing countries (Bagloee et al., 2016a).

Road construction or investment, also known as the network design problem (NDP) is not a new concept. It is the first natural and intuitive solution to the traffic congestion predicament for which there exists a plethora of research in the literature (Farahani et al., 2013). There exists a number of different definitions of the NDP. Nevertheless, the common and one of the most difficult is discussed as follows: on the one hand, there are a number of (candidate) projects associated with costs covering construction expenses including machinery, labour, land acquisition, materials, etc. and on the other hand there is a limited budget which is clearly less than the expenses. In other words, one cannot afford to build all the candidate projects. Therefore, the question to be asked is: out of the candidates, which projects should be selected for construction. Efforts to improve mobility with such costly schemes may be compromised if no proper due diligence is exercised. New roads and additional capacity are expected to improve traffic circulation, compared to the “do-nothing” or existing road scenario. In other words, new roads are expected to not to degrade overall traffic circulation. It has however been proven mathematically and observed in real practice that sometimes, by adding more roads may

counter-intuitively increase the travel time which is called Braess paradox (Braess, 1968; Braess et al., 2005).

1.2.2 Priority as an efficient soft approach for reducing traffic congestion

Efficiently managing transport networks utilising a variety of measures such as travel demand management (TDM) have been a significant alternative implemented in many cities (Nelson, 2000). Mass transit modes (public transport) face significant efficiency and effectiveness issues in situations where traffic congestion is high and growing. To this end some city jurisdictions are promoting the development of traffic priority systems for the public transport. Given the high ridership, mass transit modes deserve to receive the highest priority. Traffic priority includes a wide range of measures ranging from a full reallocation of road space to creating on-road transit lanes (exclusive bus lanes, or simply transit priority lanes) to adjustments to road layouts to remove traffic bottlenecks. Network space management involves balancing often-conflicting pressures for limited space. Reallocation of road space to give priority to bus services and the like is one of many of such measures (see Figure 1.1). It stands to reason that moving towards the mass transit or public transport also has a positive impact on relieving traffic congestion as well as improving the environment (see Figure 1.2).



Figure 1.1 The idea of transit priority lane



Figure 1.2 Public transport as a means to alleviate traffic congestion

In cities where transit use is high, the case for priority is relatively easy to justify. However, deciding the extent to which priority should be given is not so simple. In cities where transit use is low and roads are congested, economic justification for transit priority is less clear. As a result, traffic authorities are often faced with the realities of a finite resource, high car usage and low transit usage. In such cases, irrespective of the undeniable advantages of the public transport (environmental, efficiency and reliability improvements) the idea of exclusive bus lanes can be made more appealing if these lanes can also be used by freight and high occupancy vehicles (HOV). HOV are vehicles with a minimum of two occupants. Allowing HOV and freight traffic to travel in priority bus lanes can overcome the ‘empty lane syndrome’ common to lightly used bus lanes. This could provide the required justification and opportunity to introduce comprehensive transit priority lanes across the transport network rather than scarce, local and limited priorities.

Similarly, one can move towards a more sustainable and green modes of mobility using bicycles by providing a safer environment for cyclists based on the concept of bicycle priority lanes.

Generally speaking, the concept of priority has received significant momentum in recent years (Zheng and Geroliminis, 2013a) from researchers and practitioners (Basso et al., 2011; Mesbah et al., 2011b). Throughout this dissertation the problem is referred to as the Transit Priority Lane Design Problem (TPLDP).

Simply speaking the TPLDP can be stated as a two-fold question as follows. (i) Amongst congested roads which roads are deemed to be appropriate to give away one lane to be exclusively used by public transport, dubbed as a transit priority lane. An answer to this question results in a number of candidate roads yet to be further processed in the second question. (ii) Given the set of candidate roads derived from the previous question, implementation of the transit lane comes at some expense such as lane marking, signage, signals, etc. In other words, each candidate road is associated with a certain cost. Whereas, there exists a limited budget as well. Now the question to be answered is: given the candidate set, costs and budgets, which roads should be selected to designate a lane as transit lane?

1.3 Knowledge gap

A review of the literature indicates a number of shortcomings in past studies. Notably, “practicality” is relatively rare in the literature. In other words, applications of methods to large sized road networks –as is the case in real life situations- are yet to be addressed. Secondly, given the computational complexities, some aspects of the problems have been ignored or loosely treated. More precisely, any solution to the TPLDP or the NDP must be thoroughly examined based upon a reliable model to measure the traffic circulation which has been largely relaxed due to the theoretical and computational burdens.

To shed more light on the above deficiencies, in the next section the structure and different features of the problems followed by an extensive discussion on the computational difficulties of them are outlined

1.4 Structure of the problems

The aim of this research is to address the traffic congestion on two fronts, the NDP as a hard measure and the TPLDP as a soft approach. In the next section, the problems are discussed by elaborating on three fundamental technical features of them.

1.4.1 Bi-level programming

The approach towards addressing these problems must be a network-wide approach. That is, a change – even slightly - in one corner of a road network may ripple through the entire network. More precisely, people as users of the transport system react to the changes in the system. Therefore, it is important to come up with a mechanism by which the behaviour of the people with respect to the changes in the transport system is fully accounted for, which is called the traffic assignment problem (TAP). The TAP itself is an optimisation problem (Beckmann et al., 1956) to ensure Wardropian traffic flow or better known as the user-equilibrium (UE) that is the drivers choose the quickest paths or routes to get to their destinations.

It is obvious that any initiative such as the introduction of the transit priority or constructing a new bridge is supposed to improve the traffic circulation or the performance of the transport system. The consensus in the literature centres on the cost of system as an intuitive index to measure the performance of the transport system. The travel time spent by people in the transport system is widely considered as a valid surrogate to measure the performance of the transport system. Overall, the aim is to minimize the total travel time spent in the road network inferred as the system-cost.

Consequently, one must solve two problems simultaneously: the minimization of the total travel time and the TAP. In the optimisation or (Operations Research (OR) literature, the term to describe these kind of simultaneous problems is “bi-level” which are known to be of difficult problems. Accordingly, these problems are formulated as generalized bi-level programming problems to account for the behaviour of commuters in the lower level and minimization of the system-cost in the upper level.

1.4.2 Discrete or integer variables

This research can aid traffic planners, authorities and managers with their decision related to where to provide transit priority where to construct new roads. These problems involve a number of decision variables which are intrinsically discrete, integer or binary (1 to build and 0 not to build). In optimisation, integer programming (versus the continuous) are also known to be difficult problems. As the result, the discreteness of the problem is ensured by employing binary variables as decision variables. For the TPLDP, a decision variable is assigned value of 1; to allocate a lane of the roadway as a transit priority lane

and 0 otherwise. For the discrete NDP (also denoted as DNDP), decision variables are 1 to build the road and 0 not to build. Given the inevitable presence of integer variables in conjunction with some continuous variables such as travel times and traffic volumes in the objective functions, the above problems become mixed integer programming (MIP) problems. In terms of the computational burden, MIPs are known to be difficult problems to solve.

1.4.3 Multiclass and multimodal traffic assignment

The literature review underscores the significant shortcoming in past studies in which traffic congestion is loosely treated mainly because of the computational complexities of bi-level programming. It is important to consider the mutual impacts of both transit and private modes which is known as multimodal traffic assignment. In addition, within the private modes, there exists a number of distinct classes such as HOV, freight, taxis which is referred to as a multiclass traffic assignment.

1.5 Research barriers

It is essential to first gain some insights of the level of the difficulties of the problems before attempting to solve them. This will help in looking for appropriate solution methods in the quest of addressing the problems. The above-mentioned triplet features define the level of difficulty of the problems as discussed below.

In computational complexity theory, the time that is required to solve a problem is regarded as the degree to which the respective problem is “hard” or “easy”. The computational time varies with respect to the size of the problem, (note that, number of variables and parameters describe the size of a problem). Moreover, for “easy” problems such as shortest path finding algorithms, the computational time is a “polynomial” function of the size, denoted by “P” as shown in Figure 1.3.

Unfortunately both bi-level programming problems and mixed integer programming problems are proven to be of highest complexities known as NP hard (non-deterministic polynomial-time hard). Simply stated, in an NP hard problem, as the size increases the problem rapidly becomes intractable. Furthermore, the problems, as set out before, carry some nonlinear terms such as the total travel time as an objective function (note that the total travel time can be formulated as “sum over all roads (traffic volume * travel time)”,

the terms inside the parentheses is a nonlinear terms). Nonlinearity also adds extra complexities.

For bi-level programming problems, it is mathematically proven that even when all terms are linear the problem is still NP-hard (Ben-Ayed and Blair, 1990; Colson et al., 2005, 2007; Dempe, 2003). Consequently, integer versus continuous, nonlinear versus linear, bi-level versus single-level add extra burden on the computational requirements. Unfortunately, the above problems (TPLDP and DNDP) bear all the difficult ingredients: mixed integer, nonlinear and bi-level. Therefore, in the design of the solution methodologies it is necessary to develop innovative schemes to address the hardness of these problems.

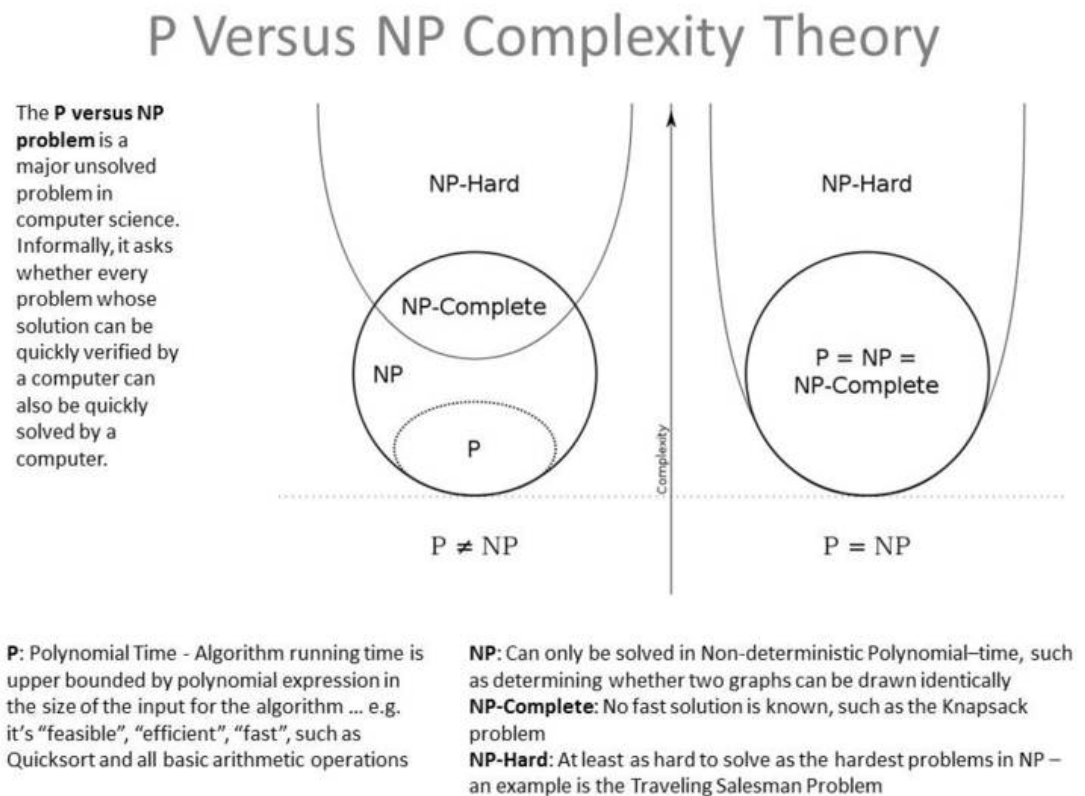


Figure 1.3 How hard is NP-hard in computational complexity theory

1.6 Strategic nature of generalised network design problems

With respect to the central position of NDP in the literature the following remarks are worth noting:

- In light of our exposure to real projects in industry, it should be acknowledged that there is a large gap between academic and engineering practice, and this problem is becoming more relevant. The main problem is that it is difficult to find research that tailors methodologies for large-sized networks. In addition, cities' traffic models are more complicated than some pedagogical networks such as Sioux-Falls, South Dakota in the United States. For instance, multiclass and multimodal features are indispensable parts of such models, which are largely simplified in the literature. So researchers have yet to offer a product meeting real needs of the industry.
- Problems such as the NDP deserve to be viewed as milestone or benchmark problems which test our knowledge and computational technologies at time. Working toward such milestone problems may bring some other advantages or by-products. For instance, in the quest to solve the discrete NDP, it was found that the Lagrangian sub-problem (of the Benders decomposition method) is in fact a capacitated traffic assignment problem which is still a live and relevant subject for both scholars as well as practitioners. Furthermore, the algorithm developed for the NDP can also be applied to other contemporary problems such as congestion pricing, facility placement, etc.
- Types of NDP have long been of great interest to scholars in operational research especially when mixed integer programming (MIP) problems are of concern. MIP has widespread applications in management science, manufacturing designs, decision making and planning, etc.

1.7 Research road map

This research began with an extensive and comprehensive review of the relevant literature in the transport science as well as mathematical and optimisation literature which involved 168 books, papers or reports. In the next chapter the literature review covering themes related to the road network design as well as the transit priority is presented. In Section 3, a variety of mathematical methods for solving mixed integer nonlinear problem as the core components of the undertaken problems are described. This includes discussion of the Benders decomposition, outer approximation as well as branch and bound algorithms. The pros and cons of each method are also investigated in detail. A methodology developed for the network design problem using a combination of the Benders decomposition and branch and bound is then described. To this end, it is needed

to first solve a capacity-constraint traffic assignment as a prerequisite of the Benders decomposition, for which a new method called inflated travel time has been developed (see chapter 4). The DNDP is tackled in chapter 5. Chapter 6 is dedicated to the transit priority lanes design problem for which an efficient branch and bound algorithm is developed. Chapter 7 concludes the thesis in which the contributions are highlighted and several themes for future studies are proposed

Before closing this chapter, an overview of the contributions as well as a report on the publications already arising from this research, communicated in the form of journal papers, book chapter and peer reviewed conference papers are provided.

1.8 Contributions of the research

The previous section described the knowledge gap, research problems, the research difficulties and various features of the research problems. The contributions of this research include:

- A network based approach for the problem of transit priority lane design (TPLDP) is developed.
- The discrete network design problem (DNDP) which is a benchmark problem in the computational complexity is solved using an exact method consisting of a Benders decomposition method and a branch and bound algorithm.
- For both problems (TPLDP and DNDP), the methodologies are tailored for real life road networks.
- A RAM-efficient and memoryless branch and bound algorithm based on an innovative concept (merit index) is developed.
- To enhance realism of the models, in the solutions provided for the two problems, the models are subjected to multiclass and multimodal traffic flow.
- A parameter-less method is developed for the capacitated traffic assignment problem.

1.9 Communication of the research results

Below is a complete list of the papers derived from this research which comprises of six peer-reviewed journal papers, one book chapter, six peer reviewed conference papers or presentations, one report as well as three papers currently under review. The genesis of the publications is described below.

This research started off with an extensive literature review and a significant effort to compile some leading algorithms proposed for the MIP which was published as a book chapter (paper # 7) as well as a conference paper recently presented at the Australasian Transport Research Forum (ATRF) in 2016, (# 13).

It turned out that applications of the Benders decompositions to the DNDP has a prerequisite better known as the capacitated traffic assignment for which a new method was developed and presented at the TRB annual conference in 2016 (paper #11) and subsequent journal publication (paper # 6). This line of research was further extended to the applications of the GAMS a leading optimisation software which resulted in paper # 9. Moreover, the idea of capacitated traffic assignment and GAMS' application were further investigated for the emerging connected vehicle technologies and autonomous vehicles (papers #12 and #2). Furthermore, the concept of capacitated traffic assignment was also applied to the famous problem of congestion pricing (papers #14 and #17).

The DNDP was extensively investigated which resulted in papers (#1, #10 and #16). In particular, paper #1 reports on the hybridised method (Benders decomposition and branch and bound) for the DNDP. As for priority lanes, first an application of this concept to bicycle lanes was published (paper #4). Paper #15 reports on the TPLDP.

Applications of the DNDP was extended to the problems pertaining to disaster management, identifying critical roads and disruption scenarios (papers #3, #8, and #5).

Peer reviewed journal papers:

1. Bagloee, S.A., Sarvi, M., Patriksson, M. (in press) A hybrid branch-and-bound and Benders decomposition algorithm for the network design problem. *Computer-Aided Civil and Infrastructure Engineering*. **IF 5.28**
2. Bagloee, S., Sarvi, M., Patriksson, M., Rajabifard, A. (in press) A mixed user-equilibrium and system-optimal traffic assignment for connected vehicles stated as a complementarity problem. *Computer-Aided Civil and Infrastructure Engineering*. **IF 5.28**
3. Bagloee, S.A., Sarvi, M., Wolshon, B., Dixit, V. (in press) Identifying critical disruption scenarios and a global robustness index in road transport networks. *Transportation Research Part E: Logistics and Transportation Review*. **IF 2.80**
4. Bagloee, S.A., Sarvi, M., Wallace, M. (2016) Bicycle lane priority: Promoting bicycle as a green mode even in congested urban area. *Transportation Research Part A: Policy and Practice* 87, 102-121. **IF 2.39**

5. Bagloee, S., Sarvi, M., Rajabifard, A., Thompson, R., (in press) Identifying Achilles-heel roads in real sized networks, *Journal of Modern Transportation*, springer. Submission code: **IF 0.73**.
6. Bagloee, S.A., Sarvi, M. (2015) Heuristic Approach to Capacitated Traffic Assignment Problem for Large-Scale Transport Networks. *Transportation Research Record: Journal of the Transportation Research Board*, 1-11. **IF 0.44**.

Book Chapter:

7. Sarvi, M., Bagloee, S.A., Bliemer, M., 2016. Network design for road transit priority in: Bliemer, M.C.J., Corinne, M., Claudine, M. (Eds.), *Handbook on Transport and Urban Planning in the Developed World*. Edward Elgar Publishing Ltd, UK.

Peer reviewed Conference:

8. Bagloee, S., Sarvi, M. (2016) Shannon entropy to measure road network redundancy and reliability *Proceedings of Traffic Flow Theory (TFT2016)*, Sydney, Australia.
9. Bagloee, S., Sarvi, M. (2016) Capacitated traffic assignment problem subject to variable demand, a nonlinear formulation cum solution code in GAMS. *Proceedings of Australasian Transport Research Forum (ATRF 2016)*, Melbourne, Australia.
10. Bagloee, S., Sarvi, M., Rajabifard, A., Thompson, R.G., Saberi, M. (2016) A solution to the road network design problem for multimodal flow *Proceedings of IEEE 19th International Conference on Intelligent Transportation Systems (ITSC 2016)*, Rio de Janeiro, Brazil.
11. Bagloee, S.A., Sarvi, M. (2015) A Heuristic Approach to Capacitated Traffic Assignment Problem Tailored to Large Scale Networks. *Proceedings of Transportation Research Board*, Washington D.C., United States.
12. Bagloee, S.A., Sarvi, M. (2016) Autonomous Vehicles: past, present and future implications. *Intelligent Transportation Systems (ITS) World Congress*, Melbourne, Australia.
13. Sarvi, M., Bagloee, S., Rajabifard, A., Thompson, R.G. (2016) Urban transport system: Large scale multiclass modelling; challenges, opportunities and future

trend. Proceedings of Australasian Transport Research Forum (ATRF 2016), Australia, Melbourne.

Report:

14. Bagloee, SA, Florian, M, Sarvi, M., Centre Interuniversitaire de Recherche sur les Réseaux d'Entreprise (2016) A New Policy in Congestion Pricing: Why only Toll? Why not Subsidy? la Logistique et le Transport (CIRRELT), Montreal, Canada.

Papers under review

15. Bagloee, S., Sarvi, M., Ceder, A., Transit priority lanes in the congested road networks, Public Transport – Springer, Submission code: PUTR-D-16-00018R2
16. Bagloee, S., Sarvi, M., Rajabifard, A., Thompson, R.G., System optimal relaxation and Benders decomposition algorithm for the large sized road network design problem, International Journal of Logistics Systems and Management (IJLSM), Submission code: IJLSM-161310.
17. Begloee, S., Sarvi, M., A modern congestion pricing policy for urban traffic: subsidy plus toll, Journal of Modern Transportation, Springer, Submission code: JMTR-D-16-00121R2

2 LITERATURE REVIEW

In this chapter an extensive review of the relevant literature in road network design as well as transit priority lanes is presented. Accordingly, this chapter is presented into two sections and the findings are summarised at the end of each section.

2.1 Literature on road network design

With respect to the types of the decision variables, the NDP is classified as discrete network design problem (DNDP) and continuous network design problem. The latter suffers from a lesser degree of realism and fidelity where the outcomes are a number of real values (not binary) begging for interpretations. Interested readers in the continuous network design problem can consult with (Lin, 2011; Unnikrishnan and Lin, 2012; Waller et al., 2006).

The NDP in general and the DNDP in particular have been studied extensively in mathematics literature as well as transportation science. A thorough discussion of the problems can be found in (Farahani et al., 2013; Magnanti and Wong, 1984; Minoux, 1989; Yang and Bell, 1998). The approaches taken in the literature can be classified as exact and heuristics. The exact methods aim to arrive at a global optimal solution but their applications to real networks are restricted. On the other hand, heuristic methods aim to render good solutions for sizeable networks within an acceptable computational time by relaxing some crucial properties of the problem (such as discreteness of the decision variables) (Wong, 1985). Furthermore, in recent years, rapid expansions in the developing and emerging economies (largely in Asia and the Middle East) have made the DNDP relevant even amongst practitioners. In spite of the time-effectiveness of the heuristic methods, they are yet to appeal to the industry. This is due in part to the fact that the heuristic methods provide supposedly good, but sometimes non-deterministic (or random) solutions. Such an uncertainty and the lack of stability of the results make them hard to sell. Moreover, as long as the best solution is not known, the degree of goodness of the solutions remains obscure.

Fortunately, ongoing enhancements in computational technology provides momentum to pursue exact methods (Wang et al., 2015). It has been estimated during the

course of a decade, optimisation methods have become a million times faster thanks to improvements in hardware as well as software (Lodi, 2010). On the other hand, although the size of the networks dealt with in the industry are large, the number of candidate projects (decision variables) is limited (say a dozen or so). It is mainly the number of binary variables (and not the size of the networks) that significantly determines the solution spaces (Bagloee et al., 2013b).

Given the above-mentioned characteristics of the problem and available computational technology, in this study an exact method for the DNDP, tailored to real-size networks is developed. Consequently, in this section, a synthesized overview of the literature with a primary focus on exact methods is provided. Then, the recent evolutions in the literature over the course of the last two decades are described

2.1.1 Literature on the discrete road network design problem

Among the pioneers, LeBlanc (1975) solved the DNDP using a branch-and-bound (BB) algorithm. Poorzahedy and Turnquist (1982) approximated the total travel time function in the upper level to a well behaved function and arrived at a single-level formulation that was then solved by a heuristic BB algorithm. Farvaresh and Sepehri (2013) proposed a more efficient BB. Generally speaking, one of the challenging dimensions of the DNDP lies in how to tune the methodology to deal with the intrinsic non-convexity that arises from the non-linear UE constraints (Wang et al., 2013) to account for traffic circulation. Gao et al. (2005) introduced the concept of a support function to include new additional projects into the traffic flow by which the bi-level DNDP was transformed into a general, mixed, non-linear problem. They then employed the generalized Benders decomposition (GBD) method as a solution algorithm.

Wang and Lo (2010) employed complementary constraints for UE traffic flow to arrive at a single-level problem for which a convex-combination based piecewise linear approximation was developed as a solution algorithm. Luatthep et al. (2011) transformed the DNDP to a single-level problem in which the variational inequality (VI) constraints represent the UE conditions, followed by a cutting plane based algorithm to seek the optimal solution. Farvaresh and Sepehri (2011) replaced the UE conditions with equivalent Karush–Kuhn–Tucker (KKT) conditions which led to a single-level, mixed-integer linear problem. Wang et al. (2013) expanded the DNDP model to find the optimal number of lanes for existing candidate roads. Fontaine and Minner (2014) employed a

piecewise linear approximation scheme to arrive at a single-level, mixed integer linear problem to be solved by Benders decomposition method. There are also heuristic approaches for the DNDP in the literature in which a variety of methods such as genetic algorithms, ant colony systems and hybrid meta-heuristics are used. A thorough review of these methods is provided by (Bagloee et al., 2013b).

As the above review shows, a general approach to bi-level programming problems such as the DNDP is to transform the problem into a single-level problem (Colson et al., 2005, 2007; Dempe, 2003). The convention is to replace the lower level decisions by an implicitly-determined function (reaction function) or by corresponding KKT conditions. The resulting single-level MINLP problem is then solved by various methods such as Benders decomposition, Lagrangian relaxation, descent methods (such as sequential quadratic programming), outer approximation, branch-and-bound, penalty function methods, or trust-region methods (Floudas, 1995; Leyffer, 1993; Li and Sun, 2006). A detailed review of the recent literature on the discrete road network design problem is discussed below to shed more light on the subject.

Gao et al. (2005) developed a methodology based on the concept of a support function to transform the bi-level NDP problem into a single-level MINLP. The resulting problem is then solved using the generalized Benders decomposition algorithm (Geoffrion, 1972). It is numerically shown that their method fails to find a global optimum solution in some cases (Farvaresh and Sepehri, 2013).

Zhang and Gao (2009) formulated a mixed, continuous and discrete NDP as a single-level mathematical programming problem with complementarity constraints to represent the UE traffic flow. Although the numerical results presented for small-scale examples are promising, due to employing a locally convergent algorithm, the capacity of the methodology to arrive at optimal solutions has yet to be investigated.

Wang and Lo (2010) developed a single-level optimisation formulation with complementary constraints for UE traffic flow that transforms the DNDP into a mixed integer linear programming (MILP) problem. The resulting MILP model is based on enumerating the paths between origins and destinations (OD) and a piecewise linear approximation of the link travel time functions with binary decision variables. Thus, the outcomes are significantly dependent upon the linearization scheme. With respect to the path enumeration component, its application to sizable networks has yet to be investigated.

Similarly, Luathep et al. (2011) formulated the DNDP as a single-level optimisation problem with a variational inequality constraint representing the UE conditions. The VI constraint efficiently obviates the need for path enumeration based on the accumulated number of extreme points. Nonetheless, the results are significantly dependent upon the linearization scheme used for the delay functions. The network of Sioux-Falls (of 24 nodes) is used for numerical analysis. In view of the number of extreme points used for representing the UE condition, application to large sized networks has yet to be investigated.

Farvaresh and Sepehri (2011) developed a single-level MILP by representing the UE conditions as KKT conditions and employing linearization schemes. To do so, the non-linear delay function is replaced (or approximated) by some linear segments. Hence the linearization scheme refers to the location and number of segments. The final results vary over different linearization parameters.

Farvaresh and Sepehri (2013) address the bi-level aspect of the problem explicitly by developing a branch-and-bound algorithm based on the seminal work of LeBlanc (1975) while a more efficient lower bound is sought. Given a feasible binary solution, the UE traffic assignment problem (UE-TAP) is solved to obtain an upper bound of the total travel time. A valid lower bound is also obtained by solving a system optimal (SO) version of the NDP. Due to the gap that usually exists between these bounds (Roughgarden and Tardos, 2002), applications have been limited to cases in which the difference in traffic flow under UE and SO conditions is negligible.

Wang et al. (2013) expanded the NDP to consider the number of additional lanes as decision variables. They first relaxed the bi-level programming model by formulating a single-level problem in which a SO (not UE) network design problem is solved. Two methods based on the relationship between UE and SO principles are developed. The first method, termed SO relaxation, takes advantage of the property that an optimal network design decision under SO traffic flow condition can be regarded as an approximate solution with UE traffic flow. The second method, termed UE reduction, reduces the gap between the bi-level programming model and the single-level model by adding convex inequalities based on the UE model's (objective) function (Beckmann et al., 1956) to the constraints of the relaxed problem. Similar to the work of Farvaresh and Sepehri (2013),

the efficiency of the proposed methodology hinges on the assumption that the UE and SO solutions are close to each other.

Fontaine and Minner (2014) developed a scheme by which the upper-level objective function as well as the objective function of the lower-level problem are approximated by piecewise linear functions as well as the Beckmann function of the UE traffic flow. The bi-level problem is then transformed into a single-level problem by representing the UE conditions through its corresponding KKT conditions. Benders decomposition is then employed to solve the resulting problem. Similarly, the quality of the solution as well as the efficiency of the methodology is highly dependent upon the linearization scheme. Furthermore, a linearization of the Beckmann function compromises arriving at a global optimum solution. Numerical results for a medium sized example with 36 zones are reported.

2.1.2 A summary of the review of the road network design

As can be seen even among the most recent studies, addressing large sized networks or considering multiclass and multimodal traffic flow is rarely reported in the literature. Furthermore, a clear majority of the algorithms utilize simple approximations of the upper and/or lower-level objective functions. The conventional wisdom in the literature is to primarily move away the intractable elements of the problem with a view to arriving at a more simplified and well-behaved problem (Poorzahedy and Turnquist, 1982). Such approaches may deprive the problem from the dimensions based on which the DNDP stands, to the extent the solutions become overly simplified and unreliable. To this end, in order to enhance the realism and fidelity of the model, this study aims to establish a solid exact foundation embracing the DNDP in its full capacity.

2.2 Literature on the transit priority lane design

Given factors such as cost efficiency, environmental concerns, equity and public support, promoting transit even at the cost of a private mode is gaining momentum (Ceder, 2015). It is intuitively conceivable to assert that the priority of mobility has to be given to people rather than vehicles. Hence, providing priority to transit modes in terms of road space (Mesbah et al., 2011b) and signal timing is of interest to practitioners as well as academia (Bagherian et al., 2015; Diab and El-Geneidy, 2013; Guler and Menendez, 2015; Mirchandani et al., 2010).

The concept of priority lanes has been introduced in many cities (Smith and Hensher, 1998). It also has been adopted in special traffic access plans for large scale crowd gathering events such as sports games, concerts, New Year's events, etc. (Cova and Johnson, 2003). A number of studies have investigated the impact of priority initiatives extensively, while little is devoted to the strategic design of such schemes (Bagloee et al., 2016b). For evaluation, a variety of methods based on statistical analysis and simulations have been employed (Eichler and Daganzo, 2006; Li and Ju, 2009; Liu et al., 2006; Sakamoto et al., 2007; Tse et al., 2014; Viegas and Lu, 2004).

As discussed before, for strategic design, the subject of transit priority, or in general, the subject of transit network design/planning can be viewed as a general network design problem, which is proven to be a NP-hard problem. Among the variety of methods available in the literature, some try to reach an optimum solution but cannot scale to handle large-sized networks, whilst others aim to address large-sized networks at the cost of compromising the quality of the solution.

In the following section a comprehensive review of the relevant literature on the TPLDP is provided.

2.2.1 Literature on the transit priority lane design problem

Due to computational complexity involved in the TPLDP, a majority of the previous studies have taken a localized approach for a specific road or region. In contrast, in a few studies, a network based approach has been developed by Mesbah et al., (Mesbah et al., 2008; Mesbah et al., 2011b). They developed algorithms based on the Benders decomposition and genetic algorithms (GA) and applied them to a grid-structured network consisting of 38 nodes and 49 links. Since then the literature is still leaning toward a more localized focus on the subject of transit priority. Basso et al. (2011) studied congestion management policies through numerical analysis of a local pilot area centred on a short-length road and drew conclusions in favour of dedicated bus lanes as the best ad hoc policy. Guler and Cassidy (2012) investigated traffic operations of exclusive transit lanes at traffic bottlenecks occurring at intersections. They suggested some operational measures for sharing space at the bottlenecks temporarily. Xie et al. (2012) studied the intermittent priority of bus lanes using simulation methods. Zheng and Geroliminis (2013) proposed a macroscopic method for solving the problem of road space allocation based on a fundamental diagram, applicable to the cases of an aggregate size area such as

districts. Geroliminis et al. (2014) studied the dynamism of bimodal traffic flow (public and private modes) using the notion of fundamental diagrams which can then be utilized for traffic management schemes including transit priority.

Yao et al. (2012) developed a bi-level programming method similar to that of Mesbah et al. (2011a) to address the optimisation of exclusive bus lanes at the network level. The GA has been used as a solution algorithm and the arrivals' headways of the buses were optimised for a network consisting of 13 nodes, 19 links. At the lower level of the formulation, a path-based traffic assignment using the method of successive averages was employed. Hadas and Ceder (2014) addressed the TPLDP at network level through a single level optimisation formulation while approximating the impact of congestion. Khoo et al. (2014) also adopted a bi-level programming approach, while a GA was employed as a solution algorithm. For the lower (traffic assignment) problem, simulation software was used at the cost of higher computational time. Some studies have also employed a GA to solve the transit priority lane problem (Chen, 2015; Sun et al., 2014). In the work of Sun et al. (2014) the concerns of the transit enterprise in operating the transit fleet efficiently as a business has also been taken into consideration. The algorithm was applied to a network consisting of 29 origin-destination (OD) pairs and 52 nodes. Wang et al. (2016) also modelled the design of exclusive bus lanes as a bi-level problem for which they proposed a heuristic method to find the priority lanes as well as the transit assignment for a network of 24 nodes; 76 links. Zhang et al. (2014) addressed the simultaneous design of road expansion and a transit system in which the formulation was able also to find appropriate transit lanes. The problem was transformed to a single level mixed integer problem. Given the number of binary variables involved in the formulation, application to large-sized networks was left for further investigation.

The adverse impacts of transit priority lanes on private modes has also been a subject of research (Fang et al., 2014; Wu et al., 2013; Wu et al., 2015; Wu et al., 2014; Yingfeng and NaiQi, 2010; YunFei et al., 2011). In this respect, the key point is obviously, to take the interaction of both public and private modes into account which is a nontrivial task. In the above-mentioned works, the adverse impact of the transit priority lanes on private modes is considered a priori (i.e. a given exogenous input). Hence, there exist significant space for improvement to consider the mutual impacts of public and private modes. Nevertheless, the formulations and models developed in the above-mentioned articles can be used in uncongested networks such as, finding reserved lanes for transport

of hazardous materials on the outskirts of cities where traffic congestion is not a source of concern (Zhou et al., 2014; Zhou et al., 2012). In a similar fashion, Fang et al. (2013) sought applications of priority lanes for freight transport in which the freight demand was considered exclusively.

Recently Yao et al. (2015) have taken the uncertainty in the travel times and the capacity of the roads as well as travellers' risk adverse behaviour into account. Their results underscore the importance of incorporating combinatorial optimisation of the exclusive bus lanes.

The concept of exclusive bus lanes has also been introduced at signalized intersections along with an exclusive bus phase (in the timing of the signals) to reduce the delay to buses. A recent review on this subject is provided by (Guler et al., 2016).

2.2.2 A summary of literature on the TPLDP

The literature underscores the importance of adopting a network wide approach to the problem of transit priority lane design. This has come with extensive computational expense as well as theoretical complexities. To this end, a large body of research has been made towards the application of heuristic methods. Similar to what has been seen for the DNDP, practical applications of the proposed methods has been rarely presented in the existing literature. Moreover, a vast majority of the past studies suffer from a proper model to consider the interaction between private and public transport.

2.3 Summary and conclusion

Network design has recently witnessed a renewed interest in seeking methods that have more analytical strength (Fontaine and Minner, 2014; Gao et al., 2005; Wang et al., 2013; Zhang et al., 2014). Such new trends can be partly motivated by the great interest in better deterministic and not stochastic solutions. In the one hand, a number of (meta) heuristic methods have been developed addressing the scalability of the problems, their applications are extensively criticised based on some of the properties of the solutions such as the stochasticity of the solutions and the degree of goodness of the solutions. Moreover, the random elements in some of the heuristic methods are conducive to instability of the final solutions (known as stochasticity of the solutions). Even though heuristic methods are designed to yield good solutions, it is not clear how good the solutions are.

In the other hand, the recent rise in the use of such methods which come at the cost of greater computational times owes much to the recent advances in computational technology and knowledge of optimisation in both hardware and software (Achterberg and Wunderling, 2013; Bixby, 2012; D'Ambrosio and Lodi, 2013). It is noteworthy to highlight the fact that problems such as the DNDP or the TPLDP are of strategic nature (due to their pervasive impact), so that, one can afford greater computation in the scale of hours or even days.

In summary, a salient shortcoming in the previous studies can be attributed to the lack of practical exact methods tailored to large-sized networks which is addressed in this research. Moreover, in the vast majority of the past studies, the interaction between the private and public transport is largely simplified or relaxed which casts some doubt on the fidelity of the outcomes.

3 MATHEMATICAL METHODS

In Chapter 1, the discrete network design problem as well as the transit priority lane design problem (TPLDP, DNDP) were described as mixed integer nonlinear bi-level programming problems. In these problems the decision variables are binary (1 or 0). In the upper level the performance of the system is improved by minimizing the total costs of the system. The lower level accounts for multiclass and multimodal traffic assignment to thoroughly consider the impact of the changes in the road network.

As noted before, the conventional approach to the bi-level problems is to transform it to a single-level problem which results in a mixed integer nonlinear problem (MINLP). The MINLP is then articulated as seeking a subset of the candidates subject to some budget constraints. In order to efficiently solve the MINLP three leading algorithms have been reviewed: generalized Benders decomposition (GBD) or simply Benders decomposition, outer approximation (OA) and branch and bound (BB). Mathematical principles associated with them are presented. Then a step-by-step numerical practice based on a pedagogical example is provided

3.1 Solving mixed integer bi-level programming

A bi-level programming problem itself even with only linear constraints and objective functions is a NP-hard problem (Colson et al., 2007). An additional facet of being mixed integer and nonlinearity makes the problem more difficult. The general consensus to solving a bi-level problem is first to unify the levels and reach at a single-level programming problem (Colson et al., 2007). To this end there are two general methods suggested in the literature:

- the objective functions in both levels are interlinked with support functions (Gao et al., 2005; Mesbah et al., 2011b),
- the binary structure of the problem is laid down over a tree structure of a “branch and bound”. Each node of the tree represents a subarea of the solution space, for which the lower level problem is solved.

In these methods the common denominator is the need to solve a MINLP. This section is dedicated to the solution methods for the MINLP problems by covering two general methods: enumeration and decomposition (See Figure 3.1). With enumeration

methods, all the combinations of the integer decision variables are implicitly evaluated and hence the global optimum solution is guaranteed. The most notable method of enumeration is the BB which uses a tree structure to processing all the combinations (note that the “enumeration” is a jargon in the optimisation to enumerate all possible combinations of a subject). In minimization problems, as the tree expands, at each node representing a subset of the solution space, a lower bound is calculated and branching at the respective node is frozen when the lower bounds are found to be greater than that of the best found solution. It is evident that as the size of the problem (number of decision binary variables) increases, the method becomes computationally prohibitive.

Alternatively, the decomposition methods aim to address the problem efficiently by splitting it to easy and hard parts. Many hard problems are in fact easy problems complicated by a relatively small number of difficult constraints (Fisher, 2004). Such observations are greatly exploited by decomposition methods such that the decision variables are split into two sets, easy and hard. The two prominent decomposition methods are the GBD and OA which are investigated in this section.

Underlying mathematical principles and the above mentioned algorithms for MINLP problems followed by a numerical example are provided in this chapter. Further details can be found in two text books (Floudas, 1995; Li and Sun, 2006).

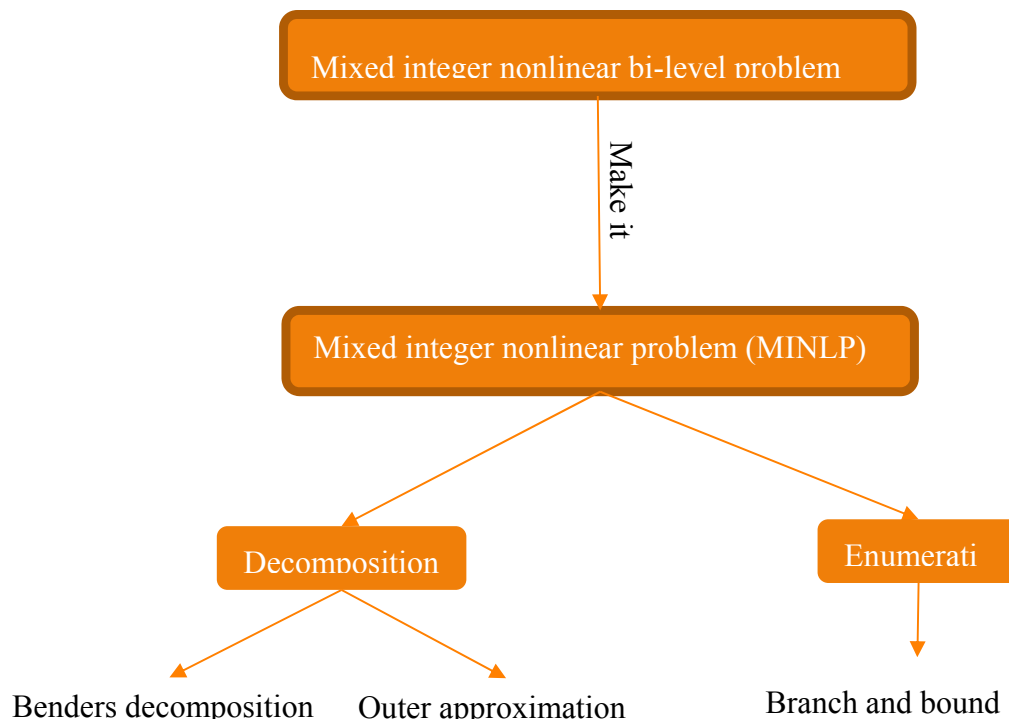


Figure 3.1, Outlook of solutions to the bi-level MINLP

3.2 Solution algorithms for the MINLP

The general formulation of MINLP problems is (Li and Sun, 2006):

$$(MINP) \quad \min f(x, y) \quad (3.1)$$

$$s.t. \quad g(x, y) \leq 0, \quad (3.2)$$

$$x \in X \subseteq R^n, y \in Y \subseteq Z^m \quad (3.3)$$

where X is a nonempty convex set in R^n (continuous variables) and Y is a finite integer set in Z^m (in case of binary variable $Y = \{1,0\}^m$; f, g are convex in the space of (x, y)).

3.2.1 Generalized Benders decomposition (GBD)

Consider MINLP (3.1) to (3.3), and let S, V to be solution space and feasible solution of the decision variable respectively as:

$$S = \{(x, y) \in X \times Y \mid g(x, y) \leq 0\} \quad (3.4)$$

$$V = \{y \in Y \mid \exists x \in X, g(x, y) \leq 0\} \quad (3.5)$$

The algorithm starts with a feasible solution for decision variables at iteration $i = 1$. Hence the MINLP becomes a nonlinear programming (NLP) problem as follows:

$$(NLP(y^i)) \quad v = \min f(x, y^i) \quad (3.6)$$

$$s.t. \quad g(x, y^i) \leq 0, \quad (3.7)$$

$$x \in X \subseteq R^n \quad (3.8)$$

The $NLP(y^i)$ is solved and renders continuous variables x^i . Given the newly found x^i , the algorithm proceeds to find a new set of decision variables for the next iteration. Let's relax the $NLP(y)$ problem from the constraints using Lagrangian multipliers $\lambda \geq 0$: $d_y(\lambda) = \min_{x \in X} L(x, y, \lambda) = f(x, y) + \lambda \cdot g(x, y)$. Then the Lagrangian dual problem of $NLP(y)$ becomes: $\max_{\lambda} d_y(\lambda)$. Furthermore, since $NLP(y)$ is a feasible solution to the respective MINLP, its optimal value $v(NLP(y))$ yields an upper bound to MINLP, therefore:

$$\min_{(x,y) \in S} f(x, y) = \min_{y \in V} v(NLP(y)) = \min_{y \in V} (\max_{\lambda} \min_x L(x, y, \lambda)) \quad (3.9)$$

$$= \min_z \quad s.t. \quad z \geq \min_x L(x, y, \lambda), y \in V \quad (3.10)$$

where z is a lower bound to the optimal value of the original problem. The only constraints in equation (3.10) are called Benders cuts to the solution space which are being accumulated as the algorithm proceeds in the iterations. Therefore, the above mixed integer problem known as Master Generalized Benders Decomposition, MGBD can be rewritten as follows:

$$MGBD^i \quad \min z \quad (3.11)$$

$$s.t \quad z \geq L(x^k, y^k, \lambda^k) + \nabla_y L(x^k, y^k, \lambda^k) \cdot (y - y^k) \quad k \in T^i \quad (3.12)$$

$$y \in Y \quad (3.13)$$

where T^i represents the set of solutions (x^k, λ^k) to $NLP(y^i)$ found up to the current iteration (i) . Now the GBD algorithm can be described as follows:

Step 0. Initialization: Set iteration $i=1$ and choose a feasible solution for decision variables $y^i \in Y$. Initialize lower bound and upper bound as $lb^0 = -\infty$, $ub^0 = +\infty$.

Step 1. Calculate the upper bound: Solve $NLP(y^i)$ to obtain x^i, λ^i . Update the value of best solution found so far by setting $ub^i = \min\{ub^{i-1}, f(x^i, y^i)\}$. Save it as best solution (x^*, y^*) if it is found to be the best solution so far.

Step 2. Calculate the lower bound: Given x^i, λ^i solve the master problem $MGBD^i$ to obtain optimal solutions of z^i, y^{i+1} .

Step 3. Termination: Set $lb^i = z^i$, if $lb^i \geq ub^i$ stop and (x^*, y^*) is the optimal solution, otherwise set $i:=i+1$ and go to Step 1. ■

Note that hereinafter, the GBD is sometimes referred to as Benders Method

3.2.2 Outer approximation (OA)

Similar to the GBD, the Outer Approximation (OA) alternates between solving a nonlinear programming sub-problem and a mixed integer linear programming master problem. The difference lies in how to derive the master problem. To this end, the OA exploits the gradient property of the problem both the objective function and the constraints. In fact the objective functions and the constraints are represented by their linear approximation as follows:

$$MOA^i \quad \min z \quad (3.14)$$

$$s.t. \quad z \geq f(x^k, y^k) + \nabla f(x^k, y^k) \cdot (x - x^k, y - y^k) \quad k \in T^i \quad (3.15)$$

$$0 \geq g(x^k, y^k) + \nabla g(x^k, y^k) \cdot (x - x^k, y - y^k) \quad k \in T^i \quad (3.16)$$

$$x \in X, y \in Y \quad (3.17)$$

where T^i represents set of solution (x^k, y^k) found up to current iteration i , in other words:

$$T^i = \{k | y^k \in V \text{ and } x^k \text{ solves } NLP(y^k), k = 1..i\} \quad (3.18)$$

Step 0. Initialization: Set iteration $i=1$ and choose a feasible solution for decision variables $y^i \in Y$. Initialize lower bound and upper bound as $lb^0 = -\infty$, $ub^0 = +\infty$.

Step 1. Calculate the Upper bound: Solve $NLP(y^i)$ to obtain x^i . Update the value of best solution found so far by setting $ub^i = \min\{ub^{i-1}, f(x^i, y^i)\}$. Save it as best solution (x^*, y^*) if it was found the best solution so far.

Step 2. Calculate the Lower bound: Given x^i solve the master problem MOA^i to obtain optimal solutions of z^i, x^{i+1}, y^{i+1} .

Step 3. Termination: Set $lb^i = \alpha^i$, if $lb^i \geq ub^i$ stop and (x^*, y^*) is the optimal solution, otherwise set $i:=i+1$ and go to Step 1. ■

3.2.3 Branch and bound (BB)

The basic idea of the BB is to partition the discrete solution space and discard non-promising parts. To this end, the BB algorithm for the MINLP is made on the continuous relaxation of the integrality of variable y over space of $\alpha \leq y \leq \beta$ where α, β are lower bound and upper bound of the respective discrete variable:

$$(CNLP) \quad \min f(x, y) \quad (3.19)$$

$$s.t. \quad g(x, y) \leq 0, \quad (3.20)$$

$$x \in X \subseteq R^n, y \in R^m, \alpha \leq y \leq \beta \quad (3.21)$$

The continuous solution space is split into subsets and each subset is represented by a node on the tree structure of the BB method. The algorithm starts with a feasible solution as the best solution found so far, known as the incumbent solution. As the algorithm proceeds, the value of the objective function of the best feasible solution found is an upper

bound to the original MINLP problem. At each node, the corresponding continuous nonlinear programming CNLP problem is solved in which the value of the objective function is a lower bound to the original MINLP problem (note that the CNLP is a relaxed version of the MINLP). If this lower bound is found to be greater than the value of the objective function of the incumbent solution, the respective subset is discarded. This process is called fathoming and it carries on until no subset is left. Subsequently a formal description of the BB can be written as follows:

Step 0. Initialize the upper bound as $ub^* = +\infty$, find a feasible solution for the MINLP as the incumbent solution to be represented by the root node of the tree. Set the root node as the current node identified by $c=1$.

Step 1. Solve the relaxed $CNLP^c$ problem, obtain continuous variables $(y_1^c, \dots, y_j^c)$ corresponding to integer variables and $v(CNLP^c)$ the value of the objective function.

Step 2. If $v(NLP_c) \geq ub^*$ or the current node represents a feasible solution (all y are integer) then fathom the current node. Consider the current node as incumbent solution if the current node is a feasible solution and it renders better solution than the incumbent solution (i.e. the value of its objective function is lower than of the incumbent solution). Update the upper bound $ub^* = \min(v(NLP_c), ub^*)$.

Step 3. If there is no unfathomed node left, stop, the incumbent solution is the optimal solutions. Otherwise, select an unfathomed node as the current node. Then choose a y whose value in the current node is not integer ($y \neq [y]$) and split the solution space in two domains one by adding $y \leq [y]$ and the other one by $y \geq [y] + 1$ in the constraints ($[y]$ returns the first integer value before y). Represent these two subareas by adding two branches at the end of the current node of the tree. Go to Step 1.

3.3 A numerical example for GBD, OA and BB

In this section using the above discussed algorithm a simple MINLP problem analogous to a network design problem is solved. Consider the network consisting of a single road or link (#4) connecting an origin-destination pair with travel demand of $q_{od} = 10$. The plan is to construct additional roads up to maximum three separate roads (#1, #2, #3). The available budget can afford maximum two roads ($c_1 = c_2 = c_3 = 1; B = 2$) (note that B

represents the limited budget and c_1, c_2, c_3 denote the construction costs of roads 1, 2 and 3 respectively. Figure 3.2 depicts the problem as well as the delay functions (cruise and waiting times combined) associated with the routes. Establish the objective function and the constraints. Find the best selection of the projects.

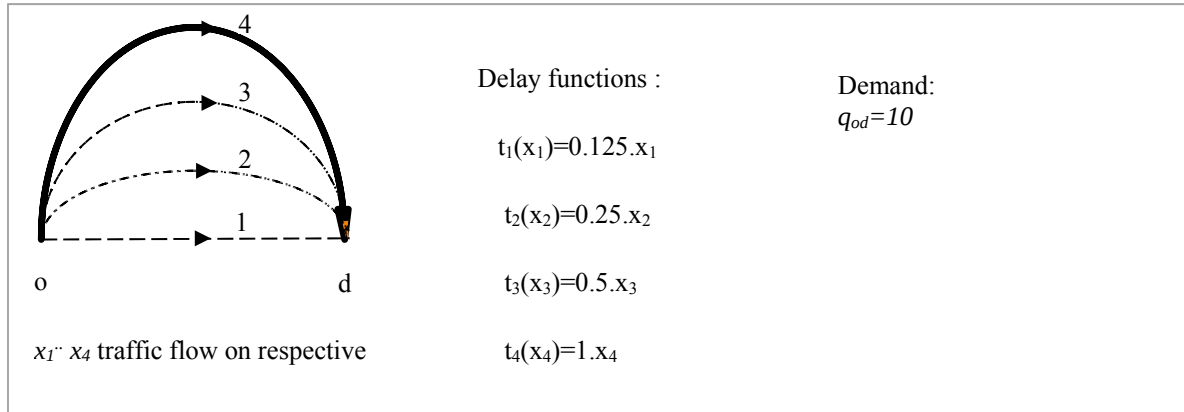


Figure 3.2, A numerical example

The objective function is defined as minimizing the total time spent on the system subject to budget and discrete constraints formulated as follows:

$$\min f(x, y) = .125x_1x_1 + .25x_2x_2 + .5x_3x_3 + x_4x_4 \quad (3.22)$$

$$\text{s.t. :} \quad x_1 + x_2 + x_3 + x_4 = 10 \quad (3.23)$$

$$x_1 - M.y_1 \leq 0 \quad \text{----- Lagrangian Coefficient} \quad \lambda_1 \quad (3.24)$$

$$x_2 - M.y_2 \leq 0 \quad \text{----- Lagrangian Coefficient} \quad \lambda_2 \quad (3.25)$$

$$x_3 - M.y_3 \leq 0 \quad \text{----- Lagrangian Coefficient} \quad \lambda_3 \quad (3.26)$$

$$y_1 + y_2 + y_3 \leq 2 \quad (3.27)$$

$$x_1 \dots x_4 \geq 0; \quad y_1 \dots y_3 \in \{0,1\} \quad (3.28)$$

where x_i represents the traffic flow on the respective roads, constraint (3.23) ensures that the traffic flow meets the travel demand. The y_i is the binary decision variable, it is 1 if the respective candidate project is decided to be constructed and 0 otherwise. Given M as a sufficiently large value, constraints (3.24), (3.25) and (3.26) ensure zero traffic volume for a candidate project if the respective project is decided not to be constructed ($y_i = 0$). Constraint (3.27) is the budget constraint.

Furthermore, given a feasible solution for the y_i , constraints (3.24), (3.25) and (3.26) can be written either as $x_i \leq 0$ or $x_i \leq 10$ which can be considered as capacity

constraints. In other words, the x_i traffic volume of road i must be less than or equal to the right hand side value: either 0 or 10. Hereafter, constraints (3.24), (3.25) and (3.26) are referred to as capacity constraints.

The optimal solution of the this example is $(y_1, y_2, y_3) = (1, 1, 0)$, $(x_1, x_2, x_3) = (6.1, 3.1, 0.0, 0.8)$ and $\arg f(x, y) = 7.7$. The computations provided in below were made using GAMS (2014) a leading optimisation software.

GBD:

The objective function of $NLP(y^i)$ can be defined by the Lagrangian multiplier method as follows:

$$L(x, y, \lambda) = \min x_1 (.125x_1 + \lambda_1) + x_2 (.25x_2 + \lambda_2) + x_3 (.5x_3 + \lambda_3) + x_4 .x_4 - 10(\lambda_1 y_1^i + \lambda_2 y_2^i + \lambda_3 y_3^i) \quad (3.29)$$

It is evident that $x_k \leq 10, k = 1..4$, as the result, for $y_k = 0$ the respective constraint is binding which means: $\lambda_k \geq 0$ otherwise $y_k = 1$, $\lambda_k = 0$ hence $\lambda_k . y_k = 0$ always holds. Therefore, the last term in the objective function is always zero. The value of $NLP(y^i)$ subject to $x_1 + x_2 + x_3 + x_4 = 10; x_1 \dots x_4 \geq 0$ becomes:

$$v(NLP(y^i)) = \min x_1 (.125x_1 + \lambda_1) + x_2 (.25x_2 + \lambda_2) + x_3 (.5x_3 + \lambda_3) + x_4 .x_4 \quad (3.30)$$

The Benders cuts at iteration i in the $MGBD^i$ are derived as:

$$z \geq v(NLP(y^k)) - 10\lambda_1^k . (y_1 - y_1^k) - 10\lambda_2^k . (y_2 - y_2^k) - 10\lambda_3^k . (y_3 - y_3^k) \text{ where } k = 1..i \quad (3.31)$$

Step 0. Set $lb^0 = -\infty$, $ub^0 = +\infty$, $i = 1$ choose feasible solution for binary variables

$$(y_1^0, y_2^0, y_3^0) = (0, 0, 0)$$

Step 1. Solve $NLP(y^1)$ and obtain optimal solutions: $(x_1^1, x_2^1, x_3^1, x_4^1) = (0, 0, 0, 10)$,

$$(\lambda_1^1, \lambda_2^1, \lambda_3^1) = (20, 20, 20) \text{ and update the upper bound: } ub^i = \min\{-\infty, v(NLP(y^i))\} = 100\}$$

Step 2. Solve the relaxed problem:

$$\min\{z; s.t. z \geq 100 - 10 y_1^1 20 - 10 y_2^1 20 - 10 y_3^1 20; y_1^1 + y_2^1 + y_3^1 \leq 2\}$$

Also obtain the lower bound $lb^1 = -300$, and optimal solution $(y_1^1, y_2^1, y_3^1) = (1, 0, 1)$.

Step 3. Evaluate the termination condition $lb^1 = -300 \neq ub^1 = 100$ hence $i = i + 1$ and go to Step

1. ■

The algorithm carries on and the optimum solution is found at the end of second iteration but the algorithm continues further to fill the gap between the upper and lower bounds till the forth iteration. The details of the calculations are shown in Table 3.1.

OA:

Given y^i a feasible solution for the binary variables, the $NLP(y^i)$ can be solved using Lagrangian multipliers similar to what discussed in the GBD. It is obvious that in MOA^i the linear approximation problem at iteration i , the linearized constraints pertaining to $g(x^k, y^k)$ would reproduce the constraints in the example. Linearization of the constraints pertaining to the objective function $f(x^k, y^k)$ leads to:

Table 3.1 Results of generalized Benders decomposition method

i	Y^{i-1}	X^i	λ^i	$v(NLP(y^i))$	ub^i	lb^i	Y^i
1	0,0,0	0.0,0.0,0.0,10.0	20.0,20.0,20.0	100.0	100.0	-300	1,0,1
2	1,0,1	7.3,0.0,1.8,0.9	0.0,1.8,0.0	9.1	9.1	-9.1	1,1,0
3	1,1,0	6.1,3.1,0.0,0.8	0.0,0.0, 1.5	7.7	7.7	-7.7	0,1,1
4	0,1,1	0.0,5.7,2.9,1.4	2.9,0.0,0.0	14.3	7.7	9.1	1,0,0
$v(NLP(y^1)) = \min \{x_1^1(.125x_1^1 + 20) + x_2^1(.25x_2^1 + 20) + x_3^1(0.5x_3^1 + 20) + x_4^1.x_4^1\}$ $v(NLP(y^2)) = \min \{.125x_1^2.x_1^2 + x_2^2(.25x_2^2 + 1.8) + .5x_3^2.x_3^2 + x_4^2.x_4^2\}$ $v(NLP(y^3)) = \min \{.125x_1^3.x_1^3 + .25x_2^3.x_2^3 + x_3^3(.5x_3^3 + 1.5) + x_4^3.x_4^3\}$ $v(NLP(y^4)) = \min \{x_1^4(.125x_1^4 + 2.9) + .25x_2^4.x_2^4 + .5x_3^4.x_3^4 + x_4^4.x_4^4\}$							
Benders cuts as constraints accumulated in successive iterations $i=1: z \geq 100 - 10y_1^1 20 - 10y_2^1 20 - 10y_3^1 20$ $i=2: z \geq 9.1 - 10y_2^2 1.8$ $i=3: z \geq 7.7 - 10y_3^3 1.5$ $i=4: z \geq 14.3 - 10y_1^4 2.9$							

min z

$$s.t. z \geq v(NLP(y^k)) + 2x_1^k .125(x_1 - x_1^k) + 2x_2^k .25(x_2 - x_2^k) + 2x_3^k .5(x_3 - x_3^k) + 2x_4^k (x_4 - x_4^k) | k = 1..i \quad (3.31)$$

(3.23)..(3.28)

Step 0. Set $lb^0 = -\infty$, $ub^0 = +\infty$, $i=1$ choose feasible solution for binary variables

$$(y_1^1, y_2^1, y_3^1) = (0, 0, 0)$$

Step 1. Solve $NLP(y^1)$ and obtain optimal solutions: $(x_1^1, x_2^1, x_3^1, x_4^1) = (0, 0, 0, 10)$, and update the upper bound: $ub^1 = \min\{-\infty, v(NLP(y^1)) = 100\}$.

Step 2. Given $x_j^1, j = 1..4, y_j^1, j = 1..3$ and $v(NLP(y^1))$ solve the linearized problem (3.31) and obtain the lower bound $lb^1 = z = -100$ and optimal solutions of the relaxed problem and obtain the lower bound $lb^1 = -300$, and optimal solution $(y_1^{1+1}, y_2^{1+1}, y_3^{1+1}) = (1, 0, 0)$ (and by-product of $(x_1^{1+1}, x_2^{1+1}, x_3^{1+1}) = (10, 0, 0)$ which will be superseded in step 1).

Step 3. Evaluate the termination condition $lb^1 = -100 \not\geq ub^1 = 100$, hence $i = i + 1$ and go to Step 1. ■

The algorithm carries on and the optimum binary solution is found at the end of the second iteration. Details of the calculations are shown in Table 3.2.

Table 3.2 Results of outer approximation method

i	Y^i	X^i	$v(NLP(y^i))$	ub^i	lb^i	Y^{i+1}	X^{i+1}
1	0,0,0	0.0,0.0,0.0,10.0	100.0	100.0	-100	1,0,0	10.0,0.0,0.0,0.0
2	1,0,0	8.9,0.0,0.0,1.1	11.1	11.1	-11.1	1,1,0	0.0,10.0,0.0,0.0
3	1,1,0	6.1,3.1,0.0,0.8	7.7	7.7	-7.7	0,0,1	0.0,0.0,10.0 0.0
4	0,0,1	0.0,0.0,6.7,3.3	33.3	7.7	7.7	0,1,0	0.0,10.0,0.0 0.0
Linear approximation of the last iteration MOA^i : $\min z$ s.t. $z \geq 20x_4 - 100$ $z \geq 2.2x_1 + 2.2x_4 - 11.1$ $z \geq 7.7x_1 + 1.5x_2 + 1.5x_4 - 7.7$ $z \geq 6.7x_3 + 6.7x_4 - 33.3$ (3.23)...(3.28)							

BB:

Step 0: set $ub^* = +\infty$ and the initial feasible solution as the incumbent solution:

$(y_1^0, y_2^0, y_3^0) = (0, 0, 0)$, $z^* = 100$ in the root node of the BB's tree. Set the current node

$c = 1$

Step 1: Solve the relaxed (continuous) $CNLP^1$ which renders optimal value of

$ub^1 = v(CNLP^1) = 6.7$ and solution $(y_1^1, y_2^1, y_3^1) = (0.5, 0.3, 0.1)$.

Step 2: since $v(CNLP^1) = 6.7 \neq ub^* = +\infty$ the current node cannot be fathomed. Update the best upper bound $ub^* = \min(6.7, +\infty)$. Since the current node does not represent a feasible solution hence it is considered as a unfathomed node.

Step 3: if there is no unfathomed node, consider the incumbent solution as the optimal and terminate. Select the current node which is the only unfathomed node (so far) for branching. Then select y_1 which has the maximum value for branching one with additional constraint $y_1 \geq 1$ and the other with $y_1 \leq 0$. This leads to two new nodes. Select the former as the current node and go to step 1.

The configuration of the tree structure as well as the detail of calculations are shown in Table 3.3

Table 3.3 Results of Branch and Bound method

i	Y^i	X^i	$v(NLP^i)$
0	0,0,0	0.0,0.0,0.0,10.0	100.0
1	0.5,0.3,0.1	5.3,2.7,1.3,0.7	6.7
2	1,0.3,0.1	5.3,2.7,1.3,0.7	6.7
3	1,1,0	6.1,3.1,0.0,0.8	7.7
4	1,1,0.2	7.3,0.0,1.2,0.9	9.1
5	0,0.6,0.3	0.0,5.2,2.9,1.4	14.3


```

graph TD
    1[1] -- "1 ≤ y1" --> 2[2]
    1 -- "y1 ≤ 0" --> 5[5]
    2 -- "1 ≤ y2" --> 3[3]
    2 -- "y2 ≤ 0" --> 4[4]
    style 3 stroke-dasharray: 5 5
    subgraph Optimal
    3
    end
  
```

3.4 Conclusion

In terms of solution algorithm for the mixed integer nonlinear bi-level programming:

- The BB algorithm has a simple structure but it becomes RAM intensive for large scale networks. Furthermore, the key to obtaining an efficient BB algorithm is the

size of discarded solution spaces due to comparison between lower bounds and incumbent value. There is no guarantee for such cuts. In case of not many cuts, the BB would have no superiority over an exhaustive enumeration. In fact, the BB adds additional computation burden for computing the lower bounds.

- The GBD is an effective method for a variety of mixed integer problems. The main issue dwells on the Lagrangian coefficients of the binary constraints. For example, for the network design problem, the traffic assignment as a nonlinear programming problem is widely solved using the famous Frank-Wolfe algorithm (Patriksson, 1994) which does not render the Lagrangian coefficients. Furthermore, the traffic assignment problem (TAP) has to be treated as a capacitated TAP. To this end the Lagrangian coefficients of the capacity constraints are treated as side constraints to the NLP for which some customised solution algorithms such as augmented Lagrangian methods (ALM) or Inner Penalty Methods (IPM) are employed (Larsson and Patriksson, 1995; Nie et al., 2004). The major stumbling blocks of the application of these methods are as follows: (i) the computational expense is significant, some studies have shown it to be four times that of a non-capacitated TAP (Bagloee and Sarvi, 2015a; Larsson and Patriksson, 1995), (ii) in some of the above-mentioned methods, there exists a number of parameters for which the calibration is a non-trivial task. (iii), In addition, arriving at an initial feasible solution at the outset of the algorithms' computation is also a non-trivial challenge. It is worth noting that, given the Lagrangian values of the capacity constraints, the respective mixed integer sub-problem can be solved efficiently which is the main selling point.
- The OA compared to the GBD doesn't require any element of the Lagrangian values of the side constraint. Instead, the respective mixed integer sub-problem comes with a high number of constraints and variables to linearize the traffic assignment. As the result the computational time becomes a prohibitive factor.

In summary, the BB has a simple structure to be embedded in any application, nevertheless the efficacy and computational burden are the main factors to take into account. The GBD is highly efficient but comes at the cost of Lagrangian values, for which arriving at an efficient method is a worthy effort. To this end, in the next chapter a method dubbed inflated travel time to solve a capacitated TAP as well as the Lagrangian

values of the constraint is developed and described. The OA does a very good job as long as the size of the problem is not significant.

4 CAPACITATED TRAFFIC ASSIGNMENT PROBLEM

The Benders decomposition method calls on the Lagrangian values of the capacity constraints in the traffic assignment. In addition, the capacity constraints can represent many realistic features but they are largely ignored in practice due to mathematical complexities in the application of the methods proposed in the literature. In this chapter, such complexities are relaxed by adopting an intuitive interpretation for the Lagrange values of the capacity constraints. Given an over-saturated road (traffic volume greater than capacity), its travel time is artificially increased at a gradual pace such that the excessive traffic volume decreases to zero. This artificially travel time is interpreted as the Lagrange values of the capacity constraint. In other words, the Lagrangian value is in fact a penalty added to the travel time of the over-saturated links to discharge the excessive flow. This penalty term bears some similarity to the marginal cost of the system optimal. Hence the capacitated traffic assignment problem (TAP) becomes a normal uncapacitated TAP in which the aforementioned additional penalty is updated iteratively. The proposed method is flexible enough to be embedded in solution algorithms for the TAP such as Frank-Wolfe.

4.1 Introduction

In the traffic analysis, the TAP is referred to as the problem of calculating traffic flow on a network for a given origin-destination travel demand. The widely recognised model for traffic flow is based on Wardrop principals ensuring commuters seeking shortest (least cost) paths that leads to user-equilibrium traffic flow (Boyce, 2013, 2014; Marcotte and Patriksson, 2007). Aggregation of travel times of the roads constituting a path is considered as the travel time of the respective path. The cost is considered as a collection of dis-utilities faced in making a trip such as travel time, fuel costs and parking fees - commonly referred to as (general) travel cost/time. Travel times on a road are considered as a non-decreasing functions of traffic volume –namely delay function- which need to be calibrated based on a field data (notably traffic count/survey). In order to maintain the

TAP to be mathematically and computationally amenable, no capacity constraints are considered for the delay functions. As such one may find over-saturated links in the equilibrium solution of the TAP. In other words, the issue of queues building up in the over-saturated roads is overlooked. Consideration of queue in traffic flow is however addressed in dynamic traffic assignment (DTA) which is still an active field of research (Nie et al., 2004; Shahpar et al., 2008). Capacity constraints are also studied under a broader umbrella, referred to as side constraints in the literature. In addition to the true meaning of capacity which is the physical capacity of the road to process a certain amount of traffic, many realistic features that are left-out can also be brought into the problem, embodied as side constraints such as: (i) refinement of the traffic equilibrium (Ferrari, 1997; Larsson and Patriksson, 1999), (ii) environmental constraints (Chen et al., 2011a), (iii) replicating traffic counts (Bell et al., 1997), (iv) traffic control (Yang and Bell, 1997), (v) congestion pricing (Yang and Bell, 1997), (vi) queuing effects (Larsson et al., 2004), (vii) combined/integrated modelling (Ryu et al., 2014). Moreover, as noted before, the main motive is to solve the DNDP. In particular, in the Lagrangian based algorithms for DNDP such as Benders decomposition or Lagrangian relaxation, one needs to solve an capacitated traffic assignment (Bagloee and Ceder, 2011; Bagloee et al., 2013b; Gao et al., 2005; Mesbah et al., 2011a; Mesbah et al., 2011b). Therefore, given the widespread applications of conventional (static) traffic assignment models in the industry, improvement to the method to make it more realistic, is a worthwhile endeavour (FHWA, 2002; Larsson and Patriksson, 1999).

The efforts to address capacitated traffic assignment problems (CTAP) can be classified in two general groups: (i) Lagrangian multipliers and (ii) penalty functions. In addition to mathematical complexities, the main challenge of these methods is the number of parameters that need to be calibrated. As such, despite the interests in the industry and among practitioners, the commercial software has not yet met this demand.

Alternatively, in this study the mathematical complexity of the CTAP is overcome by adopting an intuitive interpretation of capacity: the Lagrangian multipliers are interpreted as additional delay incurred, up to the level at which the traffic volume does not exceed capacity. Hence the CTAP becomes a conventional un-capacitated TAP in which the aforementioned additional delay is updated iteratively. The proposed concept is flexible enough to be implemented in commercial transport planning packages. As such it can be coded as an open-source “macro” to be easily used in EMME 3 – a leading

transport planning software. The proposed algorithm is compared with previous studies using the Hearn benchmark example. Furthermore, the practical merits of the proposed algorithm are also examined using a real dataset from the city of Winnipeg, Canada.

Throughout this chapter, it is assumed that: (i) travel demand is fixed (ii) the users have full understanding of the travel time and (iii) neither demand nor travel time change over time. Hence it is a deterministic and static traffic assignment.

The relevant studies in the literature are reviewed in the next section. The concepts and underlining mathematical features of the methodology are elaborated on section 4.3. Numerical results are presented in Section 4.4 followed by a conclusion.

4.2 Literature review

The TAP is traditionally solved by the Frank-Wolfe (FW) method based on a proposition by which the problem can be decomposed to linearized subproblems, equivalent to single-commodity shortest path finding problems. Explicit injection of the side constraints will obliterate the aforementioned proposition and turns the problem to a “multi-commodity least cost flow” which is by far more difficult to solve (Marcotte and Patriksson, 2007). Various aspects of the complexities involved in the CTAP have been comprehensively discussed in the literature (Larsson and Patriksson, 1995, 1999; Patriksson, 1994). The first mathematical remedy for the CTAP was to consider an asymptotical road delay function at capacity (Daganzo, 1977a, b), that is, as the traffic volume gets close to the capacity, the travel time tends to infinity. Such methods come at unbearable costs of numerical disorder near capacity flows, unrealistic high travel time and strange rerouting of trips (Boyce et al., 1981; Chen et al., 2011a; Ferrari, 1997; Patriksson, 1994).

Another method is to maintain the TAP as an uncapacitated problem by moving the side constraints to the objective function using two general methods: Lagrangian multipliers (Hearn and Ribera, 1980; Larsson and Patriksson, 1992, 1995; Larsson et al., 2004; Nie et al., 2004) or penalty function (Hearn, 1980; Inouye, 1987; Morowati-Shalilvand and Mehri-Tekmeh, 2013; Nie et al., 2004; Prashker and Toledo, 2004; Ryu et al., 2014; Shahpar et al., 2008; Yang and Yagar, 1994, 1995). For the former, the objective function is augmented to accommodate the Lagrangian exterior penalty terms representing the capacity constraints which is called the augmented Lagrangian method (ALM). At every iteration, a new solution is found and the corresponding Lagrange multipliers are updated until a termination criterion is met. For the latter, which is also

referred to as inner penalty function (IPF), the iterative solution algorithms for the TAP such as FW can still be used, but the links exceeding their capacities are penalized in the objective function. Nie et al. (2004) concluded that achieving a feasible solution using the ALM is a challenge, instead the IPF methods always render feasible solutions but at a cost of longer computation time.

Larsson et al. (2004) proposed a column generation method for the CTAP with linear side constraints which in essence is a IPF method. The computational time is improved largely from the adoption of a dual stabilization scheme but the algorithm needs to start from a strictly feasible solution which itself is a challenging. The proposed algorithm has yet to be examined in real size networks.

Shahpar et al. (2008) developed a method based on a dynamic penalty function (DPF) for which the traffic assignment is solved with a complementarity method. The proposed algorithm is enable to solve two types of side constraints: link constraints and node constraints.

Further to the mathematical complexities involved, the main challenge with these methods is the number of parameters to be calibrated. For instance, the initial choice of Lagrangian values plays a crucial role in the overall performance of the ALM (Bertsekas, 1982; Nie et al., 2004). The same is the case for IPF in setting up penalty parameters to be used in objective functions. Therefore, one has to resort to a trial-error effort to come up with appropriate parameters which varies from scenario to scenario.

4.3 Mathematical features

In this section, the CTAP is first formulated and the Lagrangian values of the capacity constraints are discussed and interpreted.

4.3.1 Formulation for capacitated traffic assignment problem (CTAP)

Consider $G(N, A)$ a traffic network as a graph consisting of N , A sets of nodes and links respectively on which $O, D \subset N$ are sets of origins and destinations. The CTAP can be formulated as a non-linear programming problem as follows (throughout this chapter, all terms are non-negative unless otherwise stated):

$$[\text{CTAP}]: \min \quad z(x) = \sum_{a \in A} \int_0^{x_a} t_a(x) dx \quad (4.1)$$

$$\text{s.t.}: \sum_p f_p^{od} = q_{od} \quad \forall o \in O, d \in D \text{ --Lagrangian multipliers-- } w_{od} \quad (4.2)$$

$$f_p^{od} \geq 0 \quad \forall p \in P_{od}, o \in O, d \in D \quad (4.3)$$

$$x_a = \sum_o \sum_d \sum_p f_p^{od} \cdot \delta_{a,p}^{od} \quad \forall a \in A, \forall p \in P_{od}, o \in O, d \in D \quad (4.4)$$

$$x_a \leq C_a \quad \forall a \in A \text{ -----Lagrangian multipliers --- } \beta_a \quad (4.5)$$

where z : the objective function to be minimized; x_a : traffic flow on link a ; $t_a(x)$: the delay functions are assumed to be non-decreasing, convex and separable; q_{od} : travel demand from o to d ; f_p^{od} : the flow on path from o to d ; P_{od} : set of all possible paths from o to d ; $\delta_{a,p}^{od}$: the link-path incidence (1: if link a belongs to path p from o to d , and 0 otherwise); C_a : the capacity of link a .

It has been proven that the CTAP subject to linear constraints is a strictly convex problem which renders a unique global optimal solution of link flows (Hearn, 1980; Inouye, 1987; Larsson and Patriksson, 1995; Marcotte and Patriksson, 2007; Nie et al., 2004; Patriksson, 1994). Let us consider w_{od}, β_a as Lagrangian multipliers for travel demand and capacity constraints respectively, hence the Karush-Kuhn-Tucker (KKT) conditions are established as:

$$f_p^{od} (u_p^{od} + \sum_{a \in A} \delta_{a,p}^{od} \cdot \beta_a - w_{od}) = 0 \quad \forall p \in P_{od}, o \in O, d \in D \quad (4.6)$$

$$\beta_a (C_a - x_a) = 0 \quad (4.7)$$

$$u_p^{od} + \sum_{a \in A} \delta_{a,p}^{od} \cdot \beta_a - w_{od} \geq 0 \quad (4.8)$$

$$C_a - x_a \geq 0 \quad (4.9)$$

$$f_p^{od} \geq 0 \quad \forall p \in P_{od}, o \in O, d \in D \quad (4.10)$$

$$\beta_a \geq 0 \quad \forall a \in A \quad (4.11)$$

$$\sum_p f_p^{od} = q_{od} \quad \forall o \in O, d \in D \quad (4.12)$$

where $u_p^{od} = \sum_{a \in A} \delta_{a,p}^{od} \cdot t_a$ is total travel time of the respective path. Let $\hat{u}_p^{od}$ be the “inflated” travel time of the respective path as:

$$\hat{u}_p^{od} = \sum_{a \in A} \delta_{a,p}^{od} \cdot (t_a + \beta_a) \quad (4.13)$$

Introduction of (4.13) into (4.6) and (4.8) results in:

$$f_p^{od} (\hat{u}_p^{od} - w_{od}) = 0 \quad \forall p \in P_{od}, o \in O, d \in D \quad (4.14)$$

$$\hat{u}_p^{od} - w_{od} \geq 0 \quad (4.15)$$

With respect to equations (4.14) and (4.15), it is proven that w_{od} is the travel time of the shortest path from o to d , hence, if path p takes some traffic volume ($f_p^{od} > 0$), it is certainly the shortest path ($\hat{u}_p^{od} - w_{od}$). In other words the first principle of Wardrop holds

and the global optimum solution of the CTAP is user-equilibrium traffic flow (Larsson and Patriksson, 1995; Nie et al., 2004; Patriksson, 1994). According to equations (4.7), (4.9) and (4.11), if the capacity constraint is binding which means saturation ($x_a = C_a$), then the corresponding langrage multipliers are non-zero ($\beta_a > 0$), otherwise it is zero. There are two terms contributing to the inflated travel time (equation (4.13)): normal or cruise travel time and beta. Beta has been interpreted as additional delay or waiting time caused by the queue built up on the over-saturated links (Larsson and Patriksson, 1995; Marcotte et al., 2004; Marcotte and Patriksson, 2007; Nie et al., 2004; Patriksson, 1994; Shahpar et al., 2008; Yang and Yagar, 1994, 1995).

Alternatively beta can be interpreted as a deterrence penalty imposed on the over-saturated roads to keep them at their respective capacities. This interpretation is utilized to develop a straightforward methodology to easily solve the CTAP.

4.3.2 Mathematical features

Let us consider the delay function $t_a = t_a^0 \cdot (1 + f(x_a))$ where t_a^0 is free flow travel time on link a and $f(x_a) \geq 0$ is a non-decreasing and convex function of x_a such that $f(x_a) = 0 \mid x_a = 0$. The aforementioned function can accommodate a variety of known delay functions including the widely used function proposed by US Bureau of Public Roads (BPR delay function) (Spiess, 1990). As such, the message is intuitive and simple: as long as the road is empty or uncongested, the travel time is the free flow travel time ($x_a \approx 0 \Rightarrow t_a \approx t_a^0$); as the traffic builds up, the travel time increases, certainly higher than the free flow travel time, one can consider it to be a factor of t_a^0 , where the factor is greater than 1 (i.e.: $(1 + f(x_a)) \geq 1$). According to equation (4.13), $\hat{t}_a$ the inflated travel time for each “saturated” link a can be formulated as:

$$\hat{t}_a = (t_a^0 + b_a) \cdot (1 + f(x_a)) \quad (4.16)$$

$$\beta_a = \hat{t}_a - t_a = b_a \cdot (1 + f(x_a)) \quad (4.17)$$

where b_a is an additional penalty in the free flow time in equation (4.16). In other words, it is the value of beta at $x_a = 0$, hence referred to it as “initial beta”. The travel time in the CTAP is replaced by the inflated travel time ($t_a \leftarrow \hat{t}_a$) and the capacity constraint is also dropped since the beta-Lagrangian multipliers of the corresponding capacity constraint does now contribute in the travel time. Therefore, the CTAP is transformed to a (uncapacitated) TAP. If the global optimum value of beta (in the CTAP), is already known

one just needs to simply solve the TAP using any known algorithm such as FW or any commercial transport planning software. Of course it is not the case, hence, values of the betas are updated iteratively in the course of solving the TAP as explained in the following section. Moreover, the value of beta is zero unless the corresponding road is saturated. In other words equations (4.16) and (4.17) apply only to the saturated links. In the case of saturation flow, a non-zero value is assigned to beta and it is set to be updated in the next iteration.

4.3.3 Heuristic method to update the (initial) beta

The concept embedded in equations (4.16) and (4.17) is to uplift the delay function of the saturated links until the traffic volume stabilizes at capacity. Hence the value of the initial beta in the delay functions is iteratively updated for which the main challenge is to progressively stride towards convergence and user equilibrium. The amount of additional delay to be added to obtain the inflated travel time is derived from the concept of marginal cost proposed as follows (Beckmann et al., 1956; Patriksson, 1994; Sheffi, 1985) :

$$\tilde{t}_a(x_a) = t_a(x_a) + x_a \cdot \partial t_a(x_a) / \partial x_a \quad (4.18)$$

where $\tilde{t}_a(x_a)$ is the marginal cost or travel time experienced by an additional commuter added to already, x_a commuters on link a , and $\partial t_a(x_a) / \partial x_a$ is the additional travel time experienced by each driver among x_a . The marginal cost enforces system optimal flow, with better and more uniformly distributed traffic across the network such that underutilized roads will take additional traffic off the saturated links. This notion is exploited to enforce a capacitated traffic assignment. Hence $x_a \cdot \partial t_a(x_a) / \partial x_a$ the additional delay imposed on the respective link is considered as a template to update the initial betas as follows:

$$t_{a,C}^{(i)} = (t_a^0 + b_a^{(i)}) \cdot (1 + f(C_a)) \quad (4.19)$$

$$\nabla b_a^{(i)} = (x_a^{(i)} - C_a) \cdot \frac{t_a^{(i)} - t_{a,C}^{(i)}}{C_a} \quad (4.20)$$

$$b_a^{(i+1)} = b_a^{(i)} + \nabla b_a^{(i)} \quad (4.21)$$

where superscripts i and a denote the current iteration and respective saturated link; $t_{a,C}^{(i)}$ is the travel time at capacity ($x_a = C_a$) over inflated delay function; $b_a^{(i)}$ is initial beta or an additional penalty to free flow travel time ($b_a^{(i)}$); $\nabla b_a^{(i)}$ is the pace of the initial beta computed at current iteration while $b_a^{(i+1)}$ is the updated initial beta computed for the next

iteration, initialized to zero ($b_a^{(1)} = 0$). It is evident that equation (4.20) follows the aforementioned template where the pace value is proportional to excessive traffic volume as well as the difference of current travel time and travel time at capacity normalised by the capacity. In Figure 4.1, exhibition-1, the above formulations of equations (4.19) to (4.21) for three iterations on the (inflated) delay function are shown graphically.

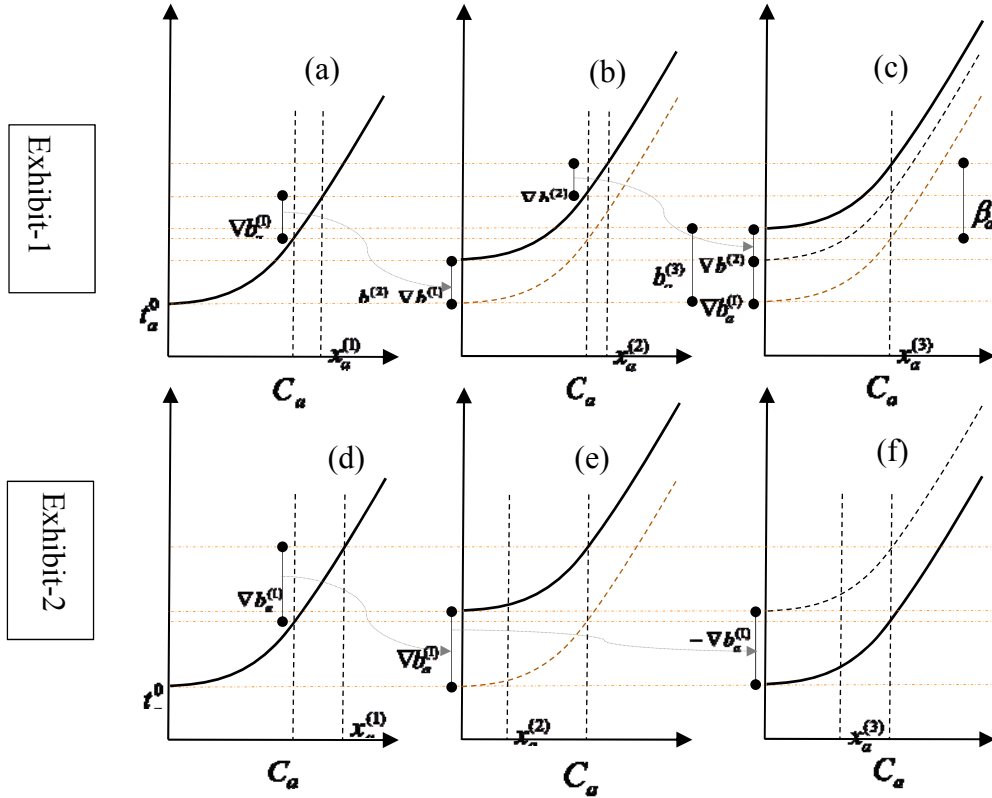


Figure 4.1 Conceptual representation of the proposed methodology on the road delay functions

In the first iteration when there is no initial-beta ($b_a^{(1)} = 0$), the volume stands at $x_a^{(1)} > C_a$, and $\nabla b_a^{(1)} > 0$ the pace is computed as shown graphically in Figure 4.1 (a) which shifts up the delay function for the next iteration ($b_a^{(2)} = 0 + \nabla b_a^{(1)}$). In the second iteration the volume still stands above capacity ($x_a^{(2)} > C_a$) (see Figure 4.1 (b)), hence value of the pace ($\nabla b_a^{(2)} > 0$) is corrected followed by updating the initial-beta to be carried over to the next iteration to uplift the delay function ($b_a^{(3)} = b_a^{(2)} + \nabla b_a^{(2)}$). The third iteration is executed and the volume stands at capacity ($x_a^{(3)} = C_a$) (see Figure 4.1 (c)). Three key words or components of the proposed algorithm are shown in the figure which is beta (β_a), initial-beta (b_a) and the pace (∇b_a). During this progressive approach, in an intermediate iteration,

if a saturated link is found unsaturated (see Figure 4.1 (d) versus Figure 4.1 (e)), its corresponding penalty is nullified ($b_a^{(i+1)} = 0 \mid x_a^{(i)} < C_a$) (see Figure 4.1. (f)). This process is shown in Figure 4.1, exhibit-2.

The above measures to compute the initial-beta at each iteration can be summarized by two rules:

$$b_a^{(i+1)} = \begin{cases} b_a^{(i)} + \nabla b_a^{(i)} & \text{from (20)} & x_a^{(i)} \geq C_a \\ 0 & & x_a^{(i)} < C_a \end{cases} \quad (4.22)$$

4.3.4 Termination conditions

The proposed algorithm is developed based on a solution algorithm for TAP such as FW. Hence, given fixed rates of the initial-betas, the solution algorithm itself needs to converge and meet its own termination criteria which is mainly driven by a relative gap. Boyce et al. (2004) recommended a relative gap of 0.01% to ensure convergence to link flow stability.

It is also expected that the initial-betas show convergence behaviour over the successive iterations. Given the gradual built up of the initial-beta, one convergence criteria can be considered as reaching at enough small values for the pace in descending manner which is defined as follows:

$$\max_a |\nabla b_a^{(i)} / \hat{t}_a| \leq \varepsilon \quad (4.23)$$

where ε is a small value called the relative pace value. Numerical results show that 1% as relative pace value is sufficient to obtain reliable results. Consequently, the algorithm does not terminate unless both criteria, the relative gap and the relative pace values have been met.

4.3.5 Capacity feasibility

In case where travel demand is higher than the capacity of the network, the problem becomes infeasible. It is important for any algorithm to have some mechanism to detect and address these infeasibility cases. To this end one can introduce a dummy node connected with all zones via uncapacitated links associated with fixed and high travel times. Therefore, problems always become feasible with solutions at the end and those left with residual traffic on the dummy link are detected as capacity-infeasible cases.

Similar to the terminology used in the literature, for ease of reference, the proposed methodology can be referred to as the “inflated travel time” (ITT) method.

Dummy links not only obviate any feasibility concerns, they are also used to replicate reality. Therefore, consider a single origin-destination (OD) connected via a single road with capacity of 10 vehicles. Faced with 15 vehicles demand, the algorithm allows 10 vehicles to enter on the road while 5 vehicles have no chance to get on the road and instead they use the dummy links as leftover flow. In reality these excess 5 vehicles have no choice except changing their departure times which is also studied separately in the literature. In other words, the residual demand remains off the network until the next available traffic assignment interval, which is discussed under dynamic traffic assignment literature (Zhong et al., 2011).

4.4 Numerical results

This methodology has been coded into a “macro”, that is the programming language of EMME 3 (INRO, 2009). A desktop PC with a 3.70GHz CPU and 64 GB of RAM” is employed. In order to provide a comparative analysis with literature, the benchmark network of Hearn is used for numerical evaluation. The ITT algorithm has been further applied to large-scale benchmark network, the city of Winnipeg, Canada.

4.4.1 Hearn’s benchmark case study

The Hearn benchmark problem consists of 4 ODs, 9 nodes and 18 links associated with BPR delay functions $(t_a^0(1 + 0.15.(x_a/C_a)^4))$ which has been previously employed to comparatively analysing the IPF (Nie et al., 2004) and the ALM (Larsson and Patriksson, 1995; Nie et al., 2004) as well as the DPF (Nie et al., 2004; Shahpar et al., 2008). The ITT algorithm is applied to Hearn’s benchmark problem and the results are shown in Table 4.1. The global optimal solution of the problem has also been reported in the literature (Shahpar et al., 2008) and is presented in Table 4.1. Since these methods have been coded in different software and implemented with different computers, the reported computational time cannot be used for comparisons. Alternatively number of attempts to solve the shortest paths which consumes a significant CPU time (Sheffi, 1985) is considered as a relatively fair comparison basis (Shahpar et al., 2008). Moreover, compared to attempts required by the FW algorithm which is also used in the ITT, the other methods endure extra effort. For instance in the DPF which is a path based method each iteration comprises some inner iterations. Hence, each iteration in the ALM, the IPF and the DPF is equivalent to 2 or 3 (or a factor bigger than 1) iterations in the FW or the

ITT. Accordingly, for the purpose of fair comparisons, the ITT algorithm was run using 156 FW iterations (identical to that of the ALM).

Table 4.1 Hearn network: comparisons results.

				Optimal (Shahpar et al., 2008)	IPF(Nie et al., 2004)		ALM(Nie al., 2004)		et DPF(Shahpar et al., 2008)		ITT (this study)								
n	i	j	t^0	C	x	β	x	β	X	β	x	β	x	β	t	$\hat{t}$	∇b	%	b
1	1	5	5	12.02	12.02	0.13	12.02	0.14	12.02	0.22	12.02	0.16	12.02	0.55	5.75	6.30	0.00	0.41	
2	1	6	6	18.02	17.98	0.00	17.98	0.00	17.98	0.00	17.98	0.02	17.98	0.00	6.89	6.89	0.00	0.00	
3	2	5	3	43.59	43.59	5.85	43.59	5.92	43.59	5.94	43.59	5.9	43.70	7.07	3.45	10.52	0.03	5.32	
4	2	6	9	26.59	26.41	0.00	26.42	0.00	26.42	0.00	26.41	0.01	26.30	0.00	10.29	10.2_9	0.00	0.00	
5	5	6	1	50.00	0.00	0.00	0.14	0.00	1.21	0.00	0.31	0.00	0.36	0.00	1.00	1.00	0.00	0.00	
6	5	7	5	25.00	20.61	0.00	20.46	0.00	19.4	0.00	20.30	0.00	20.27	0.00	5.32	5.32	0.00	0.00	
7	5	9	2	35.00	35.00	0.78	35.00	0.77	35.00	0.68	35.00	0.74	35.09	0.45	2.30	2.75	0.10	0.33	
8	6	5	1	50.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	
9	6	8	5	25.00	20.41	0.00	20.46	0.00	20.66	0.00	20.30	0.00	20.51	0.00	5.34	5.34	0.00	0.00	
10	6	9	2	35.00	23.99	0.00	24.07	0.00	24.95	0.00	24.40	0.00	24.12	0.00	2.07	2.07	0.00	0.00	
11	7	3	3	25.00	25.00	5.58	25.00	5.76	25.00	5.59	25.00	5.55	25.04	6.39	3.45	9.84	0.06	4.81	
12	7	4	6	24.00	24.00	0.54	24.00	0.54	24.00	0.54	24.00	0.54	24.04	0.57	6.91	7.48	0.00	0.43	
13	7	8	1	50.00	5.60	0.00	5.57	0.00	5.34	0.00	5.70	0.00	5.46	0.00	1.00	1.00	0.00	0.00	
14	8	3	8	39.00	15.00	0.00	15.00	0.00	15.00	0.00	15.00	0.00	14.96	0.00	8.03	8.03	0.00	0.00	
15	8	4	6	43.00	36.00	0.00	36.00	0.00	36.00	0.00	36.00	0.00	35.96	0.00	6.44	6.44	0.00	0.00	
16	8	7	1	50.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00	1.00	1.00	0.00	0.00	
17	9	7	2	35.00	33.99	0.00	34.07	0.00	34.95	0.00	34.4	0.00	34.18	0.17	2.27	2.44	-0.13	0.13	
18	9	8	2	25.00	25.00	0.97	25.00	0.99	25.00	0.99	25.00	0.98	25.04	1.30	2.30	3.60	0.01	0.98	
No of iterations					N.A.		156		84		26		156						
obj function					1572.27		1572.31		1572.36		1572.29		1571.45						
obj fn error%					0.000		0.002		0.006		0.001		0.053						
total flow error					0.00		0.54		4.86		1.65		1.86						
max (x/C)%					100.0000		N.A.		N.A.		99.9992		100.5589						

Note: ALM: Augmented Lagrangian Method; IPF: Inner Penalty Function; DPF: Dynamic Penalty Function; ITT: Inflated Travel Time; n: Link no; i: start node; j: end node; t^0 : free flow travel time; C: link capacity; x: traffic volume; β : beta or additional delay due to queue; t: cruise travel time = $t^0(1 + 15.(x/C)^4)$; $\hat{t}$: inflated travel time = $t + \beta$; ∇b %: pace value at the last iteration; b: initial beta or additional penalty added to free flow travel time; total flow error: sum of difference of the links volumes of the respective method versus the global optimal in absolute values

As per Table 4.1 all methods including the ITT renders similar traffic flows close to the global solution. In addition to the traffic flow and beta values, the cruise travel time, inflated travel time, pace values and the initial values are also reported in Table 4.1. The

maximum pace values were found to be 0.0013, equivalent to relative pace value of 0.05% = $0.0013/2.44$ which is far below the maximum accepted level of 1%.

Across all the results, there are consistently six saturated links (highlighted in the table with bold font). It is evident that compared to other methods the capacity constraints are not strictly held in the ITT. However, it is trivial and negligible in practice, such that the maximum volume per capacity stands only slightly higher than the capacity (1.005589). Such slight relaxation results in reaching a value for the objective function slightly lower than the optimal value (0.053%). In terms of total differences of the traffic volumes versus the optimal volumes, ITT is as good as other studies reported here. Figure 4.2 depicts the convergence behaviour of the proposed algorithm graphically.

In Figure 4.2(a) the Beckmann value consistently converges to the optimal value with some trivial fluctuations. The same behaviour is also seen for the total amount of violating capacity constraints which converges to zero. Methods such as the IPF try very hard to abide by the capacity constraints during the successive iterations at excessive computational costs (Nie et al., 2004). In other words these algorithms converge to the capacity level strictly from one side of the constraints (below capacity). But in the method developed here the capacity constraints are relaxed across all iterations in such a way, the volumes on the saturated links approach the capacities from both sides of the constraints. That is why the Beckmann value converges in an ascending manner. Figure 4.2(b) also demonstrates how the six saturated links converge to their respective capacities. It is important to note that the ITT gains much of its convergence in earlier iterations (say iteration 50) compared with other methods. In the light of being essentially a heuristic method such efficient convergence behaviour is interesting.

4.4.2 Large sized Winnipeg case study

Large scale transport data of the city of Winnipeg, Canada which is widely used as a benchmark network in the literature (Bar-Gera, 2016) was used for numerical tests (it was also provided by INRO (INRO, 2009) as part of EMME 3 application software). The case study comprises of 154 zones, 903 nodes, 2,995 directional links with an hourly travel demand of 56,219. The delay functions comply with the general format of BPR functions. First, given a relative gap of 0.0001, the traffic assignment without any capacity constraints (TAP) is solved which elapses 18 seconds and it terminates at iteration 561 with a Beckmann value of 798,531. Second, the ITT algorithm is carried out to solve the

capacitated traffic assignment problem (CTAP). Based on the relative pace of 1% the algorithm terminated at iteration number 2,164 within almost 3 minutes computation time, in which a Beckmann value of 1,186,670 was obtained. It is worth noting that that number of iterations required to solve the CTAP versus that of the TAP ($3.8=2164/561$) is close to 4 which was experienced by Larsson and Patriksson (Larsson and Patriksson, 1995) in applications of the ALM. The residual flow on the dummy links was found to be 144 (equivalent to $72 = 144/2$ out of 56,219 trips). The algorithm is run for further iterations up to 2,500 to investigate the convergence behaviour of the algorithm as shown in Figure 4.3.

Figure 4.3(a) shows that the flows initially taken by the dummy links are discharged until iteration around 1200, and it stabilizes around a solid level, hence the residual flow is the excess demand. As it is evident from Figures 4.3(b) and 4.3(c), the algorithm shows chaotic behaviour in most the first half of the iterations and it then stabilizes and converges. Similar chaotic behaviour in earlier iterations has been reported in literature (Shahpar et al., 2008). Such chaotic behaviour ought to be caused by excessive demand which leads to heavy flow on the dummy links in the earlier iterations. This hypothesis was examined in two scenarios by running the algorithm under all things being equal except the demand: one with half and the other with two-thirds of the travel demand. No excessive demand was found, the dummy links were spared from taking any traffic flow and chaotic behaviour never happened. Figure 4.3(c) illustrates the convergence of the Beckmann value. It is interesting to see that contrary to the results of the Hearn case study, the Beckmann value of the Winnipeg case study converges in an ascending fashion. One reason could be the fact that a significant number of the links in the Hearn case are saturated or capacity binding (6 out of 18 = 33%). As discussed before, our approach of relaxing the capacity constraints enables the algorithm to converge to the capacity from both sides of the constraints. But, for the Winnipeg case, only 162 links out of 2995 links (5%) were found to be capacity binding. Hence the capacity relaxation does not significant change the convexity of the convergence curve.

Furthermore, the first 4 top saturated links (the ones with highest volume-per-capacity ratio) in the last iteration are singled out and shown in Figure 4.4 (links are identified by “start node-end node”). As such the maximum volume-per-capacity ratio was found to be 1.0039. Figure 4.4(a) shows the way those links converge to their respective capacities. For instance, link 759-760 takes no flow until around iteration 400,

from that point onward, it starts to attract discharged flow from the saturated links. Figure 4.4(b) also shows variation of initial betas for the aforementioned links which are very high in the earlier iterations of the chaotic period and then stabilize from iteration 1500 onward.

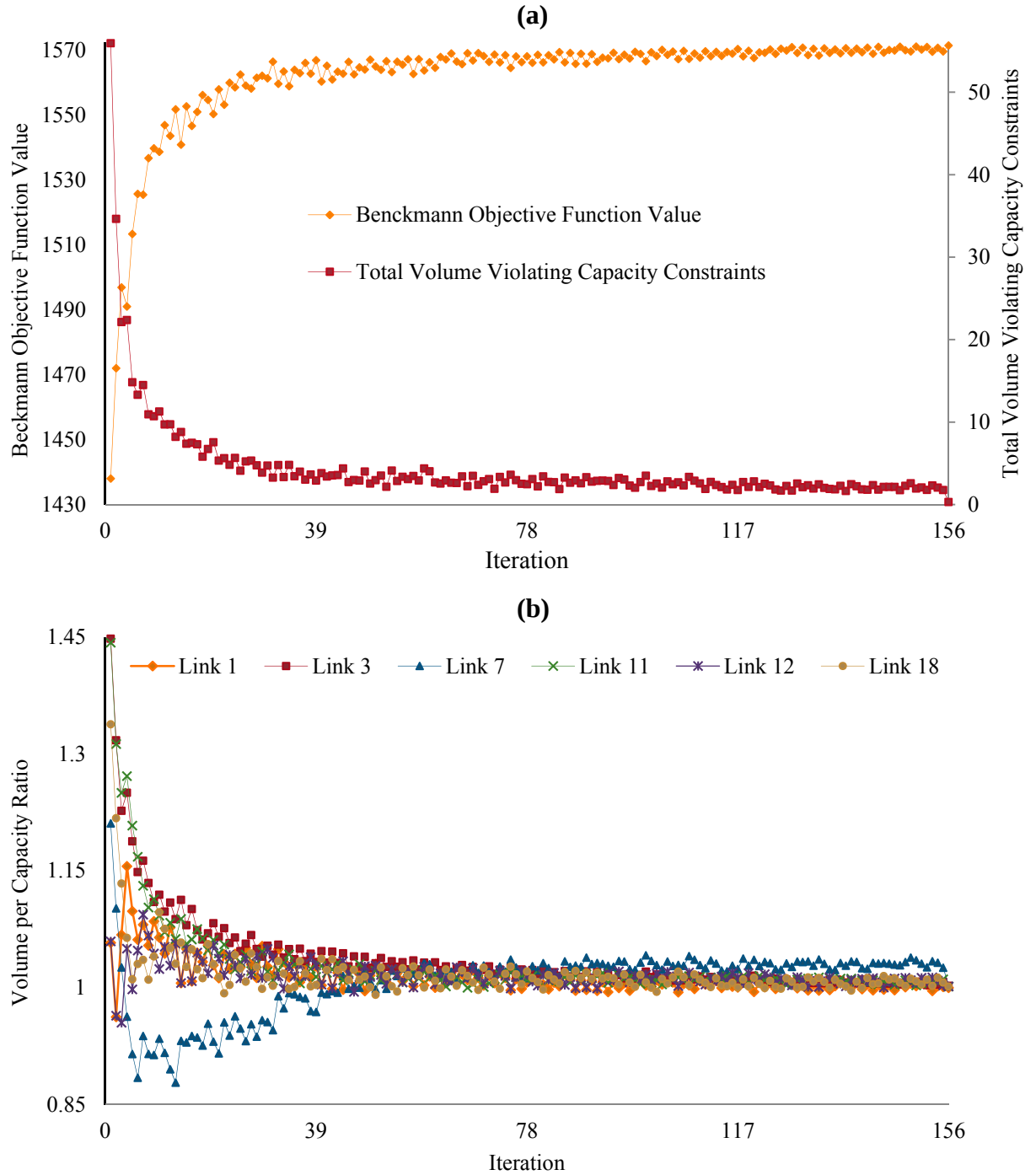


Figure 4.2 Hearn network: (a) convergence, (b) fluctuation of flows on the saturated links.

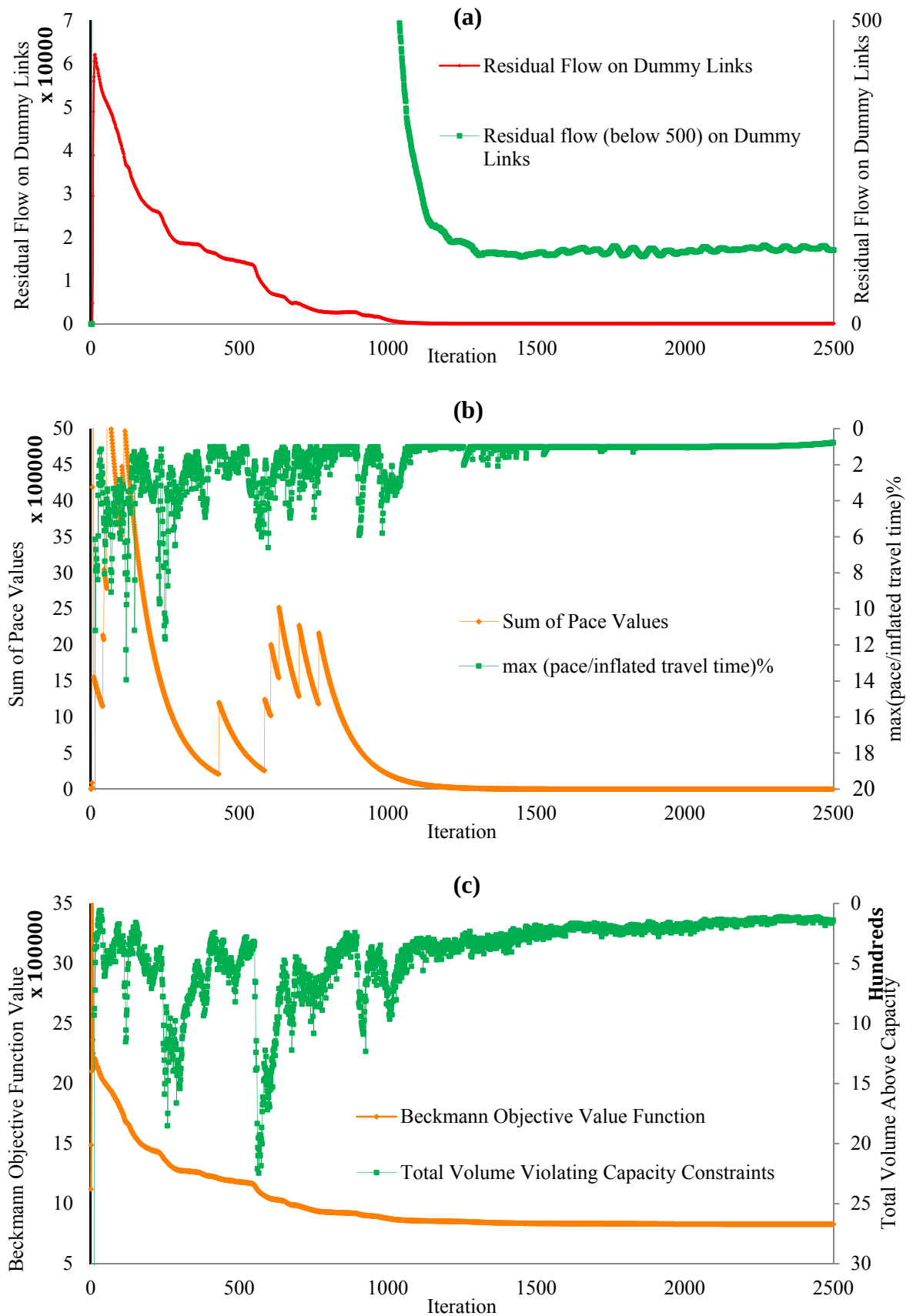


Figure 4.3 Winnipeg network results: (a) residual flows, (b) pace values, (c) convergence.

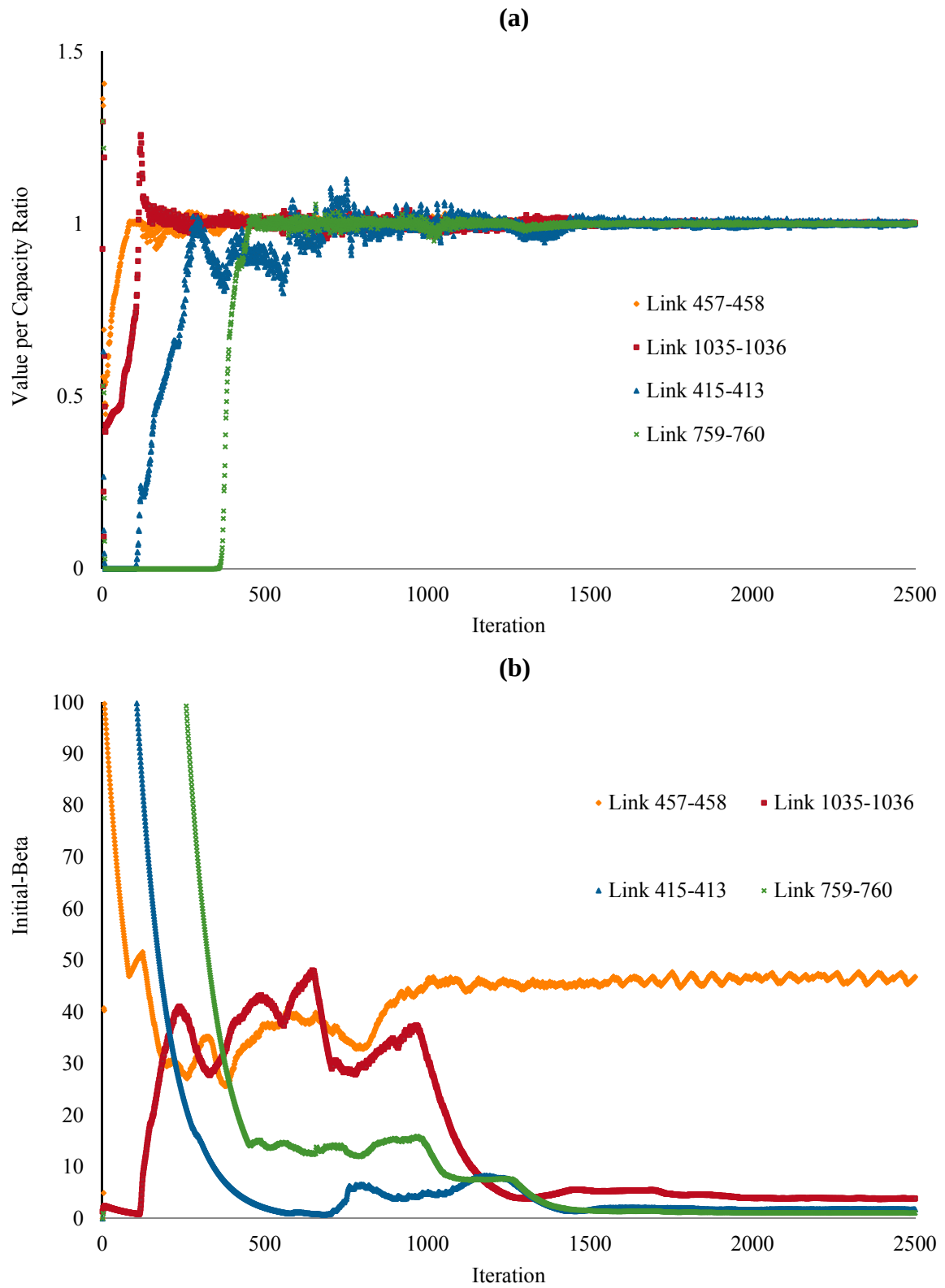


Figure 4.4 Winnipeg network results, 4 top saturated links: (a) fluctuation of flow, (b) Initial-Beta.

The impact of the excessive demand on traffic assignment was further studied. First it is needed to derive the corresponding demand matrix out of the excessive flow accumulated over the dummy links to be deducted from the total travel demand. This is a popular exercise among practitioners known as the “select link analysis” (Boyce and Xie, 2013). An EMME3 macro was prepared to conduct the “select link analysis” which resulted in a sub-matrix of total excessive demand of 72.6 vehicle trips at the end of 2,500th iteration. This sub-matrix was taken off the original matrix and the resultant matrix was assign on the network. Subsequently the results are graphically shown in Figure 4.5. The total traffic flow accumulated on dummy links over 2,500 iterations for both scenarios (with and without excessive demand) are illustrated in Figure 4.5 (a). It is clear that the excessive demand of 72.6 (versus total 56,219) causes a significant shift of volume to dummy links in early iterations. Nonetheless the excessive demand does not change the ultimate traffic flow on the real network. This is shown in Figure 4.5(b) in which the Beckmann values of the real network (excluding dummy links) for the two scenarios over progressive iterations are depicted. As such the Beckmann values are almost identical as the algorithm converges.

4.5 Conclusion

A heuristic approach was developed to address the capacitated traffic assignment problem. The approach was made based on a new interpretation of the Lagrange values of the capacity constraints, that is, amount of penalty added to the travel times of the over-saturated links to discharge the excessive flow. The penalties are specified up to the level at which the over-saturated links become saturated. This penalty term bears some similarity to the concept of the marginal cost of the system optimal flow. The additional penalty is added to the free flow time of the delay functions to be updated in the successive iterations of solving a normal (uncapacitated) traffic assignment problem. The benchmark network of Hearn is used for comparative analysis with other methods. The main motivation for this study was to address the needs of the industry hence the large scale network of the city of Winnipeg was also used in the numerical tests. The results found are promising. The advantages of the proposed method are: (i) The capacitated Traffic Assignment Problem (CTAP) is transformed to an uncapacitated TAP for which no new solution algorithm for traffic assignment is needed. (ii) The only requirement is to amend the way the travel times are required to be updated iteratively in every normal solution

algorithm for the TAP such as Frank-Wolfe (FW). To this end a set of rules were devised (see equations (4.23)). (iii) In contrast to other methods such as the Augmented Lagrangian Method (ALM) or the Inner Penalty Function (IPF), there are no parameters that need to be calibrated. Therefore, no setup preparation is needed. (iv) The proposed algorithm is intuitively conceivable and straightforward, and it can easily be implemented even in commercial transport planning software. Hence it has potential to appeal to the already accumulated interests in the industry and amongst practitioners. To this end, the proposed algorithm was encoded in EMME 3 a leading commercial software for transport planning. (v) In conventional solution algorithms for TAP such as FW, a variety of link-specific parameters and variables pertaining to delay function as well as data of gradient descent are saved in the RAM which can amount to a dozen of attributes per link. The proposed algorithm only adds two attributes (current and previous penalties) hence it is not RAM intensive. Considering everything being equal, the additional two variables in the delay functions for the Winnipeg case study resulted in 3% increase in the computational time. (vi) In terms of the properties of the delay function, the algorithm does not require anything more than what is required by the solution algorithm for the TAP. (vii) The problem would always have a solution. In case of over-saturated network where the total demand is higher than the supply (capacity), the excessive demand is collected on dummy links. That can also assist planners to identify where the supply shortage occurs. The main shortcoming of the method is the fact that it is a heuristic method. Despite of showing promising results, the convergence of the algorithm has yet to be mathematically proven. Other extensions to the traffic assignment problem such as multi-class, multi-modal and consideration of non-separable delay functions are left for further studies.

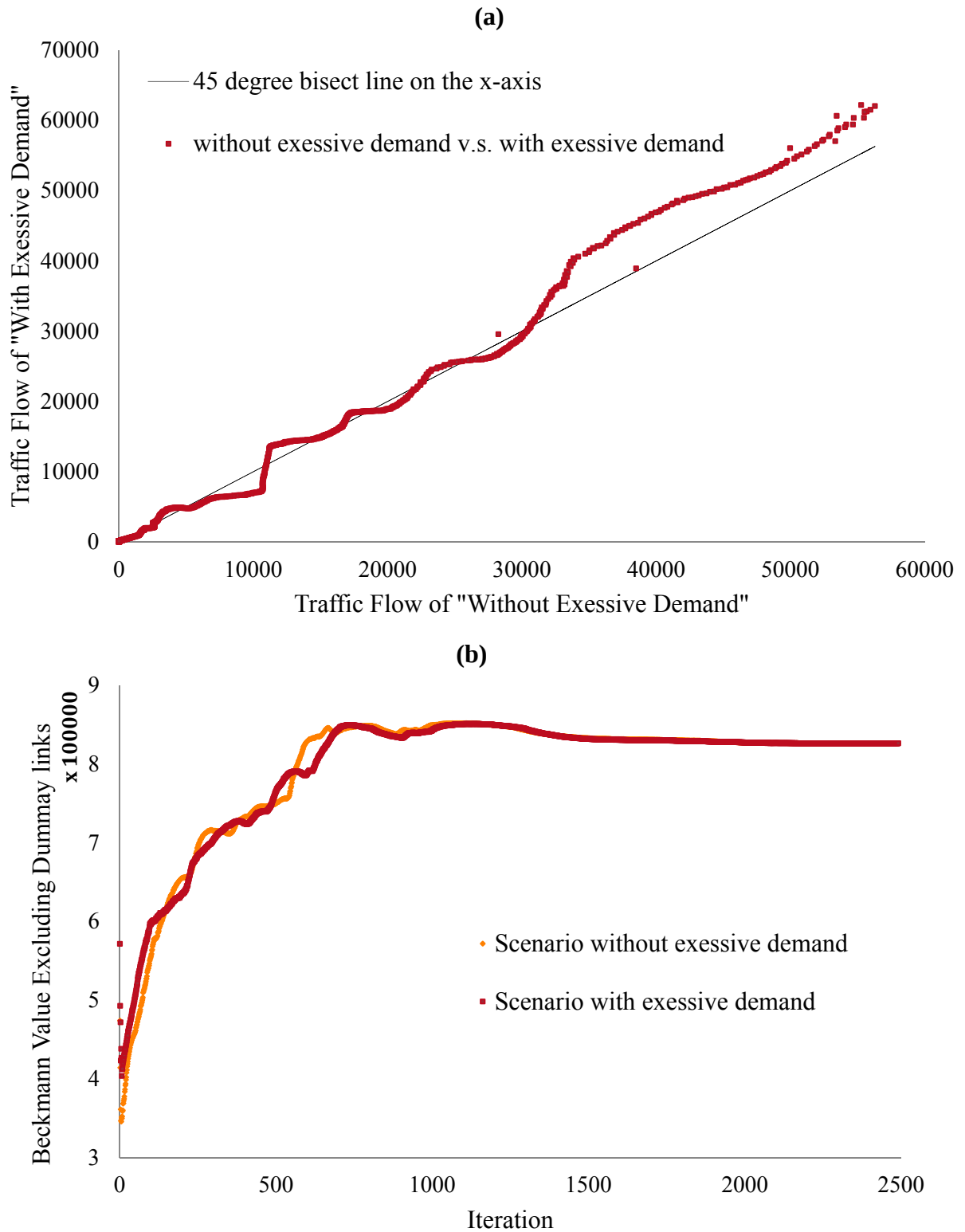


Figure 4.5 Winnipeg network results, sensitivity analysis between with/without excessive demand scenarios: (a) total flow on dummy links, (b) variations of Beckmann values (excluding dummy links) over iterations

5 THE ROAD NETWORK DESIGN PROBLEM

Road investment also known as the network design problem (NDP) is discussed in this chapter. The previous chapter described the development of a method to solve the capacitated traffic assignment as well as the Lagrangian values of the capacity constraint. Accordingly, in this chapter the Benders decomposition is applied to the NDP. Given a set of candidate road projects and associated costs, this problem involves identifying the best subset with respect to a limited budget. The NDP is expressed as a bi-level programming problem. In this study, a special case of the NDP where the decision variables are integers known as the discrete network design problem (DNDP) is tackled. Although a variety of exact solution methods have been proposed for the DNDP, due to the combinatorial complexity, the literature has yet to address the problem for large-sized networks, and accounting for the multimodal and multiclass traffic flows. To this end, the bi-level problem is solved by the branch-and-bound algorithm. At each node of the search tree, a valid lower bound based on the system optimal (SO) traffic flow is calculated. The SO traffic flow is formulated as a mixed integer, non-linear programming (MINLP) problem for which the Benders decomposition method (Benders, 1962) is used.

5.1 Introduction

Traffic congestion is a chronic challenge for cities. In addition to the demand management, making wise investments into expanding the supply side is inevitable. Such investments have to be efficient, and this motivates formulating and solving a bi-level form of the DNDP. The exposition of the DNDP in the literature is as follows: There are a number of candidate road extension projects with associated costs and a limited budget. Hence the problem is postulated as finding the best choice of affordable candidate projects while accounting for the way the users (drivers) utilize the network. Such a premise is formulated as a bi-level programming problem (Magnanti and Wong, 1984) in which the total cost (travel time) incurred by vehicles is minimised. The problem is subject to the drivers' behaviour obeying the principles of user-equilibrium (UE) which in itself is a

programming problem known as the traffic assignment problem (TAP). It has already been established that the NDP is NP-hard (Balakrishnan et al., 1997; Magnanti and Wong, 1984); that is, as the network becomes bigger, the problem in general becomes computationally prohibitive.

The complexity of the DNDP is rooted in two features: the bi-level and discrete nature of the problem. Any bi-level programming problem, even in its simplest configuration (i.e. objective functions and constraints being linear while no integer variable exists) is NP-hard (Ben-Ayed and Blair, 1990; Colson et al., 2005, 2007; Dempe, 2003). The decision variables are binary (1: to build; and 0: not to build the project) and the objective functions are non-linear which make the DNDP a bi-level mixed integer non-linear programming (B-MINLP) problem. In the literature, the phrase “discrete network design problem (DNDP)” is used to emphasize the inclusion of binary elements (rather than continuous decision variables) into the general NDP. The continuous network design problem has been investigated by a number of researchers (Lin, 2011; Unnikrishnan and Lin, 2012; Waller et al., 2006). Despite such complexity, the NDP in general and the DNDP in particular have been studied extensively in many disciplines such as computer science, electrical engineering and mathematics.

As noted before, in this study an exact method for the DNDP, tailored to real-size networks is developed. Throughout this study, it is assumed that travel demand is fixed, deterministic and exogenous. The methodology proposed here includes two important features of the real networks that have been largely neglected in the literature: (i) multimodal: consideration of private traffic flow as well as public or transit flow and (ii) multiclass: various distinct classes of private traffic flow including cars, trucks, and HOV, etc. After examining the literature, it appears that no such attempt (employing exact methods for real networks subject to multimodal and multiclass traffic flow) has been made before.

Although some scholars may prefer a simpler and more parsimonious model based on single class traffic flow, by subjecting the formulation to a multiclass and multimodal traffic assignment this research appeals to the industry as well. Of course, these aspects add to the complexities of the calculations, but it is the cost incurred to address a real life problem and to close the gap between science and practice.

To solve the problem, an efficient branch-and-bound (BB) algorithm hybridized by the Benders decomposition method (dubbed as BB-B) has been devised. The BB

method is employed to address the discrete nature of the DNDP (Boyce et al., 1973; Chen and Alfa, 1991; LeBlanc, 1975), and the Benders decomposition method is used to find tight lower bounds at the nodes of the BB tree. Hybridization refers to the fact that the Benders decomposition method is used to calculate a lower bound for a newly-generated node in the tree.

To tailor the methodology for large-size networks, a variety of innovative techniques have been developed. Node selection and the branching rules of the BB are made based on a merit index computed for each of the candidate projects. To exploit the fact that projects are wisely chosen a term called, budget consumption has also been added in the formulation. Hence the optimum solution is intuitively supposed to utilize the allocated budget as much and effectively as possible. Also, a term, α (varying between 0 and 1) has been added, devised to speed up the algorithm when dealing with large-sized networks. It is embedded in the lower bound calculations as the search on BB proceeds, which is specified upon the employed computational technology. A memoryless search mechanism for the BB algorithm was also developed, that is, the algorithm does not need to remember the entirety of the tree which rapidly expands, otherwise it makes the RAM a serious cause of concern. The objective function is the total travel time spent in the network. The algorithm is launched by an intuitively good solution for which the multimodal, multiclass user equilibrium traffic assignment problem (MMMC-UE-TAP) is computed. It is worth noting that the sub-problem MMMC-UE-TAP can also be replaced with any other traffic assignment model such as quasi-dynamic, dynamic or stochastic etc. Nonetheless, the UE principles are widely used and recognized among scholars due to their widespread applications in research as well as practice.

An extensive review of the relevant literature has already been provided in Chapter 2. In the remainder of this chapter, Section 5.2 provides the general formulation of the problem; in Section 5.3, a customised BB algorithm for the DNDP is introduced; Section 5.4 is dedicated to the Benders decomposition method; numerical results are provided in Section 5.5; and finally Section 5.6 concludes the chapter.

5.2 Formulation of the discrete equilibrium network design problem

Define:

A, A' : Sets of existing roads (or arcs), and candidate road projects (projects in short), respectively

N : set of nodes

B : budget

y_a : binary decisions variable of project $a \in A'$; 1: to build and 0: no to build

c_a : construction cost of project $a \in A'$

$x_a, \bar{x}_a$: auto and transit traffic flows (both in passenger car unit-PCU) on link $a \in A \cup A'$ respectively

$t_a(x_a + \bar{x}_a)$: travel cost or time or delay of link $a \in A \cup A'$, defined by a non-decreasing BPR function (Spiess, 1990) of link flow $x_a + \bar{x}_a$ (called delay function). Some studies have illustrated the highly nonlinear nature of travel time (Lo et al., 2006). Nonetheless the delay functions can be of any form other than BPR as long as they are non-decreasing and differentiable.

A_n^-, A_n^+ : set of links starting and ending at node $n \in N$ respectively; $A_n^-, A_n^+ \subset A \cup A'$

M : set of distinct user classes

b_a^m : additional delay (constant bias) perceived by auto class $m \in M$

x_a^m : traffic volume of auto class $m \in M$ of link $a \in A \cup A'$, in other words: $x_a = \sum_{m \in M} x_a^m$ (see equation (5.7))

O, D, Q : set of origins, destinations and origin-destination pairs respectively, $Q = O \times D$.

$q_i, \bar{q}_i$: auto and transit travel demand in PCU for origin-destination $i \in I$ respectively.

P_i : set of paths between origin-destination, $i \in I$.

h_k, h_k^m : Total flow of all auto classes and flow pertaining to class $m \in M$ on path $k \in P_i$, respectively: $h_k = \sum_{m \in M} h_k^m$ (combination of all traffic flow of different classes constitutes total volume on path k)

$\delta_{a,p}^m$: link-path incident index, 1 if link $a \in A \cup A'$ belongs to path p pertaining to class $m \in M$ and 0 otherwise

$\bar{h}_k$: transit flow in PCU on path $k \in P_i$

$\bar{\delta}_{a,p}$: it is 1 if link $a \in A \cup A'$ belongs to path p pertaining to transit network

w_n : average waiting time at node $n \in N$ pertaining to transit system

f_a : sum of frequency of service for all transit lines on link $a \in A \cup A'$

U : is a sufficiently large value, total demand $\sum_i (q_i + \bar{q}_i)$.

The bi-level DNDP may be written as follows:

$$\underset{y}{\text{Minimize}} \quad \sum_{a \in A \cup A'} (x_a + \bar{x}_a) \cdot t_a(x_a + \bar{x}_a) \quad (5.1)$$

subject to

$$\sum_{a \in A'} c_a \cdot y_a \leq B, \quad a \in A', \quad (5.2)$$

$$y_a \in \{0,1\}, \quad a \in A', \quad (5.3)$$

$$\underset{x, \bar{x}}{\text{Minimize}} \quad \sum_{a \in A \cup A'} \int_0^{x_a} t_a(x_a + \bar{x}_a) dx + \sum_{m \in M} \sum_{a \in A \cup A'} x_a^m \cdot b_a^m, \quad (5.4)$$

subject to

$$\sum_{m \in M} \sum_{k \in P_i^m} h_k^m = q_i, \quad i \in Q, \quad (5.5)$$

$$x_a^m = \sum_{i \in I} \sum_{k \in P_i^m} h_k^m \delta_{a,p}^m, \quad a \in A \cup A', \quad \forall m \in M, \quad (5.6)$$

$$x_a = \sum_{m \in M} x_a^m, \quad a \in A \cup A', \quad \forall m \in M, \quad (5.7)$$

$$x_a \leq U \cdot y_a, \quad a \in A', \quad (5.8)$$

$$\left\{ \begin{array}{l} \bar{x}_a \in \arg \min_{a \in A \cup \bar{A}} \sum_{a \in A \cup \bar{A}} \bar{x}_a \cdot t_a(x_a + \bar{x}_a) + \sum_{n \in N} w_n, \\ \text{subject to} \\ \sum_{a \in A_i^+} \bar{x}_a - \sum_{a \in A_i^-} \bar{x}_a = \bar{q}_i \quad i \in N \\ \bar{x}_a \leq f_a \cdot w_n, \quad a \in A_n^+, n \in N, \\ \bar{x}_a \geq 0, \quad a \in A \cup \bar{A} \end{array} \right\} \quad (5.9)$$

$$x_a^m \geq 0, \quad a \in A \cup A', m \in M. \quad (5.10)$$

Equation (5.1) describes the upper-level goal of minimising total travel time. Mathematical expressions (5.2) and (5.3) ensure the feasibility of projects with respect to their construction costs and available budget. At the lower level, (mathematical expressions (5.4) - (5.7)) the Beckmann formulation of UE flow consists of m distinct auto classes are computed. Constraint (5.8) ensures that projects corresponding to no-build decisions ($y_a = 0$) are excluded from the traffic assignment (U is a sufficiently large value, total demand $\sum_i (q_i + \bar{q}_i)$. Equation (5.9) carries out transit assignment based on an optimal strategy (Spiess, 1993; Spiess and Florian, 1989)) and it returns $\bar{x}_a$ as additional or background traffic volume in PCU to be considered in the traffic assignment. The multiclass facet of the traffic assignment is embedded in the interpretation of the bias term b_a^m in which all distinct auto classes using link a are subject to a same congestion level

(based on the total traffic volume of all classes) plus an additional term (the bias term) exclusive to each class (i.e. $t_a^m(x_a) = t(x_a) + b_a^m$). It is worth noting that the Beckmann formulation is convex w.r.t to x_a . For each class, the shortest path computations of each class take into account the class-specific bias as well as the travel time given by the volume-delay. Therefore, it is not necessary to store the class specific volumes explicitly (x_a^m), whereas, the total volumes are sufficient (x_a) (Spiess, 1984) (INRO, 2009).

The way that a multimodal traffic assignment is conducted in EMME is as follows: based on the headway and transit demand a prior estimation for the transit volume is made (i.e. $\bar{x}_a$). This $\bar{x}_a$ is then treated as a background volume for the auto traffic assignment followed by conducting a transit assignment to get a more precise assignment result. Accordingly $\bar{x}_a$ in equation (5.4) is treated as a constant term derived from transit assignment (sub-problem (5.9)). Since our primary intention was to make use of commercial software for the traffic assignment (i.e. EMME 3) interested readers are referred to the software's manual (INRO, 2009) and (Boyce, 2014).

Such formulations define a simplified way to consider the multiclass aspect of traffic flow. The roads' delay functions are calibrated based on traffic survey data for which the bias term is the intersection value of the non-linear regression. In real practice for cities with traffic analysis models, the delay functions are already calibrated and provided.

5.2.1 Treatment of multiclass and multimodal traffic

A comprehensive consideration of the multiclass user equilibrium traffic assignment problem (UE-TAP) leads to an asymmetric and non-monotone user equilibrium model. For multiclass UE-TAP, a variety of methods such as variational inequality, complementarity method, fixed-points and entropy maximization have been proposed (Aashtiani, 1979; Bar-Gera and Boyce, 1999; Chen et al., 2011b; Dafermos, 1972; Florian and Morosan, 2014; Nagurney, 2000; Nagurney and Dong, 2002; Zhang and Chen, 2010). Nevertheless, the literature has yet to come to a consensus on how to address the multiclass UE-TAP which is still the subject of ongoing debate (Boyce, 2014).

It is worth noting that the evolving knowledge in the state-of-the-art "bush"-based and origin-based algorithms such as algorithm B and TAPAS present a "precious" means of considering the multiclass feature as well. A recent review of the latest advances in the solution algorithms of the UE-TAP is provided by (Xie and Xie, 2014, 2015). Since these

algorithms decompose the UE-TAP to the origins (i.e. one-origin to all-destinations), one can further decompose the UE-TAP at each origin to the number of vehicle classes. Nevertheless, in theory, the computational effort required to solve the multiclass UE-TAP is multiplied by a factor of the number of distinct classes. Usually, practitioners deal with a relatively large number of classes (say a dozen), hence the computation time becomes a significant concern. Alternatively, the bias term (Spiess, 1984) is adopted to turn any multiclass case (no matter how many vehicle classes are involved) into a single-class TAP. Similarly, the combination of (private) traffic and transit assignment (multimodal) results in a nonconvex programming problem for which uniqueness and stability of the solutions with respect to the inputs are not guaranteed (Florian and Morosan, 2014). The relevant studies either fail to fully consider the simultaneous interaction of private and transit modes, or suffer from high computation time (De Cea et al., 2005; Liu and Meng, 2012).

Given the complexities involved as described, the above formulation (equations (5.4) - (5.9)) is empirically proven to be acceptable to addressing the MMMCUE-TAP (Spiess, 1984) such that it has been adopted in some transport planning software (INRO, 2009). In this study, equations (5.4) - (5.9) have been coded as a module in EMME 3 (INRO, 2009) and is summoned when needed.

In the next section the methodology developed to solve the DNDP is presented.

5.3 Branch-and-bound algorithm

A general description of the branch and bound algorithm is presented in chapter 3. In this section a BB algorithm is customised for the DNDP. In order to make this section self-contained, some of the previously defined terminology is introduced again.

5.3.1 Discreteness of the DNDP over the BB

The bi-level DNDP expressed in equations (5.1)-(5.9) is a mixed integer nonlinear programming problem with $|A'|$ the set of binary decision variables. The discreteness of the problem is laid over a tree-shaped structure where each node of the tree represents a sub-area of the solution space, delineated by a sub-problem. The algorithm is first launched with a feasible solution for which the MMMC-UE-TAP is solved, and the corresponding objective function value is labelled as an incumbent value (as an iterative algorithm proceeds, the best solution found is labelled the “incumbent solution” and the

corresponding objective value is called the incumbent value. In other words, the incumbent value is the least total travel time found as the algorithm proceeds through the iterations). Since the problem is of a minimisation nature, the incumbent value is an upper bound value denoted by UB^* . The algorithm can be initiated with any feasible solution. Perhaps the obvious one is the “do-nothing” scenario in which all binary variables are set to zero. Nevertheless, one can seek a more informed initial solution rather than the simple do-nothing, hoping that it facilitates the rest of the algorithm’s process.

A tree is then planted upon a root node representing the entire solution space. Once a new node is generated, a local lower bound is also calculated and tagged on the respective node. In case the lower bound is found to be above the incumbent value, the corresponding node is frozen (or fathomed) and consequently the respective unexplored part of the tree is discarded from further exploration since a better solution (i.e. the incumbent value) has already been secured. As a result, it is very desirable to arrive at fathoming cases (lower bound $>$ incumbent value) to cut the solution spaces as much as possible, otherwise the algorithm chooses an unfathomed node and branches out two new nodes (in other words, the respective solution space is further split into two smaller sub-areas).

This three-phase process (finding no unfathomed node, a lower bound calculation and a comparison with the incumbent value) proceeds until no unfathomed node is found. In this quest, as the tree grows, the sub-areas represented by nodes deep down the tree become smaller and smaller. Sometimes, a sub-area or a sub-problem (represented by a node of the tree) becomes too small, in such a way that it contains only one feasible solution. When the expansion of the tree reaches a feasible solution, the corresponding MMMC-UE-TAP is solved and accordingly the incumbent value is updated. At the end, the incumbent value and the corresponding feasible solution is the final solution. The questions remaining to be answered are as follows: (1) Node selection: which node must be chosen for branching? (2) Branching rule: once a node is chosen, how should the sub-area be split? And (3) Lower bound: how to calculate the lower bounds? The first two questions are discussed in the next sub-section as navigation on the tree.

5.3.2 Navigation in the BB’s tree

Each node in the tree represents either a sub-area or a feasible solution. A sub-area is encoded as a string of binary values (“1” and “0”) to indicate whether the respective

project is to be constructed or not constructed, and “2” is yet to be decided (either 1 or 0). For instance, the string “01022” depicts a sub-area consisting of five candidate projects where the first three components are to be build/no-build (0/1) and the last two (represented by “2”) are unspecified (yet to be decided). In other words, subarea “01022” encompasses all feasible solutions in which the first three projects are fixed as “010” and the last two are either of the following combinations: “00”, “01”, “11”, or “10”.

As the tree grows, at each iteration, a node “z” representing a sub-area needs to be chosen for further branching. To this end, an undecided project (a project represented by value “2”) has to be selected to be replaced by 0 (no-build) and 1 (to-build) on two new branches. The result is that it will create two new nodes.

During this mitotic phase the tree continues to expand, the algorithm is constantly faced with the decision of which node with which partial solution to be chosen for branching. Once a node is selected, the algorithm must still pick one undecided project of the corresponding partial solution to complete the next phase of branching. To this end, there are some methods that require to solve additional problems subject to retrieve the entire database with a view to finding the best node for branching. As the size of the network increases, such methods become computationally intensive. The roles described below are designed to overcome such pitfalls.

5.3.3 Branching rule based on merit index:

Alternatively, via a novel approach, the order of the projects placed in the string from left to right as priority for node selection and branching is considered. Therefore, a merit index is first defined and calculated for each project. The projects are then sorted from the highest to the lowest to be placed accordingly in the string.

The merit index aims to find the most likely projects among the candidates. As such, the merit index is defined as: $x_a/v_a/c_a$ where v_a stands for capacity of road $a \in A'$. The rationale behind the proposed merit index is based on two considerations: (i) the traffic volume alone is not a good enough indicator to be a prompt to increase the qualification or chance of a candidate, the capacity also needs to be considered. The more congested a road (i.e., higher volume-per-capacity x_a/v_a), the more demanding the road is and hence it may deserve to be put forward for the construction; and (ii) between two roads with similar traffic conditions, it is a wise choice to choose the one that would result

in the lowest construction cost; therefore, the volume-per-capacity in the merit index is normalised by the construction cost.

Consequently, the branching rule becomes very simple: first sort the projects based on their merit indices in descending order. For branching, there is only one rule: choose the very next undecided project in the corresponding string.

5.3.4 Node selection rule

As for the node selection, the algorithm abides by two simple rules: (i) choose the deepest node of the tree; and (ii) in the case of two nodes at the same level, choose the one located on a branch associated with $y_a = 1$.

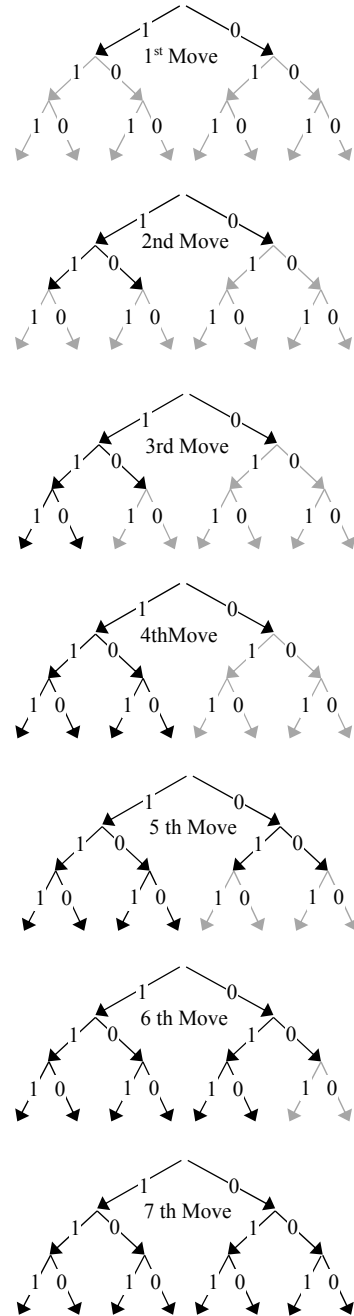
There are two advantages with such a convention: (i) given the fact that the projects are sorted on a merit basis, it makes sense to go deep into the tree to select the next best project for branching, and (ii) the algorithm needs not to save/retrieve/process the information for the entire tree. Once a new node is made, it is yet to be further processed for lower bound/fathoming. At each node, the algorithm just needs to move forward as much as possible on the paths that consist of $y_a = 1$ branches. In cases where there is no space for such movement, the algorithm moves only one node back to the previous node and then moves through the $y_a = 0$ and then follows a $y_a = 1$ branch (if possible). This process carries on until the termination criterion is met. Figure 5.1 illustrates the gradual build-up of the tree based on these rules.

The algorithm does not need to remember the paths already traversed nor the paths ahead. As shown in Figure 5.1 as the structure expands; it just needs to know the lower bounds of the nodes on the current path plus the best solution found so far which is a string of binary values (0/1) and the corresponding incumbent value. This can be called a memoryless search mechanism. For example, if the current node is (11002), the next move is to process node (11001) followed by the node represented by (11000). For the third move, the algorithm moves three nodes back to reach node (10222) and carries on from there.

5.4 Benders decomposition

In this section, first a discussion how to place the Benders decomposition method in the BB's tree structure to elicit a valid and tight lower bound value is presented. Although a basic introduction of the Benders decomposition was provided in chapter 3, here an in-

depth elaborate on the mathematical principles underlying the Benders decomposition method relevant the DNDP is provided.



- Starting solution at root node: “222”
- Node selection rules: (i) choose the deepest node of the tree (ii) in case of having two nodes at the same level choose the one that was made of a branch representing $y_a = 1$.
- Branching rule: choose the very next undecided project in the corresponding string solution.
- No budget restriction

Figure 5.1 Proposed node selection and branching in the branch-and-bound algorithm

5.4.1 Formulation of the lower bound

One of the primary concerns in any application of the BB algorithm is the method of calculating lower bounds at the tree's nodes. For the sub-area represented by "01022", it is necessary to compute a lower bound on the objective function of (equation 5.1) evaluated at all possible and feasible combinations: "01000", "01001", "01010", and "01011". An apparent method of computing a valid lower bound is to set all undecided variables equal to one and solve the MMMC-UE-TAP which results in the UE flows on the resulting network. According to Braess Paradox (Braess, 1968; Braess et al., 2005), such measure may result in worse-off traffic or a higher lower-bound. To this end, replacing the user equilibrium (UE) flow with system optimal (SO) traffic flow (MMMC-SO-TAP) ensures valid and decreasing successive lower bounds (LeBlanc, 1975). Nevertheless, in theory, there might be a significant gap between the UE flow and the SO flow which would lead to a very loose lower bound. This means that it is unlikely to truncate the solution space due to arriving at a lower bound above the incumbent value (upper bound).

In such cases, the algorithm has to process every feasible solution and as a result, the algorithm will have no superiority over an exhaustive enumeration. In fact, it becomes much worse because in addition to calculating the upper bounds as required in the enumeration, the algorithm has to calculate the lower bounds as well.

In order to obtain a tighter lower bound for a (new) node, one can seek a solution out of the following problem (LeBlanc, 1975):

$$\min \sum_{a \in A \cup A'} x_a^m t_a^m(x_a) \text{ s.t (5.2), (5.3), (5.5), (5.6), (5.7), (5.8) and (5.10)} \quad (5.11)$$

In the above problem (5.11), the non-linear convex objective function is subject to linear constraints and the output consists of binary decision variables (y_a) as well as traffic flows which are continuous variables (x_a). Hence, it is a mixed integer nonlinear programming (MINLP) problem for which a solution algorithm based on the Benders decomposition method (Benders, 1962; Lasdon, 2013) is developed. In fact, the lower bound formulated in (5.11) is a system-optimal discrete network design problem (SO-DNDP). In some studies, the SO-DNDP itself is treated as an approximation approach to address the DNDP (while employed it as a lower bound). The main advantage of such formulation rests in the fact that the bi-level DNDP is easily dissolved into a single-level problem which is easier to solve.

As discussed before (see equation (5.9)), in the above formulation, the output of the transit assignment is $\bar{x}_a$ which is viewed as additional or background traffic volume to be considered in the UE traffic assignment. Therefore, the travel time function can be first updated to accommodate the $\bar{x}_a$ before carrying out the traffic assignment. Hence, the notation $\bar{x}_a$ as well as equation (5.9) can be omitted from the formulation.

5.4.2 Benders decomposition method for a MNLP

Consider the original problem (5.11) in the general form of MINLP as follows:

$$OP : \underset{x,y}{Min} \ f(x,y) \ s.t. \ S(x,y) \leq 0, \ x \in X \subset \Re^n, \ y \in Y = \{0,1\}^q \quad (5.12)$$

where f is the objective function ($\sum_{a \in A \cup A'} x_a t_a(x_a)$), x is the vector of links flows, y is the vector of binary decision variables along with set S delineate solution space to which the problem is subject. Note that x, y denotes the continuous traffic volumes and binary decision variables respectively. Analogous to problem (5.11) consider $S(x,y) \leq 0$ represents constraints (5.2), (5.3), (5.5), (5.6), (5.7), (5.8) and (5.10) and $f(x,y) = \sum_{a \in A \cup A'} x_a^m t_a^m(x_a)$.

Consider eliminating the binary decision variables by fixing them to some feasible values (y^i ; i is iteration counter), hence the problem changes to searching over feasible x . It is referred to as the “primal sub-problem (PSP)”:

$$PSP(i) : \underset{x}{Min} \ f(x, y^i) \ s.t. \ S(x, y^i) \leq 0, \ x \in X \subset \Re^n \quad (5.13)$$

Once it is solved, the corresponding traffic volume x^i and Lagrange multipliers of the constraints ω^i are obtained.

In order to solve the PSP and finding the traffic volumes and the Lagrangian values a method known as inflated travel time has been developed (Bagloee and Sarvi, 2015a) which is discussed in the previous chapter.

The partial dual (Lagrange) format of the objective function can be written as $L(x^i, y, \omega^i) = f(x^i, y) + \omega^i \cdot S(x^i, y)$. According to “weak duality theorem”, each feasible solution (y) to the dual problem is a lower bound to the original problem. Therefore, given (x^i, ω^i) , the algorithm seeks a new set of feasible binary variables for the next iteration (y^{i+1}) by solving the following problem which is called “relaxed master problem (RMP)”:

$$RMP(i): \quad \text{Min } V \quad \text{s.t. } V \geq \min_{y \in \bar{Y}, V} L(x^k, y, \omega^k); \quad k = 1..i \quad (5.14)$$

The problem (5.14) is a mixed integer linear programming (MILP) problem which is easier to solve (compared to MINLP), even for sizable problems. The right hand side of the constraint is a succinct way of representing a series of linear constraints to which “v” must be found greater than the minimum of them. Note that $y \in \bar{Y}$ ensures feasibility of the binary variables where $\bar{Y}$ is feasibility observing two points: (i) it entails all combinations of binary variables that satisfy the budget constraint (constraint 5.3); and (ii) the linear constraints in the relaxed problem are actually viewed as Benders cuts (cutting planes) to the solution space collected from the first iteration to current iteration i , (i.e. $k = 1..i$). In order to always obtain a new y^{i+1} at the current iteration i , the solutions that fail to satisfy the following constraints are discarded from $\bar{Y}$ (Balas and Jeroslow, 1972):

$$\sum_{a \in Y1} y_a - \sum_{a \in Y0} y_a \leq |Y1| - 1, Y1 = \{a \mid y_a^k = 1\}; \quad Y0 = \{a \mid y_a^k = 0\}, \quad (5.15)$$

where $Y1, Y0$ represents the candidate projects that have taken a value of 1 and 0 respectively in the respective solution k .

On the one hand, the result of the primal sub-problem $PSP(i)$ is a feasible solution, so the value of $f(x, y^i)$ out of $PSP(i)$ is an upper bound on the optimal value of the original problem. On the other hand, as mentioned before, the relaxed master problem is in fact a relaxed dual problem to the original problem, hence v out of the $RMP(i)$ is a lower bound on the original problem’s optimal value. Consequently, the solution algorithm is set out to solve the primal and relaxed problems ($PSP(i)$, $RMP(i)$) alternatively until the lower and upper bounds get within a close enough proximity to each other.

5.4.3 Benders decomposition for a system optimal DNDP

Let’s rewrite the original problem given in formulation (5.11) by expanding on the traffic volumes:

$$OP \text{min} \sum_{a \in A} x_a \cdot t_a(x_a) + \sum_{a \in A'} x_a \cdot t_a(x_a) + \sum_{m \in M, a \in A \cup A'} x_a^m \cdot b_a^m \quad \text{s.t.} \quad (5.2), (5.3), (5.5), (5.6), (5.7), (5.8), (5.10) \quad (5.16)$$

5.4.3.1 Establishing the primal sub-problem

Given a feasible initial binary solution (y_a) to start with, the original problem (5.16) for the first iteration ($i=1$) is solved to return traffic volume (x_a) as well as ω_a ; $a \in A'$ Lagrange multipliers associated with the inequality (5.8). Hence the Lagrangian objective function can be written as follows (Note that problem (5.17), is derived based on problem (5.16) via a feasible binary solution. Constraints (5.2 & 5.3) contain binary variables which stand for the feasibility, hence both are automatically dissolved when a feasible binary solution is arrived at. As such there is no point to bring them into the dual problem):

$$L(x_a, y_a, \omega_a) = \sum_{a \in A} x_a t_a(x_a) + \sum_{a \in A'} x_a t_a(x_a) + \sum_{m \in M, a \in A \cup A'} x_a^m \cdot b_a^m + \sum_{a \in A'} \omega_a (x_a - U y_a) \quad (5.17)$$

The Lagrange function can be rearranged as follows:

$$L(x_a, y_a, \omega_a) = \sum_{a \in A} x_a t_a(x_a) + \sum_{a \in A'} x_a (\omega_a + t_a(x_a)) + \sum_{m \in M, a \in A \cup A'} x_a^m \cdot b_a^m - \sum_{a \in A'} \omega_a U y_a \quad (5.18)$$

where U is a sufficiently large value hence, if constraint (5.8) is found binding (or $\omega_a = 0$), $y_a = 1$, otherwise $\omega_a > 0$ for $y_a = 0$. Therefore, a complementarity constraint always holds $y_a \cdot \omega_a = 0$, or equivalently $x_a \cdot \omega_a = 0$. Accordingly, the last term in equation (5.18) vanishes. Given a feasible solution y_a :

$$UB_b^i = \text{Min } L(x_a, \omega_a) = \sum_{a \in A} x_a t_a(x_a) + \sum_{a \in A'} x_a (\omega_a + t_a(x_a)) + \sum_{m \in M, a \in A \cup A'} x_a^m \cdot b_a^m \quad (5.19)$$

s.t. (5.5), (5.6), (5.7), (5.10)

The above problem has simply become a capacitated MMMC-SO traffic assignment while the only additional component is omega ($\omega_a \geq 0$). As can be seen, the omega needs to be added to the travel time of the respective project. This problem is still convex and can be solved by the augmented Lagrangian method (ALM) (Larsson and Patriksson, 1995; Patriksson, 1994) or inner penalty function (IPF) (Nie et al., 2004). There are some challenges in both methods such as the number of parameters involved.

Alternatively, as discussed in the previous chapter, a method based upon an intuitive interpretation of the omegas (Lagrangian values) for the general capacitated traffic assignment problem (CTAP) is developed. In the objective function of the above problem, the omegas sit next to the delay term. Thus the omegas can be treated as a penalty term to be imposed on the candidate projects (roads), those who have been decided as no-build in order to block them. This interpretation leads to a method dubbed “inflated travel time” which bears none of the aforementioned shortcomings in the AFW

and IPF methods (Bagloee and Sarvi, 2015b). As a result, the CTAP is transformed into a normal, uncapacitated TAP for which any conventional methods such as Frank-Wolfe are applicable.

The value of the objective function in formulation (5.19) denoted by UB_b^i is the total travel time of the MMMC-SO traffic flow, which renders an upper bound at iteration i (subscript b refers to the Benders Method to distinguish this upper bound from the upper bound of the BB).

So far, the primal sub-problem which returns (x_a^i, ω_a^i) is solved. Now the relaxed master problem can be established to seek new binary decision values for the next iteration (y^{i+1}) .

5.4.3.2 Establishing the relaxed master problem

Considering the feasible solution that resulted from solving the previous primal sub-problem, let's now rewrite the original problem as follows. As discussed before, the dual (Lagrangian) format of the original problem (OP) is established based upon the solution that results from the primal Sub-problem $(x_a) : L(x^i, y, \omega^i) = f(x^i, y) + \omega^i \cdot S(x^i, y)$. Note that the objective function of the OP (equation (5.16)) is free from any binary decision variables (y) , hence $f(x^i, y) = f(x^i)$ becomes the total travel time of the MMMC-SO traffic flow, which is already computed and denoted by UB_b^i . As for the second term $(\omega^i \cdot S(x^i, y))$, given the already specified values of x_a^i , the only constraint left in the OP is constraint (5.8). Using ω_a^i as the dual variable for constraint (5.8), it can be brought up to the objective function using the penalty term: $\omega_a^i \cdot (x_a^i - U \cdot y_a)$. Note that it has already been shown the following complementarity relationship: $\omega_a^i \cdot x_a^i = 0$. Hence, by summing up over $a \in A'$, the second term is obtained as $-U \cdot \sum_{a \in A'} \omega_a^i \cdot y_a$. Consequently, the dual lower bound to the OP (Benders cuts) can be finalised as: $V \geq UB_b^i - U \cdot \sum_{a \in A'} \omega_a^i \cdot y_a$ where V is the value of the objective function of the relaxed master problem. The above cut, along with the other cuts stacked up in the previous iterations are combined and are included in the relaxed master problem as follows:

$$LB_b^i = \min_{y_a} V, \quad (5.20)$$

$$S.t. \quad V \geq UB_b^k - U \cdot \sum_{a \in A'} \omega_a^k \cdot y_a, \quad k = 1..i, \quad (5.21)$$

$$\sum_{a \in Y1^k} y_a - \sum_{k \in Y0^k} y_a \leq |Y1^k| - 1, \text{ where } Y1^k = \{a \mid y_a^k = 1\}; Y0^k = \{a \mid y_a^k = 0\}, k = 1..i, \quad (5.22)$$

$$\sum_{a \in A'} c_a \cdot y_a \leq B, \quad (5.23)$$

where constraints (5.22) and (5.23) ensure rendering a new and feasible binary variable (y_a^i) solution at each iteration. So far, both primal and relaxed problems are established and the Benders decomposition algorithm can be established:

Step 0 (Initialization) - set iteration counter $i := 1$; $UB_b^i := +\infty$; set initial solution for binary variables ($y_a^i = 0$); set convergence gap ε .

Step 1- Given y_a^i solve the primal problem (5.19) to find $UB_b^i, x_a^i, \omega_a^i$. Set $UB_b^* = \min(UB_b^i, UB_b^*)$. Note that the algorithm starts with an initial solution for the binary variables, based on which a traffic assignment is calculated and values of x_a are found. If a project is decided not to be constructed then $x_a = 0$.

Step 2- Give x_a^i, ω_a^i solve the Relaxed Master Problem (5.20) to (5.23) to find LB_b^i, y_a^i

Step 3- (termination) if $(UB_b^* - LB_b^i) / LB_b^i \leq \varepsilon$ then the convergence is achieved so it returns UB_b^* as the final solution. Otherwise $i := i + 1$; go to Step1. ■

In order to better understand how Benders decomposition works, a simple and pedagogical example is undertaken as presented in chapter 3. The merit index in the BB algorithm is an educated guess so as to quickly arrive at the optimum solution. It is based on an intuitive view on the optimum solution; that is, if a road is really necessary, upon completion of construction, it must become highly congested. In order to show the significance of the merit index, the aforementioned pedagogical example is also solved with a merit index and the results are compared against that of without merit index. The results are presented in the next section which shows that the search for the optimal solution is highly decreased.

5.4.4 Evaluation of the merit index in the performance of the BB algorithm

With respect to the pedagogical example presented in chapter 3 which is a mixed integer quadratic programming problem, Figure 5.2 depicts a tableau of the example solved by the BB in two scenarios: with and without the merit index (x_i is continuous and y_i is binary variables). From the coefficient of the variables in the objective function, it is intuitively perceivable that the merit order of the binary variables is as y_1, y_2, y_3 . The optimal solution was found as follows: $(y_1, y_2, y_3) = (1, 1, 0)$, $(x_1, x_2, x_3) = (6.1, 3.1, 0.0, 0.8)$ and $f(x, y) = 7.7$.

A quick comparison between the two scenarios highlights the significance and constructive role of the merit index in efficacy of the BB, such that the number of attempts to reach the global solution and total computational time increases almost three-fold should no merit index be considered.

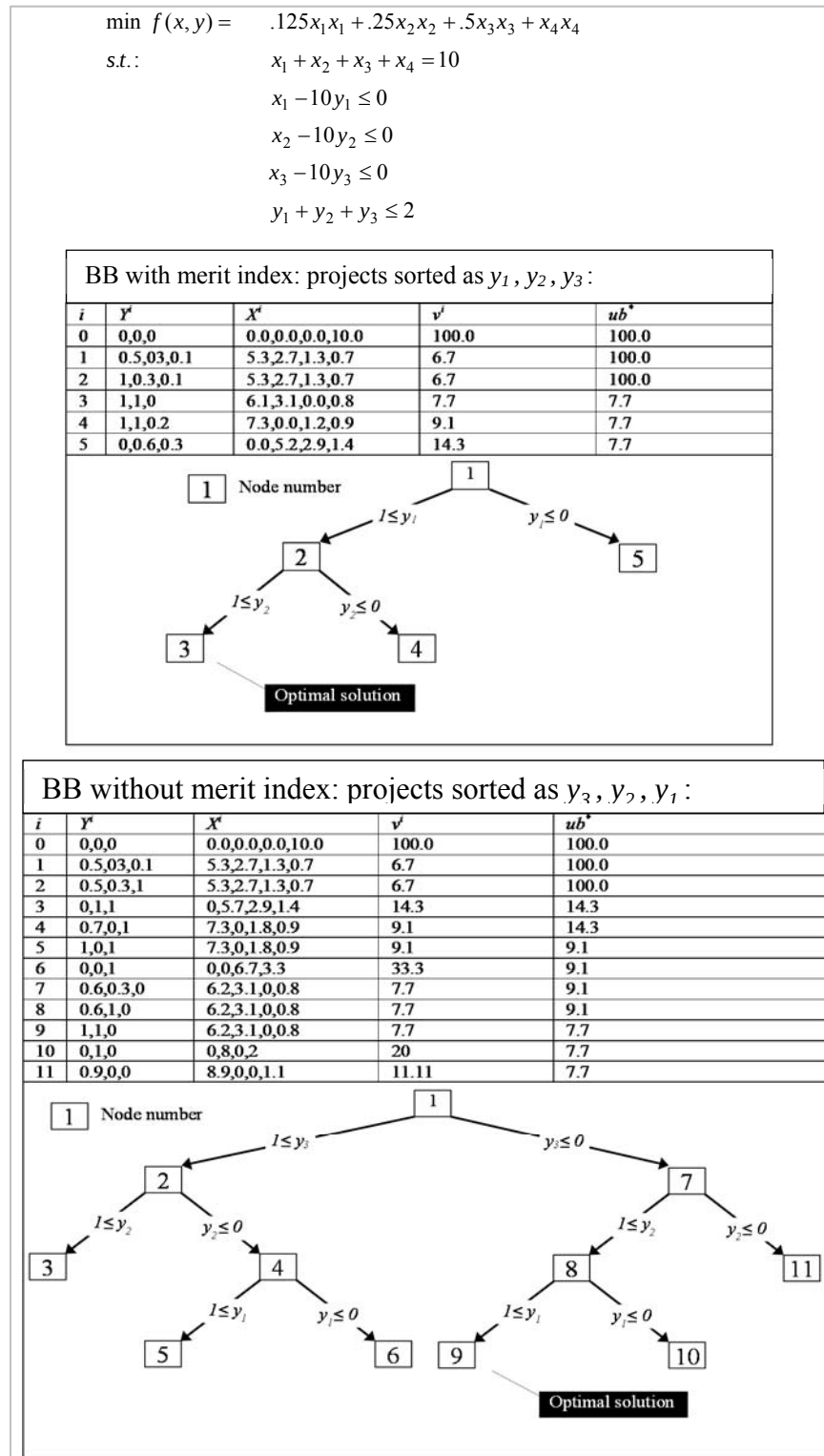


Figure 5.2 BB's performance with/without the merit index (note: v^i is value of the objective function at iteration i , and ub^* is the best upper bound or the incumbent value)

The BB initiates on the candidate projects that are already sorted in order of their merit indices. Then, the algorithm takes $y_a = 1$ branches on the tree as deep as possible because it is believed that the projects are wisely selected, hence the more obtained of $y_a = 1$, the better it would be. Therefore, the proposed algorithm is a combination of the best-first-node and depth-first-node which results in a memoryless search algorithm; that is, no need to keep track of the whole of the tree structure.

Before proceeding to the next section dedicated to numerical evaluations, there are some remarks which are discussed in the next section.

5.4.5 Some remarks on the methodology

Efficacy of the tree structure in the lower bound's calculation: At each iteration a (parent) node renders two new (offspring) nodes, but it is only required to calculate the lower bound for one of the newly generated nodes. Because the other offspring node inherits the lower bound from the parent node, this makes the computation more efficient. Figure 5.3 depicts this observation graphically. The parent node corresponds to sub-area “1122...22”, has a lower bound corresponding to string “1110...01” with an objective function value of 85. Branching is made at the third project (the very next project with value of “2”) which results in offspring nodes “1112...22” in the left hand and “1102...22” in the right hand. The third project in the string corresponding to the lower bound of the parent node is found “1”, so if the lower bound for the left hand offspring node (which has “1” at the third project) is calculated, no better solution than what has already been found, will be found. For the right hand node, the lower bound was found not better than the parent's lower bound ($88 \nless 85$). If the lower bound string for parent node happened to be “1100...01” and the right hand offspring node were to inherit the lower bound from the parent, a new lower bound would have to be calculated for the left hand node (LeBlanc, 1975).

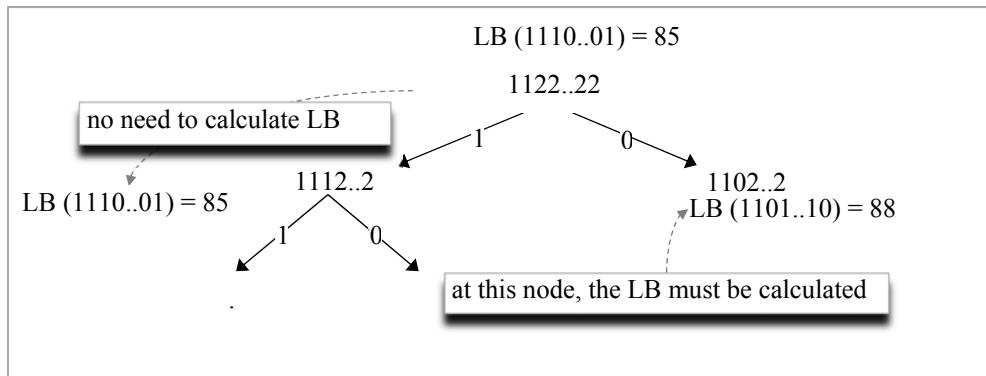


Figure 5.3 Illustration of how the lower bound values are inherited through the tree

Benders algorithm; epsilon, convergence gap: In the proposed Benders algorithm, Step 3, epsilon (ε) is a pre-specified parameter (in percentage) for which a perfect and desirable value is 0%. Preliminary results of the application of Benders on large scale MINLP problems suggest that the global solution is likely to be arrived in early iterations, while the epsilon only prolongs the computational time to close the gap between the upper and lower bounds. In order to speed up the algorithm, instead of the perfect value of 0, a meagre value for epsilon was adopted (say $\varepsilon = 0.02$).

Benders algorithm; initial solution: In Step 0, the algorithm starts with an all-out null solution, that is $(y_a^0 = 0, \forall a \in A')$. One option is to look at the inventory of the best solutions found in the preceding node of the tree to see whether there is a binary solution that complies with the requirement of the current node. If nothing is found, then there still might be a better educated guess rather than the null scenario. One possibility is to fill the blank string from left to right with projects based on the merit index until the budget is depleted.

This technique has been implemented in the final computer code and was applied in the numerical analysis.

Benders algorithm; a greedy search for better a incumbent value (I): A Benders algorithm was devised to render a tight lower bound based on the MMMC-SO traffic flow. The corresponding binary solution of the best solution emerged out of the Benders algorithm and might render a much better incumbent value too. As a result, at the end of each Benders algorithm the resulted binary solution is taken and for which the MMMC-UE-TAP is solved. The outcome is used to update the incumbent value. The rationale behind the adjustment of the algorithm is as follows: the SO version of the network design problem is a good approximation for the UE version. It is important to note that the first

lower bound is calculated for the root node of the BB's tree (where representing the entire solution space by the string "22...22". In fact, this lower bound is equivalent to solving a SO version of the NDP which is proven to be easier and this is called SO-relaxation (Wang et al., 2013). This idea is strongly reinforced in the numerical analysis (next section) such that in most cases, the global optimal solution is achieved in the first lower bound calculated at the root node.

This technique has been implemented in the final computer code and was applied in the numerical analysis.

Benders algorithm; a greedy search for a better incumbent value (II): Provided the candidate projects have been selected wisely, a good solution is expected to consume the budget to its full. The more projects that contribute to the solution, the better the solution becomes. To this end, to force more of values of "1" in the binary strings (y_a) , the objective function of the relaxed master problem (equation 5.20) is changed to: $V - \sum_{a \in A'} y_a$. The newly added term $(-\sum_{a \in A'} y_a)$ is called "budget consumption term" which again improves the results. In the next section (numerical tests), both cases with and without the budget consumption term are reported on.

A much tighter system optimal lower bound: No matter how perfect the attempt is to find the maximum possible lower bound, since the lower bound is based on the system optimal (SO) traffic flow, the gap between SO lower bound and UE incumbent value might be noteworthy. Roughgarden and Tardos (2002) proved mathematically that the incumbent value corresponding to the UE traffic flow can be as high as 2.15 times the SO lower bound in networks governed by BPR delay functions. Recently, a similar result has been reported by (Szeto and Wang, 2015). In the following discussion, the causes of the aforementioned gap are highlighted. Then a parameter to close this deep gap is proposed.

The SO flow can be easily computed (even using commercial transport planning software) by replacing $t_a(x_a + \bar{x}_a)$ the delay function in the Beckmann objective function of the UE flow (equation (5.4)) to $\tilde{t}_a(x_a + \bar{x}_a)$ (Newell, 1980; Potts and Oliver, 1972; Sheffi, 1985):

$$\tilde{t}_a(x_a + \bar{x}_a) = t_a(x_a + \bar{x}_a) + x_a \cdot \frac{\partial t_a(x_a + \bar{x}_a)}{\partial x_a} \quad (5.24)$$

If $t_a(x_a + \bar{x}_a)$ is considered as the cost of traveling on a road $a \in A$, then $\tilde{t}_a(x_a + \bar{x}_a)$ is known as the marginal cost of using the respective road. The deep gap between SO and UE emerges from the second term in the right side of equation (5.24) which is the

additional externality cost imposed on the users. The two functions $t, \tilde{t}$ show benign and similar behaviour as long as the volume is below capacity. As the volume gets close to or exceeds capacity, the externality cost increases rapidly, which results in the larger gap between the SO and UE flows. Since the delay functions are not capacity restricted, in theory, the volume and hence the delay as well as the marginal cost can increase to infinity. This results in a much larger gap between SO and UE. In order to eliminate such an unrealistic gap, alpha $0 \leq \alpha \leq 1$ a coefficient to the externality term is proposed as follows:

$$\tilde{t}_a(x_a + \bar{x}_a) = t_a(x_a + \bar{x}_a) + \alpha \cdot x_a \cdot \frac{\partial t_a(x_a + \bar{x}_a)}{\partial x_a} \quad (5.25)$$

As alpha gets close to zero, the SO gets close to UE and the gap vanishes. It is worth noting that the alpha addresses the unfortunate trade-off between the computational time (CPU time) and accuracy of the final results. Lower values of alpha lead to faster but less accurate results. As such, the value of alpha can be set as per the user's discretion depending on the computational technology at the time and how affordable the computational time is. In other words, alpha is a value in the hand of the modeller based on which the accuracy of the results along with the computational time can be adjusted depending on the size of the network, the available computational technology and the strategic value of the final solution.

In addition, adopting any positive value below 1 for alpha is a diversion from solving a full SO to a semi SO network design problem. The numerical results suggest that even a trivial value of alpha (say $\alpha = 0.005$) is enough to secure global optimal solutions (which were already identified through exhaustive enumeration).

A note on convergence of the Benders algorithm; In some circumstances the Benders decomposition does not converge when the NLP is non-convex, (Bagajewicz and Manousiouthakis, 1991). In the following exposition, these circumstances and why the proposed algorithm does not encounter such circumstances are discussed.

The important point to note is that given a feasible solution of binary variables (y), problem (5.11) is convex on the continuous variables of the roads' traffic volume (x). In fact, for every feasible solution y, problem (5.11) is a system-optimal traffic assignment problem which has a guaranteed unique solution. According to (Bagajewicz and Manousiouthakis, 1991), the key to avoid stagnation in the local optimum is a condition coined by Geoffrion (1972) known as "property (P)". Property P stands for the

situation in which the Lagrangian relaxation problem can be taken, essentially independently of y , so that the dual problem (equations (5.20)...(5.23)) can be solved on y . This is exactly what is being conducted in solving the lower bound problem. Generally speaking, convergence of the scheme is guaranteed if - whenever the integer values y are fixed to some feasible vector - the remaining continuous sub-problem is convex and is such that strong duality holds (the Lagrangian dual maximum value is equal to the primal optimum value) (Sahinidis and Grossmann, 1991).

Before proceeding to the numerical analysis as a recap, the process for implementing the Benders algorithm is shown in Figure 5.4.

5.5 Numerical evaluations

This section, firstly examines the no-frills (unimodal and single class) versions of the algorithm over Gao's 12-nodes network (Gao et al., 2005) and the Sioux Falls benchmark network (Farvaresh and Sepehri, 2013; LeBlanc, 1975) to cast the proposed methodology in the context of its peers in the literature. The algorithm is then applied in its full capacity (multiclass and multimodal) to the large scale network of Winnipeg. Exhaustive enumerations have already been carried out to find optimal solutions for all the case studies over various budget levels. For the Winnipeg case study, the enumeration entails all combinations of network scenarios with accounts for solving 1,048,576 ($= 2^{20}$) traffic assignment problems. The quality of the solutions resulting from the proposed algorithm will be compared against the optimal solutions elicited from the enumeration.

The parameters setup for the algorithm are as follows: (i) the relative gap introduced in Benders algorithm, in Step 3 is assumed: $\varepsilon = 2\%$ (ii) the value of α introduced in equation (5.25) is initiated with zero to seek the tightest possible lower bound. Since $\alpha = 0$ might compromise the quality of the solutions, should the optimal solution not found at $\alpha = 0$, more attempts with positive values of the alpha are tested.

The algorithm is first run with alpha equals to zero. The subsequent result shows that the optimal solutions are likely to be found in early iterations. Hence, in real practice, where the optimal solutions are not known and faced with an NP-hard problem – one can still run with alpha-zero and expecting to find a good solution (if not the optimal) at an affordable computational time.

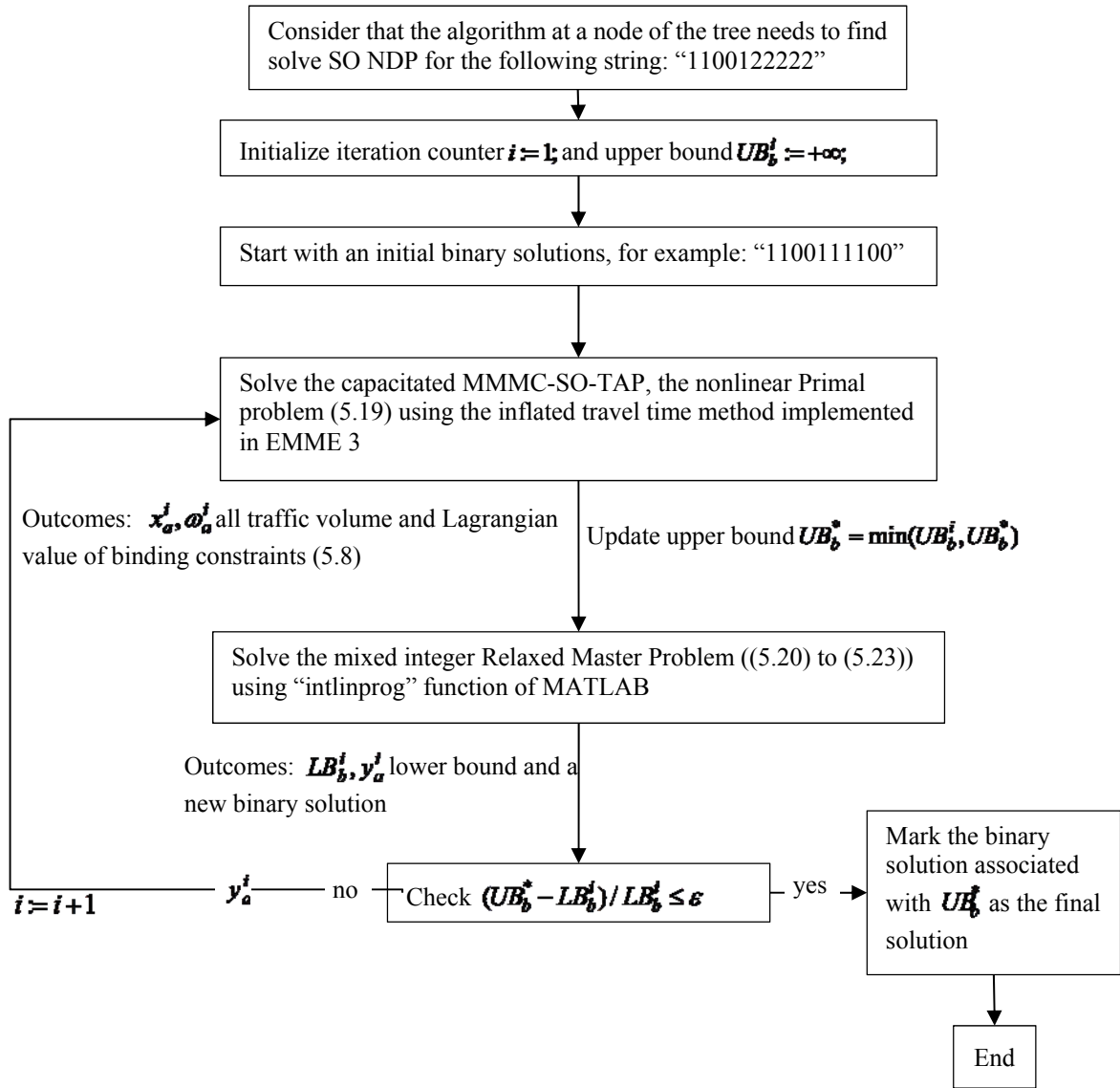


Figure 5.4 Benders algorithm: the flowchart graphically represents the steps

As discussed before (in Remark 4), the first lower bound is calculated for the root node of the BB's tree (representing the string "22...22"). In most of the numerical tests, the optimal solutions were found at the root node. As a result, in the comparative analysis for the Gao network and the Sioux-Falls benchmark network, the focus is set on the number of iterations made at the root node to arrive at the optimal solution. A comparison will be made against some of the state-of-the-art exact methods in past studies. In light of the fact that the CPU time is heavily subject to the computational technology used at the time, as well as the coding architect, the number of iterations can be regarded as a fair yardstick in comparative analysis. Furthermore, each iteration of exact methods is usually involved in alternately solving two sub-problems (the UE-TAP and an MILP) which is

analogous to the proposed method in this study. Nevertheless, the algorithm quickly terminated for the Gao's 12-nodes network and the Sioux-Falls network in a matter of a few seconds and a few minutes respectively. Furthermore, the CPU time for solving the large scale Winnipeg network will be discussed in further detail.

For all the case studies, in addition to optimal solutions and the corresponding objective function value (i.e. the total travel time dubbed as the incumbent value), the size of the solution spaces (i.e. the number of feasible solutions) are also presented. The size of the solution spaces gives us an indication of how rough and unlikely the path toward the global optimum solution may be.

As for the computational technology, a desktop computer with Intel(R) Xeon(R) 3.70 GHz and 64.0 GB RAM was used. The algorithm was coded with Visual Basic linked to MS-Excel and MS-Access as an interface and save/retrieve database. The computer code is also synchronized with EMME 3 to solve the multimodal and multiclass traffic assignment problems. The code also calls on MATLAB 14a to solve the MILP problems using the newly released module "intlinprog" (MathWorks, 2014). All delay functions associated with the links conform to the BPR type.

For the Sioux-Falls and Winnipeg case studies, the candidate projects are two-way roads that are found in the real world. Should a candidate project receive approval for construction, two directional links need to be added to the network. Instead of representing each directional link as a separate decision variable (which leads to an increase in the number of binary variables and constraints), the concept of a "directional switch link" to represent the two-way roads is developed and proposed. Figure 5.5 illustrates a two-way road between A and B at the top. The same two-way road is disconnected into two pieces (without change to any characteristics) and is then reconnected with a one-way switch link (the dashed line) at the bottom (the two-way links are shown as bending outward to better illustrate the switch link). In this way, the switch link can represent the two-way links (painted by red and green colours).

What then needs to be done is to code the disconnected two-way roads in the base network scenario. All components of the disconnected links, including the switch links, have a zero value for travel time, except the two segments that correspond with the two directions. These two segments inherit all of the specifications of the original direction such as length, travel time and the number of lanes, etc. as denoted by t_{AB} and t_{BA} in Figure 5.5. Should a decision be made to construct a (two-way road) project (i.e. $y=1$),

the corresponding one-way switch link is then added to the network to reconnect the two-way road. The concept of a switch link can also be used to represent more complicated projects such as spaghetti interchanges.

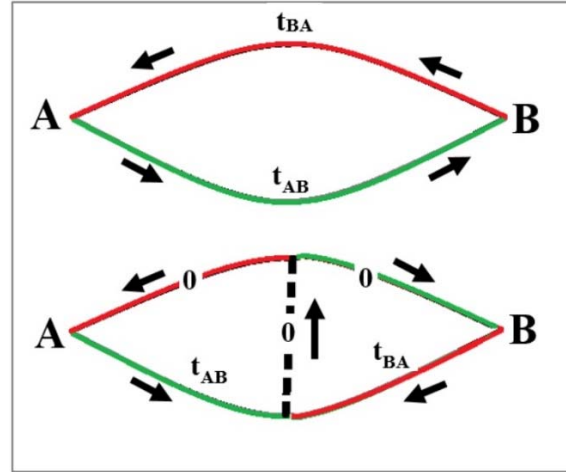


Figure 5.5, A one-way switch link (dashed line) representing a two-way road (red and green lines)

5.5.1 Example 1: Gao's network

Figure 5.6 illustrates the example network developed by Gao et al. (2005) with one OD pair (1,12) and the travel demand of $q_{1,12} = 20$. The delay function is $t_a = \bar{t}_a + .008x_a^4$, and it can be rearranged as per the BPR format: $t_a = \bar{t}_a(1 + .15(x_a/w_a)^4)$ where the links capacities are $w_a = \sqrt[4]{.15\bar{t}_a/.008}$. There are six candidate (one-way) roads with a total cost of 70. Gao et al. (2005) developed and applied the generalized Benders decomposition (GBD) to various budget levels and the results are summarized in Table 5.1.

Table 5.1 shows total number of iterations and the iteration at which the optimal solution was found in Gao's GBD method as well as the proposed BB-B algorithm. As can be seen across all budget levels, the BB-B demonstrates significantly superior performance over Gao's GBD. Furthermore, the BB-B on two avenues of the objective functions (with/without budget consumption term) showed close results.

Since the proposed algorithm (as proven before) guarantees the optimum solutions, the merit index is devised to accelerate the algorithm. For instance as Table 5.1 suggests, in terms of the number of iterations, Gao's GBD (which lacks any merit index) lags behind in all budget levels.

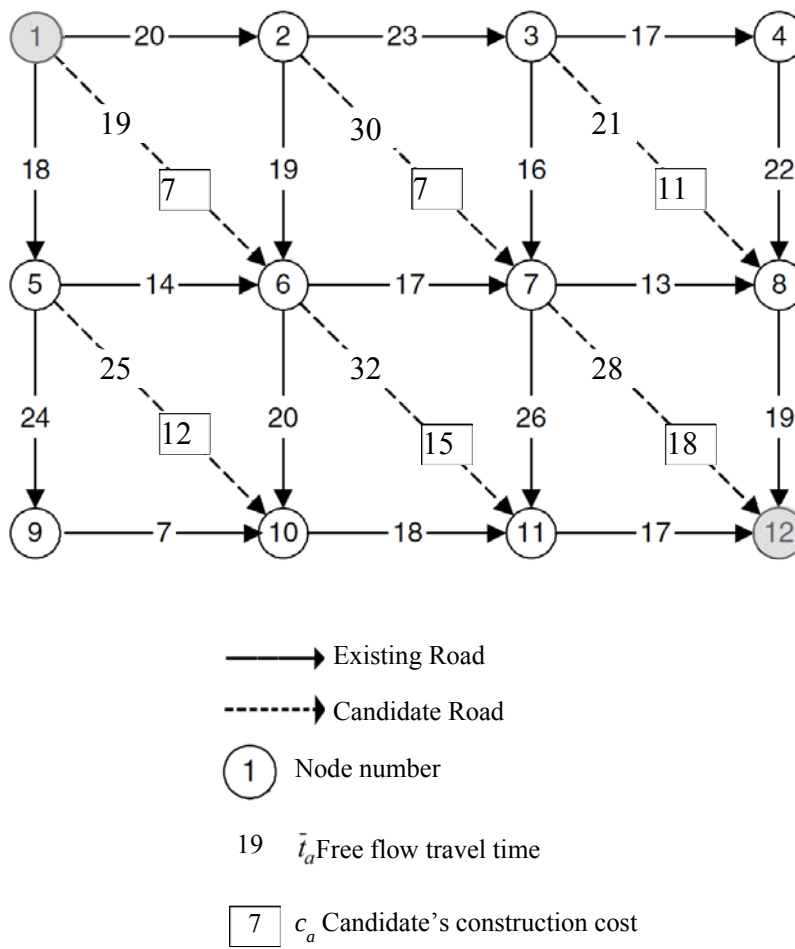


Figure 5.6 Gao's test network

Table 5.1 Example 1, Gao's Network: GBD (Gao et al., 2005) versus proposed BB-B

				GBD method ³ :	Proposed; BB_B method ⁵					
				Total iteration (Optimum solution was found at iteration)	With budget consumption term:			Without budget consumption term:		
Budget ¹	Optimal ² solution	Number of feasible solutions	Incumbent Value		No. of UE solved ⁶	No. of Benders (lower bound) solved ⁶	Benders iteration at which optimum solution was found	No. of UE solved ⁶	No. of Benders (lower bound) solved ⁶	Total iteration ⁶ (Optimum solution was found at iteration)
10	100000	3	4076	3 (2)	3	2	0 ⁴	3	2	0 ⁴
20	101000	12	3952	6 (4)	3	5	0 ⁴	3	5	0 ⁴
30	100001	26	2668	7 (6)	4	3	2	4	3	2
40	100101	41	2524	9 (4)	4	5	2	4	5	2
50	101101	52	2404	10 (4)	4	6	5	4	6	3
60	101111	61	2281	8 (5)	4	5	2	4	6	5
70	111111	64	2256	5 (5)	3	1	0 ⁴	3	1	0 ⁴

¹Total construction costs

²The digits in the binary strings represents the following one-way candidates respectively: (1,6), (5,10), (2,7), (6,11), (3,8), (7,12)

³ (Gao et al., 2005)

⁴Iteration zero refers to the intuitive (or initial) solution, the sorted projects as per the merit index is: (1,6), (2,7), (7,12), (5,10), (3,8), (6,11)

⁵epsilon = 2%, alpha = 0

⁶No. of UE solved: number of times at which the traffic assignment is solved, no. of Benders (lower bound) solved: Number of times at which a pair of nonlinear Primal (problem (5.19)) and mixed integer relaxed problem (problem (5.20) to (5.23)) are solved. The runtimes over various budget levels are less than a minute. Since computational technologies substantially vary over time, there is no point in reporting them. Instead the number of iterations broken down in number of times that Benders were solved plus number of times that a traffic assignment was solved is reported. The former is comparable with the number of iterations reported by Gao, since in both, two primal and relaxed problems are alternatively solved. So "total iteration" in the B&B_B refers to number of times Benders is solved.

5.5.2 Example 2, Sioux-Falls network

The Sioux-Falls dataset was first introduced by (LeBlanc, 1975), and a slightly modified version was recently used by Farvaresh and Sepehri (2013) employing a branch-and-bound and outer approximation (BB-OA) method. In this section, the application results of our study with the results reported by Farvaresh and Sepehri (2013) on the same Sioux-Falls network are compared. There are five two-way candidate roads with a total cost of 4,325. In a similar fashion, Table 5.2 presents the comparative results. As is evident from Table 5.2, in this case the proposed BB-B surpasses the BB-OA.

It was in only one out of three budget levels (pertaining to the "without the budget consumption") in which the BB_B was found slightly lagging behind the BB-OA. Whereas in the presence of the budget consumption term (which is the preferable method), our proposed algorithm (the BB_B) was by far leading the BB-OA.

Nevertheless, the spirit of the literature suggests that it is unlikely to arrive at an absolute and superior algorithm. The important point is that for the first time, the proposed algorithm is purposely tailored to large sized networks. So when it is applied to small and artificial benchmark networks, the proposed algorithm has also shown a relative superiority.

Table 5.2, Example 2;Sioux-Falls: BB-OA (Farvaresh and Sepehri, 2013) vs proposed BB-B

Budget ¹	Optimal solution ²	Number of feasible solutions	Incumbent Value	B&B-OA method: Optimum solution found at iteration ³	Proposed; B&B_B method ⁵	
					Without budget consumption term: Optimum solution found at iteration	With budget consumption term: Optimum solution found at iteration ⁴
2000	00101	14	158.4158	2	1	1
3000	00111	23	113.2047	3	6	1
4000	10111	32	94.1993	6	4	2

¹Total construction costs is 4325

² the digits in the binary strings represents the following two-ways candidates respectively: (6,8), (7,8), (9,10), (10,16), (13,24)

³(Farvaresh and Sepehri, 2013)

⁴ the sorted list of the candidate projects as per the merit index for the intuitive solution is (9,10), (6,8), (13,24), (7,8), (10,16)

⁵ epsilon = 2%, alpha = 0

5.5.3 Example 3: Winnipeg large-scale network

Real-size transportation data for the city of Winnipeg, Canada is widely used in the literature (Bagloee and Asadi, 2015; Bar-Gera, 2016; Ryu et al., 2015) and is undertaken for numerical tests considering multimodal and multiclass traffic assignment. This dataset has also been provided in EMME 3 (INRO, 2009). The road network is comprised of 154 zones, 943 nodes and 3075 directional links. The transit system consists of 2 transit vehicle types, 133 transit lines and 4345 transit line segments.

As with the multiclass aspect of traffic flow, in addition to the inclusion of different types of vehicles (trucks, cars, etc.), the bias term can be applied to many other real life applications such as traffic restrictions, high occupancy lanes (HOV) and toll gates, etc. (INRO, 2009). For instance, in the case of $b_a^m = \infty$, the respective user class m is prevented from entering the district denoted by link a . Accordingly, the same dataset that was used for the single-class Winnipeg case study is also used (but split between) the central business district (CBD) and non-CBD which resulted in two travel demand

matrices. The CBD matrix which accounts for 19,742 vehicle trips that can use all roads including roads in the CBD. The non-CBD commuters (36,476 vehicle trips) are prohibited in the CBD. The scale and location of the CBD is shown in Figure 5.7 (note that the roads are annotated with their respective hourly traffic volumes). The transit demand contains 18,211 passenger trips. In the existing scenario, the average travel time of the non-CBD commuters is 13.62 minutes while it is 17.10 minutes for the CBD bound commuters (CBD bound trips are the ones that have at least either their origin or their destination falling inside the CBD). As such, the average in-vehicle time experienced by passengers is 22 minutes.



Figure 5.7 Winnipeg's central business district,

Twenty two-way road projects with speeds of 50 km/h (or an equivalent free flow time of $60 \times \text{length} / 50$ minutes) and a capacity of 1700 vehicle per hour per direction (vphpd) will be considered in this case study.

Table 5.3 presents the candidate projects sorted and based on their merit index in descending order. Figure 5.8 shows the location of the candidate projects and the extent of the undertaken case study on which the MMC UE traffic volumes are depicted (all projects, irrespective of any budget constraints, are included). These projects are wisely set forth, to complement the ring roads around the CBD and over the river passing through the city. The length of roads will be considered as construction costs which amount to the total monetary cost of $C = 23.31$. With respect to the total cost (C), five levels of budget (B) are taken into account as follows: $B/C = 20\%$, 40% , 60% , 80% and 100% .

Table 5.3 Winnipeg example, two-way (candidate) road projects sorted based on the merit index

Id	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
I-node	887	595	513	420	1035	437	774	1057	551	325	304	168	177	829	335	288	299	424	330	441
J-node	889	602	595	592	301	424	739	297	610	330	423	784	853	173	449	294	1058	327	428	494
Cost ¹	0.44	0.59	0.79	0.58	0.75	0.86	0.6	0.88	1.51	1.3	1.29	1.09	1.52	1.24	0.64	2.5	1.35	1.61	1.73	2.04
Traffic volume ²	1634	1554	1554	822	1059	1138	702	1025	1688	1447	1235	941	950	753	377	1338	413	469	164	165
Merit Index ³	3713	2634	1967	1417	1411	1323	1170	1164	1118	1113	957	863	625	607	589	535	306	291	95	81

¹Total Cost: 23.31

² It is the volume accumulated on the corresponding switch link, hence it is the sum of traffic volumes on both directions

³ Provided that the capacity of the projects are same (1700 vphpd) the merit index is simply calculated as: traffic volume/Cost

The MMC-UE traffic assignment is solved with a relative gap of 0.001% where each traffic assignment lasts approximately 3 seconds.

In the previous two examples, the budget consumption term was a contributing factor in reaching a faster global solution. Accordingly, in the Winnipeg case, the analysis was carried out using the budget consumption terms. Consequently, the numerical results of the BB-B pertaining to the objective functions equipped with the budget consumption terms are reported in Table 5.4.

In Table 5.4, global optimal solutions are first introduced over various budget levels. As can be seen, the global optimal solutions have used up almost all of the budgets, such that the budget consumption varies from 89% to 99%. As a result, it is an endorsement of the introduction of the budget consumption term in the objective function. In other words, encouraging the algorithm to use up the full capacity of the available

budget results in and contributes to global solutions being arrived at quicker. A closer look at the binary string of the global solutions over the incremental budget levels suggests that the strings incline so that they are filled from left to the right. This observation underscores the validity of the merit index in an attempt to enhance the likelihood of finding the global solution.

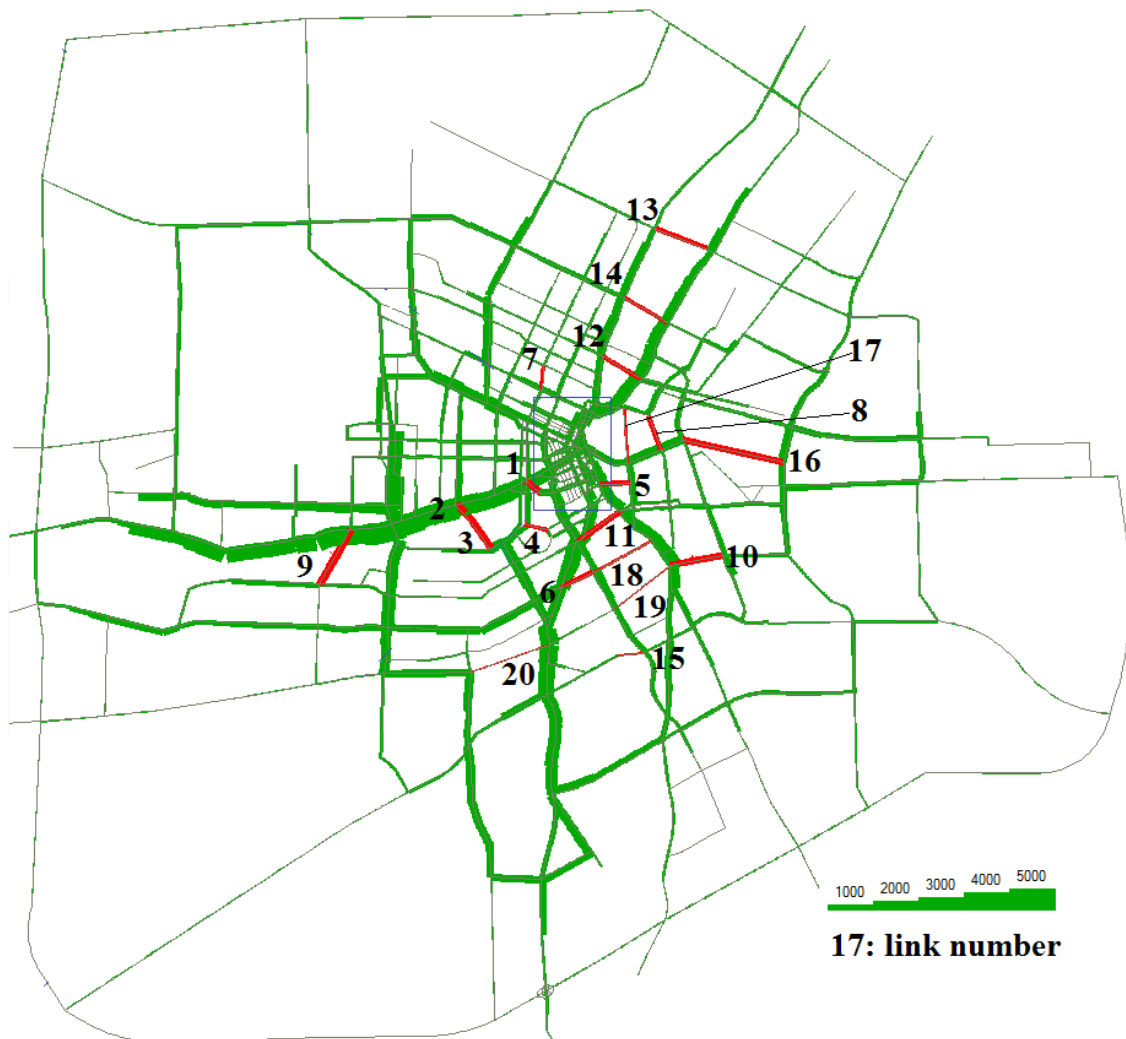


Figure 5.8 Winnipeg example, with 20 road projects.

According to Table 5.4, for each budget level starts with $\alpha=0$. In two out of the five budget levels, the global optimal solutions were found. In the remainder, good solutions ranked 4th, 6th and 3rd were achieved. However, consideration of a meagre value for the alpha (i.e. $\alpha=0.001$) secured global solutions for budget levels 40% and 100%. It was only for budget level 80% where $\alpha=0.001$ enhanced the quality of the solution from 6th to 4th. In addition, a slight push on the alpha from $\alpha=0.001$ to $\alpha=0.005$ resulted in the global optimal solution.

Table 5.4 also indicates the number of times and the depth at which fathoming occurred, both of which are a key efficacy barometer. The 11th column (“Total budget consumption before fathoming”) shows the average depth of the fathoming point as a percentage of the budget. Less depth is more desirable since it suggests that fathoming has occurred at the beginning of the tree structure, hence a big portion of the solution space has been discarded.

The computational (CPU) time has also been reported, such that in the worst case ($B/C = 40\%$, $\alpha = 0.001$), the computation lasts for almost 6.6 hours. The last column in the table represents the percentage of the total CPU time at which the best solutions are found, and this can vary from very early (2%) to almost half way through (62%). In six out of nine budget levels, optimal solutions were found in the first half of the CPU time. These percentages show that there is a high probability of reaching the best solution in early iterations.

Furthermore, the number of attempts to solve the Benders problem (the MMMC-SO-TAP with Lagrangian value of the capacity constraints as well as the MILP) and the MMMC-UE-TAP are also reported in the table. Such a breakdown in detail provides a better understanding of the sheer size of the computational burden over different sized networks. For instance, Xie and Xie (2015) have recently reported on CPU times for solving various network sizes which varies from a few seconds to a few minutes for super large-sized networks (with +1,770 zones). In dealing with large-size networks, instead of a couple of hours (which was the case for the Winnipeg network), one may need to wait a couple of days to successfully terminate the algorithm. It is worth noting that for strategic planning decisions such as network design which involves significant investments, one can afford CPU time in the scale of hours or even days.

Table 5.4 Winnipeg case study: results of the proposed BB-B

Outputs of the proposed BB-B																	
Global Optimal Solution									Best solution Found								
B/C%	Budget	Number of feasible solutions	binary string	Cost ¹	Budget consumption (%)	Obj Fnc	alpha	No. of fathomed	Total budget consumption before fathoming	Average depth of fathoming: percentage of budget consumption ²	rank	binary string	Obj Fnc	no of UE solved ²	no of Benders (lower bound) solved ³	CPU time (hr)	Time to reach at best solution (CPU%) ⁴
20	4.662	6381	01111000000001100000	4.59	98.46	854011	0.000	2	0.88	10	1	01111000000001100000	854011	4	160	1.04	62
40	9.324	222664	01110001101011100000	9.04	96.95	847252	0.000	2	0.88	5	4	'01110011001111100000	847461	4	544	3.85	22
							0.001	15	53.23	39	1	01110001101011100000	847252	18	975	6.59	23
60	13.986	825912	01111011101111110000	14	99.96	842523	0.000	2	0.88	3	1	01111011101111110000	842523	4	273	1.79	55
80	18.648	1042195	111111111111110100	18.2	97.54	840350	0.000	2	0.88	2	6	111111111111111000	840556	3	56	0.37	2
							0.001	39	323.72	46	4	0111111101111111100	840515	43	310	2.07	26
							0.005	33	231.09	38	1	111111111111110100	840350	38	162	1.11	2
100	23.31	1048576	011111111111111110	20.8	89.36	839462	0.000	2	0.88	2	3	'111111111111111111	839617	3	3	0.02	33
							0.001	37	317.49	41	1	011111111111111110	839462	41	55	0.43	53

¹ unit of the cost is assumed to be equivalent to the length of the respective road.

² it is computed as follows: "total budget consumption before fathoming"/"no of fathomed"/"Budget"*100. For example: 39% = 53.23/15/9.324.

³ no. of UE solved: number of times at which the traffic assignment is solved, no. of Benders (lower bound) solved: Number of times at which a pair of nonlinear Primal (problem (5.19)) and mixed integer relaxed problem (problem (5.20) to (5.23)) are solved.

⁴ it is the time it took to find an optimal solution, So the remaining time was used to verify that it was the optimal solution.

5.6 Conclusions

This study has approached the discrete network design problem (DNDP) via an exact algorithm tailored to the need, scale and nature of the problems dealt with in the industry. Although the DNDP has been extensively studied in academia, practitioners have yet to find anything usable in it. There is a two-fold reason for this: (i) the size of the realistic problems and the combinatorial nature of the DNDP are significant prohibitive factors; and (ii) in spite of the existence of efficient heuristic (but not exact) methods, their outcomes are treated sceptically mainly due to their non-deterministic nature.

The ongoing growth of transportation infrastructure in Asia (Estache et al., 2013) on the one hand and the enhancement in computational power as well state-of-the-art optimisation techniques on the other hand, have raised an interest in filling the gap between academia and industry via exact (not heuristic) methods. In spite of the fact that the DNDP is an NP-hard problem, the size of the problems being dealt with in the industry are intractable (say a dozen candidate projects, for more information refer to (Farahani et al., 2013), see Table 7 and listed papers therein). All together, these make addressing the DNDP on realistic scales a worthwhile endeavour. To this end, the BB-B method that is based on branch and bound (BB), hybridized with Benders decomposition to streamline achieving a global optimum solution is proposed.

Using a set of greedy rules, the tree of the BB was built upon the sorted projects on a merit basis, aiming to reaching the global optimum solution. To hedge against the Braess Paradox, the total travel time of the system optimal (SO) flow was computed as a lower bound at each new node on the tree (LeBlanc, 1975). To further reduce the computation time, a Benders Method was devised to obtain a tight lower bound. The algorithm proposed was evaluated numerically using a real data set from the City of Winnipeg, with 20 candidate projects subject to various budget levels. In the context of the past studies, a comparative analysis was also provided. Accordingly, the BB-B method developed showed superior performance in terms of computational efficiency.

The algorithm could be further improved by considering asymmetric delay functions to enhance the degree of realism of the traffic flow. Furthermore, consideration of changes in travel demand in response to changes in the network side (known as demand elasticity) is a worthwhile thread for research in the context of the NDP. Furthermore, the conventional wisdom is to represent congestion using delay functions and the Beckmann

formulation. There exist a wide area that can improve the congestion representations that is to consider the dynamic changes of the congestion. Dynamic traffic assignment (DTA) is an evolving subject that takes the above concerns into consideration. As a result, integration of the DNDP and DTA can also be of interest for future studies.

The NDP is based on a solid premise that the candidate projects are already defined. Loosening such a premise and arriving at a more sophisticated way to find candidate projects in conjunction with the NDP is worthy of further research.

Furthermore, it is sometimes necessary to arrive not only at the best (global) optimal solution (1st), but at other good solutions belonging to the top of the list (say 2nd, 3rd, 4th, etc.). Given many different stakeholders and vested interests in transport infrastructure, providing decision makers with a variety of top-performing solutions is appealing in the industry. In some cases, the difference between the 1st- and 2nd- best solutions in terms of the objective function is only mathematically marginal, but the 2nd best solution may have some other advantages which cannot be quantified in the objective function. Similar to the terminology used in the optimisation literature (k-shortest paths), here this new problem is called K-NDP, where the intention is to find the first “k” best solutions.

Extending the discussion to selecting an optimum configuration of one-way and two-way roads is also a worthwhile line of research (Drezner and Wesolowsky, 1997, 2003). Concern arose out of the vested interests involved in the road network and this may bring about a variety of objective functions in the DNDP including environmental costs (Szeto et al., 2014) and emission reduction (Ferguson et al., 2012). As such, the problem becomes a multi-objective DNDP (Xie, 2014). Changes to the network in the short and long run may affect the way commuters choose their mode of transport, their destination and departure time (if any). Hence, the DNDP subject to a combined traffic model considering these changes deserves more investigation (Boyce and Janson, 1980; Boyce and Soberanes, 1979). Similarly investigation of land use changes as the result of changes in the network infrastructure can also be noted (Szeto et al., 2010; Szeto et al., 2013). Although the consensus in the literature is to ignore the uncertainty of the travel demand, one may want to explore such uncertainties involved in the DNDP (Ukkusuri et al., 2007).

6 TRANSIT PRIORITY LANES

As discussed earlier, public and mass transport modes deserve priority in the transport system in general as well as more space. The aim is to address optimal reallocation of road space to transit modes on an existing urban transport network. In particular, this study is interested in finding a network of exclusive transit priority lanes in the heart of cities in which the congestion is a chronic stigma. In this chapter, the question of interest is: which roads can be nominated to give an exclusive lane to transit modes? Taking space away from private modes in favour of public transport may adversely affect congestion levels. To this end, inspired by the Braess Paradox, mis-utilized space used by private modes mainly on congested roads is sought to be dedicated to transit modes. To find such candidate roads, a merit index based on transit ridership and congestion level is first defined. Then based on the merit index a number of roads to be designated as transit priority lanes are selected. This problem is formulated as a bi-level mixed-integer, nonlinear programming problem in which the decision variables are binary (1: to cause the respective road to have an exclusive transit lane or 0: not). The adverse effects are minimised on the upper level represented by total travel time (public and private modes) spent on the network. The lower level accounts for a multimodal traffic assignment, to consider the impact of transit priority on private modes, an efficient low-RAM-intensity branch and bound as a solution algorithm has been developed. The search for the subset is made in such a way that improved public transport is achieved at zero cost to the overall performance of the network.

In the next section, first an introduction followed by extensive discussion on the mathematical underlying of the problem is provided in section 6.2. The methodology is elaborated in Section 6.3. Section 6.4 is dedicated to the numerical evaluations followed by the conclusion in Section 6.5.

6.1 Introduction

A typical road can be considered as a directional link consisting of one or more lanes. In general, roads and lanes can be used by all traffic modes (cars and buses). If a road is set to be a transit priority lane, at least one of its lanes is dedicated for exclusive use by buses. Turning a lane into a transit priority lane is also associated with expenses derived from

road marking, special signage/signals, lighting, etc. Hence, the transit priority lanes design problem (TPLDP) can be discussed as follows: Given a set of candidate roads, which lanes should be selected as priority lanes while accounting for a limited budget? Taking space away from private modes in favour of public transport is a delicate task, which may adversely lead to gridlock congestion. Therefore, one needs to minimise such adverse impacts. One intuitive way is to minimise the total travel time spent in the network which is set as the objective function in the TPLDP. The consensus in the literature is to model the TPLDP as a bi-level mixed-integer, nonlinear optimisation problem. The objective function is placed in the upper level, while the lower level accounts for the user equilibrium (UE) private traffic flow as well as the transit flow. It has been proven that any bi-level programming problem is NP-hard so that the problem becomes quickly intractable as the size of the problem increases (Ben-Ayed and Blair, 1990).

Among a variety of methods available in the literature, some try to reach an optimum solution but cannot scale to handle large-sized networks, others aim to address large-sized networks at the cost of compromising the quality of the solution.

To this end, a RAM-efficient branch and bound (BB) method to address the TPLDP tailored to large-sized networks has been developed. First, a set of roads with significant transit volumes is identified as a candidate set. Second, using a BB method, the possibility of selecting a subset of the candidate roads is investigated, such that the overall performance of the traffic system, including private and transit flows, is not negatively impacted. It is even possible to improve the overall performance of the transport network, as is well illustrated by the Braess paradox (BP). The BP refers to the fact that adding a new capacity to a transportation network might adversely degrade the traffic circulation. Empirical evidence, as well as mathematical theories, have shown that the presence of the BP is prevalent in real transport networks (Nagurney, 2010). That is, if there exists some Braess-tainted roads, their closure would improve traffic circulation. The main idea of this study (for the TPLDP) is instead of completely closing Braess-tainted roads, to convert them to transit priority lanes. It is a utilitarian approach to take advantage of this stigma.

Further, the TPLDP while considering the multimodal feature of the traffic flow which enhances the authenticity of the model has been solved. The real dataset for Winnipeg, Canada which is readily available in the literature as a benchmark is used to

demonstrate the numerical impact of the solutions calculated by the method developed here.

It is important to note that the structure of a transit system such as stop positions, transit routes, and fleet size remains intact in this demonstration. Changes may occur with respect to some segments of the existing transit routes currently sharing road space (lanes) with private modes, which may come to be dedicated as exclusive lanes. Hence, the challenge is to find these segments without detriment to traffic circulation.

This research contributes to the literature on three fronts: (i) A network-wide approach to the TPLDP tailored to large-sized networks of congested roads is developed. (ii) BP is utilized to nullify the adverse effect of transit priority lanes on the private mode by providing faster public and even private transport. (iii) A RAM-efficient BB algorithm tailored to a multimodal traffic model so that its simple structure can easily be embedded in any programming language.

6.2 The transit priority lanes design problem (TPLDP)

In this section, a set the mathematical definitions of the TPLDP is presented. Then, the way the multimodal aspect of the traffic flow is included is elaborated upon.

For ease of formulation, the following convention was adopted: roads considered as a candidate are denoted by ℓ (with, for instance, three lanes). Suppose that it is replaced with two new roads $\ell' \in \bar{A}$ and $\ell'' \in A$: (i) road $\ell' \in \bar{A}$ with only one lane which is to be either a mixed mode road or an exclusive transit lane or road and (ii) road $\ell'' \in A$ with two lanes for mixed mode use. Alternatively these are referred to as transit lanes or transit roads (and, by doing so, they can alternatively be called transit priority lanes or transit roads). Therefore, having: $\bar{A}$: a set of roads currently with mixed modes (transit and private modes) but considered as candidates for exclusive use by transit modes, and the rest of the roads are denoted by A (the candidate road henceforth is simply called “candidate”).

Although the exposition of the BP in the literature supports a complete closure of the BP-tainted roads, the road kept open to private cars for two reasons: (i) connectivity: in order to preserve the connectivity of the network, the roads that are closed to create space for buses must have at least two lanes; in the event they become nominated to give away one lane as a transit priority lane, they still will have at least one lane remaining (ii) optimality: it is proven mathematically that a partial closure (like closing a lane for the transit mode) is more likely to result in a better traffic circulation (this concept is highly

exploited in the congestion pricing, (Yang and Huang, 2005). The basic tenet of congestion pricing is to redistribute the traffic load evenly over the road network by enforcing a “toll” instead of by any physical restriction. It is important to note that any changes to the network or the travel demand such as adding a priority lane may change the BP’s status (Aashtiani and Poorzahedy, 2004; Nagurney, 2010). Nevertheless, in the formulation of the problem (i.e., in the objective function) these changes result in a better traffic circulation across the network. The following notation is used:

N : set of nodes,

B : budget available to cover the costs of transit lane implementations such as marking, pavement, curb raising, etc.

y_a : binary decision variable associated with candidate $a \in \bar{A}$; 1: to be used as exclusive transit lane and 0: to remain mixed use road or lane,

c_a : implementation cost associated with candidate $a \in \bar{A}$.

$\bar{x}_a$: hourly public passenger volume on road $a \in A \cup \bar{A}$,

$x_a, \bar{x}_a$: hourly private and transit traffic flow in hourly passenger car equivalent or unit (“PCE” or “PCU”) on road $a \in A \cup \bar{A}$ respectively, (Note, (i) the network available to the private and transit roads/lanes are A and $A \cup \bar{A}$ respectively, hence $x_a, \bar{x}_a \geq 0$ for $a \in A$ and $x_a = 0, \bar{x}_a \geq 0$ for $a \in \bar{A}$, (ii) $\bar{x}_a$ is the hourly volume of passenger traffic on the road while $\bar{\bar{x}}_a$ is the car equivalent value of the corresponding number of buses on the respective road $a \in A \cup \bar{A}$, (iii) the PCE reflects the physical and operational characteristics of the buses that can vary from 1.2 to 4.5. In traffic models, these values have already been assigned to the transit fleet. For instance for the Winnipeg traffic model the PCE is either 2 or 2.5 depending on the type of bus).

$t_a(x_a + \bar{\bar{x}}_a)$: general travel time of link $a \in A \cup \bar{A}$, a non-decreasing BPR function of the flow $x_a + \bar{\bar{x}}_a$ of the traffic (called the delay function (Sheffi, 1985; Spiess, 1990)). Note that the background traffic, $\bar{\bar{x}}_a$, is a fixed value. In addition, switching the delay functions between with/without priority lanes is technically a trivial task in our proposed methodology. To this end, generally speaking, the new delay function must be calibrated based on field survey data. Nevertheless, the BPR delay functions of free flow speed and capacity are used. In this formulation, as the number of lanes change, the capacities are updated accordingly. However, this keeps the free flow speed intact for the following reason. The priority lanes are sought among the congested roads (there is no point to give priority to

the mass transit in the uncongested roads). Therefore it is conceivable that a congested road (which is not governed by the free flow speed) after giving away a lane is still congested with delays remaining more or less same.

A_n^-, A_n^+ : set of links starting and ending at node n respectively, $A_n^-, A_n^+ \subset A \cup \bar{A}$,

x_a : hourly traffic volume in auto or private mode,

R : set of OD pairs $R \subset N^2$,

q_r : hourly travel demand in PCU for OD pair $r \in R$ pertaining to the auto mode.

g_{ij} : hourly transit passenger demand from node i to destination node j . In order to simplify the notation, let us define $g_j = -\sum_{i \in N - \{j\}} g_{ij}$ that is the total trip attraction to node j , see (Spiess and Florian, 1989),

P_r : set of paths between OD pair $r \in R$ available to the auto mode,

h_k : hourly traffic flow on paths $k \in P_r$, pertaining to the auto mode,

$\delta_{a,k}$: road-path incident index, 1 if road $a \in A \cup \bar{A}$ belongs to path $k \in P_r$ pertaining to the auto mode, and 0 otherwise

w_n : average waiting time at node $n \in N$ pertaining to transit system,

f_a : sum of frequency of service for all transit lines on roads $a \in A \cup \bar{A}$.

The bi-level TPLDP may be written as follows (note, all variables and parameters are considered non-negative unless otherwise stated):

$$\min \sum_{a \in A \cup \bar{A}} (x_a + \bar{x}_a) \cdot t_a(x_a + \bar{x}_a) \quad (6.1)$$

$$\text{s.t.} \quad y_a = 1 \text{ or } 0, \quad a \in \bar{A} \quad (6.2)$$

$$\sum_{a \in \bar{A}} c_a \cdot y_a \leq B \quad (6.3)$$

$$\min \sum_{a \in A \cup \bar{A}} \int_0^{x_a} t_a(x_a + \bar{x}_a) dx \quad (6.4)$$

$$\text{s.t.} \quad \sum_{k \in P_r} h_k = q_r, \quad r \in R \quad (6.5)$$

$$x_a = \sum_{r \in R} \sum_{k \in P_r} h_k \delta_{a,k} \quad \delta_{a,k} = \begin{cases} 1 & a \in k \\ 0 & a \notin k \end{cases}, \quad a \in A \cup \bar{A} \quad (6.6)$$

$$\bar{x}_a \leq U \quad a \in \bar{A} \quad (6.7)$$

$$x_a \leq (1 - y_a) \cdot U \quad a \in \bar{A} \quad (6.8)$$

$$\left\{ \begin{array}{l} \bar{x}_a \in \arg \min_{a \in A \cup \bar{A}} \sum \bar{x}_a \cdot \bar{t}_a + \sum_{n \in N} w_n, \\ \text{subject to} \\ \sum_{a \in A_i^+} \bar{x}_a - \sum_{a \in A_i^-} \bar{x}_a = g_i \quad i \in N \\ \bar{x}_a \leq f_a \cdot w_n, \quad a \in A_n^+, n \in N, \\ \bar{x}_a \geq 0, \quad a \in A \cup \bar{A} \end{array} \right\} \quad (6.9)$$

The objective function (6.1) describes the upper-level goal of minimising the total travel time. Constraints (6.2) and (6.3) ensure the feasibility of the candidates with respect to the costs and the available budget. At the lower level ((6.4), (6.5), (6.6)), the Beckmann formulation of the UE flow pertaining to the private mode is computed. Constraints (6.2), (6.7) and (6.8) ensure that private flow will not enter the dedicated transit lanes. (U is a sufficiently large value, say the total demand $\sum_r q_r$). Although constraint (6.7) is redundant, it is placed within the constraints to emphasize that buses can use candidate roads either exclusively (if it turns out to be $y_a = 1$) or mixed with private mode (i.e., $y_a = 0$). If it is decided that candidate a is to be an exclusive transit lane/road (i.e., $y_a = 1$), then constraint (6.8) ensures the respective road will be closed to the private mode (i.e., $1 - y_a = 0$). Sub-problem (6.9) carries out transit assignment based on optimal strategy (Spiess and Florian, 1989) and it returns $\bar{x}_a$ as passenger traffic volume per hour. The sub-problem also returns the effective frequency of the transit lanes (or the number of buses) on the roads (note that the roads are also associated with transit delay functions which are functions of travel times experienced by the auto mode). The equivalent value of buses in PCU (denoted by $\bar{x}_a$) is then considered as background traffic in the traffic assignment (Spiess, 1984).

At the lower level, a combination of traffic and transit assignment (multimodal) theoretically leads to a nonconvex programming problem. Such problems then require some computationally expensive methods such as variational inequality, not to mention some unresolved issues such as uniqueness and stability of the solutions (Florian and Morosan, 2014). The relevant studies either fall short of fully considering the simultaneous interaction between private and public modes or suffer from lengthy computation time (De Cea et al., 2005; Liu and Meng, 2012). Given these complexities, the above formulation ((6.4)...(6.9)) is proven to be able to solve the multimodal traffic assignment adequately, so it has been adopted in some commercial planning applications (INRO,

2009). In this study, the formulation (6.4)...(6.9) has been coded as a module in EMME 3 (INRO, 2009) and it is called upon by the BB algorithm whenever needed.

6.3 Methodology

In this section, integration of the merits index into the BB algorithm is discussed. The BB has been elaborated on the previous chapters. To offer a self-contained discussion in this chapter, however, some basic tenets of the BB are also presented.

6.3.1 Branch and bound in the context of optimisation methods

The most notable method of enumeration for the mixed integer programming problems is BB which uses a tree structure to process all the combinations. In the minimisation problems, as the tree expands, a lower bound is calculated at each node and the branching is frozen (fathomed) wherever the lower bounds are found greater than the best-found solution (which is also an upper bound value and is called incumbent value). It is evident that as the size of the problem (number of decision variables) increases, the method becomes computationally prohibitive. The special BB developed in this study can be easily coded in any application.

LeBlanc (1975) proposed a BB method for the DNDP, but due to the computational technology available at the time, it was considered inefficient. In this study, it is attempted to customise the structure of the BB to the TPLDP finely in order to achieve a more efficient algorithm. As the result, a new method for constructing the tree structure at node selection and branching based on the concept of merit index (Bagloee et al., 2016b) is proposed.

The ways the structure of the tree is formed, as well as the presence of the merit index, have a significant impact on the efficiency of the BB algorithm; it results in a less RAM intensive and a memoryless algorithm. In the next section, It is discussed how to initialize the BB based on the merit index and how to arrive at a tighter lower bound value to serve the purpose of making the BB as efficient as possible. A detailed discussion on the lower bound values, node selection and branching rules, as well as the termination conditions, have already been discussed in Chapter 5.

6.3.2 Merit index to find candidate roads

The first stage is to come up with a set of candidate roads for transit priority lanes. Looking for transit priority lanes in a suburb or uncongested roads has no point.

Accordingly, in this study, the challenging task of laying down the transit priority lane network in the congested parts of the urban road network (namely in downtown areas or CBDs) is addressed. Based on the concept of the BP, it is endeavoured to look for some roads, though congested to take away one lane for public transport, without worsening current congestion. In doing so, a merit index is defined based on which the roads are sorted in descending order as follows:

$$\bar{A} = \left\{ \frac{vc_a \cdot \frac{\bar{x}_a}{x_a} \cdot (x_a + \bar{x}_a)}{c_a} \quad \text{for } a \in A \text{ such that } vc_a > 0.85 \right\} \quad (6.10)$$

where $\bar{A}$ is the ordered list of the candidate roads (sorted in descending order), vc_a is the volume per capacity ratio of the link $a \in A \cup \bar{A}$. As noted before, it is evident that taking away space from already underutilized roads (i.e., low vc_a) to the transit bears no additional advantages. Hence, a threshold of $vc_a > 0.85$ is considered. This threshold is equivalent to level of service (LOS) E which is regarded as the “working at capacity condition” (HCM2010, 2010). This threshold or equivalently the LOS E is the approximate point at which the speed of the traffic suddenly drops (see (HCM2010, 2010): Exhibit 11-6 p. 11-8 and Exhibit 11-15, p. 11-20). As noted before, the aim is not to lay out a transit lane on uncongested roads (i.e., $vc_a < 0.85$). Moreover, in order to keep the connectivity level of the network intact, only roads with at least two lanes (per direction) are designated for conversion to a transit lane. Having said that, a two-lane road with $vc_a > 0.5$ could easily become over saturated with the implementation of a transit lane. It is important to reiterate that the mandate of priority planning is to give priority to mass transit vehicles such as buses even at the cost of more delay for private cars, to encourage them to shift to public transport. However, without any modal shift, according to BP, the example road could be found of no interest by private cars, that is, its traffic volume could come down to zero. In other words, these private cars may have found other shorter paths. This is the beauty of BP. Of course, there might be some cases in which some (uncongested) roads become highly congested, but, by minimising the objective function the overall performance of networks will not deteriorate.

According to condition (6.10), the more congested a road is, the greater the chance it has to be designated as a dedicated transit lane. To ensure a road with a high percentage of transit flow to likely receive transit priority, term $\bar{x}_a/x_a$ is added to the formulation

(6.10). Between two roads with the same volume-capacity ratios, the one that carries more traffic is more likely to be designated as a dedicated transit lane. That is why the term $x_a + \bar{x}_a$ is also added. Nonetheless, the merits of the roads are normalised by their respective costs. This is a greedy way to push more cost efficient roads to the top of the merit list (the list in descending order). The numerical result shows that the above index is effective, such that the projects selected in the final optimum solutions are among the top ones in the sorted list.

Transit lanes are just like auto lanes with passenger car equivalency (PCE) vehicles running in them. Nevertheless, the BP could still occur when the lanes are taken off existing links to be used as bus lanes. Note that the whole link is not removed, but just some lanes are removed from the private modes and are put in use in another form (to be used by the public transport modes). To address such concerns, it is worth noting that the aim of this approach is not to eradicate Braess paradox, rather the aim is to make the best out of the likely existing BP to “promote” and advocate public transport ridership. In other words, despite all efforts, the approach may end up reaching a situation in which a number of roads are designated to convert a lane to public transport while the BP still exists among them. However, improvement to the overall performance of the network compared to the existing situation may have already been achieved, even though BP may still exist.

Should Braess-tainted links be completely blocked? This subject deserves further investigation. A complete road closure is a very sensitive action (not from the traffic point of view, more from politically-vested interests, its societal consequences), implications for land use (business, outlets, shops along the respective road ought to suffer and resist) etc.

It is also important to note that, the proposed methodology provides a pro-public transport network (i.e. flagging some bus lanes throughout the road network) without compromising the integrity, connectedness and performance of the network compared to the do-nothing network. Any other good things, like finding BP over the rest of the network can be treated as boons which deserve more investigation. The reader interested to know more about BP detection is referred to (Bagloee et al., 2013a).

A similar concern may arise with respect to the way the candidate set (condition (6.23)) is derived. A better approach seems to be to detect BP automatically and remove these links from the network (even not to be used as bus lane). To this end, further to what

is discussed above, it is important to highlight two points. First, BP detection is an extremely difficult problem (Roughgarden and Tardos, 2002). Second, there is a practical advantage in the proposed methodology regarding condition (6.23), as follows. First it comes up with a set of candidate roads suspected to be Braess-tainted. The BB algorithm is then launched over this set to identify the best subset. This initial candidate set can also be altered, based on other concerns (for example, traffic authorities might be interested to practice a number of what-if scenarios). To this end, it is widely believed that transportation is largely driven by non-transportation vested interests. Furthermore, a complete road closure, as is the case in BP detection is highly controversial. That could jeopardise the whole point of promoting public transport.

6.3.3 A tight lower bound

For the mixed integer programming problem of the TPLDP, given the candidate set $\bar{A}$ (or the binary decision variables) the algorithm initiates from the existing (do nothing) scenario $z_j = (0,0,0,0,0)$ represented by the first node of the tree ($j = 0$). Each node in the tree represents either a partial or complete solution. For example, if there are five binary variables, solution $(0,1,0,2,2)$ represents the situation in which only the first three components are determined with values of 0/1 and the last two, represented by 2, are as yet unspecified; hence, it is a “partial solution”. Each time a node z is added to the tree, a lower bound based on the system optimal (SO) follows and the total travel time (objective function (6.1)) is evaluated (LeBlanc, 1975). Therefore, all the free binary variables “2” are set to “0” and the SO flow on the network is computed.

The SO flow for the respective network of z_j is computed and the total travel time corresponding to objective function (6.1) is set as a lower bound. The UE flow for the network of z_j is also computed and the corresponding total travel time is saved as the upper bound and is called the incumbent value. As the tree expands, the incumbent value takes the objective value (total travel time, the objective function (6.1)) of UE flow of the best solution found. In other words, the incumbent value is the minimum of the upper bounds.

One of the key factors contributing to the efficacy of the algorithm is rooted in how narrow the distance is between the lower bounds and the incumbent values. It is important to note that the lower bound and the incumbent values are calculated based on SO and UE flows respectively. As described in the previous chapter, the ratio of the travel

time of the UE flow to SO flow, called “price of anarchy,” can be as high as 2.15 (Roughgarden and Tardos, 2002). In this section, a recap of the heuristic procedure developed to relax the SO flow in order to bridge such a wide gap between SO and UE is presented.

The SO flow can be easily computed using commercial transport planning software by replacing the delay function in the objective function of the UE flow (objective function (6.4)) to (Sheffi, 1985):

$$\tilde{t}_a(x_a + \bar{x}_a) = t_a(x_a + \bar{x}_a) + x_a \cdot \frac{\partial t_a(x_a + \bar{x}_a)}{\partial x_a}, \quad (6.11)$$

where, if $t_a(x_a + \bar{x}_a)$ is considered as the cost of traveling on road $a \in A$, then $\tilde{t}_a(x_a + \bar{x}_a)$ is known as the marginal cost of using the respective road. As for the delay function, this is a non-decreasing multinomial BPR function. The wide gap between SO and UE emerges from the second term on the right-hand side of the equation (6.11) which is the additional externality costs imposed on the users. The two functions $t, \tilde{t}$ show similar behaviour as long as the volume is below capacity. As the volume reaches (or exceeds) the capacity, the externality costs increase rapidly, hence it results in a wide gap between the SO and UE flows. Since the capacity of the delay function is not restricted, in theory, the volume and hence the delay as well as the marginal cost can increase to infinity, which results in a much wider gap between SO and UE. In order to decrease such an unrealistic gap, alpha $0 \leq \alpha \leq 1$ as a coefficient in the externality term is proposed:

$$\tilde{t}_a(x_a + \bar{x}_a) = t_a(x_a + \bar{x}_a) + \alpha \cdot x_a \cdot \frac{\partial t_a(x_a + \bar{x}_a)}{\partial x_a} \quad (6.12)$$

As alpha reaches zero, the SO moves closer towards the UE and the gap vanishes. It is worth noting that the alpha addresses the unfortunate trade-off between computational time and the accuracy of the algorithm. With lower levels of alpha, the more quickly it leads to a less accurate solution. The value of alpha can be chosen at the user's discretion depending on the current computational technology and the affordability of the computational time. In other words, alpha is a valve in the hand of the modeller based on which, the accuracy of the results along with the computational time can be adjusted depending on the size of the network, available computational technology and the strategic value of the final solution. Though, alpha, in fact, simplifies the problem to an SO relaxation, the validity of the results especially for real-life networks are strongly upheld. Recent studies have shown that the difference the UE and SO traffic patterns

stand in close proximity such that for the case of the city of Chicago the difference does not exceed sex percentage across the entire network (Boyce and Xiong, 2004; Zheng and Boyce, 2011). In other words, the alpha reflects on the observation that, for the real life road networks, the SO traffic pattern stands close to the UE traffic pattern. As the result, alpha can be considered as an engineering way to take advantage of this revelation enabling real-sized networks to be addressed. In the next section, values of alpha for the case study undertaken are discussed.

6.4 Numerical demonstration

Real-size transportation data for Winnipeg, Canada which is widely available in the literature (Bar-Gera, 2016) was used to evaluate the proposed methodology. The case study is comprised of 154 zones, 903 nodes, 2995 directional links and 133 transit lines. Total hourly car and transit passenger demands are 56,219 and 18,211 respectively. As for the computational technology a desktop computer with Intel(R) Core(TM) 3.40 GHz and 16.0 GB RAM was used. The algorithm was written using Visual Basic linked with MS-Excel as an interface and MS-Access to handle the data efficiently. It is also synchronized to EMME3 to carry out bimodal traffic assignments.

Table 6.1 shows candidate roads sorted according to their merit indices (condition (6.10)), adding up to 15.27 km of roads to be considered in the analysis. The length of the roads are considered as the corresponding costs ($c_a, a \in \bar{A}$) and the total budget is $B = 10.00$ (in units of length). As for convergence of private traffic assignment a relative gap of 1% is used which is proven empirically to be close to an acceptable level for equilibrium assignment (INRO, 2009). Note that the traffic assignment problem is solved iteratively and in each iteration the relative gap as a termination condition is computed as the difference between the total travel times calculated, based on the currently used paths and current shortest paths, divided by the former. In each iteration, the traffic assignment including private and transit vehicles quickly converges in less than 3 seconds. The total travel time of the private cars and the transit passengers in the do-nothing scenario is calculated as 978,634.6 car-minutes and 466,665.7 passenger-minutes respectively.

The algorithm was started with a meagre value for alpha $\alpha = 0.010$. As shown in Table 6.2 the algorithm terminates within almost two hours and the best solution found entails three links (links 1, 3 and 11 of Table 6.2) with a total length of 0.24 km out of a total budget of 10.00 km. It is worth noting that the transit priorities are sought over the

Braess-tainted roads, and the available budget of 10.00 km roads does not imply that potentially there are 10 km of Braess tainted roads available. That is the reason why less than the entire budget has been used. The total travel time (private and public modes) of this solution became 972,671.1 car-minutes and 463,865 respectively, which was 0.61% improvement compared with those of the do-nothing scenario. The algorithm was able to shrink the search domain over the branch-and-bound tree on 1,290 occasions in which the lower bound stood above the incumbent values. In the next runs, the algorithm was tested for two additional values of alpha ($\alpha = 0.015$ and $\alpha = 0.020$) as shown in Table 6.2. The computational times for each value of the alpha have been reported based on which the one pertaining to the alpha of $\alpha = 0.020$ lasted almost three days. In these two runs of the alphas, the same solution was obtained, in which the total travel time (private and public modes) became 971,429.3 car-minutes, and 463,256.1 respectively equivalent to approximately 0.74% improvement compared with that of the do-nothing scenario.

In terms of the roads identified for transit priority lanes (roads 1 to 15 and 29 of Table 6.2), it can be observed, that they are chosen consistently from the top of the sorted list in Table 6.1. It indicates that the notion of sorting the candidate roads on a merit basis proposed in condition (6.10) is obviously working. Furthermore, the depleted budget is 1.61 km which accounts for 16% of the available budget. It is evident that an increase of 1% in the value of alpha resulted in a drastic increase in the computation time. This implies that the value of alpha strongly influences the size of the solution domains. Furthermore, Figure 6.1 shows the computational cost as well as the incremental improvements of the value of the objective function over a range of the values of alpha. That is, higher values of alpha result in better solutions (i.e., lower values of the objective functions), but at the expense of a higher computational time.

Table 6.1 Candidate links to be considered as transit priority lanes, data for Winnipeg, Canada

no	I_node	J_node	length (km) *	number of lanes	free flow speed (km/hr)	capacity per lanes (PCE)	hourly car volume	hourly transit volume (in car equivalency)	volume per capacity (V/C)	sorting Index
1	1046	1045	0.07	4	35	375	1419.52	328.529	1.17	6761.98
2	1047	1050	0.07	4	35	375	1002.49	283.464	0.86	4467.29
3	1050	1047	0.07	4	35	375	1685.39	199.892	1.26	4024.8
4	1044	1043	0.1	4	35	375	1258.42	304.892	1.04	3939.12
5	1047	1046	0.09	4	35	375	1771.47	223.529	1.33	3720.08
6	937	948	0.1	4	35	375	1001.74	282.24	0.86	3111.15
7	1041	1040	0.07	4	35	375	1466.54	167.877	1.09	2913.32
8	931	937	0.1	4	35	375	1030.56	262.24	0.86	2829.15
9	1051	1050	0.11	4	35	375	1685.39	209.892	1.26	2703.63
10	1045	1044	0.16	4	35	375	1258.42	304.892	1.04	2461.95
11	917	931	0.1	4	35	375	1200.65	216.163	0.94	2397.76
12	1042	1025	0.12	4	35	375	1023.83	254.38	0.85	2249.55
13	1043	1042	0.18	4	35	375	1261.47	304.892	1.04	2187.37
14	901	917	0.1	4	35	375	1196.33	194.734	0.93	2105.82
15	1020	1019	0.05	2	25	200	286.331	83.0119	0.92	1970.24
16	1010	1009	0.06	4	35	375	1747.37	91.6667	1.23	1977.75
17	899	898	0.08	4	35	625	2056.45	129.375	0.87	1495.47
18	1008	1007	0.07	4	35	375	1528.51	91.6667	1.08	1499.1
19	606	605	0.14	4	50	625	3019.8	139.375	1.26	1312.27
20	967	980	0.1	2	25	200	250.28	98.5119	0.87	1194.4
21	947	967	0.1	2	25	200	249.761	98.5119	0.87	1195.1
22	411	410	0.1	3	40	875	2843.67	89.3647	1.12	1032.34
23	1034	1035	0.09	3	40	875	2383.62	89.3647	0.94	968.36
24	1011	1010	0.12	4	35	375	1609.37	91.6667	1.13	912.36
25	412	411	0.1	3	40	875	2631.72	77.3647	1.03	820.28
26	1037	1038	0.17	4	35	375	1429.08	114.365	1.03	748.37
27	1009	1008	0.17	4	35	375	1512.47	108.333	1.08	737.53
28	1036	1037	0.18	4	35	375	1418.92	114.365	1.02	700.3
29	1041	1042	0.12	4	35	375	1249.88	84.5119	0.89	669.18
30	605	604	0.24	4	50	625	2426.03	139.375	1.03	632.51
31	414	973	0.12	2	45	1250	2295.62	76.6667	0.95	627.21
32	607	606	0.37	4	50	625	3205	159.174	1.35	609.61
33	410	1034	0.15	3	40	875	2383.62	89.3647	0.94	581.01
34	604	603	0.25	4	50	625	2426.03	129.375	1.02	556
35	1035	1036	0.22	4	35	375	1870.51	89.3647	1.31	557.55
36	170	169	0.23	3	50	625	2478.54	85.8036	1.37	528.78
37	1053	1052	0.24	4	35	375	1180.54	129.089	0.87	519.12
38	608	607	0.22	4	50	625	2993.71	84.799	1.23	487.53
39	1012	1011	0.17	4	35	625	2384.09	76.6667	0.98	456.17
40	415	414	0.23	2	40	875	2169.41	76.6667	1.28	441.75
41	166	165	0.2	2	55	1250	2243.53	85.8036	0.93	414.25
42	603	602	0.37	4	50	625	2655.43	129.375	1.11	407.03
43	600	599	0.36	4	50	625	2331.27	129.375	0.98	371.73
44	599	600	0.36	3	50	625	1677.41	129.375	0.96	371.61
45	887	899	0.33	4	35	625	2056.45	129.375	0.87	362.54
46	167	166	0.23	2	55	1250	2243.53	85.8036	0.93	360.21
47	601	600	0.42	4	50	625	2199.21	129.375	0.93	303.33
48	601	602	0.58	3	50	625	1759.88	129.375	1.01	241.85
49	602	601	0.58	4	50	625	2248.88	129.375	0.95	224.1
50	165	1055	0.37	2	55	1250	2243.53	85.8036	0.93	223.92
51	304	412	0.38	3	40	875	2631.72	77.3647	1.03	215.86
52	973	1012	0.35	2	45	1250	2295.62	76.6667	0.95	215.05
53	437	436	0.46	3	50	625	2003.75	76.6667	1.11	192.08
54	175	174	0.45	3	50	625	1711.39	79.6271	0.96	177.77
55	441	442	0.51	3	50	625	1912.45	78.5417	1.06	169.95
56	442	441	0.51	3	50	625	1822.49	78.5417	1.01	162.25
57	1055	1059	0.75	2	55	1250	2243.53	85.8036	0.93	110.47
58	436	423	0.82	3	40	875	2311.83	76.6667	0.91	87.9
59	423	415	0.89	3	40	875	2495.26	76.6667	0.98	87.01
60	423	436	0.82	3	50	625	1526.17	76.6667	0.85	83.46

*Total length is 15.27 km

Figure 6.2 demonstrates graphically the roads identified to provide one lane dedicated to transit. Apart from a few sporadic roads, the transit priority lanes are topographically consistent.

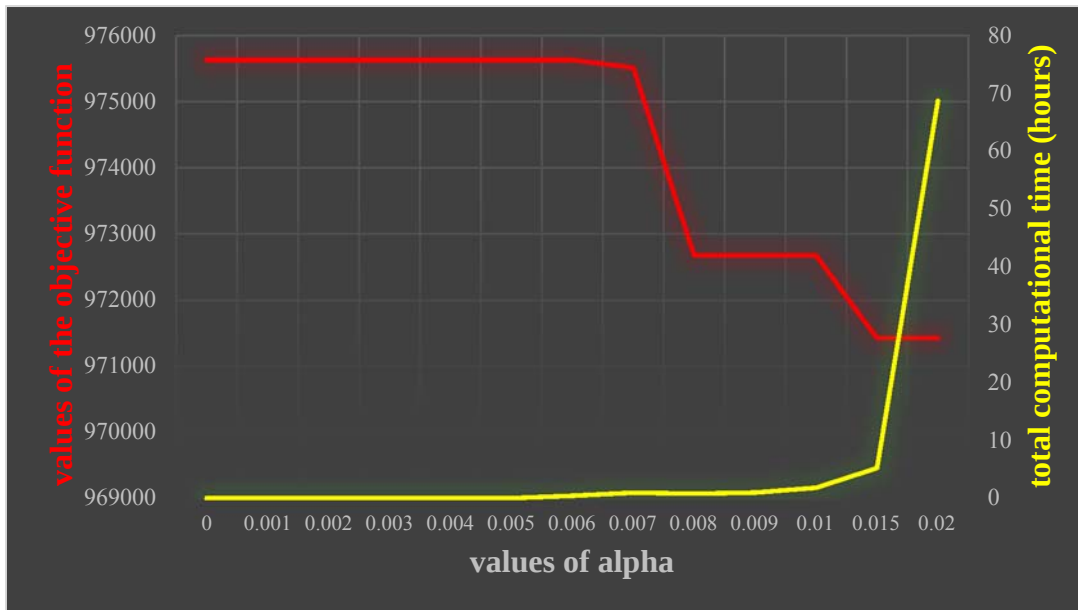


Figure 6.1 Impact of the alphas on the computational time and the objective function

With respect to BP, the algorithm sought mis-utilized capacity (even in the congested area), to be taken away from the private mode and to be used exclusively by the transit mode. Although the idea of providing priority to the transit is appealing, there is certainly a level above which the overall performance of the network (private and public) will deteriorate. The performance of the network was measured as the total travel time/cost formulated in the objective function (6.1) and was referred to as the incumbent value. Since the algorithm started with the incumbent value of the do-nothing scenario, in the end, the algorithm did not render any solution worse than that of the do-nothing scenario. For the greater cause of transit priority, should a slight deterioration of the private mode be acceptable, one can re-launch the algorithm with a slightly higher incumbent value. Therefore, more links are likely to be found as transit priority lanes. Such measures can be strongly justified in light of possible shifts in the travel demand from private to transit if greater priority and incentives are provided for public transport. This brings us to set out new areas for further studies which are discussed in the next section.

6.5 Conclusions

In this study, the aim was to enhance the attractiveness of the transit system by providing transit priority lanes at no cost, and no additional burden to the private mode. The method proposed in this chapter is motivated by Braess's paradox in seeking roads for which closure will counterintuitively improve overall traffic circulation. Accordingly, instead of

complete closure, a lane was taken away for the transit vehicles use only and left the rest of the space for the private mode so as to maintain the same level of network connectivity. This problem was formulated as a bi-level, nonlinear programming problem mixed with binary decision variables which are proven to be extremely intractable for large-sized networks. At the upper level, the total system cost (or total travel time) is minimised, while multimodal traffic assignment is taken into account at the lower level.



Figure 6.2 Winnipeg transport network and selected transit priority lanes

To address the scalability of the methodology, a greedy and RAM-efficient branch and bound algorithm tailored to large-sized networks was developed. This algorithm was coded using MS Office applications (Access, Excel) synchronized with a commercial transport planning software (EMME 3) targeting the needs of industry and practitioners.

In the first phase, a set of roads deemed appropriate for candidacy as transit priority lanes were identified. To this end, criteria such as current transit ridership, congestion levels, and even costs pertaining to implementation of a transit lane were considered (condition (6.10)). In the second phase, a subset of the candidate set was

sought using the proposed BB algorithm. In doing so, based on a number of traffic characteristics, a merit index for the candidate projects was calculated. The roads were then sorted in descending order from the most likely to be the best selection for designation as transit priority lanes. The tree structure of the BB is built on the sorted list of candidates. The branching is also done based on the sorted list from the top, descending steeply into the last possible candidate, subject to budget constraints or weak lower bounds. This would help with the use of extremely small RAM space which is a decisive factor in handling large-sized networks. Such a simple rule makes the search over the tree quite smooth, with no effort required to remember the structure of the rapidly growing tree. Subsequently, this offers an ideal leverage for dealing with large-sized networks.

As advised by LeBlanc (1975), the lower bound value was calculated based on system optimal (SO) traffic flow. In cases in which no lower bound is found to be higher than the incumbent value, the wide gap between SO and UE flow may affect the numerical affordability of the BB adversely. Therefore, inserting the alpha coefficient in the marginal delay function to bridge the gap is proposed. The value of alpha controls the trade-off between accuracy of the solution and computational time easily.

Given that the proposed algorithm is linked to Braess's paradox, it may be valid to ask how the method would perform if it was applied to a network which was already designed to prevent BP from happening? To answer this question let's first underscore the primary aim of transit priority, that is to give priority to the public transport-even at the cost of leaving less space to private cars- to encourage people to shift to public transport. Laying out the bus lanes on the shoulder of the Braess's Paradox is a boon to a primary cause that is a conservative approach to the bus priority. Nevertheless, in previous studies, bus lanes are added irrespective of maintaining the same level of service for the private mode. In the case of dealing with a BP-free network the proposed methodology can still be applied by launching the branch-and-bound with a higher incumbent value (say infinity; as noted before the initial incumbent value is the total travel time of the do-nothing or existing network). As a result of designating some lanes for the public transport, private cars are inevitably faced with longer travel times.

The algorithm was evaluated using real data for Winnipeg. The best solution comprises 1.61 km of transit priority lanes primarily located in the central business district. It is important to note changes to the transit network structure would change the bus frequencies and network, but these are ignored here. Moreover, It should be noted

that any changes in the transit network would also change the signal phases and timing. In addition, giving priority to the transit is supposed to entice more ridership (modal shift from private modes and is aimed at doing so). As the result, a worthy line of research is to develop a combined model in which modal split and traffic assignment are synchronized to fully take the mutual changes (network vs demand) into account. Accordingly, the algorithm developed can be further improved on several fronts as follows: (i) in this study the travel demand matrices for both private and transit modes were assumed to be fixed. Given the intention of transit priority to make the transit mode more appealing, the methodology presented here can be further improved to consider flexible travel demand and hence the possible shift from private mode to transit mode. (ii) The concept of transit priority lanes would work more efficiently in synchronization with transit signal priorities. To this end, road delay functions to adjust the priority signal settings become non-separable. This gives rise to path-based traffic assignment methods such as complementarity and variational inequality methods (Aashtiani, 1979; Nagurney, 1998) which is still an evolving subject in the literature. (iii) The possible spare capacity of the transit lanes can provide an opportunity for promoting car-sharing schemes or high occupancy vehicles (HOV) as well as the bicycles. The model can be further improved by considering the variation of travel demand over time, as well as in response to changes in the network.

In light of real-time data, big data, sensor revolution, and the internet of the things, some scholars advocate moving toward dynamic traffic assignment (DTA) that is to include the time variation features of the traffic. DTA is based on the fundamental diagram, a method derived from traffic flow theory, to model congestion with a greater realism and fidelity. Furthermore, one of the main drawbacks of the (static) traffic assignment method, regardless of having a priority lane, is a lack of consideration of vehicle-to-vehicle interactions (for example in a one-by-one road, if a car stops, all others should stop, too). To this end, alternative methods are DTA or a more disaggregated model such as microsimulation which are based on car following methods. Though the reward is enormous, the task is highly challenging, attributable to some theoretical hardship as well as computational costs. Nevertheless, DTA seems to be the future, as the result, integration of DTA in the proposed methodology is a thread of research deserving further investigation.

7 CONCLUSION

This section presents a summary of the research undertaken, highlights of the findings and challenges as well as the contributions made. This chapter concludes with a number of suggestions for further investigation.

7.1 Summary of the research

In response to the chronic issue of traffic congestion two approaches were undertaken in this study. Firstly, a hard approach consisting of adding more capacity to the road infrastructure including constructing new roads, bridges or widening existing roads, better known as the discrete network design problem (DNDP) was investigated. It is called a hard approach, because it involves in a number of laborious, time consuming and capital intensive (construction) projects. In contrast, a soft approach was also investigated that grants road space priority to mass transit modes (public transport) better known as the transit priority lane design problem (TPLDP). In contrast to the DNDP, the TPLDP is called a soft approach, mainly because it is not a capital or labour intensive nor time-consuming approach.

The DNDP was defined as follows, given a number of candidate projects (new roads, road widenings, grade-separated interchanges, etc.) and associated construction costs and a limited budget, which candidate projects should be selected to ease traffic congestion the most.

Similarly, the TPLDP was defined as follows, given a number of (existing) roads nominated to possibly designate a lane to be used exclusively by public transport modes, which ones should be selected to ease traffic congestion. Note that each road is associated with some (minor) expenses pertaining to the lane marking, signage, signals etc., and the final selection of the projects must respect a limited budget. An important point to note is that, despite leaving less space for private vehicles, the aim is still to ease traffic congestion. Though it seems implausible, it is theoretically possible thanks to Braess Paradox. Braess paradox stands for a counterintuitive phenomenon in which traffic circulation sometimes improves when some (Braess-tainted) roads are blocked. Therefore, the idea is to search for Braess-tainted roads that will not to be fully closed, but rather to give away a lane to public transport modes.

Considering the nature of these two approaches, three analogous traits can be observed: (i) the decisions variables are of a binary nature (1 or 0), to build or not (for the DNDP) and to designate a road as priority lane or not (for the TPLDP) (ii) the aim is always to ease traffic congestion which can be formulated as the total travel time spent on the network by all users. Hence, the total travel time becomes a nonlinear objective function which can also be considered as an index to measure the performance of the respective network subject to the decisions that were made. Accordingly, both the DNDP and the TPLDP can be expressed as optimisation problems in which the objective function (i.e. the total travel time) are minimised (iii) there should also exist a model to mimic the way the users navigate the network so as to be able to calculate the total travel time. This model is called the traffic assignment problem (TAP) which itself is an optimisation problem. Therefore, the above-mentioned problems were formulated as bi-level programming problems to minimise the total travel time in the upper level (subject to the binary decision variables, cost and budget constraints) while accounting for the users' routing as a TAP in the lower level.

Being bi-level is enough to make a problem computationally intractable known as NP-hard. In other words, as the size of the problem (number of roads, intersections, decision variables) increases which is the case in real life examples, the computational time becomes a prohibitive factor.

The above problems are found to be mathematically and computationally intractable. That is for real life road networks, finding a reliable and valid solution procedure is a significant concern. Inclusion of the nonlinearity, binary (or integer) variables make an already difficult problem more complex. These complexities call on special and innovative ideas to be able to provide effective solution methodologies.

A review of the literature indicated a number of shortcomings in past studies. Notably, "practicality" is relatively rare when dealing with NP-hard problems. In other words, applications of methods on large sized road networks, as is the case in real life situations are yet to be addressed. Secondly, given the computational complexities, some aspects of the problems have been ignored or loosely treated. More precisely, any solution to the TPLDP or the NDP must be thoroughly examined based upon a reliable model to measure the traffic circulation which has been largely relaxed due to the theoretical and computational burdens.

These shortcomings were addressed in this research. The methodologies developed in this study were subjected to multiclass and multimodal traffic assignment to keep a high level of realism and fidelity for the models. Given the initiatives developed in this research, for the numerical analysis, real life datasets were used and the results were shown to be promising.

7.2 Solution methodologies

As noted earlier, the main part of the complexities is rooted in the fact that the problems are bi-level. To this end, a branch-and-bound algorithm was developed to represent the bi-level structure. In order to expedite the quest to find global optimum solutions, a merit index was defined and calculated for candidate projects, based on which the projects were sorted in descending order. The merit index was defined based on the congestion levels, capacity of the roads, traffic load (for the DNDP) or transit ridership (for the TPLDP) in such a way to give more merit to the roads that were deemed more likely to be selected for construction (for the DNDP) or for priority lanes (for the TPLDP).

The merit index was then used to search for the solution over the branch-and-bound algorithm which made the search highly RAM-efficient.

As the BB expands over the solution space, a lower bound and an upper bound to the value of the objective function are recorded. The key to the success of the BB is to be able to quickly shrink the solution space towards the optimum solution. As the tree structure of the algorithm grows (over newly generated nodes), it stops at the nodes in which the respective lower bound is found to be higher than the upper bound. As the result, it is highly significant for the algorithm to calculate very tight lower bound values.

To this end, for the DNDP, the branch-and-bound algorithm was hybridized with a Benders decomposition method which was shown to be highly effective.

A prerequisite of the Benders decomposition is found to be related to the capacitated traffic assignment. To this end, a method dubbed inflated travel time (ITT) was developed in which the travel times of the oversaturated roads are inflated artificially to the extent they become saturated. In the context of the available literature, the main advantage of the ITT is to obviate any additional parameter and automatic mechanism to initiate a feasible solution.

In the following section the major contributions of this research are highlighted.

7.3 Contributions of the research

Contributions of this research can be summarised as follows:

- A network based approach for the problem of transit priority lane design (TPLDP) is developed.
- The discrete network design problem (DNDP) which is a benchmark problem in computational complexity is solved using an exact method consisting of a Benders decomposition method and a branch and bound algorithm.
- For the both problems (TPLDP and DNDP), the methodologies are tailored for real life road networks.
- A RAM-efficient and memoryless branch and bound algorithm based on an innovative concept (merit index) is developed.
- To enhance the realism of the models, in the solutions provided for the two problems, the models are subjected to multiclass and multimodal traffic flow.
- A parameter-less method is developed for the capacitated traffic assignment problem.

7.4 Suggestions for further research

Given the nature of the research questions a number of extensions worthy of further investigation have been identified and are described as follows:

There is ample space for improvement associated with the traffic assignment model used in this study. As noted before, it was assumed that the travel demand was fixed, given and exogenous. Furthermore, it was assumed that users have full knowledge of traffic conditions and choose paths accordingly with no ambiguity. Therefore, a deterministic traffic assignment with fixed travel demand was used. The traffic assignment model can be further improved by not including such assumptions. To this end consideration of variable demand as a function of the congestion level and a simultaneous modal choice between competing public and private modes is worth noting.

In addition, the deterministic traffic assignment can also be relaxed to a stochastic assignment in which users do not necessarily have a full understanding of the traffic congestion when they choose paths.

The delay functions associated with the roads are a functions of the respective roads which are called asymmetric and separable delay functions. Relaxing this assumption results in more comprehensive functions. To this end, the TAP cannot be

formulated as an optimisation problem in the lower level. Alternative approaches are complementarity methods, variational inequality or in a more general sense, fixed-point methods.

In the above assignment model, there is no provision for the variation of the parameters with respect to time. In other words, they are static and not dynamic models. However, in recent years dynamic traffic assignment (DTA) has gained significant momentum mainly due to its unique application to the real time simulation and modelling. DTA can take delays at signalized and unsignalized junction into account.

In all the above extensions, the computational time is a significant concern, nevertheless, for the DTA it is much worse.

For the TPLDP, inclusions of priority phase in the junctions' signal timing and transit priority lanes can result in a more positive outcome.

As alluded to before, for infrastructure investment (DNDP), it is of the highest practical value to not only arrive at the best possible solution but a number of top solutions that can be provided to decision makers. It is called k-DNDP, in which the k best investment scenarios are identified.

The DNDP was built on the tenets of a given number of candidate projects. This assumption can also be relaxed such that a set of candidate projects is identified first. Similarly, the physical characteristics of candidate projects (number of lanes and capacity) can also be relaxed to be a variable in the DNDP.

Investigation of long term impact of the transportation (such as a new road, priority lane, etc.) to land use changes is also important. It is well known that accessibility is related to economic activities and this is in turn expected to have an impact on the land use as well as real estate values. Nevertheless, transport and land use are intertwined based on mutual influences. Furthermore, partial disruption of accessibility which could be the case for transit priority lanes or during construction in road investments may adversely affect some businesses or properties. Anecdotal evidence has shown these negative consequences might be conducive to some contention from vested interests.

In the above mentioned problems, when the solutions are found (to build new roads or convert some roads to transit priority) no regard was given to implementation. More precisely, though the budget is included in the formulation, due to some other limitations, it may not be possible to undertake all the qualified projects in one go. In other words, one has to prioritise or further schedule the implementation process. As the

result, prioritisation and scheduling can be added into the framework of the DNDP and the TPLDP.

The main expectation in the DNDP and the TPLDP was to identify initiatives to ease traffic congestion. In other words, the objective function was set to be total travel time. Given the undeniable importance of the environment and societal impact of transportation, it is sometimes necessary to include them in the formulation which results in a multi-objective problem. In other words, expanding the DNDP and the TPLDP into the multi objective problem deserves further investigation.

8 REFERENCES

- Aashtiani, H.Z. (1979) The multi-modal traffic assignment problem. PhD dissertation, Massachusetts Institute of Technology.
- Aashtiani, H.Z., Poorzahedy, H. (2004) Braess' phenomenon in the management of networks and dissociation of equilibrium concepts. *Transportation planning and technology* 27, 469-482.
- Achterberg, T., Wunderling, R. (2013) Mixed integer programming: Analyzing 12 years of progress. *Facets of Combinatorial Optimization*. Springer, pp. 449-481.
- Bagajewicz, M., Manousiouthakis, V. (1991) On the generalized Benders decomposition. *Computers & chemical engineering* 15, 691-700.
- Bagherian, M., Mesbah, M., Ferreira, L. (2015) Using delay functions to evaluate transit priority at signals. *Public Transport* 7, 61-75.
- Bagloee, S.A., Asadi, M. (2015) Prioritizing road extension projects with interdependent benefits under time constraint. *Transportation Research Part A: Policy and Practice* 75, 196-216.
- Bagloee, S.A., Ceder, A. (2011) Transit-network design methodology for actual-size road networks. *Transportation Research Part B: Methodological* 45, 1787-1804.
- Bagloee, S.A., Ceder, A., Tavana, M., Bozic, C. (2013a) A heuristic methodology to tackle the Braess Paradox detecting problem tailored for real road networks. *Transportmetrica A: Transport Science* 10, 437-456.
- Bagloee, S.A., Sarvi, M. (2015a) Heuristic Approach to Capacitated Traffic Assignment Problem for Large-Scale Transport Networks. *Transportation Research Record: Journal of the Transportation Research Board*, 1-11.
- Bagloee, S.A., Sarvi, M. (2015b) A Heuristic Approach to Capacitated Traffic Assignment Problem Tailored to Large Scale Networks. *Proceedings of Transportation Research Board*, Washington D.C., United States.
- Bagloee, S.A., Sarvi, M., Patriksson, M. (2016a) A Hybrid Branch-and-Bound and Benders Decomposition Algorithm for the Network Design Problem. *Computer-Aided Civil and Infrastructure Engineering*.
- Bagloee, S.A., Sarvi, M., Wallace, M. (2016b) Bicycle lane priority: Promoting bicycle as a green mode even in congested urban area. *Transportation Research Part A: Policy and Practice* 87, 102-121.
- Bagloee, S.A., Tavana, M., Ceder, A., Bozic, C., Asadi, M. (2013b) A hybrid meta-heuristic algorithm for solving real-life transportation network design problems. *International Journal of Logistics Systems and Management* 16, 41-66.
- Balakrishnan, A., Magnanti, T.L., Mirchandani, P. (1997) Network design. *Annotated bibliographies in combinatorial optimization*, 311-334.
- Balas, E., Jeroslow, R. (1972) Canonical cuts on the unit hypercube. *SIAM Journal on Applied Mathematics* 23, 61-69.

- Bar-Gera, H. (2016) Transportation Network Test Problems. <http://www.bgu.ac.il/~bargera/tntp/>; Accessed by Dec. 1, 2016.
- Bar-Gera, H., Boyce, D. (1999) Route flow entropy maximization in origin-based traffic assignment. *Proceedings of 14th International Symposium on Transportation and Traffic Theory*, Jerusalem, Israel, pp. 397–415.
- Basso, L.J., Guevara, C.A., Gschwender, A., Fuster, M. (2011) Congestion pricing, transit subsidies and dedicated bus lanes: Efficient and practical solutions to congestion. *Transport Policy* 18, 676-684.
- Beckmann, M., McGuire, C., Winsten, C.B. (1956) *Studies in the Economics of Transportation*, Yale University Press, New Haven, CT.
- Bell, M.G., Shield, C.M., Busch, F., Kruse, G. (1997) A stochastic user equilibrium path flow estimator. *Transportation Research Part C: Emerging Technologies* 5, 197-210.
- Ben-Ayed, O., Blair, C.E. (1990) Computational Difficulties of Bilevel Linear Programming. *Operations Research* 38, 556-560.
- Benders, J.F. (1962) Partitioning procedures for solving mixed-variables programming problems. *Numerische mathematik* 4, 238-252.
- Bertsekas, D.P. (1982) Constrained optimization and Lagrange multiplier methods. *Computer Science and Applied Mathematics*, Boston: Academic Press, 1982 1.
- Bixby, R.E. (2012) A brief history of linear and mixed-integer programming computation. *Optimization Stories* ed Grötschel, M. Deutsche Mathematiker-Vereinigung, Bielefeld, pp. 107–121.
- Boyce, D. (2013) Beckmann's transportation network equilibrium model: Its history and relationship to the Kuhn–Tucker conditions. *Economics of Transportation* 2, 47-52.
- Boyce, D. (2014) Network equilibrium models for urban transport. *Handbook of Regional Science* eds Fischer, M.M., Nijkamp, P. Springer Berlin Heidelberg, pp. 759-786.
- Boyce, D., Farhi, A., Weischedel, R. (1973) Optimal network problem: a branch-and-bound algorithm. *Environment and Planning* 5, 519-533.
- Boyce, D., Janson, B. (1980) A discrete transportation network design problem with combined trip distribution and assignment. *Transportation Research Part B: Methodological* 14, 147-154.
- Boyce, D., Janson, B., Eash, R. (1981) The effect on equilibrium trip assignment of different link congestion functions. *Transportation Research Part A: General* 15, 223-232.
- Boyce, D., Ralevic-Dekic, B., Bar-Gera, H. (2004) Convergence of traffic assignments: how much is enough? *Journal of Transportation Engineering* 130, 49-55.
- Boyce, D., Xie, J. (2013) Assigning user class link flows uniquely. *Transportation Research Part A: Policy and Practice* 53, 22-35.
- Boyce, D., Xiong, Q. (2004) User-optimal and system-optimal route choices for a large road network. *Review of network Economics* 3, 371-380.

- Boyce, D.E., Soberanes, J.L. (1979) Solutions to the optimal network design problem with shipments related to transportation cost. *Transportation Research Part B: Methodological* 13, 65-80.
- Braess, D. (1968) Über ein Paradoxon aus der Verkehrsplanung. *Unternehmensforschung* 12, 258-268.
- Braess, D., Nagurney, A., Wakolbinger, T. (2005) On a paradox of traffic planning. *Transportation science* 39, 446-450.
- Ceder, A. (2015) *Public Transit Planning and Operation: Modeling, Practice and Behavior*. CRC Press.
- Chen, A., Zhou, Z., Ryu, S. (2011a) Modeling physical and environmental side constraints in traffic equilibrium problem. *International Journal of Sustainable Transportation* 5, 172-197.
- Chen, B.Y., Lam, W.H.K., Sumalee, A., Shao, H. (2011b) An efficient solution algorithm for solving multi-class reliability-based traffic assignment problem. *Mathematical and Computer Modelling* 54, 1428-1439.
- Chen, M., Alfa, A.S. (1991) A network design algorithm using a stochastic incremental traffic assignment approach. *Transportation Science* 25, 215-224.
- Chen, Q. (2015) An optimization model for the selection of bus-only lanes in a city. *PloS one* 10, e0133951.
- Colson, B., Marcotte, P., Savard, G. (2005) Bilevel programming: A survey. *4OR* 3, 87-107.
- Colson, B., Marcotte, P., Savard, G. (2007) An overview of bilevel optimization. *Annals of operations research* 153, 235-256.
- Cova, T.J., Johnson, J.P. (2003) A network flow model for lane-based evacuation routing. *Transportation research part A: Policy and Practice* 37, 579-604.
- D'Ambrosio, C., Lodi, A. (2013) Mixed integer nonlinear programming tools: an updated practical overview. *Annals of Operations Research* 204, 301-320.
- Dafermos, S.C. (1972) The traffic assignment problem for multiclass-user transportation networks. *Transportation Science* 6, 73-87.
- Daganzo, C.F. (1977a) On the traffic assignment problem with flow dependent costs—I. *Transportation Research* 11, 433-437.
- Daganzo, C.F. (1977b) On the traffic assignment problem with flow dependent costs—II. *Transportation Research* 11, 439-441.
- De Cea, J., Fernández, J.E., Dekock, V., Soto, A. (2005) Solving network equilibrium problems on multimodal urban transportation networks with multiple user classes. *Transport Reviews* 25, 293-317.
- Dempe, S. (2003) Annotated bibliography on bilevel programming and mathematical programs with equilibrium constraints. *Optimization* 52, 333–359.
- Diab, E.I., El-Geneidy, A.M. (2013) Variation in bus transit service: understanding the impacts of various improvement strategies on transit service reliability. *Public Transport* 4, 209-231.

- Drezner, Z., Wesolowsky, G.O. (1997) Selecting an optimum configuration of one-way and two-way routes. *Transportation Science* 31, 386-394.
- Drezner, Z., Wesolowsky, G.O. (2003) Network design: selection and design of links and facility location. *Transportation Research Part A: Policy and Practice* 37, 241-256.
- Eichler, M., Daganzo, C.F. (2006) Bus lanes with intermittent priority: Strategy formulae and an evaluation. *Transportation Research Part B: Methodological* 40, 731-744.
- Estache, A., Ianchovichina, E., Bacon, R., Salamon, I. (2013) Infrastructure and employment creation in the Middle East and North Africa (MENA). *Directions in development ; infrastructure. Washington D.C. : The Worldbank.* 74918.
- Fang, Y., Chu, F., Mammar, S., Che, A. (2013) An optimal algorithm for automated truck freight transportation via lane reservation strategy. *Transportation Research Part C: Emerging Technologies* 26, 170-183.
- Fang, Y., Chu, F., Mammar, S., Che, A. (2014) A cut-and-solve-based algorithm for optimal lane reservation with dynamic link travel times. *International Journal of Production Research* 52, 1003-1015.
- Farahani, R.Z., Miandoabchi, E., Szeto, W., Rashidi, H. (2013) A review of urban transportation network design problems. *European Journal of Operational Research* 229, 281-302.
- Farvaresh, H., Sepehri, M.M. (2011) A single-level mixed integer linear formulation for a bi-level discrete network design problem. *Transportation Research Part E: Logistics and Transportation Review* 47, 623-640.
- Farvaresh, H., Sepehri, M.M. (2013) A branch and bound algorithm for bi-level discrete network design problem. *Netw Spat Econ* 13, 67-106.
- Ferguson, E.M., Duthie, J., Travis Waller, S. (2012) Comparing Delay Minimization and Emissions Minimization in the Network Design Problem. *Computer-Aided Civil and Infrastructure Engineering* 27, 288-302.
- Ferrari, P. (1997) Capacity constraints in urban transport networks. *Transportation Research Part B: Methodological* 31, 291-301.
- FHWA (2002) Status of the Nation's Highways, Bridges and Transit: Conditions and Performance. . *US Department of Transportation* Washington, DC,.
- Fisher, M.L. (2004) The Lagrangian relaxation method for solving integer programming problems. *Management science* 50, 1861-1871.
- Florian, M., Morosan, C.D. (2014) On uniqueness and proportionality in multi-class equilibrium assignment. *Transportation Research Part B: Methodological* 70, 173-185.
- Floudas, C.A. (1995) *Nonlinear and mixed-integer optimization: fundamentals and applications*. Oxford: Oxford University Press,.
- Fontaine, P., Minner, S. (2014) Benders decomposition for discrete–continuous linear bilevel problems with application to traffic network design. *Transportation Research Part B: Methodological* 70, 163-172.
- GAMS (2014) GAMS Development Corporation, GAMS Development Corporation, Washington DC. .

- Gao, Z., Wu, J., Sun, H. (2005) Solution algorithm for the bi-level discrete network design problem. *Transportation Research Part B: Methodological* 39, 479-495.
- Geoffrion, A.M. (1972) Generalized Benders decomposition. *Journal of Optimization Theory and Applications* 10, 237-260.
- Geroliminis, N., Zheng, N., Ampountolas, K. (2014) A three-dimensional macroscopic fundamental diagram for mixed bi-modal urban networks. *Transportation Research Part C: Emerging Technologies* 42, 168-181.
- Guler, S.I., Cassidy, M.J. (2012) Strategies for sharing bottleneck capacity among buses and cars. *Transportation research part B: methodological* 46, 1334-1345.
- Guler, S.I., Gayah, V.V., Menendez, M. (2016) Bus priority at signalized intersections with single-lane approaches: A novel pre-signal strategy. *Transportation Research Part C: Emerging Technologies* 63, 51-70.
- Guler, S.I., Menendez, M. (2015) Pre-signals for bus priority: basic guidelines for implementation. *Public Transport* 7, 339-354.
- Hadas, Y., Ceder, A. (2014) Optimal Connected Urban Bus Network of Priority Lanes. *Transportation Research Record: Journal of the Transportation Research Board*, 49-57.
- HCM2010 (2010) Highway Capacity Manual 2010. *National Academy of Sciences. Yhdysvallat*.
- Hearn, D. (1980) Bounding flows in traffic assignment models. *Technical report Research Report* 80-4.
- Hearn, D., Ribera, J. (1980) Bounded flow equilibrium problems by penalty methods. *Proceedings of Proceedings of IEEE International Conference on Circuits and Computers*, pp. 162-166.
- Inouye, H. (1987) Traffic equilibria and its solution in congested road networks. *Proceedings of IFACIFIPIFORS CONFERENCE ON CONTROL IN*.
- INRO (2009) EMME3 v 3.2. *EMME3 User's Guide* 3.2 ed, Montreal, Quebec, Canada.
- Khoo, H.L., Teoh, L.E., Meng, Q. (2014) A bi-objective optimization approach for exclusive bus lane selection and scheduling design. *Engineering Optimization* 46, 987-1007.
- Larsson, T., Patriksson, M. (1992) Simplicial decomposition with disaggregated representation for the traffic assignment problem. *Transportation Science* 26, 4-17.
- Larsson, T., Patriksson, M. (1995) An augmented Lagrangean dual algorithm for link capacity side constrained traffic assignment problems. *Transportation Research Part B: Methodological* 29, 433-455.
- Larsson, T., Patriksson, M. (1999) Side constrained traffic equilibrium models—analysis, computation and applications. *Transportation Research Part B: Methodological* 33, 233-264.
- Larsson, T., Patriksson, M., Rydergren, C. (2004) A column generation procedure for the side constrained traffic equilibrium problem. *Transportation Research Part B: Methodological* 38, 17-38.
- Lasdon, L.S. (2013) *Optimization theory for large systems*. Courier Corporation.

- LeBlanc, L.J. (1975) An Algorithm for the Discrete Network Design Problem. *Transportation Science* 9, 183-199.
- Leyffer, S. (1993) Deterministic methods for mixed integer nonlinear programming. *PhD University of Dundee*.
- Li, D., Sun, X. (2006) *Nonlinear integer programming*. Springer, Boston.
- Li, S., Ju, Y. (2009) Evaluation of bus-exclusive lanes. *Intelligent Transportation Systems, IEEE Transactions on* 10, 236-245.
- Lin, D.Y. (2011) A Dual Variable Approximation-Based Descent Method for a Bi-level Continuous Dynamic Network Design Problem. *Computer-Aided Civil and Infrastructure Engineering* 26, 581-594.
- Liu, R., Van Vliet, D., Watling, D. (2006) Microsimulation models incorporating both demand and supply dynamics. *Transportation Research Part A: Policy and Practice* 40, 125-150.
- Liu, Z., Meng, Q. (2012) Bus-based park-and-ride system: a stochastic model on multimodal network with congestion pricing schemes. *International Journal of Systems Science* 45, 994-1006.
- Lo, H.K., Luo, X.W., Siu, B.W.Y. (2006) Degradable transport network: Travel time budget of travelers with heterogeneous risk aversion. *Transportation Research Part B: Methodological* 40, 792-806.
- Lodi, A. (2010) Mixed integer programming computation. *50 Years of Integer Programming 1958-2008* eds Jünger, M., Liebling, T.M., Naddef, D., Nemhauser, G.L., Pulleyblank, W.R., Reinelt, G., Rinaldi, G., Wolsey, L.A. Springer Berlin Heidelberg, pp. 619-645.
- Luathap, P., Sumalee, A., Lam, W.H.K., Li, Z.-C., Lo, H.K. (2011) Global optimization method for mixed transportation network design problem: A mixed-integer linear programming approach. *Transportation Research Part B: Methodological* 45, 808-827.
- Magnanti, T.L., Wong, R.T. (1984) Network design and transportation planning: models and algorithms. *Transportation Science* 18, 1-55.
- Marcotte, P., Nguyen, S., Schoeb, A. (2004) A strategic flow model of traffic assignment in static capacitated networks. *Operations Research* 52, 191-212.
- Marcotte, P., Patriksson, M. (2007) Traffic equilibrium. *Handbooks in Operations Research and Management Science* 14, 623-713.
- MathWorks (2014) MATLAB and Statistics Toolbox. *Release 2014a*, , Natick, Massachusetts, United States.
- Mesbah, M., Sarvi, M., Currie, G. (2008) New methodology for optimizing transit priority at the network level. *Transportation Research Record: Journal of the Transportation Research Board* 2089, 93-100.
- Mesbah, M., Sarvi, M., Currie, G. (2011a) Optimization of Transit Priority in the Transportation Network Using a Genetic Algorithm. *Intelligent Transportation Systems, IEEE Transactions on Intelligent Transportation Systems* 12, 908-919.

- Mesbah, M., Sarvi, M., Ouveysi, I., Currie, G. (2011b) Optimization of transit priority in the transportation network using a decomposition methodology. *Transportation Research Part C: Emerging Technologies* 19, 363-373.
- Minoux, M. (1989) Networks synthesis and optimum network design problems: Models, solution methods and applications. *Networks* 19, 313-360.
- Mirchandani, P.B., Li, J.-Q., Hickman, M. (2010) A macroscopic model for integrating bus signal priority with vehicle rescheduling. *Public Transport* 2, 159-172.
- Morowati-Shalilvand, S., Mehri-Tekmeh, J. (2013) An extended origin-based method for solving capacitated traffic assignment problem. *Acta Universitatis Apulensis*, 169-186.
- Nagurney, A. (1998) *Network economics: A variational inequality approach*. Springer.
- Nagurney, A. (2000) A multiclass, multicriteria traffic network equilibrium model. *Mathematical and Computer Modelling* 32, 393-411.
- Nagurney, A. (2010) The negation of the Braess paradox as demand increases: The wisdom of crowds in transportation networks. *EPL (Europhysics Letters)* 91, 48002.
- Nagurney, A., Dong, J. (2002) A multiclass, multicriteria traffic network equilibrium model with elastic demand. *Transportation Research Part B: Methodological* 36, 445-469.
- Newell, G.F. (1980) *Traffic flow on transportation networks*, MIT Press, Cambridge, Mass.
- Nie, Y., Zhang, H., Lee, D.-H. (2004) Models and algorithms for the traffic assignment problem with link capacity constraints. *Transportation Research Part B: Methodological* 38, 285-312.
- Patriksson, P. (1994) *The traffic assignment problem: models and methods*, VSP BV, The Netherlands. Facsimile reproduction published in 2014 by Dover Publications, Inc., Mineola, New York, NY, USA.
- Poorzahedy, H., Rouhani, O.M. (2007) Hybrid meta-heuristic algorithms for solving network design problem. *European Journal of Operational Research* 182, 578-596.
- Poorzahedy, H., Turnquist, M.A. (1982) Approximate algorithms for the discrete network design problem. *Transportation Research Part B: Methodological* 16, 45-55.
- Potts, R.B., Oliver, R.M. (1972) *Flows in transportation networks*, Academic Press, New York.
- Prashker, J.N., Toledo, T. (2004) A gradient projection algorithm for side-constrained traffic assignment. *European Journal of Transport and Infrastructure Research* 4, 177-193.
- Roughgarden, T., Tardos, É. (2002) How bad is selfish routing? *Journal of the ACM (JACM)* 49, 236-259.
- Ryu, S., Chen, A., Choi, K. (2015) Solving the stochastic multi-class traffic assignment problem with asymmetric interactions, route overlapping, and vehicle restrictions. *Journal of Advanced Transportation*, n/a-n/a.

- Ryu, S., Chen, A., Xu, X., Choi, K. (2014) A Dual Approach for Solving the Combined Distribution and Assignment Problem with Link Capacity Constraints. *Netw Spat Econ*, 1-26.
- Sahinidis, N., Grossmann, I.E. (1991) Convergence properties of generalized Benders decomposition. *Computers & chemical engineering* 15, 481-491.
- Sakamoto, K., Abhayantha, C., Kubota, H. (2007) Effectiveness of bus priority lane as countermeasure for congestion. *Transportation Research Record: Journal of the Transportation Research Board*, 103-111.
- Sarvi, M., Bagloee, S.A., Bliemer, M.C. (2016) Network design for road transit priority. *Handbook on Transport and Urban Planning in the Developed World* eds Bliemer, M., Mulley, C., Moutou, C. Edward Elgar Publishing Ltd., Institute of Transport and Logistics Studies, University of Sydney, Australia pp. 355–374.
- Shahpar, A.H., Aashtiani, H.Z., Babazadeh, A. (2008) Dynamic penalty function method for the side constrained traffic assignment problem. *Applied Mathematics and Computation* 206, 332-345.
- Sheffi, Y. (1985) *Urban transportation networks: equilibrium analysis with mathematical programming methods*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey
- Smith, N., Hensher, D. (1998) The future of exclusive busways: the Brazilian experience. *Transport Reviews* 18, 131-152.
- Spiess, H. (1984) Contributions à la théorie et aux outils de planification des réseaux de transport urbain. Montréal: Université de Montréal, Centre de recherche sur les transports.
- Spiess, H. (1990) Technical note—Conical volume-delay functions. *Transportation Science* 24, 153-158.
- (1993) *Transit equilibrium assignment based on optimal strategies: an implementation in EMME/2*. Haldenstrasse 16, CH-2558 Aegerten, Switzerland.
- Spiess, H., Florian, M. (1989) Optimal strategies: a new assignment model for transit networks. *Transportation Research Part B: Methodological* 23, 83-102.
- Sun, X., Lu, H., Fan, Y. (2014) Optimal Bus Lane Infrastructure Design. *Transportation Research Record: Journal of the Transportation Research Board*, 1-11.
- Szeto, W.Y., Jaber, X., O'Mahony, M. (2010) Time-Dependent Discrete Network Design Frameworks Considering Land Use. *Computer-Aided Civil and Infrastructure Engineering* 25, 411-426.
- Szeto, W.Y., Jiang, Y., Wang, D., Sumalee, A. (2013) A sustainable road network design problem with land use transportation interaction over time. *Netw Spat Econ*, 1-32.
- Szeto, W.Y., Wang, A.B. (2015) Price of anarchy for reliability-based traffic assignment and network design. *Transportmetrica A: Transport Science* 11, 603-635.
- Szeto, W.Y., Wang, Y., Wong, S.C. (2014) The Chemical Reaction Optimization Approach to Solving the Environmentally Sustainable Network Design Problem. *Computer-Aided Civil and Infrastructure Engineering* 29, 140-158.

- Tse, L.Y., Hung, W.T., Sumalee, A. (2014) Bus lane safety implications: a case study in Hong Kong. *Transportmetrica A: Transport Science* 10, 140-159.
- Ukkusuri, S.V., Mathew, T.V., Waller, S.T. (2007) Robust Transportation Network Design Under Demand Uncertainty. *Computer-Aided Civil and Infrastructure Engineering* 22, 6-18.
- Unnikrishnan, A., Lin, D.Y. (2012) User equilibrium with recourse: continuous network design problem. *Computer-Aided Civil and Infrastructure Engineering* 27, 512-524.
- Viegas, J., Lu, B. (2004) The intermittent bus lane signals setting within an area. *Transportation Research Part C: Emerging Technologies* 12, 453-469.
- Waller, S.T., Mouskos, K.C., Kamaryiannis, D., Ziliaskopoulos, A.K. (2006) A linear model for the continuous network design problem. *Computer-Aided Civil and Infrastructure Engineering* 21, 334-345.
- Wang, D.Z., Liu, H., Szeto, W. (2015) A novel discrete network design problem formulation and its global optimization solution algorithm. *Transportation Research Part E: Logistics and Transportation Review* 79, 213-230.
- Wang, D.Z.W., Lo, H.K. (2010) Global optimum of the linearized network design problem with equilibrium flows. *Transportation Research Part B: Methodological* 44, 482-492.
- Wang, J., Liu, H., Xie, C. (2016) Transit Network Design with Exclusive Bus Lanes. *Proceedings of Transportation Research Board 95th Annual Meeting*.
- Wang, S., Meng, Q., Yang, H. (2013) Global optimization methods for the discrete network design problem. *Transportation Research Part B: Methodological* 50, 42-60.
- Wong, R.T. (1985) Transportation network research: Algorithmic and computational questions. *Transportation Research Part A: General* 19, 436-438.
- Wu, P., Che, A., Chu, F. (2013) A quantum evolutionary algorithm for lane reservation problem. *Proceedings of Networking, Sensing and Control (ICNSC), 2013 10th IEEE International Conference on*, pp. 264-268.
- Wu, P., Chu, F., Che, A. (2015) Mixed-integer Programming for a New Bus-lane Reservation Problem. *Proceedings of Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*, pp. 2782-2787.
- Wu, P., Chu, F., Che, A., Shi, Q. (2014) A bus lane reservation problem in urban bus transit network. *Proceedings of Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*, pp. 2864-2869.
- Xie, C. (2014) Bicriterion discrete equilibrium network design problem. *Networks* 63, 286-305.
- Xie, J., Xie, C. (2014) An improved TAPAS algorithm for the traffic assignment problem. *Proceedings of Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*, pp. 2336-2341.
- Xie, J., Xie, C. (2015) Origin-Based Algorithms for Traffic Assignment: Algorithmic Structure, Complexity Analysis, and Convergence Performance. *Proceedings of Transportation Research Board 94th Annual Meeting*.

- Xie, X., Chiabaut, N., Leclercq, L. (2012) Improving Bus Transit in Cities with Appropriate Dynamic Lane Allocating Strategies. *Procedia-Social and Behavioral Sciences* 48, 1472-1481.
- Yang, H., Bell, M.G. (1997) Traffic restraint, road pricing and network equilibrium. *Transportation Research Part B: Methodological* 31, 303-314.
- Yang, H., Bell, M.G.H. (1998) Models and algorithms for road network design: a review and some new developments. *Transport Reviews* 18, 257-278.
- Yang, H., Huang, H.-J. (2005) *Mathematical and economic theory of road pricing*.
- Yang, H., Yagar, S. (1994) Traffic assignment and traffic control in general freeway-arterial corridor systems. *Transportation Research Part B: Methodological* 28, 463-486.
- Yang, H., Yagar, S. (1995) Traffic assignment and signal control in saturated road networks. *Transportation Research Part A: Policy and Practice* 29, 125-139.
- Yao, J., Shi, F., An, S., Wang, J. (2015) Evaluation of exclusive bus lanes in a bi-modal degradable road network. *Transportation Research Part C: Emerging Technologies* 60, 36-51.
- Yao, J., Shi, F., Zhou, Z., Qin, J. (2012) Combinatorial Optimization of Exclusive Bus Lanes and Bus Frequencies in Multi-Modal Transportation Network. *Journal of Transportation Engineering* 138, 1422-1429.
- Yingfeng, W., NaiQi, W. (2010) An approximate algorithm for the Lane Reservation Problem in Time Constrained Transportation. *Proceedings of Advanced Computer Control (ICACC), 2010 2nd International Conference on*, pp. 192-196.
- YunFei, F., Feng, C., Mammar, S., Che, A. (2011) Iterative algorithm for lane reservation problem on transportation network. *Proceedings of Networking, Sensing and Control (ICNSC), 2011 IEEE International Conference on*, pp. 305-310.
- Zhang, G., Chen, J. (2010) Solving multi-class traffic assignment problem with genetic algorithm. *Proceedings of Computational Intelligence and Natural Computing Proceedings (CINC), 2010 Second International Conference on*, pp. 229-232.
- Zhang, H., Gao, Z. (2009) Bilevel programming model and solution method for mixed transportation network design problem. *Journal of Systems Science and Complexity* 22, 446-459.
- Zhang, L., Yang, H., Wu, D., Wang, D. (2014) Solving a discrete multimodal transportation network design problem. *Transportation Research Part C: Emerging Technologies* 49, 73-86.
- Zheng, J., Boyce, D. (2011) Comparison of User-Equilibrium and System-Optimal Route Flow Solutions under Increasing Traffic Congestion, 11-0581. *Proceedings of Transportation Research Board 90th Annual Meeting*.
- Zheng, N., Geroliminis, N. (2013) On the distribution of urban road space for multimodal congested networks. *Transportation Research Part B: Methodological* 57, 326-341.
- Zhong, R.X., Sumalee, A., Friesz, T.L., Lam, W.H.K. (2011) Dynamic user equilibrium with side constraints for a traffic network: Theoretical development and numerical solution algorithm. *Transportation Research Part B: Methodological* 45, 1035-1061.

Zhou, Z., Che, A., Chu, F., Chu, C. (2014) Model and Method for Multiobjective Time-Dependent Hazardous Material Transportation. *Mathematical Problems in Engineering* 2014.

Zhou, Z., Chu, F., Che, A., Mammar, S. (2012) A multi-objective model for the hazardous materials transportation problem based on lane reservation. *Proceedings of Networking, Sensing and Control (ICNSC), 2012 9th IEEE International Conference on*, pp. 328-333.