OPTIMAL LIFE-CYCLE MANAGEMENT OF TRANSPORTATION ASSET SYSTEMS

LE ZHANG

PhD

The Hong Kong Polytechnic University

2019

The Hong Kong Polytechnic University

Department of Electrical Engineering

OPTIMAL LIFE-CYCLE MANAGEMENT OF TRANSPORTATION ASSET SYSTEMS

LE ZHANG

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

August 2018

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

ZHANG LE (Name of student)

Abstract

Transportation assets play a vital role in the national economy and in people's daily life. In this thesis, we develop new models of optimal life-cycle management for two common types of transportation asset systems: highway pavement systems and truck fleet systems. The pavement system models can be tailored for the optimal management of other infrastructure systems, such as bridges and tunnels; and the truck fleet models can be applied with modification to other vehicle asset systems, such as bus, airplane, and maritime vessel fleets.

For the optimal management of highway pavement systems, we first present a novel optimization model with **deterministic** pavement deterioration. The model jointly optimizes the schedules of three common types of pavement management activities (treatments), i.e., preventive maintenance, rehabilitation, and reconstruction/replacement (MR&R), for a system of pavement segments under budget constraints. The objective is to minimize the total costs incurred by both highway users and pavement management agencies. We propose a Lagrange multiplier approach to relax the budget constraints, and employ derivative-free quasi-Newton algorithms to find the optimal solution. By relaxing the budget constraints, the solution approach decomposes the optimization problem for a pavement system into optimization subproblems for each pavement segment. Hence, it can be applied to pavement systems with any models of segment-level cost and treatment effectiveness, as long as the segment-level optimization subproblems are solvable. This approach also ensures a bounded optimality gap for a system-level solution, as long as the segment-level solutions have bounded optimality gaps. (In other words, it guarantees the near-optimality of the system-

level solution if the segment-level subproblems can be solved to near-optimality.) Finally, the approach exhibits linear complexity with the number of pavement segments, which ensures computational efficiency, especially for optimizing large-scale systems.

We further extend this work to the **stochastic** case, which accounts for uncertainties in the pavement deterioration process. Inspection activities whose aim is to reveal the actual pavement state are included in the menu of management activities for scheduling. We formulate a semi-continuous model for selecting optimal inspection scheduling and management policies; and propose a statistical learning approach to update the model parameters sequentially after inspections. We employ approximate dynamic programming to solve the segment-level problem, which has a great advantage in terms of computational efficiency for solving large-size problems. Managerial insights are unveiled in numerical case studies, which can help highway agencies formulate more cost-effective inspection and management policies and budget allocation plans.

The vehicle fleet management problem is different from the infrastructure management problem, mainly because: i) the fleet size can vary over time, while an infrastructure system is usually fixed in size; and ii) a vehicle's maintenance cost and replacement schedule are affected by its utilization, which can be controlled by the fleet manager (the utilization of highway pavement by private vehicles, in contrast, is difficult to control). In this thesis, the truck fleet management problem is formulated as a mixed-integer nonlinear program, which jointly optimizes the purchase, replacement, and utilization policies for the trucks under deterministic, time-varying demand. Our contribution is mainly

methodological. We first approximate the original discrete formulation with a continuous-time one, where the numerous decision variables in the discrete-time model are replaced by a few decision *functions* in continuous time. This continuous approximation technique allows us to derive the analytical conditions for optimal utilization plans using the calculus of variations. These optimality conditions are then converted back to the discrete-time kind, and are used to develop a demand allocation rule, which greatly reduces the solution space of the original problem. The reduced program can then be solved efficiently using heuristic methods, such as tabu search. Based on extensive numerical case studies, we verify that this novel approach can produce better solutions with much lower computational cost than previous solution algorithms proposed in the literature.

The thesis concludes with a discussion on potential extensions under the proposed methodological frameworks to account for more realistic features in real-world transportation asset management problems, and to model other types of infrastructure and fleet asset systems.

Keywords: transportation asset management; pavement system; truck fleet; optimization; Lagrange multiplier method

Acknowledgements

The process of pursuing a doctorate degree is long and arduous, but this has also been the best period in my life. I feel I am a better person after these four years of studies, training and research. I really appreciate the precious guidance and help I have received from many people. Without their support, I would not have been able to complete this thesis.

First and foremost, I want to express my sincere gratitude to my chief supervisor, Dr. Weihua Gu for his valued guidance and encouragement during my past four years at PolyU, and for his patience, understanding and tolerance. I have been greatly impressed by his knowledge and research attitude. He set a good example for me and taught me what is required to be an excellent researcher. Aa a highly responsible supervisor, Dr. Gu has done everything he can to supervise my research and help me improve my writing skills. He has helped me develop my research skills and my scholarliness. He is not just my advisor, but also my life mentor. It has been my honour and good fortune to study under his guidance. My gratitude to him goes beyond words.

I also want to give special thanks to my co-supervisors, Dr. Jinwoo Lee and Professor Weinong Fu. Dr. Lee gave me a lot of inspiring ideas and invaluable feedback during the development of my thesis. This thesis would not be complete without his valuable advice and guidance. He is also a true friend. I am grateful for everything he has done for me. I also want to express my sincere appreciation to and respect for Professor Fu for his genuine caring and help in my life. I still remember the enjoyable and memorable hiking experience with him and his team.

His optimistic attitude to life encourages me. I would like to thank him for helping me not only as an advisor, but also like a family member.

My gratitude also goes to my good friends at PolyU. First, to Larry, for all the discussions we had whenever I encountered bottlenecks in my research. He is like an elder brother and encourages me all the time. The same gratitude goes to Christine, for her concern and friendship. I will never forget the times we encouraged and accompanied each other when adapting to the new environment, especially in our first year. I sincerely appreciate and treasure all the moments we experienced together during the past four years. I am also grateful to Dr. Liangliang Fu for his patience and help during his stay in PolyU. I learned a lot about optimization and programming from him. I also want to thank my fellows, Samuel, Dr. Xin Li and Dr. Wenbo Fan, for their thoughtfulness and friendship. I am lucky to have spent my graduate school life with them.

Last, but not least, I would like to thank my dear parents, for their selfless and endless love, for their unconditional support and encouragement, and for always believing in me and encouraging me to do whatever I enjoy.

Table of contents

Chapter 1. Introduction

1.1 Background

Transportation asset systems consist of a wide variety of physical infrastructure (e.g., roads, bridges, railways, pipelines, ports and tunnels) and vehicles (e.g., trucks, buses, ships and airplanes), and play a vital role in the global economy and people's daily lives. The healthy operation of transportation asset systems is key to satisfying consumer demand for mobility, safety, comfort, and prosperity. For example, the United States alone has over 4 million miles of roads, which provides vital links among communities, cities and ports. In 2015, the roads recorded over 3 trillion vehicle-miles (CBO, 2016; ASCE, 2017). The trucking sector in the US transported 10.42 billion tons of freight and generated $676.2 billion in revenue in the same year, accounting for 70.6% of domestic freight tonnage and 79.8% of the nation's freight billings. Because of the importance of transportation asset systems to the economy, it is vital to manage them effectively to sustain a desirable level of service in the ever-changing transportation business environment. However, although capital investment in transportation assets should be increased to meet the ever-growing demand, in the US, investment has consistently risen slower than that needed (ASCE, 2017). This has had undesirable adverse outcomes. For example, because of mismanagement caused by budgetary concerns, the I-35W bridge in Minnesota collapsed on August 1, 2017, causing 13 deaths and 145 serious injuries (Gray Plant Moody, 2008). These casualties made it Minnesota's worst bridge accident in the history and one of the nation's worst as well. During the past few decades, disasters like this have resulted in increasing awareness of the importance of optimal life-cycle management of transportation asset systems.

Transportation asset management, broadly defined, is a strategic and systematic process of operating, maintaining, upgrading, and expanding physical assets effectively throughout their life cycles (Transportation Officials, 2011). Take transportation infrastructure as an example, increasing use creates ongoing deterioration and aging, and the deteriorating assets, in turn, incur higher costs, in terms of vehicle repair, traffic congestion, and extra fuel consumption and emissions, among others. To sustain the integrity and reliability of asset systems and extend their service lives, asset renewal, maintenance and preservation is a critical part of the responsibility of infrastructure management agencies. Specifically, the management agencies of infrastructure systems seek to optimally plan maintenance, rehabilitation, and reconstruction (MR&R) activities with a limited budget, so that well-defined goals and objectives are achieved efficiently and effectively. For vehicle fleets, management seeks optimal strategies regarding when to purchase new vehicles and retire old ones, and how to operate and maintain the fleet to meet demand at minimum cost. However, traditional asset management models often exhibit significant limitations: e.g., the use of oversimplified, sometimes unrealistic assumptions, and the lack of computationally efficient solutions. These limitations are mainly due to the complexities in life-cycle cost optimization models for various types of assets. Meanwhile, the multifarious uncertainties that arise in actual practice (e.g. uncertainties in the asset deterioration process, demand forecasting and model calibration) are even more difficult to incorporate into cost optimization. Therefore, there is an urgent need for better models and solutions for the optimal management of transportation assets.

The aim of this thesis is to advance the present research frontiers by developing more general and realistic models and more efficient solutions for jointly optimizing the schedules of various management-related activities (inspection, maintenance, replacement, and utilization) for large-scale transportation asset systems. Specifically, we develop new models of optimal life-cycle management for two of the most common and important types of transportation asset systems: highway pavement systems and truck fleet systems. Moreover, our pavement system models can be tailored for the optimal management of other infrastructure systems, such as bridges and tunnels; and our truck fleet models can be modified to optimize other vehicle fleet systems, such as bus, airplane and maritime vessel fleets. We next examine the strengths and deficiencies in the literature in the realm of transportation asset management.

1.2 Literature review

1.2.1 Transportation asset management problems

The objective of transportation asset management is to identify the optimal management policy for an asset system over a given planning horizon. In the literature, it is often formulated as a resource allocation problem, where the resource may refer to money, equipment or materials. Resource allocation problems focus on the allocation of limited resources among competing activities with the intention of optimizing a well-defined objective function (Luss, 2012). Starting from the seminal paper on resource allocation by Koopman (Koopman, 1953), researchers have studied various problems in this realm (e.g. Thomas, 1990; Bretthauter and Shetty, 1995; Powell et al., 2002). Specific to the asset management problem, the operation of the assets in a system (whether they are

homogeneous or heterogeneous) is often intercorrelated due to the economic interdependence of the assets. This economic interdependence can be expressed as one of the following constraints: i) joint capital budget constraints, meaning that the total agency cost for managing the asset system is capped by a financial budget (Karabakal et al., 1994; Lee and Madanat, 2014a, b); or ii) joint demand constraints, meaning that a given total demand has to be satisfied by the assets jointly (Vander Veen, 1985).

In the literature on management science, a class of problems with economic interdependence constraints are defined as *parallel replacement problems*. (Note that replacement/reconstruction is the most common type of management activities studied in the literature.) This class of problems have been studied for many different types of assets, including textile machines (Williamson, 1971), farm tractors (Chisholm, 1974; Reid and Bradford, 1983), trucks (Ahmed, 1973; Guerrero et al., 2013), buses (Keles and Hartman, 2004), ships (Evans, 1989; Nicholson and Pullen, 1971; Wijsmuller and Beumee, 1979), escalators (Scarf et al., 2007) and pavements (Lee and Madanat, 2015; Zhang et al., 2017). The classic asset replacement problem involves determining when to replace existing assets by new assets (thus the size of the asset system is fixed) to minimize the total discounted cost or maximize profit over a given planning horizon, which can be either finite (Oakford et al., 1984) or infinite (Bean et al., 1984). Management of the two common types of transportation assets, i.e. highway infrastructure and vehicle fleets, is more complicated. The complication in the infrastructure systems is created by the variety of treatment options that have different effects on the assets. In addition to reconstruction/replacement, one can also apply preventive maintenance, which is a low-cost option to prevent or slow down infrastructure

deterioration, and rehabilitation, which is a medium-cost option to restore the infrastructure's serviceability (while deterioration will still grow faster in the future). For vehicle fleet systems, the major difference from infrastructure systems lies in the fact that the fleet size can vary over time in response to changing market demand, and that vehicle utilization can be jointly optimized to account for its impact on the vehicles' maintenance and replacement schedules. We next review the literature on the management of highway pavement systems and vehicle fleets separately.

1.2.1.1 Studies on pavement management

Studies in the area of pavement management commenced by optimizing only the planning of rehabilitation activities for a single pavement segment (Friesz and Fernandez, 1979; Fernandez and Friesz, 1981; Markow and Balta, 1985). Segment-level models are important because they formed the basis of the models developed for a system of pavement segments (i.e. the system-level models), which is the focus of this thesis. A variety of segment-level optimization models were subsequently developed, which are characterized by the pavement deterioration process (memoryless or history-dependent; deterministic or stochastic), the number of treatments, and whether the time and/or pavement states were discrete or continuous variables. Table 1.1 summarizes the modelling features and solution approaches of selected segment-level studies. Note that the table shows a general trend of evolution from simpler models (deterministic with a memoryless deterioration process, single treatment, discrete variables) to more complicated but realistic ones (stochastic with a history-dependent deterioration process, multiple treatments, continuous variables). This is partly due to the

development of more sophisticated approaches to finding globally optimal solutions: e.g. the calculus of variations (Ouyang and Madanat, 2006; Lee and Madanat, 2014b). The most complicated (and realistic) segment-level model so far seems to be that of Lee and Madanat (2014a), which optimized the planning of all three treatments (preventive maintenance, rehabilitation and reconstruction) with a history-dependent deterioration process. However, the preventive maintenance effectiveness model used in segment-level MR&R optimization are unrealistic. For example, the maintenance model used by Gu et al. (2012) was hypothesized with ungrounded parameter values. As a result, the optimal MR&R plan obtained by Lee and Madanat (2014a) showed that a greater deterioration rate reduction occurred when maintenance was applied to a pavement near the end of its expected lifecycle (see Figure 4a of the cited work), which contradicts with the common understanding in practice.

In reality, however, a highway agency often manages hundreds of pavement segments. Thus, they are more interested in models that can jointly optimize for a system of pavement segments under certain budget constraints, which can be incorporated into their pavement management systems. However, system-level problems are by nature more complicated than segment-level ones. This is why a smaller number of studies are found in this category, including some works that relied on the highly idealized "top-down" approaches (Kuhn and Madanat, 2005; Durango-Cohen and Sarutipand, 2007). These top-down models assume homogeneous pavement segments in a system, and are thus unrealistic and unsuitable for real-world implementation. The more realistic, "bottom-up" approaches that appreciate the heterogeneity in pavement segments have also been

applied to system-level MR&R planning optimization. A number of selected bottom-up studies are summarized in Table 1.2.

Table 1.1. Selected studies on segment-level optimization of MR&R planning

Study	Deterioration process	Number of treatments	Discrete/Continuous time or pavement state	Solution approach
Golabi et al. (1982)	memoryless deterministic	single	discrete	linear programming
Carnahan et al. (1987)	memoryless deterministic	single	discrete	dynamic programming
Fwa et al. (1994)	memoryless deterministic	single	discrete	genetic algorithm
Friesz and Fernandez (1979)	memoryless deterministic	single	continuous	optimal control
Fernandez and Friesz (1981)	memoryless deterministic	single	continuous	optimal control
Tsunokawa and Schofer (1994)	memoryless deterministic	single	continuous	optimal control with trend curve approximation
Li and Madanat (2002)	memoryless deterministic	single	continuous	using the memoryless property
Ouyang and Madanat (2006)	memoryless deterministic	single	continuous	calculus of variation
Gu et al. (2012)	memoryless deterministic	multiple	continuous	numerical method based on the results of Ouyang and Madanat (2006)
Rashid and Tsunokawa (2012)	memoryless deterministic	multiple	continuous	optimal control with trend curve approximation
Madanat (1993)	memoryless stochastic	multiple	discrete	dynamic programming
Madanat and Ben-Akiva (1994)	memoryless stochastic	multiple	discrete	dynamic programming
Tsunokawa et al. (2006)	history-dependent deterministic	single	discrete	gradient search
Bai et al. (2015)	history-dependent deterministic	single	hybrid	dynamic programming
Miyamoto et al. (2000)	history-dependent deterministic	multiple	discrete	genetic algorithm
Lee and Madanat (2014a)	history-dependent deterministic	multiple	hybrid	dynamic programming
Lee and Madanat (2014b)	history-dependent deterministic	multiple	continuous	calculus of variation

Table 1.2. Selected studies on bottom-up system-level optimization of MR&R planning

Study	Deterioration process	Number of treatments	Discrete/Continuous time or pavement state	Solution approach
Chan et al. (1994)	memoryless deterministic	single	discrete	genetic algorithm
Ouyang and Madanat (2004)	memoryless deterministic	single	hybrid	branch and bound; greedy heuristic
Ouyang (2007)	memoryless deterministic	single	hybrid	approximate dynamic programming
Hajibabai et al. (2014)	memoryless deterministic	single	hybrid	Lagrangian relaxation
Sathaye and Madanat (2011)	memoryless deterministic	single	continuous	Lagrange method
Sathaye and Madanat (2012)	memoryless deterministic	single	continuous	Lagrange dual method
Fwa et al. (1996)	memoryless deterministic	multiple	discrete	genetic algorithm
Durango-Cohen and Madanat (2002)	memoryless stochastic	multiple	discrete	dynamic programming and adaptive control
Ohlmann and Bean (2009)	memoryless stochastic	multiple	discrete	Lagrangian relaxation
Yeo et al. (2013)	memoryless stochastic	multiple	discrete	mateheuristic
Medury and Madanat (2013)	memoryless stochastic	multiple	discrete	approximate dynamic programming
Chu and Chen (2012)	history-dependent deterministic	multiple	hybrid	tabu search
Lee and Madanat (2015)	history-dependent deterministic	multiple	hybrid	genetic algorithm
Lee et al. (2016)	history-dependent deterministic	multiple	discrete	Lagrange dual method

As the tables show, most studies in the literature have focused on deterministic formulations; while only few models address the intrinsic uncertainties in pavement deterioration and management. This is because stochastic models are by nature much more difficult to solve, especially for system-level problems. However, real pavement deterioration cannot be accurately predicted because of multiple stochastic factors, including pavement utilization levels (traffic loadings), environmental conditions, and limited inspections. The deterministic models may

therefore result in inappropriate MR&R policies and higher total lifecycle costs for the infrastructure system. Like the deterministic models, the stochastic models can be classified depending on whether the pavements in the system are homogeneous or heterogeneous. The former class of problems were usually solved by top-down approaches, and the solutions feature a common probability distribution of the actions applied to each segment. Hence, the individual segments are not distinguishable and no segment-specific recommendations can be derived. To better capture segment-level specifics, heterogeneous systems with the stochastic deterioration model have been studied, and more realistic, *deterministic* (nonrandomized) and segment-specific policies have been developed via bottom-up approaches (Ohlmann and Bean, 2009; Yeo et al., 2013; Medury and Madanat, 2013).

1.2.1.2 Studies on vehicle fleet management

Classical studies on vehicle fleet management have focused on optimizing vehicle replacement schedules under given utilization plans or constant utilization levels for each vehicle (e.g. Karabakal et al., 2000; Redmer, 2009; Parthanadee et al., 2012). However, when multiple vehicles jointly serve a given demand, the individual vehicle's utilization plan (or the plan of demand allocation among the vehicles) should be jointly optimized with the replacement schedule, because a vehicle's utilization will affect its state, and in turn affect its maintenance cost and retirement schedule. For example, a well-known fact is that when a vehicle's cumulative mileage reaches a certain level, maintenance becomes too expensive, so the vehicle has to be retired or replaced (Drinkeater and Hastings, 1967; Ghellinck and Eppen, 1967). Meanwhile, the purchase of new vehicles and the

retirement of old vehicles may be planned at different times, so that the fleet size can vary over time. However, the joint optimization of vehicle (or more generally, asset) purchase, retirement, and utilization schedules greatly increases the size of the solution space, and thus oftentimes renders the optimization problem intractable. This explains why only a handful of studies have jointly optimized asset replacement and utilization, including Vander Veen (1985), and Vander Veen and Jordan (1989) for the management of machines, Hartman (1999) for general assets, Jin and Kite-Powell (2000) for ships, and Guerrero et al. (2013) for trucks.

Although deterministic optimization models have been formulated in the majority of the literature, there are also some works that have incorporated uncertainties in vehicle management. Stochastic models can capture a variety of random factors, for example in demand (Hartman, 2004; List et al., 2003), operating conditions (List et al., 2003), and occasional vehicle breakdowns (Stasko and Gao, 2012; Childress and Durango-Cohen, 2005). But efficient solution approaches to large-size stochastic models are still missing.

1.2.2 Solution methods

Early versions of resource allocation problems were formulated as linear programs (Lasdon, 2002; Dantzig, 2016). However, the general forms of resource allocation problems are usually nonlinear, and there is no general solution method. Various exact solution methods have been developed for resource allocation models with special mathematical structures. These methods include dynamic programming (Cooper 1980; Morin and Marsten, 1976b), branch and bound (Gupta and Ravindran, 1985), a combination of these two (Marsten and Morin, 1978; Morin and Marsten, 1976a), and so on. For example, Bretthauer and Shetty (1995)

developed a branch-and-bound algorithm for a set of resource allocation problems, whose formulation is convex. A commonly used technique for solving resource allocation problems is the Lagrange multiplier method (Patriksson, 2008). This method can be used to relax some constraints and produce a Lagrangian problem that is often much easier to solve. A solution to the relaxed problem is a lower bound of the solution to the original problem (assuming that the original problem is a minimization one). The bound can then be used in branch-and-bound algorithms for finding the exact solution. There is also a special type of resource allocation problem, termed the weakly coupled dynamic problem (Hawkins, 2003; Adelman and Mersereau, 2008). A problem of this type consists of multiple subproblems that are independent of each other, except for a set of constraints linking the decision variables of each subproblem. Thus, the problem can be decoupled into lower dimensional subproblems by relaxing these linking constraints through the Lagrange method, and the solution time will be greatly reduced in consequence. In the past decades, the weakly coupled dynamic problems found its applications in many areas, including asset management (Adelman and Mersereau, 2008; Lee et al., 2016; Lee and Madanat, 2017; Zhang et al., 2018), restless bandit problems (Whittle, 1988), online marketing (Bertsimas and Mersereau, 2007), and supply chain management (Hawkins, 2003). The solution technique to the weakly coupled dynamic problems is also used in this thesis.

Various methods have also been proposed to solve parallel replacement problems. For example, Karabakal et al. (1994) modeled a parallel asset replacement problem under budget constraints as a 0-1 integer program and developed a branch-and-bound solution algorithm integrated with the Lagrangian relaxation approach.

However, only medium-sized problems can be solved, and an exact solution is not always attained. For large, real-sized instances of problems, only heuristic solutions have been proposed (Karabakal et al., 2000). Others have sought to develop analytical properties of the problem to reduce the size of the solution space (e.g. Jones et al., 1991; Childress and Durango-Cohen, 2005). Unfortunately, these useful analytical properties are often built upon idealized or limited assumptions. As a result, the general applicability of these properties is questionable.

To include uncertainties in the model, Markov decision process (MDP)-based methods have been widely used in pavement management, where the pavement deterioration process and management performance are characterized by a state transition matrix (Madanat, 1993; Durango-Cohen, 2004; Kun and Madanat, 2005; Madanat et al., 2006). Depending on whether the pavements in the system are homogeneous or heterogeneous, the problems can be classified as single- or multi-dimensional MDP problems. Problems in the former class can be formulated as a linear program (LP), whose solution can be obtained in polynomial runtimes (Smilowitz and Madanat, 2000; Kun and Madanat, 2005; Madanat et al., 2006). The LP-based approach provides a randomized optimal policy to accommodate budget constraints when making MR&R decisions (Beutler and Ross, 1985; Ross and Varadarajan, 1989, 1991; Feinberg and Shwartz, 1996, 1999; Smilowitz and Madanat, 2000), wherein the optimal action for a segment in a given state is defined by a probability distribution of a set of actions. Multi-dimensional MDP models, in contrast, focus on more realistic, deterministic policies (Ohlmann and Bean, 2009; Yeo et al., 2013; Medury and Madanat, 2013), and most of them rely on solution techniques for the weakly coupled dynamic problems we mentioned above. For the vehicle fleet management problem, the uncertainty may come from

market demand (Hartman, 2004; List et al., 2003), operating conditions (List et al., 2003), and occasional vehicle breakdowns (Stasko and Gao, 2012; Childress and Durango-Cohen, 2005). The resulting optimization problems under uncertainty have been solved by stochastic (approximate) dynamic programming (Stasko and Gao, 2012; Harman, 2004), two-stage robust optimization (List et al., 2003), fuzzy set theory (Usher and Whitfield, 1993; Chang, 2005) and the minmax approach (Tan and Hartman, 2010).

Still, the solution methods proposed in the literature have a number of limitations. The limitations in solution methods for pavement management problems are summarized as follows (see Tables 1.1 and 1.2 for the solution approaches used in representative studies in this area):

i) Many studies relied on metaheuristic methods (e.g. genetic algorithm and tabu search) to solve the complicated optimization models. Metaheuristic methods cannot guarantee the global optimality of the solution (Blum and Roli, 2003). Moreover, the studies listed in Table 1.2 often cannot assess how close the heuristic is to the global optimum. Other works sought to optimize Lagrangian and Lagrangian dual functions of their original problems (Sathaye and Madanat, 2011, 2012; Lee et al., 2016; Lee and Madanat, 2017). However, the effectiveness of their solution approaches is contingent on the convexity of the problem formulation.

ii) For most of the studies cited in Table 1.2, their solution approaches are highly dependent upon the segment-level empirical models; i.e., they cannot be directly applied to another system-level problem with different segment-level models. This is undesirable since there are many variants of

segment-level models (see Table 1.1), and new empirical segment-level models may arise in the future to replace the present ones.

iii) Most of the studies optimized only one or two treatments jointly, i.e., rehabilitation and/or reconstruction, because their approaches are insufficient to find optimal solutions within an acceptable computation time when more treatments are considered. For example, incorporating preventive maintenance into the optimal MR&R planning would add a great deal of complexity to the problem.

iv) Operational uncertainties were often ignored, and as a result, the scheduling of inspection activities was not incorporated into the MR&R decision-making process. This is because accounting for uncertainties and jointly optimizing the inspection schedules will greatly increase the size of the problem, and thus render the optimization problem much more difficult to solve, especially at the system level. At present, only a handful of segment-level and small-scale system-level studies have incorporated inspection decisions in pavement management (Madanat, 1993; Madanat and Ben-akiva, 1994; Durango-Cohen and Madanat, 2002).

The limitations in the literature on vehicle fleet management problems are summarized as follows:

i) Utilization levels were often assumed to be exogenous, and the asset size was often assumed to be fixed, i.e., a retired vehicle is always immediately replaced by a new one (Tang and Tang, 1993; Hopp et al., 1993; McClurg and Chand, 2002). Thus, the operating conditions with time-varying demand cannot be modelled properly.

ii) The cost models are often over-simplified to make the optimization problem solvable. Most existing studies assume that a vehicle's state is a function of its age (Simms et al., 1984; Parthanadee et al., 2012). However, empirical evidence has shown that the vehicles' deterioration is more a function of its utilization (i.e. cumulative mileage) than age (CARB, 2008a; Guerrero et al., 2013).

iii) The solution methods in the literature often exhibit poor computational scalability or unreliable solution quality.

1.3 Summary of research contributions of the thesis

Chapter 2 examines the joint optimization of maintenance and replacement schedules for pavement systems in both deterministic and stochastic scenarios. In the deterministic scenario, a general formulation of the system-level pavement management problem is proposed, which is independent of any specific segment-level models. A general solution approach is also developed to decompose the system-level problem into a number of segment-level subproblems. This is done by relaxing the budget constraints via Lagrange multipliers. The optimization program is then converted to a bi-level one, where the segment-level subproblems are solved for given Lagrange multiplier values at the lower level by model-specific algorithms, and the Lagrange multiplier values are found at the upper level. We show that when only a total budget constraint is applied, global optimality is retained at the system level via certain derivative-free iterative methods; i.e., if the segment-level subproblems are solved at or near optimality, then the global optimality or near-optimality of the system-level problem is guaranteed. Next, in the stochastic scenario, we present a pavement management framework for

selecting optimal inspection and MR&R policies jointly under model uncertainty. The model specifies that the inspection decision is made contingent on the pavement condition (i.e., a condition-based inspection scheme). Meanwhile, we use a statistical learning approach to improve the accuracy of the deterioration models during the planning horizon, wherein the parameters of deterioration models are updated sequentially using the inspection results. Consequently, the model uncertainty will gradually decrease as the size of inspection data increases, and how fast the uncertainty decreases is traded off with the inspection cost in a comprehensive optimization framework.

Chapter 3 develops an efficient solution approach for a general truck fleet management model, which is a generalization of the trucking sector optimization model proposed by Guerrero et al. (2013). The model specifies that the unit maintenance cost of a truck per mile of travel is a function of the truck's cumulative mileage (CARB, 2008) and considers multiple truck types with various fuel-saving technologies (i.e., more fuel-efficient trucks have a lower unit operating cost per mile, but a higher purchase cost). To solve the problem, we employ a continuous approximation (CA) of the original discrete model, for which some optimality conditions are derived. The derived optimality conditions greatly reduce the dimensions of the problem. The reduced problem can then be solved by some metaheuristic algorithms (in this thesis the tabu search method is used) more efficiently. A comparison with commercial solvers shows that the CA approach can reduce the computation time by 98% without compromising solution quality. A further comparison with a recent study (Guerrero et al., 2013) shows that the solution obtained by the new approach can reduce the total cost by an additional 13-21%.

Chapter 4 concludes the thesis by summarizing the key contributions and limitations. Potential research opportunities that build upon the present work are also discussed.

Chapter 2. Joint optimization of maintenance and replacement schedules for pavement systems

In this chapter, we present the optimal management policies for pavement systems in both deterministic and stochastic scenarios. They are presented in Sections 2.1 and 2.2, respectively. In the deterministic scenario, a general bottom-up model for the joint optimization of MR&R schedules under budget constraints is developed and solved. In the stochastic scenario, a joint optimization of inspection schedules and MR&R plans is formulated for large-scale pavement systems under model uncertainty, wherein the evolution of pavements is random due to uncontrolled utilization levels, environmental conditions, and limited empirical observations. Numerical case studies are discussed at the end of the above two sections. Section 2.3 presents a summary of the findings.

2.1 Joint optimization of preventive maintenance, rehabilitation and reconstruction planning for pavement systems in the deterministic scenario

In this section we develop a computationally efficient and non-problem-specific approach to find globally optimal or near-optimal MR&R policies for large-scale pavement systems. The section is organized as follows. Section 2.1.1 presents the general formulation of the system-level problem (i.e., the upper-level problem) and a general solution approach. A specific segment-level (i.e. the lower-level) model and its solution are described in Section 2.1.2. Section 2.1.3 provides numerical case studies.

2.1.1 A general system-level MR&R model and its solution approach

A general formulation of the system-level MR&R planning problem, regardless of the segment-level models, is presented in Section 2.1.1.1. A derivative-free iterative solution approach built upon the Lagrange multiplier method is described in Section 2.1.1.2. The description of the solution approach assumes that a solution of the segment-level problem (an example of which is presented in Section 2.1.2) is ready for use.

2.1.1.1 A general formulation

The objective of the problem is to minimize the sum of the discounted user and agency costs, $\sum_{k=1}^{K} Z_k$, for all the pavement segments $k \in \{1,2,\dots,K\}$ over a given planning horizon T ($T = \infty$ denotes an infinite-horizon problem), as shown in (2.1a) below. For each segment k, Z_k is a function of a vector of state variables (e.g. pavement roughness level and age), denoted by $\boldsymbol{q}_k$, and a vector of management decision variables (e.g., timing and intensity of MR&R activities), $\boldsymbol{x}_k$. Note that the elements $\boldsymbol{q}_k$ and $\boldsymbol{x}_k$ can be discrete or continuous functions of time. The Z_k consists of the costs incurred by the users, C_k^U, and by the management agency, $\sum_{p=1}^{P} C_{kp}$, where $p \in \{1,\dots,P\}$ is the index of a planned treatment (i.e., preventive maintenance, rehabilitation, or reconstruction).

Segment-specific constraints are divided into two classes: equality constraints (2.1b) and inequality constraints (2.1c), where $\boldsymbol{\Phi}_k$ and $\boldsymbol{\Psi}_k$ are again vectors of discrete or continuous functions of time. These constraints specify the pavements' initial conditions, how each pavement's state evolves over time (i.e. the deterioration process), and how each treatment may change the pavement's state,

depending on the type, time and intensity of the treatment (i.e. the treatment effectiveness models). Finally, we present two versions of budget constraints in (2.1d-e): i) a combined budget, which applies to the sum of agency costs for all treatments across all segments, and ii) a number of separate budgets that each applies to a specific treatment. B and B_p denote the annual combined budget and separate budget for treatment p, respectively; and r is the annual discount factor. Note here that we assume the budget can be transferred across years over the entire planning horizon so that the number of budget constraints is small (1 for the combined-budget-constraint case and 3 for the separate-budget-constraint case). Similar assumptions were adopted in a number of previous studies (e.g. Sathaye and Madanat, 2011, 2012). Our system-level approach, however, can be applied to a more general case, in which the money can be transferred only within a given budget period (e.g., of 5-10 years). Also note that the total budgets in (2.1d) and (2.1e) are discounted to the present. This is done in the interest of simplicity for model formulation and the segment-level solution procedure, because now the agency cost terms in the Lagrangian (after relaxing the budget constraints) can be easily combined, as we see below.

$$\min \sum_{k=1}^{K} Z_k(\boldsymbol{q}_k, \boldsymbol{x}_k) = \sum_{k=1}^{K} \left(C_k^U(\boldsymbol{q}_k, \boldsymbol{x}_k) + \sum_{p=1}^{P} C_{kp}(\boldsymbol{q}_k, \boldsymbol{x}_k) \right) \quad (2.1a)$$

$$\text{subject to: } \boldsymbol{\Phi}_k(\boldsymbol{q}_k, \boldsymbol{x}_k) = 0, \text{for } k = 1, \dots, K \quad (2.1b)$$

$$\boldsymbol{\Psi}_k(\boldsymbol{q}_k, \boldsymbol{x}_k) \leq 0, \text{for } k = 1, \dots, K \quad (2.1c)$$

$$\begin{cases} \text{combined budget:} & \dfrac{r}{1 - e^{-rT}} \displaystyle\sum_{k=1}^{K} \sum_{p=1}^{P} C_{kp}(\boldsymbol{q}_k, \boldsymbol{x}_k) \leq B & (2.1d) \\ \text{separate budgets:} & \dfrac{r}{1 - e^{-rT}} \displaystyle\sum_{k=1}^{K} C_{kp}(\boldsymbol{q}_k, \boldsymbol{x}_k) \leq B_p \text{ for } p = 1, \cdots, P & (2.1e) \end{cases}$$

We next present an iterative approach for solving this mathematical program

2.1.1.2 An iterative approach using Lagrange multipliers

Corresponding to the above system-level formulation, a general formulation of the segment-level problems is given in (2.2a-c). In the following discussion, we assume that the solution to this segment-level problem has been developed a priori. This segment-level solution is used as a building block in our proposed approach.

For each $k = 1, \dots, K$

$$\min\ Z_k(\boldsymbol{q}_k, \boldsymbol{x}_k) = C_k^U(\boldsymbol{q}_k, \boldsymbol{x}_k) + \sum_{p=1}^{P} C_{kp}(\boldsymbol{q}_k, \boldsymbol{x}_k) \quad (2.2a)$$

$$\text{subject to: } \boldsymbol{\Phi}_k(\boldsymbol{q}_k, \boldsymbol{x}_k) = 0 \quad (2.2b)$$

$$\boldsymbol{\Psi}_k(\boldsymbol{q}_k, \boldsymbol{x}_k) \leq 0 \quad (2.2c)$$

To be accurate, we describe the solution approaches for problems with the combined budget constraint and separate budget constraints one by one. However, they follow the same logic: first, the system-level problem is decomposed into K segment-level subproblems, each having the form of (2.6a-c); and second, built upon the solutions to the segment-level subproblems, a gradient-free iterative algorithm is used to solve the system-level optimization problem.

We first introduce a Lagrange multiplier, λ, to relax the combined budget constraint (2.1d). The relaxed optimization is presented as follows:

$$\min\ L(\boldsymbol{q}, \boldsymbol{x}, \lambda) = \sum_{k=1}^{K} Z_k(\boldsymbol{q}_k, \boldsymbol{x}_k) + \lambda \left(\sum_{k=1}^{K} \sum_{p=1}^{P} C_{kp}(\boldsymbol{q}_k, \boldsymbol{x}_k) - \frac{B}{r}(1 - e^{-rT}) \right) = \sum_{k=1}^{K} H_k(\boldsymbol{q}_k, \boldsymbol{x}_k, \lambda) - \lambda \frac{B}{r}(1 - e^{-rT}) \quad (2.3a)$$

$$\text{subject to: } \boldsymbol{\Phi}_k(\boldsymbol{q}_k, \boldsymbol{x}_k) = 0, \text{ for } k = 1, \dots, K \quad (2.3b)$$

$$\boldsymbol{\Psi}_k(\boldsymbol{q}_k, \boldsymbol{x}_k) \leq 0, \text{ for } k = 1, \dots, K \quad (2.3c)$$

$$\begin{cases} \text{either } \lambda = 0 \text{ and } V(0) \equiv V(0|\boldsymbol{q}_k^*, \boldsymbol{x}_k^*) = \sum_{k=1}^{K} \sum_{p=1}^{P} C_{kp}(\boldsymbol{q}_k, \boldsymbol{x}_k) - \frac{B}{r}(1 - e^{-rT}) \leq 0 \\ \text{or } \lambda > 0 \text{ and } V(\lambda) \equiv V(\lambda|\boldsymbol{q}_k^*, \boldsymbol{x}_k^*) = \sum_{k=1}^{K} \sum_{p=1}^{P} C_{kp}(\boldsymbol{q}_k, \boldsymbol{x}_k) - \frac{B}{r}(1 - e^{-rT}) = 0 \end{cases} \quad (2.3d)$$

where L is the partial Lagrange function, and $H_k(\boldsymbol{q}_k, \boldsymbol{x}_k, \lambda) \equiv Z_k(\boldsymbol{q}_k, \boldsymbol{x}_k) + \lambda \sum_{p=1}^{P} C_{kp}(\boldsymbol{q}_k, \boldsymbol{x}_k)$. The constraint (2.3d) is the complementary slackness condition of optimality: $\lambda > 0$ when the budget constraint is binding, and $\lambda = 0$ otherwise. One can easily verify that the optimal solution of (2.3a-d) is always optimal to (2.1a-d); i.e., the relaxed program (2.3a-d) constructs a sufficient condition for the optimality of (2.1a-d).

Without constraint (2.3d), the remaining mathematical program (2.3a-c) can be broken down by segment number k as follows:

For each $k = 1, \dots, K$,

$$\min H_k(\boldsymbol{q}_k, \boldsymbol{x}_k, \lambda) = Z_k(\boldsymbol{q}_k, \boldsymbol{x}_k) + \lambda \sum_{p=1}^{P} C_{kp}(\boldsymbol{q}_k, \boldsymbol{x}_k)$$
$$= C_k^U(\boldsymbol{q}_k, \boldsymbol{x}_k) + (1+\lambda) \sum_{p=1}^{P} C_{kp}(\boldsymbol{q}_k, \boldsymbol{x}_k)$$
$$= C_k^U(\boldsymbol{q}_k, \boldsymbol{x}_k) + \sum_{p=1}^{P} \bar{C}_{kp}(\boldsymbol{q}_k, \boldsymbol{x}_k, \lambda) \quad (2.4a)$$

$$\text{subject to: } \boldsymbol{\Phi}_k(\boldsymbol{q}_k, \boldsymbol{x}_k) = 0 \quad (2.4b)$$

$$\boldsymbol{\Psi}_k(\boldsymbol{q}_k, \boldsymbol{x}_k) \leq 0 \quad (2.4c)$$

where $\bar{C}_{kp}(\boldsymbol{q}_k, \boldsymbol{x}_k, \lambda) = (1+\lambda) C_{kp}(\boldsymbol{q}_k, \boldsymbol{x}_k)$ can be considered as a "weighted" agency cost for treatment p applied to segment k (where the weight is $1+\lambda$). Note that for a given λ, H_k has the same form as Z_k save for only a different weight for agency costs. Thus, the solution to the segment-level problem (2.2a-c) can be readily applied to (2.4a-c) for each k with a given λ. Also note that if the global optimality of segment-level solutions is guaranteed, then the global optimality of the system-level problem is attained if a λ is found to satisfy the complementary slackness condition (2.3d). Further, the following lemma ensures that if the segment-level solution is near-optimal (i.e., its relative cost gap from the optimal solution is bounded by a small fraction), then the resulting system-level solution is also near-optimal.

Lemma 2.1. For a given λ, assume $\boldsymbol{x}_k^*(\lambda)$ is the exact optimal solution to the subproblem of segment k ($k = 1,2,\cdots,K$), and $\boldsymbol{x}_k^H(\lambda)$ is a heuristic solution that satisfies:

$$\begin{cases} \left|C_k\left(\boldsymbol{x}_k^*(\lambda)\right) - C_k\left(\boldsymbol{x}_k^H(\lambda)\right)\right| \le \delta_1 \\ \left|Z_k\left(\boldsymbol{x}_k^*(\lambda)\right) - Z_k\left(\boldsymbol{x}_k^H(\lambda)\right)\right| \le \delta_2 \end{cases}, \forall k = 1,2,\cdots,K, \lambda \ge 0 \quad (2.5)$$

where $C_k(\boldsymbol{x}_k) = \sum_{p=1}^{P} C_{kp}(\boldsymbol{q}_k, \boldsymbol{x}_k)$. Further assume λ^* and λ^H are the Lagrange multiplier values when the exact and heuristic solutions are used, respectively; i.e.,

$$\lambda^* \cdot \left(\sum_{k=1}^{K} C_k\left(\boldsymbol{x}_k^*(\lambda^*)\right) - B\right) = 0 \quad (2.6a)$$

$$\lambda^H \cdot \left(\sum_{k=1}^{K} C_k\left(\boldsymbol{x}_k^H(\lambda^H)\right) - B\right) = 0 \quad (2.6b)$$

Then we have:

$$\left|\sum_{k=1}^{K} Z_k\left(\boldsymbol{x}_k^*(\lambda^*)\right) - \sum_{k=1}^{K} Z_k\left(\boldsymbol{x}_k^H(\lambda^H)\right)\right| \le K \cdot (\max\{\lambda^*, \lambda^H\}\,\delta_1 + \delta_2) \quad (2.7)$$

A sketched proof of Lemma 2.1 is furnished in Appendix B. Note (2.7) ensures that the percentage cost gap of the system-level problem is in the same magnitude of the percentage cost gaps of the segment-level heuristics, given that λ^* and λ^H are small.[1]

Finally, the following lemma specifies that as long as such a λ exists, we can always find it using an appropriately designed Quasi-Newton algorithm. The proof of this lemma is furnished in Appendix C.

[1] In our numerical case studies, λ^H is always less than 3. The λ^* is usually comparable to λ^H since $|V(\lambda^H) - V(\lambda^*)| \le K\delta_1$. Exceptions may arise only when B is near the maximum annual budget needed, where $V(\lambda)$ becomes flat.

Lemma 2.2. $V(\lambda)$ is a (strictly) decreasing function of λ if each segment-level problem furnishes a unique optimal solution (which is usually true).

An immediate corollary of this lemma is that there exists a unique solution of λ to (2.3d) (as long as program (2.3a-d) is feasible), and this solution can be attained by a number of iterative methods, including Newton and quasi-Newton methods (which presumably converge much faster than the methods of bisection, golden-section, etc.). Since the calculation of derivatives is often difficult and computationally inefficient due to the complicated mathematical forms of MR&R cost and effectiveness models, we next present an algorithm using a derivative-free method (termed the "modified secant method"). In the following algorithm, δ denotes a tolerance level that is sufficiently small to guarantee the algorithm converges. The convergence of the algorithm is proved in Appendix D.

Algorithm 2.1:

Step 1. Set $\lambda = \lambda^0 = 0$; solve the segment-level subproblems (2.4a-c) for each k. Evaluate $V(\lambda^0)$. If $V(\lambda^0) \leq 0$, end; otherwise go to *Step 2*.

Step 2. Select another initial value $\lambda = \lambda^1 > 0$, solve (2.4a-c) for each k and evaluate $V(\lambda^1)$. If $|V(\lambda^1)| < \delta$, end; otherwise set $n = 1$ and go to *Step 3*.

Step 3. Set $\lambda = \lambda^{n+1} = \lambda^n - V(\lambda^n)\frac{\lambda^n - \lambda^{n-1}}{V(\lambda^n) - V(\lambda^{n-1})}$. Solve (2.4a-c) for each k and evaluate $V(\lambda^{n+1})$. If $|V(\lambda^{n+1})| < \delta$, end; otherwise, go to *Step 4*.

Step 4. If $V(\lambda^n) \cdot V(\lambda^{n+1}) > 0$ and $V(\lambda^{n-1}) \cdot V(\lambda^{n+1}) < 0$, set $\lambda^n = \lambda^{n-1}$. Set $n = n + 1$ and repeat *Step 3*.

For the separate-budget-constraint problem, similarly, we use a Lagrange multiplier, λ_p, to relax each of the P budget constraints in (2.1e). The Lagrange function becomes:

$$\min\ L(\boldsymbol{q},\boldsymbol{x},\lambda)=\sum_{k=1}^{K}Z_k(\boldsymbol{q}_k,\boldsymbol{x}_k)+\sum_{p=1}^{P}\lambda_p\left(\sum_{k=1}^{K}C_{kp}(\boldsymbol{q}_k,\boldsymbol{x}_k)-\frac{B_p}{r}(1-e^{-rT})\right)=\sum_{k=1}^{K}H_k(\boldsymbol{q}_k,\boldsymbol{x}_k,\boldsymbol{\lambda})-\sum_{p=1}^{P}\lambda_p\frac{B_p}{r}(1-e^{-rT}) \quad (2.8)$$

where $H_k(\boldsymbol{q}_k,\boldsymbol{x}_k,\boldsymbol{\lambda})\equiv C_k^U(\boldsymbol{q}_k,\boldsymbol{x}_k)+\sum_{p=1}^{P}(1+\lambda_p)C_{kp}(\boldsymbol{q}_k,\boldsymbol{x}_k)$, and $\boldsymbol{\lambda}=[\lambda_1,\cdots,\lambda_P]^T$. The corresponding segment-level problem can be written as follows:

For each $k=1,\ldots,K$

$$\min\ H_k(\boldsymbol{q}_k,\boldsymbol{x}_k,\boldsymbol{\lambda})=C_k^U(\boldsymbol{q}_k,\boldsymbol{x}_k)+\sum_{p=1}^{P}(1+\lambda_p)C_{kp}(\boldsymbol{q}_k,\boldsymbol{x}_k) \quad (2.9a)$$

$$\text{subject to: } \boldsymbol{\Phi}_k(\boldsymbol{q}_k,\boldsymbol{x}_k)=0 \quad (2.9b)$$

$$\boldsymbol{\Psi}_k(\boldsymbol{q}_k,\boldsymbol{x}_k)\le 0 \quad (2.9c)$$

The complementary slackness conditions are:

For $p=1,\cdots,P$,

$$\begin{cases}\text{either: } \lambda_p=0 \text{ and } V_p(\boldsymbol{\lambda})=\sum_{k=1}^{K}C_{kp}(\boldsymbol{q}_k,\boldsymbol{x}_k)-\frac{B_p}{r}(1-e^{-rT})\le 0\\ \text{or:}\quad \lambda_p>0 \text{ and } V_p(\boldsymbol{\lambda})=\sum_{k=1}^{K}C_{kp}(\boldsymbol{q}_k,\boldsymbol{x}_k)-\frac{B_p}{r}(1-e^{-rT})=0\end{cases} \quad (2.10)$$

Similar to the combined-budget-constraint problem, the optimal solution of the relaxed program above is also optimal to the original problem under separate budget constraints. Here we propose a modified Broyden's method to formulate the following algorithm to solve the relaxed program.[2]

[2] The original Broyden's method is a multivariate version of the secant method; see an introduction in Jorge and Stephen (2006). One can also show that the relaxed program has a unique optimum, given that each segment-level problem has a unique optimal solution (similar to Lemma 2.2). However, unlike the combined-budget case, Algorithm 2.2 cannot guarantee global convergence to the optimum. Nevertheless, we believe this algorithm did converge to the global minimum in the numerous numerical experiments presented in Section 2.1.3.4, since the solutions under the combined budget constraint (which are guaranteed to be optimal) are consistent with the corresponding solutions under the separate budget constraints.

Algorithm 2.2:

Step 1. Set $\boldsymbol{\lambda} = \boldsymbol{\lambda}^0 \equiv [\lambda_1^0, \lambda_2^0, \dots, \lambda_P^0]^T = \mathbf{0} \equiv [0,0, \dots, 0]^T$; solve the segment-level subproblems (2.9a-c) for each k. Evaluate $\boldsymbol{V}(\boldsymbol{\lambda}^0) = [V_1, \dots, V_P]^T$. If $\boldsymbol{V}(\boldsymbol{\lambda}^0) \leq \mathbf{0}$, end; otherwise go to *Step 2*.

Step 2. Calculate the initial $P \times P$ Jacobian matrix J^0. For each $p = 1, \dots, P$, define $\boldsymbol{\lambda}^{p,0}$ as a P-dimensional vector whose p-th element is a small positive number δ_p and all the other elements are 0. The J^0 is calculated by setting its element on the i-th row and the j-th column as: $J_{i,j}^0 = \frac{V_i(\boldsymbol{\lambda}^{j,0}) - V_i(\mathbf{0})}{\delta_j}$.

Step 3. Set $\boldsymbol{\lambda}^1 = \boldsymbol{\lambda}^0 - (J^0)^{-1}\boldsymbol{V}(\boldsymbol{\lambda}^0)$. End the search if $\boldsymbol{V}(\boldsymbol{\lambda}^1)$ satisfies the complementary slackness conditions (2.10); i.e., for each $p = 1, \dots, P$, $V_p(\boldsymbol{\lambda}^1) \leq 0$ if $\lambda_p^1 = 0$, and $|V_p(\boldsymbol{\lambda}^1)| < \delta$ if $\lambda_p^1 > 0$. Otherwise set $n = 1$ and go to *Step 4*.

Step 4. Set $(J^n)^{-1} = (J^{n-1})^{-1} + \frac{(\boldsymbol{\lambda}^n - \boldsymbol{\lambda}^{n-1}) - (J^{n-1})^{-1} * (\boldsymbol{V}(\boldsymbol{\lambda}^n) - \boldsymbol{V}(\boldsymbol{\lambda}^{n-1}))}{(\boldsymbol{\lambda}^n - \boldsymbol{\lambda}^{n-1})^T * (J^{n-1})^{-1} * (\boldsymbol{V}(\boldsymbol{\lambda}^n) - \boldsymbol{V}(\boldsymbol{\lambda}^{n-1}))} * (\boldsymbol{\lambda}^n - \boldsymbol{\lambda}^{n-1})^T * (J^{n-1})^{-1}$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{n+1} = \boldsymbol{\lambda}^n - (J^n)^{-1}\boldsymbol{V}(\boldsymbol{\lambda}^n)$. End the search if $\boldsymbol{V}(\boldsymbol{\lambda}^{n+1})$ satisfies the complementary slackness conditions (2.10). Otherwise, go to *Step 5*.

Step 5. Define vector operator $\otimes$ as $\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \otimes \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} a_1 b_1 \\ a_2 b_2 \\ a_3 b_3 \end{bmatrix}$ for any $\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$ and $\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$. If the number of negative elements in vector $V(\lambda^{n-1}) \otimes V(\lambda^{n+1})$ is larger than that in $V(\lambda^n) \otimes V(\lambda^{n+1})$, set $\lambda^n = \lambda^{n-1}$. Set $n = n + 1$ and return to *Step 4*.

2.1.2 A specific segment-level formulation and its solution approach

This section presents a typical formulation and its solution approach to the segment-level subproblem (i.e., a special case of (2.2a-c)), which jointly optimizes all three treatments: i.e., preventive maintenance (chip seal), rehabilitation and reconstruction. While the framework mentioned above applies to almost all segment-level subproblems, to stay focused, we present here only a segment-level formulation that is discrete in time but continuous in the pavement condition (i.e.,

the roughness index) for an infinite planning horizon. Most of the problem formulation, except for the preventive maintenance model, is similar to that presented in Lee and Madanat (2014a). We develop an efficient greedy heuristic, and compare its solution quality and computation efficiency against a benchmark dynamic programming algorithm.

The state variables are $\boldsymbol{q}_k = (q_k(t)|t = 0,1,2,\cdots) = (s_k(t), h_{kt}|t = 0,1,2,\cdots)$, where $s_k(t)$ and h_{kt} are the pavement's roughness index and age (number of years since the last reconstruction), respectively, for segment k in year t. The decision variables are $\boldsymbol{x}_k = \left(v_{kt}, \omega_{kt}, x_{kt,1}, x_{kt,2}, x_{kt,3}\middle| t = 0,1,2,\cdots\right)$, where the binary variable $x_{kt,p}$ $(p = 1,2,3)$ is equal to 1 if a rehabilitation (corresponding to $p = 1$), reconstruction $(p = 2)$ or preventive maintenance $(p = 3)$ activity is executed in year t for segment k, respectively, and 0 otherwise; v_{kt} and ω_{kt} represent the maintenance and rehabilitation intensities in year t for segment k, respectively. The full formulation is presented as follows:

$$\min\ Z_k(\boldsymbol{q}_k, \boldsymbol{x}_k) = C_k^U(\boldsymbol{q}_k, \boldsymbol{x}_k) + \sum_{p=1}^{3} C_{kp}(\boldsymbol{q}_k, \boldsymbol{x}_k) \tag{2.11a}$$

subject to:

$$C_k^U(\boldsymbol{q}_k, \boldsymbol{x}_k) = \sum_{t=0}^{\infty} \int_t^{t+1} l_k(c_k^1 s_k(u) + c_k^2)\, e^{-ru} du \tag{2.11b}$$

$$C_{k,1}(\boldsymbol{q}_k, \boldsymbol{x}_k) = \sum_{t=0}^{\infty} x_{kt,1}\, (m_k^1 \omega_{kt} + m_k^2) e^{-rt} \tag{2.11c}$$

$$C_{k,2}(\boldsymbol{q}_k, \boldsymbol{x}_k) = \sum_{t=0}^{\infty} x_{kt,2}(\, z_k^1 + z_k^2 l_k) e^{-rt} \tag{2.11d}$$

$$C_{k,3}(\boldsymbol{q}_k, \boldsymbol{x}_k) = \sum_{t=0}^{\infty} x_{kt,3}\, (\gamma_k^1 v_{kt} + \gamma_k^2) e^{-rt} \tag{2.11e}$$

$$\bar{b}_k - b_{kt} = x_{kt,3} E_k\big(v_{kt}, s_k(t)\big), \forall t \tag{2.11f}$$

$$E_k\big(v_{kt}, s_k(t)\big) = \frac{\alpha_k v_{kt}}{\left(s_k(t)\right)^{\beta_k}} \tag{2.11g}$$

$$0 \le v_{kt} \le D_{kt} = \min\left\{\bar{v}_k, \frac{(\bar{b}_k - b_k^*)\left(s_k(t)\right)^{\beta_k}}{\alpha_k}\right\}, \forall t \tag{2.11h}$$

$$s_k(t) - s_k^+(t) = x_{kt,1} G_k\big(\omega_{kt}, s_k(t)\big) + x_{kt,2}(s_k(t) - s_k^{new}), \forall t \tag{2.11i}$$

$$G_k\big(\omega_{kt}, s_k(t)\big) = \frac{g_k^1}{g_k^2 s_k(t) + g_k^3} \omega_{kt} \tag{2.11j}$$

$$0 \le \omega_{kt} \le R_{kt} = \left(\frac{g_k^2}{g_k^1} + \frac{g_k^3}{g_k^1 s_k(t)}\right) \max(0, \min\{s_k(t) - s_k^*, g_k^1 s_k(t)\}), \forall t \quad (2.11k)$$

$$s_k(u) = F_k(s_k^+(t), u - t, h_{kt}^+, b_{kt}), \forall u \in (t, t+1], \forall t \quad (2.11l)$$

$$F_k(s_k^+(t), u - t, h_{kt}^+, b_{kt}) = s_k^+(t) e^{b_{kt}(u-t)} + f_k l_k (u - t) e^{b_{kt}(h_{kt}^+ + u - t)} \quad (2.11m)$$

$$\sum_{p=1}^{3} x_{kt,p} \le 1, \forall t \quad (2.11n)$$

$$h_{kt}^+ = h_{kt}(1 - x_{kt,2}), \forall t \quad (2.11o)$$

$$h_{k(t+1)} = h_{kt}^+ + 1 \quad (2.11p)$$

$$s_k^{new} \le s_k(t) \le s_k^{max}, \forall t \quad (2.11q)$$

$$T_k^{min} x_{kt,2} \le h_{kt} x_{kt,2} \le T_k^{max} x_{kt,2}, \ \forall t \quad (2.11r)$$

$$q_k(0) = (s_k(0), h_{k0}) \quad (2.11s)$$

The models for the user cost C_k^U, rehabilitation cost $C_{k,1}$, reconstruction cost $C_{k,2}$ and maintenance cost $C_{k,3}$ are described in (2.11b-e), respectively, where l_k is the annual traffic loading on segment k, which is assumed to be constant (Sathaye and Madanat, 2011, 2012; Lee and Madanat, 2017); and c_k^1, c_k^2, γ_k^1, γ_k^2, m_k^1, m_k^2, z_k^1 and z_k^2 are (non-negative) cost coefficients.

Note that the maintenance cost model (2.11e) is for chip seal only, which is one of the most commonly used preventive maintenance activities (Labi and Sinha, 2003). Here the maintenance intensity variable v_{kt} is defined as the average least dimension (ALD) of chip seal in year t for segment k. The non-negative cost coefficients γ_k^1 and γ_k^2 depend on oil prices, the location of the pavement, labour costs, etc. Our maintenance effectiveness model is shown in (2.11f-g). The model is built upon the following two facts: i) the pavement roughness before and after the chip seal is approximately the same, but the deterioration rate diminishes, which is consistent with the findings of Mamlouk and Dosa (2014), among others; and ii) the reduction in the deterioration rate is a non-increasing function of the pavement roughness level (see Table 2 and Figures 4-7 of Mamlouk and Dosa,

2014). The latter means maintenance (e.g. chip seal) is less effective when applied to pavement in poor condition. Note that previous maintenance cost and effectiveness models (Gu et al., 2012; Lee and Madanat, 2014a, b; Lee and Madanat, 2017) resulted in predictions that were at odds with this simple and obvious fact. For example, Lee and Madanat (2014a) predicted a non-monotonic trend between deterioration rate reduction and the pavement's roughness level (see Figure 4a of that paper), where a larger deterioration rate reduction may occur when the roughness level is higher. This mistake has been corrected using our maintenance model. The $\bar{b}_k$ and b_{kt} in (2.11f) are the deterioration rates before and after applying chip seal. The mathematical form of (2.11g) is selected to fit the real test data of chip seal from Mamlouk and Dosa (2014), where parameters $\alpha_k > 0$, $\beta_k \geq 1$. In addition, there should be a technical upper bound for the ALD, $\bar{v}_k$. Also, the deterioration rate has a lower bound b_k^*, at which any additional maintenance has no effect. Thus, the effective maintenance intensity is bounded by D_{kt}, which is defined in (2.11h).

Other parts of the segment-level formulation are borrowed from previous studies, mostly from Ouyang and Madanat (2004; 2006) and Lee and Madanat (2014a). Constraints (2.11i) indicate the roughness index reduction caused by a rehabilitation or reconstruction activity, where function G_k represents the rehabilitation effectiveness as defined in (2.11j); $s_k(t)$ and $s_k^+(t)$ denote the roughness indices right before and after the activity, respectively; s_k^{new} is the roughness index immediately after a reconstruction; and g_k^1, g_k^2, and g_k^3 are coefficients. Constraints (2.11k) stipulate the upper bound, R_{kt}, for the rehabilitation intensity, where s_k^* is the best possible roughness level after rehabilitation. Constraints (2.11l) indicate how the pavement state is updated at the

moment $u \in (t, t+1]$, where F_k is a history-dependent deterioration model, shown in (2.11m); and h_{kt}^+ is the pavement age after the activity. Constraints (2.11n) ensure that at most one activity is performed every year. Constraints (2.11o) reset the pavement age to 0 after reconstruction. Constraints (2.11q-r) specify the upper and lower bounds of the roughness level and the pavement's lifecycle length. Constraint (2.11s) defines the initial pavement state.

We first decompose the infinite-horizon optimization problem (2.11a-s) into two finite-horizon subproblems. Each subproblem has fewer decision variables and is thus easier to solve. We present two algorithms to solve the subproblems: a dynamic programming algorithm similar to the one used by Lee and Madanat (2014a) and a greedy heuristic. The heuristic can achieve the same solution quality as the dynamic programming approach with only a small fraction of the computation time, as is validated later.

With the augmented state $q_k(t) = (s_k(t), h_{kt})$, the optimal MR&R decisions from year t onwards (and the future pavement states) depend only on the present state $q_k(t)$. Based on the Principle of Optimality (Bellman, 1957), the optimal roughness trajectory after the first reconstruction enters a periodic steady state, since every reconstruction resets the pavement to $(s_k^{new}, 0)$. The steady-state solution is thus characterized by a fixed lifecycle duration, denoted by T_k. The period prior to the first reconstruction is termed as the transient period, which is optimized separately. Therefore, the objective function (2.11a) is reformulated as follows:

$$\min Z_k(\boldsymbol{q}_k, \boldsymbol{x}_k) = Z_k^T(\boldsymbol{q}_k, \boldsymbol{x}_k) + \frac{e^{-rt_k^T}}{1-e^{-rT_k}} Z_k^S(\boldsymbol{q}_k, \boldsymbol{x}_k) \quad (2.12)$$

where Z_k^T is the discounted cost for the transient period, and Z_k^S is the cost for one steady-state cycle (with a reconstruction activity at the beginning), discounted to the beginning of the cycle, and t_k^T is the time of the first reconstruction. The Z_k^T and Z_k^S are given by the following equations.

$$Z_k^T(\boldsymbol{q}_k, \boldsymbol{x}_k) = \sum_{t=0}^{t_k^T-1} \left(\int_t^{t+1} l_k(c_k^1 s_k(u) + c_k^2)\, e^{-ru} du + x_{kt,3}(\gamma_k^1 v_{kt} + \gamma_k^2) e^{-rt} + x_{kt,1}(m_k^1 \omega_{kt} + m_k^2) e^{-rt} \right) \quad (2.13)$$

$$Z_k^S(\boldsymbol{q}_k, \boldsymbol{x}_k) = \sum_{\tau=0}^{T_k-1} \left(\int_\tau^{\tau+1} l_k(c_k^1 s_k(u) + c_k^2)\, e^{-ru} du + x_{k\tau,3}(\gamma_k^1 v_{k\tau} + \gamma_k^2) e^{-r\tau} + x_{k\tau,1}(m_k^1 \omega_{k\tau} + m_k^2) e^{-r\tau} \right) + z_k^1 + z_k^2 l_k \quad (2.14)$$

In equation (2.14), we use τ to denote the "age" in a steady-state lifecycle (counted from 0 starting from the previous reconstruction), and $q_k(0) = (s_k^{new}, 0)$.

Note that Z_k^S is independent of the transient period duration t_k^T and the MR&R schedule during that period. We can thus decompose this problem into two subproblems: the first subproblem for optimizing $\frac{Z_k^S}{1-e^{-rT_k}}$, and the second for optimizing Z_k given the optimal solution of the first one. We further note that Lee and Madanat (2014a) proved ω_{kt} is either 0 or R_{kt} at optimality. One can easily verify this by applying Lee and Madanat's method that the same optimality condition is true for our model. Thus, ω_{kt} can be eliminated from the list of decision variables. Now the two subproblems are summarized as follows:

Subproblem 1: Minimize $\frac{Z_k^S}{1-e^{-rT_k}}$ subject to constraints (2.11f-r) with decision variables $v_{k\tau}, x_{k\tau,3}, x_{k\tau,1}$ $(\tau = 0,1,2,\dots,T_k-1)$ and T_k.

Subproblem 2: Minimize $Z_k = Z_k^T + e^{-rt_k^T}\left(\frac{Z_k^S}{1-e^{-rT_k}}\right)^*$ subject to constraints (2.11f-r) with decision variables $v_{kt}, x_{kt,3}, x_{kt,1}$ $(t = 0,1,2,\dots,t_k^T - 1)$ and t_k^T, where $\left(\frac{Z_k^S}{1-e^{-rT_k}}\right)^*$ is the optimal value found in subproblem 1.

We first formulate a dynamic programming algorithm, which is modified from that developed by Lee and Madanat (2014a, 2015). The algorithm is relegated to Appendix E in the interest of brevity. To apply the algorithm, we discretize both the maintenance intensity $v_{k\tau}$ and the pavement roughness level $s_k(\tau)$ into $d + 1$ and $N + 1$ points, respectively, where d and N are integers. As d and N approach infinity, the dynamic programming solution converges to the global optimum. Thus, solutions of the dynamic programming approach can be used as benchmarks for verifying the solution quality of a much faster greedy heuristic. Next, we describe the details of this heuristic algorithm.

The heuristic is based upon the assumption that preventive maintenance is much cheaper than rehabilitation, which is true for most prevailing preventive maintenance treatments, including chip seal (Labi and Sinha, 2003). Thus, we start by seeking solutions where maintenance is performed more frequently, while rehabilitation is adopted only when it becomes a must. For the same reason, we postulate that a maintenance activity is always executed with the maximum intensity $D_{k\tau}$. (This postulation was verified by our extensive numerical tests.) To further avoid solutions with high frequency of rehabilitation, we specify a lower bound of roughness level, W_k, below which rehabilitation should not be executed. Different values of W_k are used in the algorithm to balance the solution quality

and computational efficiency. The algorithm for subproblem 1 is presented as follows:

Algorithm 2.3:

For each W_k, do the following and record the lowest-cost solution:

Step 1. Initialize $\tau = 1, cost_2 = \infty$.

Step 2. If $\tau < T_k^{max}$, find the action in year τ from the action set: $\{\text{Do-nothing } (x_{k\tau,3} = x_{k\tau,1} = 0),\ \text{Maintenance } (x_{k\tau,3} = 1, x_{k\tau,1} = 0),\ \text{Rehabilitation } (x_{k\tau,3} = 0, x_{k\tau,1} = 1)\}$, which minimizes the objective function $\frac{Z_k^S}{1-e^{-rT_k}}$ for the MR&R plan generated by the following *Steps 2.1-2.3*. Record the minimum objective value as $cost_1$:

Step 2.1. Keep the recorded MR&R plan before year τ and execute the selected action in year τ.

Step 2.2. For each year $y > \tau$, execute maintenance with the maximum intensity D_{ky}; and if $s_k(y+1) > s_k^{max}$, replace this maintenance in year y with rehabilitation.

Step 2.3. Among year T_k^{max} and all the years of rehabilitation between T_k^{min} and T_k^{max}, find the year in which reconstruction minimizes the objective function $\frac{Z_k^S}{1-e^{-rT_k}}$.

The selected action in year τ should also satisfy the following conditions: $s_k(\tau) > W_k$ if the selected action is rehabilitation; and $s_k(\tau+1) < s_k^{max}$ if the selected action is executed in year τ.

Step 3. If $T_k^{min} \le \tau \le T_k^{max}$, calculate the objective function $\frac{Z_k^S}{1-e^{-rT_k}}$ associated with the following MR&R plan: keep the recorded plan before year τ and execute reconstruction in year τ. Set $cost_2 = \frac{Z_k^S}{1-e^{-rT_k}}$.

Step 4. If $\tau = T_k^{max}$ or $cost_2 < cost_1$, record the reconstruction in year τ, end; otherwise, set $\tau = \tau + 1$ and go to *Step 2*.

Only minor changes are made to the above algorithm when it is applied to subproblem 2. Specifically, τ is initialized by 0 instead of 1; the objective

function is changed to $Z_k = Z_k^T + e^{-rt_k^T}\left(\frac{Z_k^S}{1-e^{-rT_k}}\right)^*$; and finally, the time range for reconstruction is replaced by $\left[T_k^{min'}, T_k^{max'}\right]$, where $T_k^{min'} = \max\{0, T_k^{min} - h_{k0}\}$ and $T_k^{max'} = \max\{0, T_k^{max} - h_{k0}\}$.

2.1.3 Numerical case studies in the deterministic scenario

Most of the numerical experiments presented in this section are for a pavement system with 100 heterogeneous segments. Although our approach is able to optimize for pavement systems that are 10 times larger in a reasonable computation time, we chose this medium-sized system for analysis simply because it is easier to run for a large batch of numerical experiments with various parameter values. We are thus able to discuss the general findings and insights revealed by these results. Section 2.1.3.1 describes the parameter values. Section 2.1.3.2 examines the solution quality and computation efficiency of the segment-level heuristic. The system-level case studies under the combined and separate budget constraints are discussed in Sections 2.13.3 and 2.1.3.4, respectively. The computational efficiency of our solution method is examined in Section 2.1.3.5. All the numerical instances presented in this section were carried out via Matlab R2014a on a PC with Inter® Xeon® 3.60 GHz CPU, 32.0GB RAM, and Windows 10 Pro 64-bit.

2.1.3.1 Parameter values

Most parameter values used in our numerical cases are summarized in Table 2.1. The cost parameters γ_k^1, γ_k^2 are derived from the empirical cost model of chip seal in Labi and Sinha (2003); the parameters for the chip seal effectiveness model $(\alpha_k, \beta_k, \bar{v}_k)$ are obtained by fitting the model to the data in Mamlouk and Dosa (2014); the other parameter values are borrowed from Lee and Madanat (2014a).

To account for heterogeneous segments, we specify that the initial pavement states, traffic loading, and some cost coefficients follow certain uniform distributions, which are denoted by the form of $U[a, b]$ in the table.

Table 2.1. Parameter values

Parameter	Value	Unit	Parameter	Value	Unit
c_k^1	U[20500, 22500]	$/IRI/km/lane/ million ESAL	l_k	U[0.4,0.9]	million ESAL/year/lane
c_k^2	0	-	$\bar{b}_k$	0.04	-
m_k^1	U[10000, 12000]	$/mm/km/lane	z_k^2	917000	$·year/km/ million ESAL
m_k^2	U[140000, 170000]	$/km/lane	f_k^*	0.093	IRI·lane·year/ million ESAL
g_k^1	0.66	-	b_k^*	0.025	-
g_k^2	7.15	mm/IRI	s_k^*	0.8	IRI
g_k^3	18.3	mm	s_k^{new}	0.75	IRI
γ_k^1	130	$/mm/lane/km	s_k^{max}	6	IRI
γ_k^2	300	$/lane/km	z_k^1	900000	$/km/lane
α_k	0.002	-	T_k^{min}	20	year
β_k	1.483	-	T_k^{max}	60	year
$\bar{v}_k$	14	mm	s_k^0	U[1,3]	IRI
r	0.07	-			

2.1.3.2 Validation of the segment-level greedy heuristic

To verify the quality of the segment-level greedy heuristic, we tested 216 numerical cases with varying values of λ_p ($p = 1,2,3$), $q_k(0)$, l_k, and r: $\lambda_p \in \{0,4\}$, $q_k(0) = (s_k(0), h_{k0}) \in \{(1,3), (2,8), (4,15)\}$, $l_k \in \{0.5,0.8,1.2\}$, $r \in \{0.05,0.07,0.1\}$. Note that the agency cost of treatment p in the objective function is multiplied by the weight $1 + \lambda_p$. The other parameters take the values listed in Table 2.1.

We compare the greedy heuristic against two instances of the dynamic programming algorithm: where $N = d = 3$ (denoted as DP1), and where $N =$

300, $d = 5$ (denoted as DP2). The solutions generated from DP2 are treated as the global optima because no meaningful improvement in the solutions was observed by further increasing N or d. The runtimes and the cost gaps of the heuristic and the DP algorithms are summarized in Table 2.2, where the cost gaps are defined as:

$$\frac{\text{cost of the greedy heuristic or DP1} - \text{cost of DP2}}{\text{cost of DP2}}$$

Both the averages and the maxima of all 216 cases are presented.

Table 2.2. Runtimes and cost gaps for the greedy heuristic and dynamic programming algorithms

	Greedy heuristic	**DP1**	**DP2**
Average runtime (second)	1.20	49.21	1439.31
Maximum runtime (second)	1.47	73.43	1981.32
Average total cost gap	0.37%	0.41%	-
Maximum total cost gap	3.56%	2.05%	-
Average maintenance cost gap	0.29%	0.28%	-
Maximum maintenance cost gap	4.17%	3.83%	-
Average rehabilitation cost gap	0.36%	0.52%	-
Maximum rehabilitation cost gap	4.35%	2.69%	-
Average reconstruction cost gap	0.42%	0.40%	-
Maximum reconstruction cost gap	3.98%	2.46%	-

The tabulated values confirm that our heuristic algorithm produces solutions that are very close to the global optima. Note that the average gap in the total cost is only 0.37%. A comparison between the greedy heuristic and DP1 shows that both algorithms furnished solutions of similar quality, but our heuristic had much shorter runtimes (about 97% less). We therefore use the greedy heuristic in the following sections for the purpose of computation efficiency. Recall that our system-level approach preserves solution quality as long as the segment-level subproblems are solved near the optimality.

2.1.3.3 Under the combined budget constraint

First, we randomly generate a 100-segment pavement system, and optimize the total discounted cost for a range of combined annual budgets: $B \in [4 \times 10^6, 5 \times 10^6]$ \$/year. The optimal total discounted cost and the cost components are plotted against B as the solid curves in Figure 2.1. These curves start from $B = 4.02 \times 10^6$ \$/year on the left because this value of B represents the minimum budget required to find a feasible MR&R plan. This minimum required budget can be calculated by optimizing the decomposed problems (2.4a-c) with a sufficiently large λ. Figure 2.1 shows that the optimal total cost (the solid curve with dot markers) decreases as B grows, until it reaches a threshold of 4.61×10^6 \$/year, marked by the arrow. This threshold represents the maximum budget needed for the pavement system; i.e., any additional budget would be redundant, and the optimal total cost would stay the same (11.73×10^7\$).

The figure also shows that the user cost (the triangle-marked solid curve) decreases as B increases, which is as expected. Meanwhile, the rehabilitation cost (the "x"-marked solid curve) generally diminishes, while the reconstruction cost (the square-marked solid curve) increases with B. This means that if there is room in the budget, the agency should apply more reconstruction but less rehabilitation to reduce user cost. If the budget is highly limited, however, more rehabilitation should be performed to extend the pavement's service life. The maintenance cost (the diamond-marked curve near the bottom of the figure) is much lower than the other cost components and is insensitive to B. This is because there is no incentive to trade off the maintenance activities: they are very cheap, but have a considerable effect on the pavement.

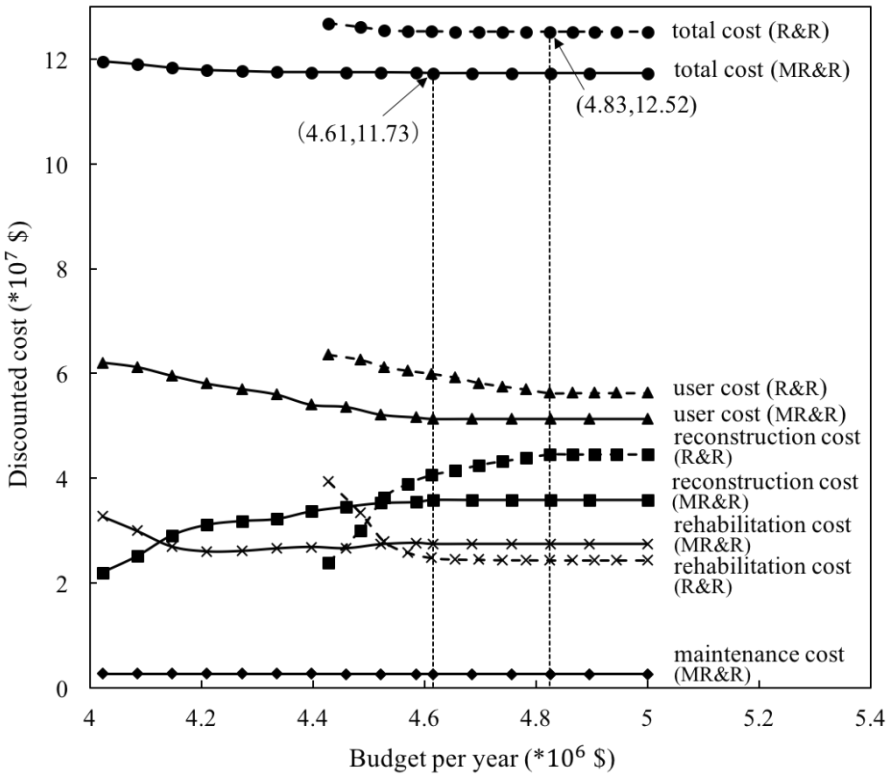
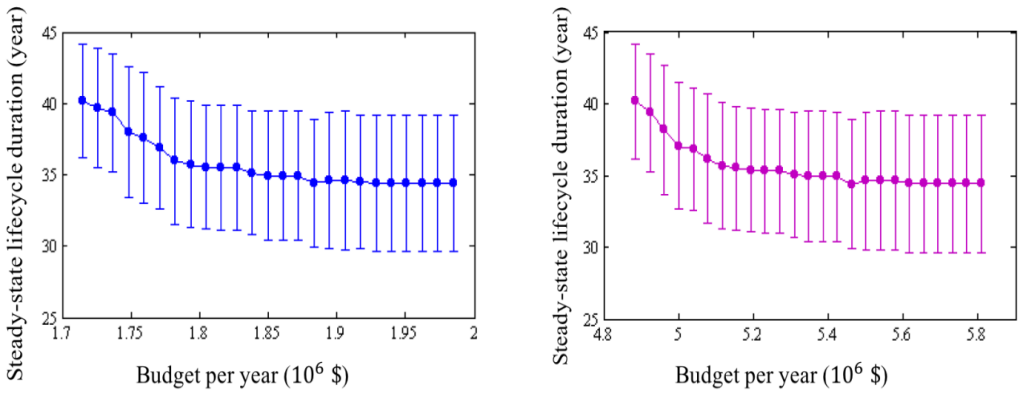


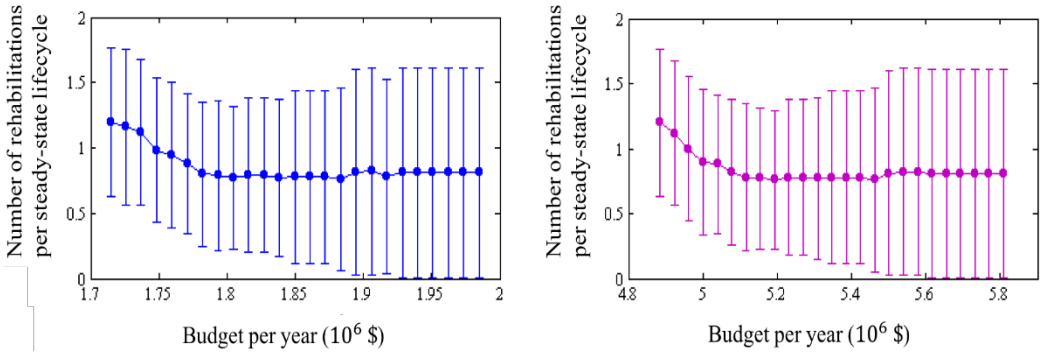
Figure 2.1. Effects of the combined agency budget on the system-level optimal costs.

To examine how adding maintenance affects the optimal MR&R plan, we compared the above total cost and cost components against those for the optimal R&R plans (i.e., without maintenance). The R&R costs are plotted as the dashed curves in Figure 2.1. A comparison reveals total cost savings of 6.3-7.5% from applying maintenance for $B \in [4.43 \times 10^6, 5 \times 10^6]$ \$/year. The minimum annual budget required is also reduced by 9.3% (from 4.43 to 4.02×10^6 \$/year). A comparison between the cost components reveals that adding maintenance usually results in lower reconstruction cost but higher rehabilitation cost. This means maintenance extends the pavement's service life, which in turn entails more rehabilitation activities.



(a) (b)

Figure 2.2. Distribution of steady-state lifecycle duration versus the budget constraint: (a) a case with good initial conditions; (b) a case with poor initial conditions.

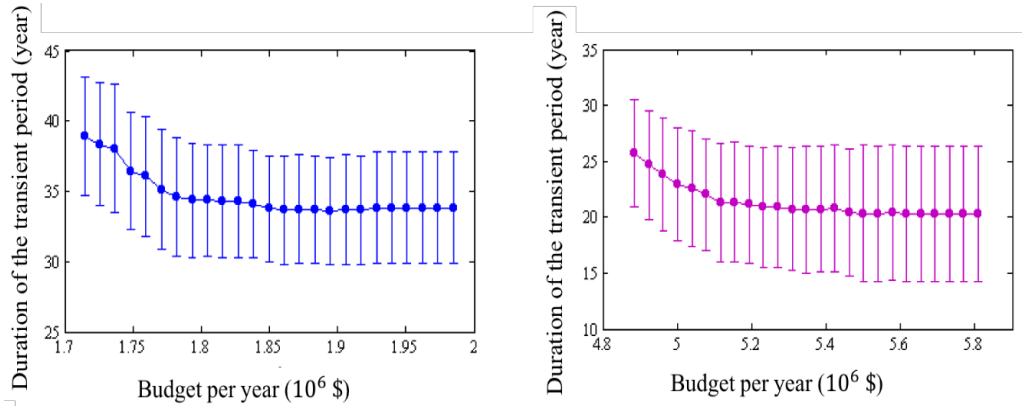


(a) (b)

Figure 2.3. Distribution of rehabilitation counts per steady-state lifecycle versus the budget constraint: (a) good initial conditions; (b) poor initial conditions.

The effects of initial pavement conditions on the optimal MR&R plans for individual segments were often overlooked in previous studies (e.g. Lee and Madanat, 2015). Here we plot against the budget constraint the distributions of i) steady-state lifecycle duration (Figure 2.2a and b), and ii) the number of rehabilitation activities per steady-state lifecycle (Figure 2.3a and b) of the 100 segments. Figure 2.2a and Figure 2.3a are for a system with good initial conditions ($s_k(0) \sim U[0,1], \forall k$). Figure 2.2b and Figure 2.3b are for the same system but with poor initial conditions ($s_k(0) \sim U[2,3], \forall k$). In each figure, the large dots indicate the mean value (lifecycle duration or rehabilitation count) of the 100 segments,

and the error bars describe the interval of two standard deviations centered at the mean. As expected, both the mean lifecycle duration and the mean rehabilitation count decrease as B increases until the constraint becomes unbinding, which is consistent with the findings in Figure 2.1. Smaller standard deviations are observed for smaller B, indicating that a tighter budget tends to "homogenize" the segment-level MR&R plans.

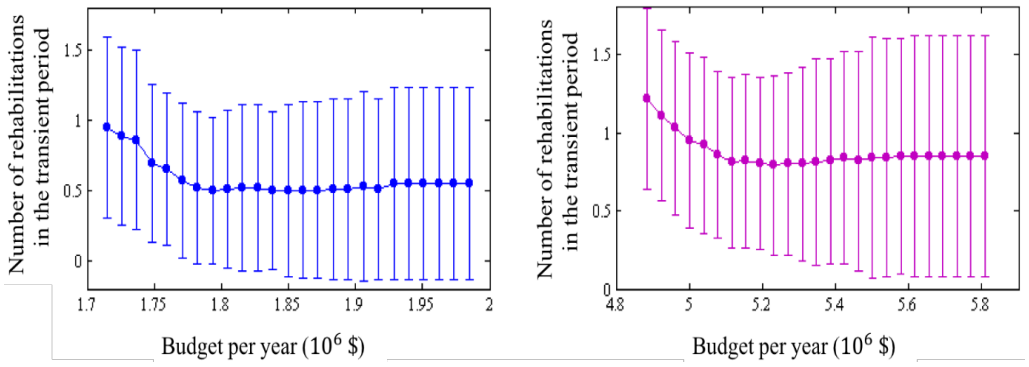


(a) (b)

Figure 2.4. Distribution of the first lifecycle's duration versus the budget constraint: (a) good initial conditions; (b) poor initial conditions.

A comparison between Figure 2.2a and b reveal that the mean and standard deviation of steady-state lifecycle durations vary along very similar paths as B increases, despite the largely different initial pavement conditions. A strong similarity is also observed between Figure 2.3a and b for the distribution of rehabilitation counts per steady-state lifecycle. This means the steady-state MR&R plans of individual segments are *almost* independent of their initial conditions. Scrutinization of the numerical results shows that most of the pavement segments have nearly (but not exactly) the same steady-state MR&R plans between the two cases. However, the optimal MR&R plans during the transient periods are significantly affected by the initial pavement conditions; note the large differences

between the distributions of the transient period durations (Figure 2.4a and b) and the rehabilitation counts in the transient periods (Figure 2.5a and b) for the cases with good and poor initial conditions.

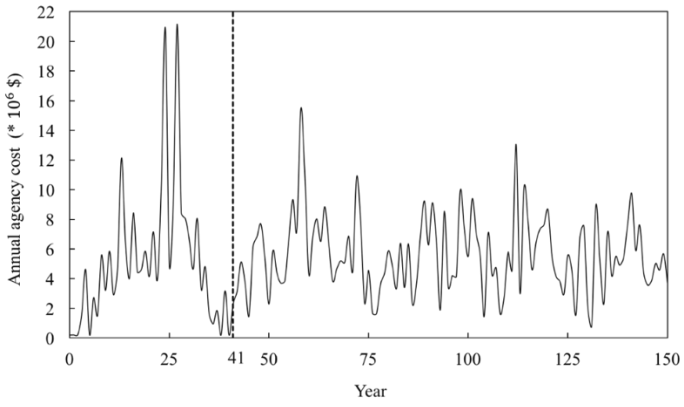


(a) (b)

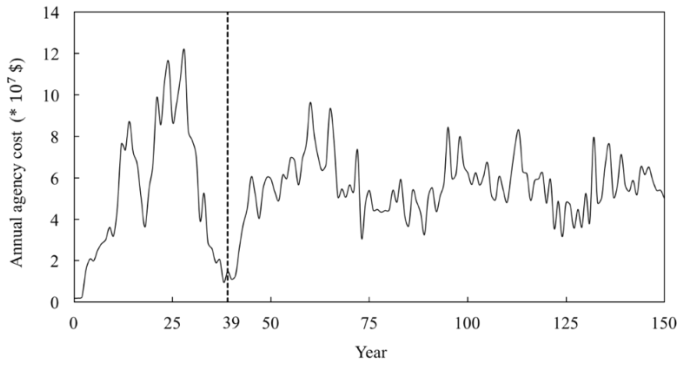
Figure 2.5. Distribution of rehabilitation counts in the first lifecycle versus the budget constraint: (a) good initial conditions; (b) poor initial conditions.

There is also the question of how the assumption that the budget can be transferred across the years affects the optimal solution. A rough look into this question can be made by examining the actual annual expenditure for the optimal MR&R plan when the budget is allowed to be transferred across years. We plotted the actual annual agency cost from year 1 to year 150 under the optimal MR&R plan for a 100-segment system, with $B = 4.42 \times 10^6$ \$/year (Figure 2.6a), and a 1000-segment system, with $B = 4.38 \times 10^7$ \$/year (Figure 2.6b). Both figures show a large variation in annual agency expenditures. The variation is especially large for a smaller-sized system (see Figure 2.6a) and during the transient period of the pavement system (i.e. before the dashed vertical line in both figures, which marks the time when the last segment enters a steady state). Also, no periodic pattern is observed in either figure. This is expected because each segment has a different lifecycle duration. These findings imply that if a constant, non-transferable budget is set each year, the optimal MR&R plan will be very different, and the optimal

cost will likely be much higher than what we obtained in this paper. In reality, an agency may have the freedom to decide how to allocate the budget over a short planning period, e.g. five years (Lee and Madanat, 2015). However, this would also be suboptimal, as Figure 2.6a and b reveal. (Interestingly, this result is at odds with the claim made by Lee and Madanat, 2015.) One solution is for the agencies to borrow and lend money across planning periods, using financial tools, for example.



(a)



(b)

Figure 2.6. Annual agency costs under optimal MR&R plans with budget transfers allowed: (a) a system of 100 segments; (b) a system of 1000 segments.

2.1.3.4 Under separate budget constraints

In reality, an agency often manages separate budgets for different treatments (Lee and Madanat, 2015). In this section we examine how this suboptimal practice

affects the performance of the optimal MR&R plan. We examine the same 100-segment pavement system analyzed in Figure 2.1, but now under separate budget constraints. For clarity of illustration, we present the results of a reduced problem with only two budget constraints: one for reconstruction and the other for maintenance and rehabilitation combined.[3] Figure 2.7a plots a contour map of the optimal total cost for an annual reconstruction budget in the range of $[0, 5.5 \times 10^6]$ \$/year, and an annual maintenance and rehabilitation budget of $[0, 5 \times 10^6]$ \$/year. Each thin, solid curve in the figure represents a contour line with the total discounted cost marked on the curve (in units of $\$10^7$). An examination of this figure unveils interesting findings that complement those in the literature.

First of all, no contour line is present in the region in the lower-left part of Figure 2.7a (labelled "***INF***"), because the MR&R optimization problem is infeasible in this region due to insufficient budgets. Note that all of the area on the left side of the vertical dashed line at 0.36×10^6 \$/year (the minimum reconstruction budget associated with $T_k^{max}, \forall k$) belongs to region ***INF***, regardless of the maintenance and rehabilitation budget. In contrast, there are no contour lines in the rectangular region in the upper-right corner of Figure 2.7a (labelled "***A***"), because in this region both budget constraints are unbinding, and the optimal total cost remains constant at 11.73×10^7 \$. Note that the bottom-left corner of

[3] We choose to present the results of this reduced problem simply for the sake of clarity. Note that now the effects of the two budget constraints can be clearly illustrated by two-dimensional contour maps (like Figure 2.7a-d). A three-budget-constraint problem can also be solved by our approach, but the effects of the three budget constraints cannot be presented in a similar way in the paper. The analysis of the reduced problem does not compromise our findings, since the maintenance cost is always small and easy to accommodate; see again Figure 2.1.

region ***A*** indicates the maximum budgets needed: 2.51×10^6 \$/year for reconstruction and 2.10×10^6 \$/year for maintenance and rehabilitation combined.

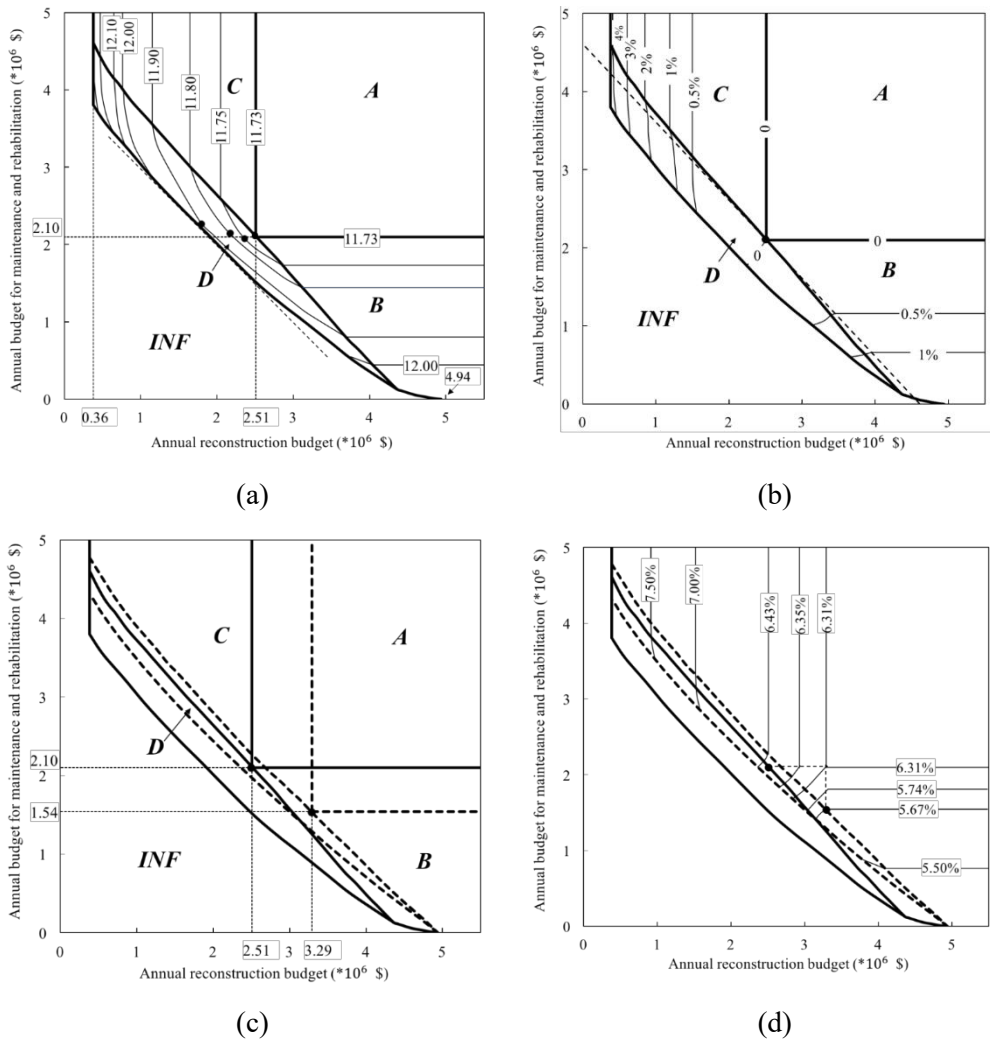


Figure 2.7. Solutions under separate budget constraints: (a) contours of the optimal total cost and solution regions; (b) percentage of cost savings from optimally allocating a combined budget; (c) comparison of the solution regions with and without maintenance; (d) percentage of cost savings from adding maintenance.

The remaining part of the figure is divided into three regions: ***B***, ***C***, and ***D***, as demarcated by the thick solid lines. Region ***B*** refers to a set of cases in which the reconstruction budget constraint is unbinding, and the maintenance and rehabilitation budget constraint is binding. Hence, the contours in this region are

horizontal lines. In region ***C***, the maintenance and rehabilitation budget constraint is unbinding, but the reconstruction constraint is binding. Finally, region ***D*** is where both budget constraints are binding. Note that each contour line that extends from the top-left to the bottom-right corner of the figure is tangent to a line with slope -1; the tangent point indicates the optimal solution under the combined budget constraint. Some of these combined-budget-constraint solutions are shown as black dots on the contour lines of 11.90, 11.80, and 11.75×10^7 \$ in Figure 2.7a. The lower boundary of ***D*** is also tangent to a line with slope -1 (the dashed line shown in Figure 2.7a); this dashed line specifies the minimum budget required for the combined budget scenario (4.02×10^6), which is consistent with Figure 2.1. This is also intuitive: if a feasible MR&R plan is found for a given pair of separate budget constraints, then the corresponding problem when all the budgets are combined is also feasible.

Figure 2.7b shows the contour map of the percentage of cost savings by optimizing for a combined budget for all three treatments. The figure shows cost savings of up to 4% when the reconstruction budget is highly limited. On the other hand, if only the maintenance and rehabilitation budget is limited, the cost savings are below 2%. The dashed line with slope -1 indicates the maximum required combined budget, and the contour lines above this dashed line should overlap with the contours of the optimal total cost shown in Figure 2.7a.

To further illustrate the effectiveness of maintenance, Figure 2.7c compares the five solution regions defined above (***A***, ***B***, ***C***, ***D***, and ***INF***) against those for the optimal R&R plan (i.e. without preventive maintenance). The solution regions for the R&R case are demarcated by thick dashed lines in Figure 2.7c. The figure

shows that when preventive maintenance is included, region ***D*** expands and moves downward, while region ***INF*** diminishes. This means including maintenance can appreciably reduce the budget needed to keep the pavements workable.

Finally, the percentage of cost savings between MR&R and R&R is plotted in Figure 2.7d. This shows that including maintenance can result in a reduction of over 5% in the optimal total cost in most cases. The highest cost savings (almost 8%) are achieved when the reconstruction budget is most limited, because maintenance can extend the pavements' lifecycle and thus reduce the need for reconstruction.

Note that the solution regions shown in Figure 2.7a-d are different from those presented by Lee and Madanat (2015). Specifically, in Lee and Madanat, the right boundary in region ***INF*** is a vertical line, and the lower boundary in region ***D*** is the horizontal axis (see Figure 4 in the cited paper). The difference is due to the different input parameters used in our case studies. For more details, please refer to Appendix F, which presents all the possible patterns of the solution regions that may arise from real-life pavement systems.

2.1.3.5 Computational efficiency

The dots in Figure 2.8 present the computation times of 110 randomly generated numerical instances under combined budget constraint against the number of pavement segments (ranging from 50 to 1000). They were carried out via Matlab R2014a on a PC with Inter® Xeon® 3.60GHz CPU, 32.0GB RAM, and Windows 10 Pro 64-bit. The dots exhibit a clear linear relationship between the

computational time and the size of the problem given the exact segment-level models. A similar linear relationship was found in cases under separate budget constraints. This is because the number of iterations needed for the Lagrange multiplier(s) λ (or λ_p) to converge is uncorrelated with the size of the system. With our selected error tolerance level (1% of the budget), this number of iterations is usually 4-5 under the combined budget constraint, and 20-30 under three separate budget constraints. Note too that a 1000-segment system takes about 1.5 hours to solve under the combined budget constraint, and about 8-10 hours under three separate budget constraints. The runtime is very reasonable for real-world implementation.

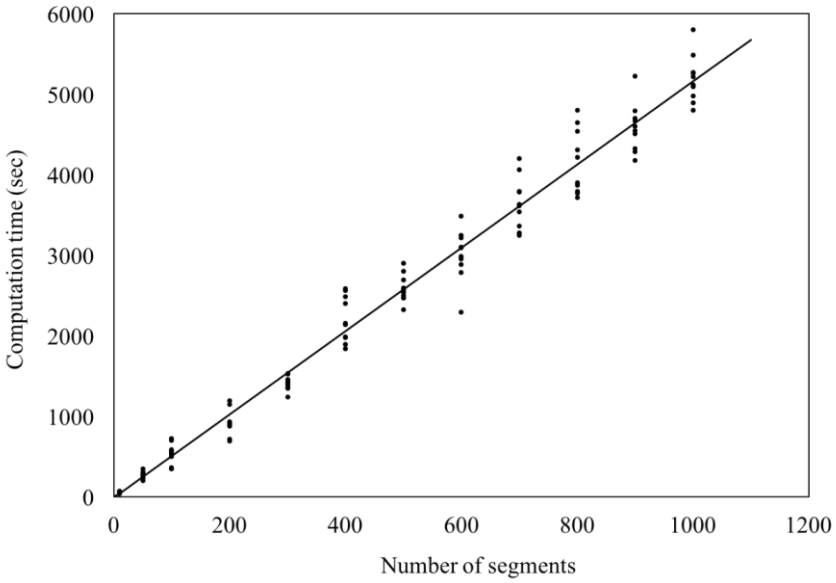


Figure 2.8. Computation times for the numerical instances under the combined budget constraint.

In comparison, the GA algorithm developed by Lee and Madanat (2015) for solving joint R&R optimization (i.e. without maintenance) seems to exhibit a polynomial complexity (see Figure 6 of the cited work); i.e. the computation time increases much faster than the linear trend. Thus, our approach is more

computationally efficient than the GA algorithm, especially for larger-scale systems.

2.2 Joint optimization of inspection and MR&R policies for pavement systems under model uncertainty

In this section, we present an infrastructure management framework for selecting optimal inspection and MR&R policies under model uncertainty. The model specifies that the inspection and MR&R decisions are made on the basis of the facility condition states (i.e., condition-based management, CBM). Following the same logic as in the deterministic scenario, this section is organized as follows. Section 2.2.1 presents a problem formulation at the system level (i.e., the upper-level problem) and a bottom-up solution method. Section 2.2.2 describes the segment-level model (i.e., the lower-level problem) and its solution. Section 2.2.3 furnishes numerical case studies. Note that our system-level stochastic model is also a weakly coupled dynamic problem and solved by decoupling into segment-level problems through the Lagrange multiplier method.

2.2.1 System-level stochastic model and its solution approach

We first formulate a model for the joint optimization of inspection and MR&R policies for a heterogeneous pavement system in Section 2.2.1.1. This joint optimization model guarantees that an inspection is made only if its cost is offset by its benefit in reducing expected future costs. Then a bottom-up solution approach built upon the Lagrange multiplier method is described in Section 2.2.1.2. The description of the solution approach assumes that the solution approach of the segment-level problems is ready for use.

2.2.1.1 Problem formulation

Generally, the decision-makers are more interested in immediate actions than those planned for future, since the latter are associated with high uncertainty. However, the optimal policies applied in the current year should be determined by considering expected future costs, which depend on future action plans (Chu and Huang, 2018). We denote the current time as τ (in the unit of year), and the time elapsed from the current time as t; $t = 0$ indicates the current time (see Figure 2.9). For descriptive purpose, two decision-making time points (τ and $\tau + 1$) are used to present the rolling-horizon procedure as shown in Figure 2.9 (Sethi and Sorger, 1991). At decision-making time point τ, we aim to find the optimal inspection and MR&R policies implemented at the current time τ, where a finite planning horizon, $t \in [0,1,..,T]$, is considered in the optimization process ($t = 0$ indicates the current time τ). For the optimal policies applied at the next time point $\tau + 1$, a new optimization process is required where both the start and the end of horizon are rolling ahead ($t = 0$ indicates the current time $\tau + 1$); i.e., the policies applied at each year should be optimized year by year. This is because the deterioration model and available budget are updated at each year due to the feedback optimization schematics presented in Figure 2.10, where the shaded and unshaded blocks describe the decision-making process at time τ and $\tau + 1$, respectively.

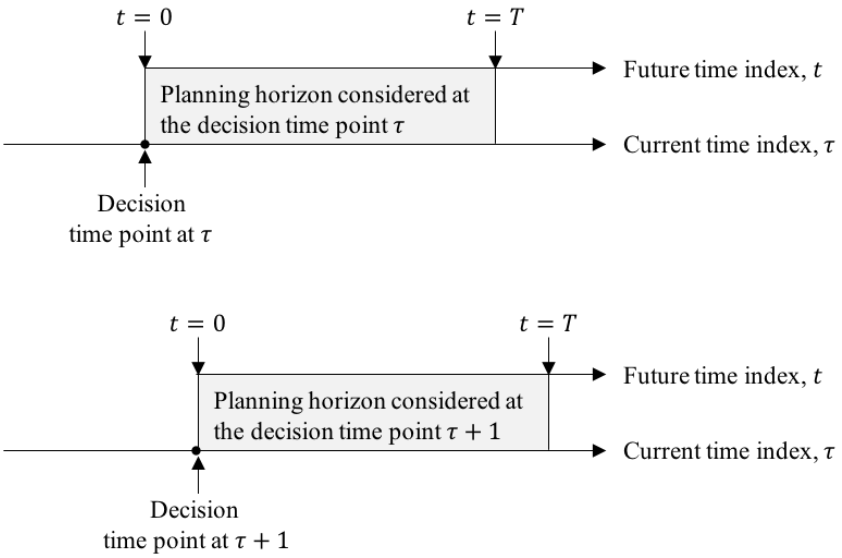


Figure. 2.9. Graphical representation of the current decision-making time index, τ, and the time elapsed from the current time (i.e., the future time index), $t \in [0,1,\cdots,T]$.

As the optimal inspection and MR&R polices applied at each year can be determined using the same solution approach, the rest of Section 2.2.1.1 is devoted to formulating the decision-making process at the current decision-making point τ. Specifically, Section 2.2.1.1.1 and 2.2.1.1.2 present the objective function and budget constraint of the optimization problem respectively. The details of the stochastic pavement deterioration model are given in Section 2.2.1.1.3.

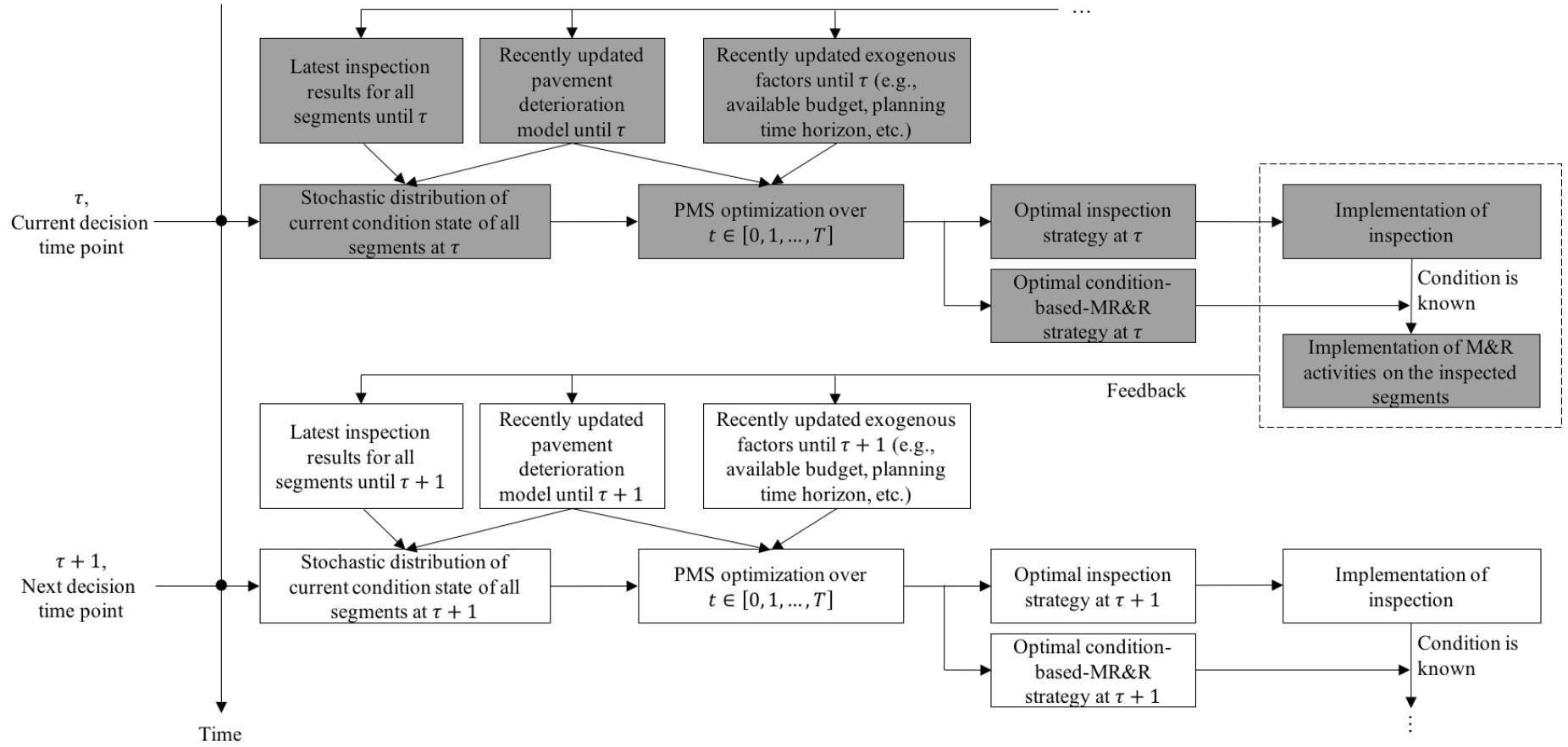


Figure 2.10. Schematic of the proposed discrete-time feedback optimization.

2.2.1.1.1 Objective function

The objective of the optimization problem at the current decision-making time τ (please refer to the shaded block in Figure 2.10) is to minimize the sum of the discounted user costs J, for all the pavement segments $k \in \{1,2,\cdots,K\}$ over a given planning horizon T (i.e, $t \in [0,1,..,T]$), as shown in (2.15) below.[4] The condition of pavement system at time t is denoted by $\boldsymbol{q}(\boldsymbol{t}) = \{q_k(t)|k = 1,\cdots,K\} = \{s_k(t), h_{kt}|k = 1,\cdots,K\}$, where $s_k(t)$ and h_{kt} are the pavement surface roughness index and the pavement's age (defined as the number of years since the last reconstruction or construction), respectively, for segment k in time t. For each segment, as the pavement has deteriorated in a stochastic way since the latest inspection performed before $t = 0$, the current roughness index of the network at $t = 0$ (i.e., $\boldsymbol{s}(\mathbf{0}) = \{s_k(0)|k = 1,\cdots,K\}$) is a random variable. Consequently, we use the expected value of the objective function with respect to $\boldsymbol{s}(\mathbf{0})$ as shown in model (2.15a). The decision variables of the problem include the network-level pavement inspection and MR&R policies applied at the current time point τ, defined as $\boldsymbol{I}(\mathbf{0}^-)$ and $\boldsymbol{M}(\mathbf{0}^+)$, respectively, which is a set of inspection and MR&R policies for all the segments in the system, i.e., $\boldsymbol{I}(\mathbf{0}^-) = \{I_k(0^-)|k = 1,\cdots,K\}$ and $\boldsymbol{M}(\mathbf{0}^+) = \{M_k(0^+)|k = 1,\cdots,K\}$.[5] Specifically, at

---

[4] Here the objective is to minimize the user cost instead of the sum of user and agency cost used in Section 2.1. Generally, the form of objective is up to the decision makers. However, our methodology also applies to the model with the objective defined as the generalized cost (i.e., the sum of user and agency cost).

[5] Note that, what we care the most is the optimal deterministic inspection and MR&R policies applied at the current time τ (i.e., $t = 0$), $\boldsymbol{I}(\mathbf{0}^-)$ and $\boldsymbol{M}(\mathbf{0}^+)$. Note that $\boldsymbol{I}^*(\boldsymbol{t}^-)$ and $\boldsymbol{M}^*(\boldsymbol{t}^+)$ for all $t > 0$ in (2.15b) are stochastic while $\boldsymbol{I}^*(\mathbf{0}^-)$ and $\boldsymbol{M}^*(\mathbf{0}^+)$ are deterministic. From $\tau + 1$

the beginning of each period t, $I_k(t^-)$ specifies whether to perform an inspection at t (i.e., $I_k(t^-) = 1$) or not (i.e., $I_k(t^-) = 0$) for segment k. Here, t^- stands for the start time of period t before the decision and application of inspection. Once an inspection is performed, we need to make a MR&R decision $M_k(t^+)$ with regards to the options for MR&R activities, as well as their treatment intensity, such as rehabilitation thickness, where t^+ is the time immediately after the application of inspection. The inherent assumption is that knowledge of the segment condition state is required for MR&R decision-making, or so-called condition-based-management, which is consistent with the joint decision models with temporal linkage proposed by Madanat (1993). In other words, no MR&R activity can be performed unless an inspection precedes it; i.e., $M_k(t^+)$ must be 'do-nothing' if $I_k(t^-) = 0$. The optimal stochastic long-term planning for $t \in [1, T)$ is defined in (2.15b). The terminal value is given in (2.15c).

$$J^*(0^-) = \min_{\substack{\boldsymbol{I}(\mathbf{0}^-)\in\mathcal{L},\\ \boldsymbol{M}(\mathbf{0}^+)\in\mathcal{M}}} \underset{\boldsymbol{s}(\mathbf{0}^-)}{E}[J(\boldsymbol{q}(\mathbf{0}^-), \boldsymbol{I}(\mathbf{0}^-), \boldsymbol{M}(\mathbf{0}^+))]$$

$$= \min_{\substack{I_k(0^-)\in\mathcal{L}_k,\\ M_k(0^+)\in\mathcal{M}_k,\forall k}} \underset{\boldsymbol{s}(\mathbf{0}^-)}{E}\left[\sum\nolimits_{k=1}^{K} J_k(q_k(0^-), I_k(0^-), M_k(0^+))\right]$$

$$= \min_{\substack{I_k(0^-)\in\mathcal{L}_k,\\ M_k(0^+)\in\mathcal{M}_k,\forall k}} \underset{\boldsymbol{s}(\mathbf{0}^-)}{E}\left[\left\{\sum\nolimits_{k=1}^{K} C_k^U\left(q_k(0^-), M_k(0^+)\right)\right\} + \alpha \cdot J^*\left(1^-|\boldsymbol{q}(\mathbf{0}^-), \boldsymbol{I}(\mathbf{0}^-), \boldsymbol{M}(\mathbf{0}^+)\right)\right] \quad (2.15a)$$

$$J^*\left(t+1^-|\boldsymbol{q}(\boldsymbol{t}^-), \boldsymbol{I}(\boldsymbol{t}^-), \boldsymbol{M}(\boldsymbol{t}^+)\right)$$
$$= \min_{\substack{I_k(t+1^-)\in\mathcal{L}_k,\\ M_k(t+1^+)\in\mathcal{M}_k,\forall k}} \underset{\boldsymbol{s}(\boldsymbol{t}+\mathbf{1}^-)|\boldsymbol{q}(\boldsymbol{t}^-),\boldsymbol{I}(\boldsymbol{t}^-),\boldsymbol{M}(\boldsymbol{t}^+)}{E}\left[\left\{\sum\nolimits_{k=1}^{K} C_k^U\left(q_k(t+1^-), M_k(t+1^+)\right)\right\} + \alpha \cdot J^*\left(t+2^-|\boldsymbol{q}(\boldsymbol{t}+\mathbf{1}^-), \boldsymbol{I}(\boldsymbol{t}+\mathbf{1}^-), \boldsymbol{M}(\boldsymbol{t}+\mathbf{1}^+)\right)\right] \quad t = 0,1,\cdots,T-2 \quad (2.15b)$$

---

on, model (2.15a-b) will be updated as described in Figure 2.10, and this updated model will be used to re-optimize the inspection and MR&R decisions applied at $\tau + 1$.

$$J^*(T^-) = \underset{\boldsymbol{s}(\boldsymbol{T}^-)}{E}\left[f^1\left(\boldsymbol{q}(\boldsymbol{T}^-)\right)\right] \quad (2.15c)$$

where

$J^*(0^-)$ is the optimal total discounted user cost for the pavement system over planning horizon T;

$J(\boldsymbol{q}(\boldsymbol{0}^-), \boldsymbol{I}(\boldsymbol{0}^-), \boldsymbol{M}(\boldsymbol{0}^+))$ is the total discounted user cost for the pavement system, given system initial condition state $\boldsymbol{q}(\boldsymbol{0}^-)$ under inspection policy $\boldsymbol{I}(\boldsymbol{0}^-)$ and MR&R policy $\boldsymbol{M}(\boldsymbol{0}^+)$, over planning horizon T, assuming that the discounted costs from the next year (i.e., $t = 1$) to the end of planning horizon is optimal given $\boldsymbol{q}(\boldsymbol{0}^-)$, $\boldsymbol{I}(\boldsymbol{0}^-)$ and $\boldsymbol{M}(\boldsymbol{0}^+)$;

$J_k(q_k(0^-), I_k(0^-), M_k(0^+))$ is the total discounted user cost for segment k, given initial condition $q_k(0^-)$ under inspection policy $\boldsymbol{I_k}(\boldsymbol{0}^-)$ and MR&R policy $\boldsymbol{M_k}(\boldsymbol{0}^+)$, over planning horizon T;

$J^*\left(t+1^-|\boldsymbol{q}(\boldsymbol{t}^-), \boldsymbol{I}(\boldsymbol{t}^-), \boldsymbol{M}(\boldsymbol{t}^+)\right)$ is the optimal total discounted user cost for the pavement system from $(t+1)^-$ to T, given the information at t, $\boldsymbol{q}(\boldsymbol{t}^-), \boldsymbol{I}(\boldsymbol{t}^-)$, and $\boldsymbol{M}(\boldsymbol{t}^+)$, for all $t \in [0, 1, \dots, T-2]$;

$C_k^U\left(q_k(t^-), M_k(t^+)\right)$ is the sum of users' vehicle operating costs for segment k during period t, given the period's initial condition state $q_k(t^-)$ and users' traffic disruption cost due to the MR&R option $M_k(t^+)$, $k \in \{1,2,\cdots,K\}$, $t \in [0,1,\cdots,T-1]$;

$J^*(T)$ is the expected salvage value, which is an expected value of the terminal costs, $f^1\left(\boldsymbol{q}(\boldsymbol{T}^-)\right)$;

$f^1\left(\boldsymbol{q}(\boldsymbol{T}^-)\right)$ is the terminal costs, a function of the terminal condition state, $\boldsymbol{q}(\boldsymbol{T}^-)$;

$\mathcal{L}$ is a set of all possible inspection policies;

$\mathcal{L}_k$ is a set of all possible inspection policies for segment k, $\mathcal{L}_k \in \{0,1\}$;

$\mathcal{M}$ is a set of all possible MR&R policies;

$\mathcal{M}_k$ is a set of all possible M&R policies for segment k;

t is the time period index;

t^- is the start time of period t, i.e. the time immediately before inspection;

t^+ is the time immediately after inspection and immediately before MR&R activities in period t;

α is the discounted factor;

T is the duration of the planning horizon;

K is the number of pavement segments;

$E[\cdot]$ is the expected value operator.

2.2.1.1.2 System-level budget constraints

We assume that flexible budget allocation and transfer is allowed during the planning horizon under a given average annual budget.

$$\underset{\boldsymbol{s}(\mathbf{0}^-)}{E}[A(\boldsymbol{q}(\mathbf{0}^-),\boldsymbol{I}(\mathbf{0}^-),\boldsymbol{M}(\mathbf{0}^+))] = \underset{\boldsymbol{s}(\mathbf{0}^-)}{E}[\textstyle\sum_{k=1}^{K} A_k(q_k(0^-),I_k(0^-),M_k(0^+))]$$

$$= \underset{\boldsymbol{s}(\mathbf{0}^-)}{E}\Big[\Big\{\textstyle\sum_{k=1}^{K}\Big\{\sum_{p=1}^{P} C_{kp}\big(q_k(0^+),M_k(0^+)\big) + cm_k \cdot I_k(0^-)\Big\}\Big\} + \alpha \cdot$$

$$A^*\big(1^-|\boldsymbol{q}(\mathbf{0}^-),\boldsymbol{I}(\mathbf{0}^-),\boldsymbol{M}(\mathbf{0}^+)\big)\Big] \le \frac{1-\alpha^T}{1-\alpha}B \qquad (2.16a)$$

where

$$A^*\big(t+1^-|\boldsymbol{q}(\boldsymbol{t}^-),\boldsymbol{I}(\boldsymbol{t}^-),\boldsymbol{M}(\boldsymbol{t}^+)\big) =$$

$$\underset{\boldsymbol{s}(\boldsymbol{t}+\mathbf{1}^-)|\boldsymbol{q}(\boldsymbol{t}^-),\boldsymbol{I}(\boldsymbol{t}^-),\boldsymbol{M}(\boldsymbol{t}^+)}{E}\Big[\Big\{\textstyle\sum_{k=1}^{K}\Big\{\sum_{p=1}^{P} C_{kp}\big(q_k(t+1^+),M_k(t+1^+)\big) + cm_k \cdot I_k(t+$$

$$1^-)\Big\}\Big\} + \alpha \cdot A^*\big(t+2^-|\boldsymbol{q}(\boldsymbol{t}+\mathbf{1}^-),\boldsymbol{I}(\boldsymbol{t}+\mathbf{1}^-),\boldsymbol{M}(\boldsymbol{t}+\mathbf{1}^+)\big)\Big] \quad t = 0,1,\cdots,T-2$$

(2.16b)

$$A^*(T^-) = 0 \qquad (2.16c)$$

$A(\boldsymbol{q}(\boldsymbol{0}^-), \boldsymbol{I}(\boldsymbol{0}^-), \boldsymbol{M}(\boldsymbol{0}^+))$ is the total discounted agency cost over T for the pavement system, given initial condition $\boldsymbol{q}(\boldsymbol{0}^-)$ under inspection policy $\boldsymbol{I}(\boldsymbol{0}^-)$ and MR&R policy $\boldsymbol{M}(\boldsymbol{0}^+)$;

$A_k(q_k(0^-), \boldsymbol{I_k}(\boldsymbol{0}^-), \boldsymbol{M_k}(\boldsymbol{0}^+))$ is the total discounted agency cost over T for segment k, given initial condition $q_k(0^-)$ under inspection policy $I_k(0^-)$ and MR&R policy $M_k(0^+)$, for all $k \in \{1, \cdots, K\}$;

$A^*\left(t+1^-|\boldsymbol{q}(\boldsymbol{t}^-), \boldsymbol{I}(\boldsymbol{t}^-), \boldsymbol{M}(\boldsymbol{t}^+)\right)$ is the discounted agency cost from $(t+1)^-$ to T under $\boldsymbol{q}(\boldsymbol{t}^-), \boldsymbol{I}(\boldsymbol{t}^-)$, and $\boldsymbol{M}(\boldsymbol{t}^+)$, corresponding to $J^*\left((t+1)^-|\boldsymbol{q}(\boldsymbol{t}^-), \boldsymbol{I}(\boldsymbol{t}^-), \boldsymbol{M}(\boldsymbol{t}^+)\right)$ for all $t \in [0,1, \ldots, T-1]$;

$A^*(T^-)$ is the terminal value;

$C_{kp}\left(q_k(t^+), M_k(t^+)\right)$ is the agency cost for segment k, treatment p during period t, given the period's initial state $q_k(t^+)$ under MR&R option $M_k(t^+)$, $k \in \{1, \cdots, K\}$, $t \in \{0, \cdots, T-1\}$ (note that $q_k(t^-) = q_k(t^+)$ because an inspection activity does not influence the condition state);

$p \in \{1, \cdots, P\}$ is the index of the treatment (i.e., 1 for rehabilitation, 2 for reconstruction and 3 for preventive maintenance);

cm_k is the unit inspection cost for segment k, varying according to facility-specific characteristics (e.g. length of the segment);

B is the average annual budget.

2.2.1.1.3 Deterioration model

In the optimization model presented in Section 2.2.1.1.1 and Section 2.2.1.1.2, it is necessary to predict the progression of the pavement condition to calculate the

user and agency costs. In this section, we propose a stochastic pavement deterioration model as shown below:

$$s_k(t+1^-) = \tilde{\mathcal{F}}_\tau\left(s_k(t^{++})|\overline{\boldsymbol{\theta}}_{\boldsymbol{\tau}}, \overline{\boldsymbol{\Psi}}_{\boldsymbol{k,\tau}}, \sigma_\tau^2\right) \quad (2.17)$$

where the deterioration model between $s_k(t^{++})$ and $s_k(t+1^-)$ is denoted as $\tilde{\mathcal{F}}_\tau(\cdot)$; and the set of segment-specific and non-segment-specific model parameters, are denoted as $\overline{\boldsymbol{\Psi}}_{\boldsymbol{k,\tau}}$ and $\overline{\boldsymbol{\theta}}_{\boldsymbol{\tau}}$ respectively. The model uncertainty is summarized as $\sigma_{\boldsymbol{\tau}}^2$. The t^{++} indicates the time after the application of MR&R activity. We propose this model (2.17) based on the following assumptions:

Assumption 2.1. Except for the considered segment-specific factors coupled with the model parameters, $\overline{\boldsymbol{\Psi}}_{\boldsymbol{k,\tau}}$, such as structural design and traffic loading, all of the segments included in the target pavement network were built in a homogenous natural environment with common physical conditions, such as regional climate and other construction-related specifications, including materials. Therefore, there exist some common model parameters across the segments, and they are denoted by $\overline{\boldsymbol{\theta}}_{\boldsymbol{\tau}}$. In addition, we assume the uncertainty parameter, $\sigma_{\boldsymbol{\tau}}^2$, is also common for all the segments.

Assumption 2.2. The pavement system has a historical record of past inspection results (i.e., before τ), and the model parameters can be estimated based on the historical data at the current time τ.

Equation (2.18) describe the deterioration model we adopt in this section. It has the same functional form as in Paterson (1990), which has been commonly used in the previous literature.

$$s_k(t+1^-) = s_k(t^{++})e^{\theta_\tau^1} + \theta_\tau^2 l_k(1+SN_k)^{\theta_\tau^3} e^{\theta_\tau^1(\mathit{\Lambda}_k(t^{++})+1)} + \varepsilon_\tau \quad (2.18a)$$

$$h_k(t+1^-) = h_k(t^{++}) + 1 \quad (2.18b)$$

where l_k and SN_k represent the traffic loading and structure number of segment k, respectively. The error term, ε_τ, for the time interval $[t^{++}, t+1^-]$ has zero mean and variance of σ_τ^2. We assume that the error terms for different periods are i.i.d. random variables. The deterioration model (2.18) can be written in a more general form:

$$s_k(t+u^-) = s_k(t^{++})e^{u\Theta_\tau^1} + u\Theta_\tau^2 l_k(1+SN_k)^{\Theta_\tau^3} e^{\Theta_\tau^1(h_k(t^{++})+u)} + u\varepsilon_\tau \quad (2.19a)$$

$$h_k(t+u^-) = h_k(t^{++}) + u \quad (2.19b)$$

The model parameters, Θ_τ^1, Θ_τ^2 and Θ_τ^3, as well as σ_τ^2, are updated sequentially at every decision-making time step, τ, based upon the inspection data collected before τ. The resulting model is used to determine the optimal inspection and MR&R policies at the current stage, τ. More details on cost and performance models are presented in Section 2.2.3.

2.2.1.2 A bottom-up approach using Lagrange multipliers

In this section, we propose a bottom-up solution method to solve the stochastic system-level optimization problem, which allows for segment-specific features and the influence of future budgets on current inspection and MR&R decisions.

Corresponding to the above system-level formulation (2.15-16), we first use the Lagrange relaxation method to decompose the stochastic system-level problem into K segment-level subproblems by relaxing the budget constraints (2.16). This decomposition is independent of any specific segment-level models. With the non-

negative Lagrange multipliers, λ, the corresponding Lagrange function is shown below:

$$L(\boldsymbol{q}(\mathbf{0}^-), \boldsymbol{I}(\mathbf{0}^-), \boldsymbol{M}(\mathbf{0}^+)|\lambda) = J(\boldsymbol{q}(\mathbf{0}^-), \boldsymbol{I}(\mathbf{0}^-), \boldsymbol{M}(\mathbf{0}^+)) + \lambda \cdot \left(A(\boldsymbol{q}(\mathbf{0}^-), \boldsymbol{I}(\mathbf{0}^-), \boldsymbol{M}(\mathbf{0}^+)) - \frac{1-\alpha^T}{1-\alpha} B\right) \quad (2.20)$$

The original system-level problem can be converted to solving the corresponding Lagrange dual problem, defined as:

$$L^*\left(\boldsymbol{q}(\mathbf{0}^-)\right) = \sup_{\lambda} L^*(\boldsymbol{q}(\mathbf{0}^-)|\lambda) = \sup_{\lambda} \min_{\boldsymbol{I}(\mathbf{0}^-), \boldsymbol{M}(\mathbf{0}^+)|\lambda} \underset{\boldsymbol{s}(\mathbf{0}^-)}{E}[L(\boldsymbol{q}(\mathbf{0}^-), \boldsymbol{I}(\mathbf{0}^-), \boldsymbol{M}(\mathbf{0}^+)|\lambda)] \quad (2.21)$$

In problem (2.21), the decision variables are $\{\lambda; \boldsymbol{I}(\mathbf{0}^-), \boldsymbol{M}(\mathbf{0}^+)\}$. Then a two-step approach can be used to find the optimal solution for problem (2.21). In step 1 (the lower-level problem), for a given λ, we optimize the Lagrange function with respect to management policy $\{\boldsymbol{I}(\mathbf{0}^-), \boldsymbol{M}(\mathbf{0}^+)\}$; in step 2 (the upper-level problem), we find the optimal Lagrange multipliers, λ^*, for maximizing $\underset{\boldsymbol{s}(\mathbf{0}^-)}{E}[L^*(\boldsymbol{q}(\mathbf{0}^-)|\lambda)]$. Note that in step 1, the interdependence among segments has been removed by relaxing the budget constraints. Consequently, the system-level problem is separable, and for a given λ, the optimal policy $\{\boldsymbol{I}^*(\mathbf{0}^-), \boldsymbol{M}^*(\mathbf{0}^+)\}$ is the set of the segment-level optimal management policy $\{I_k^*(0^-), M_k^*(0^+)|\forall k\}$:

$$\{\boldsymbol{I}^*(\mathbf{0}^-), \boldsymbol{M}^*(\mathbf{0}^+)|\lambda\} = \{I_k^*(0^-), M_k^*(0^+)|\lambda, \forall k\} = \left\{ \underset{I_k(0^-)\in\mathcal{L}_k, M_k(0^+)\in\mathcal{M}_k}{\operatorname{argmin}} \underset{s_k(0^-)}{E}[H_k(q_k(0^-), I_k(0^-), M_k(0^+)|\lambda)], \forall k \right\} \quad (2.22a)$$

$$H_k^*(q_k(0^-)|\lambda) = \min_{I_k(0^-)\in\mathcal{L}_k, M_k(0^+)\in\mathcal{M}_k} \underset{s_k(0^-)}{E}[H_k(q_k(0^-), I_k(0^-), M_k(0^+)|\lambda)] \ \forall k \quad (2.22b)$$

where

$$H_k(q_k(0^-), I_k(0^-), M_k(0^+)|\lambda) = J_k(q_k(0^-), I_k(0^-), M_k(0^+)) + A_k(q_k(0^-), I_k(0^-), M_k(0^+)) \quad (2.22c)$$

We assume the lower-level problem is solved (an example is described in Section 2.2.1.3). Then the optimization problem (2.2.1) is reduced to searching for the optimal Lagrange multiplier, satisfying:

$$L^*\left(\boldsymbol{q}(\mathbf{0}^-)\right) = \sup_{\lambda} L^*(\boldsymbol{q}(\mathbf{0}^-)|\lambda) = \sup_{\lambda}\left[\sum_{k=1}^{K} H_k^*(q_k(0^-)|\lambda) - \lambda\frac{1-\alpha^T}{1-\alpha}B\right] \quad (2.23)$$

Since calculating the derivatives of $L^*(\boldsymbol{q}(\mathbf{0}^-)|\lambda)$ with regards to λ is often computationally intensive due to the complicated mathematical form of MR&R performance and cost models, we next present a numerical algorithm using the gradient approximation method to update the Lagrange multipliers. Each iteration is indexed by $\tilde{n}$, and the Lagrange multiplier vector in iteration $\tilde{n}$ is denoted by $\lambda^{\tilde{n}}$. The numerical algorithm is summarized as Algorithm 2.4. Note that this algorithm guarantees global convergence under the assumption that the Lagrange function $L^*(\boldsymbol{q}(\mathbf{0}^-)|\lambda)$ is a pseudo-convex envelope with respect to λ.

Algorithm 2.4:

Step 1. Initialization:

Step 1.1. Select $\lambda^0 \leftarrow 0$ and find $\{\boldsymbol{I}^{\mathbf{0}}(\mathbf{0}^-), \boldsymbol{M}^{\mathbf{0}}(\mathbf{0}^+)\} \leftarrow \{\boldsymbol{I}^*(\mathbf{0}^-), \boldsymbol{M}^*(\mathbf{0}^+)\}|\lambda^0$;

Step 1.2. If $\{\boldsymbol{I}^{\mathbf{0}}(\mathbf{0}^-), \boldsymbol{M}^{\mathbf{0}}(\mathbf{0}^+)\}$ satisfies the budget constraints, then stop. Otherwise, go to *Step 2*.

Step 2. Randomly select $\lambda^1 > 0$ and find $\{\boldsymbol{I}^{\mathbf{1}}(\mathbf{0}^-), \boldsymbol{M}^{\mathbf{1}}(\mathbf{0}^+)\} \leftarrow \{\boldsymbol{I}^*(\mathbf{0}^-), \boldsymbol{M}^*(\mathbf{0}^+)\}|\lambda^1$; set $\tilde{n} = 1$;

Step 3. If the terminal condition is not satisfied, do the following:

Step 3.1. $\tilde{n} \leftarrow \tilde{n} + 1$;

Step 3.2. Update $\lambda^{\tilde{n}}$ using the gradient approximation method:

$\lambda^{\tilde{n}} = \lambda^{\tilde{n}-1} - \bar{\alpha}^{\tilde{n}-1} \cdot \varrho^{\tilde{n}-1} / \left(\lambda^{\tilde{n}-1} - \lambda^{\tilde{n}-2}\right)$, where

$\varrho^{\tilde{n}-1} = L^*\left(\boldsymbol{q}(\mathbf{0}^-), \boldsymbol{I}^{\tilde{\boldsymbol{n}}-\mathbf{1}}(\mathbf{0}^-), \boldsymbol{M}^{\tilde{\boldsymbol{n}}-\mathbf{1}}(\mathbf{0}^+)\middle|\lambda^{\tilde{n}-1}\right) -$
$L^*\left(\boldsymbol{q}(\mathbf{0}^-), \boldsymbol{I}^{\tilde{\boldsymbol{n}}-\mathbf{2}}(\mathbf{0}^-), \boldsymbol{M}^{\tilde{\boldsymbol{n}}-\mathbf{2}}(\mathbf{0}^+)\middle|\lambda^{\tilde{n}-2}\right)$ and $\bar{\alpha}_{\tilde{n}-1}$ is a positive step size.

Step 3.3. Find $\{\boldsymbol{I}^{\tilde{\boldsymbol{n}}}(\mathbf{0}^-), \boldsymbol{M}^{\tilde{\boldsymbol{n}}}(\mathbf{0}^+)\} \leftarrow \{\boldsymbol{I}^*(\mathbf{0}^-), \boldsymbol{M}^*(\mathbf{0}^+)\}|\lambda^{\tilde{n}}$.

Now we can find the unique optimal inspection and MR&R policies applied to all segments at τ. Then the original optimization problem defined by (2.15-2.16) is updated. At the next decision time $\tau + 1$, the updated problem can then be solved by adopting the same approach as above, but with the new deterioration model, $\tilde{\mathcal{F}}_{\tau+1}$ and available budget. In this thesis, we use maximum likelihood estimation to update $\tilde{\mathcal{F}}_{\tau}$ with all the inspection data collected until τ. The details are furnished in Appendix G.

2.2.2 A general segment-level solution approach

To find the best policies for segment k at the current time point 0^-, i.e., $I_k^*(0^-)$ and $M_k^*(0^+)$, we need to consider the lifecycle analysis along the horizon between 0^- and T. The solution process to find $I_k^*(0^-)$ comprises two steps:

Step 1. Given λ, find the best MR&R activity $M_k^*(0^+)$ and the best timing of the next inspection, $\Delta t_k^*(0^+)$, assuming that the inspection is undertaken at the current decision-making time, $t = 0^-$ (i.e., $I_k(0^-) = 1$ at τ).

Step 2. Decide whether to perform an inspection at the current decision-making time, $t = 0^-$, based on the comparison between the results of *Step 1* and the results if the inspection is not conducted (i.e., $I_k(0^-) = 0$ at τ).

If we apply an inspection at the current decision-making time, 0^-, we know the true condition. Then the optimal decisions after the inspection, $\{M_k^*(0^+), \Delta t_k^*(0^+)|\lambda, I_k(0^-) = 1\}$, can be obtained by solving the following Bellman equations backward from $t = T - 1$ to $t = 0$:

$$H_k^*(q_k(t^+)|\lambda, I_k(t^-)=1) = \min_{\substack{M_k(t^+)\in\mathcal{M}_k \\ \Delta t\in[1,\dots,\mathbb{t}]}} \{\lambda\cdot\sum_{p=1}^{P} C_{kp}\big(q_k(t^+), M_k(t^+)\big) +$$

$$\sum_{\mu=0}^{\Delta t-1}\alpha^{\mu} \underset{s_k(t+\mu^-)|q_k(t^+),M_k(t^+)}{E}\big[C_k^U\big(q_k(t+\mu^-)\big)\big] + \alpha^{\Delta t}\cdot$$

$$\underset{s_k(t+\Delta t^-)|q_k(t^+),M_k(t^+)}{E}[H_k^*(q_k(t+\Delta t^+)|\lambda, I_k(t+\Delta t^-)=1)] + \alpha^{\Delta t}\cdot\lambda\cdot$$

$$cm_k\}, \qquad \forall\, t=T-1,\cdots,0 \tag{2.24a}$$

$$\{M_k^*(t^+), \Delta t_k^*(t^+)|\lambda, I_k(t^-)=1\} = \underset{\substack{M_k(t^+)\in\mathcal{M}_k, \\ \Delta t\in[1,\dots,\mathbb{t}]}}{\operatorname{argmin}} \{\lambda\cdot\sum_{p=1}^{P} C_{kp}\big(q_k(t^+), M_k(t^+)\big) +$$

$$\sum_{\mu=0}^{\Delta t-1}\alpha^{\mu} \underset{s_k(t+\mu^-)|q_k(t^+),M_k(t^+)}{E}\big[C_k^U\big(q_k(t+\mu^-)\big)\big] + \alpha^{\Delta t}\cdot$$

$$\underset{s_k(t+\Delta t^-)|q_k(t^+),M_k(t^+)}{E}[H_k^*(q_k(t+\Delta t^+)|\lambda, I_k(t+\Delta t^-)=1)] + \alpha^{\Delta t}\cdot\lambda\cdot$$

$$cm_k\}, \qquad \forall\, t=T-1,\cdots,0 \tag{2.24b}$$

$$H_k^*(q_k(T)|\lambda) = J_k^*\big(q_k(T)\big), \forall q_k(T) \tag{2.24c}$$

where

$H_k^*(q_k(t^+)|\lambda, I_k(t^-)=1)$is the value function associated with state $q_k(t^+)$ for segment k at t^+, assuming an inspection is performed at t^-, which represents the minimum expected cost-to-go from t^+ to the end of the planning horizon T, given Lagrange multipliers λ;

$\mathbb{t}$ is the maximum duration between two consecutive inspections;

Δt is the duration until the next inspection;

$\{M_k^*(t^+), \Delta t_k^*(t^+)|\lambda, I_k(t^-)=1\}$ is the optimal MR&R option and the duration until next inspection for segment k at the decision time t^+, assuming the inspection is performed at t^-, given Lagrange multipliers λ.

The main advantage of solving (2.24a-c) backward recursively is that it guarantees a globally optimal solution. However, traditional dynamic programming suffers from the curse of dimensionality, wherein the computation time increases exponentially with the dimension of the problem. Consequently, such an approach

is not suitable for a large-scale system-level problem. In this thesis, we also propose an approximate dynamic programming approach, *q-learning*, to solve the segment-level problem. The details are furnished in Appendix H.

In Step 2, we choose an action between the two options: performing an inspection at the current decision-making time point, $t = 0^-$ or doing nothing by (2.25). Here we denote the time when the latest inspection was conducted on segment k before the current time 0^- as ξ_k.

$$I_k(0^-) = \underset{I_k(0^-)}{\operatorname{argmin}} \Big\{ \big(1 - I_k(0^-)\big) \cdot \min_{\Delta t \in [1,\ldots,\mathbb{t}-\xi_k]} \Big[\sum_{\mu=0}^{\Delta t-1} \alpha^{\mu} \underset{s_k(\mu^-)|q_k(-\xi_k^+)}{E} \big[C_k^U\big(q_k(\mu^-)\big)\big] + \alpha^{\Delta t} \underset{s_k(\Delta t^-)|q_k(-\xi_k^+)}{E} [H_k^*(q_k(\Delta t^+)|\lambda, I_k(\Delta t^-) = 1)] + \alpha^{\Delta t} \cdot \lambda \cdot cm_k \Big] + I_k(0^-) \cdot \Big(\underset{s_k(0^-)|s_k(-\xi_k^+)}{E} [H_k^*(q_k(0^+)|\lambda, I_k(0^-) = 1)] + \lambda \cdot cm_k \Big) \Big\} \quad (2.25)$$

Equation (2.25) shows that for segments with unknown condition states, the decision-maker needs to decide whether to conduct an inspection at the current time or later according to the latest deterioration model.

2.2.3 Numerical case studies in the stochastic scenario

We consider a pavement system consisting of 50 heterogeneous segments. The system includes a wide range of segments with various characteristics, including structure number, traffic loading, and initial condition states. All the numerical cases presented in this section are carried out via Matlab R2016a on a PC with Inter® Xeon® 3.60 GHz CPU, 32.0GB RAM, and Windows 10 Pro 64-bit. Section 2.2.3.1 presents the cost and performance models and the parameter values. The validation of segment-level approximate dynamic programming algorithm is

described in Section 2.2.3.2. System-level numerical cases are furnished in Section 2.2.3.3.

2.2.3.1 Cost models and parameter values

For simplicity, only two MR&R activities are considered in our numerical cases: rehabilitation and reconstruction. However, our model and method can be applied to the joint optimization of MR&R activities. The decision variables are $M_k(0^+) = \{\omega_{kt}, \chi_{kt,1}, \chi_{kt,2} | t = 0\}$, where the binary variable $\chi_{kt,p}$ ($p = 1,2$) is equal to 1 if a rehabilitation (corresponding $p = 1$) or reconstruction ($p = 2$) activity is executed in period t for segment k, respectively, and 0 otherwise. The ω_{kt} represents the rehabilitation intensity in period t for segment k. Thesegment-level cost and performance models are borrowed from Section 2.1.2. They are repeated as follows for the readers' convenience:

$$C_k^U\big(q_k(t)\big) = l_k(c_k^1 s_k(t) + c_k^2) \quad (2.26a)$$

$$C_{k,1}(q_k(t), M_k(t)) = \chi_{kt,1}(m_k^1 \omega_{kt} + m_k^2) \quad (2.26b)$$

$$C_{k,2}(q_k(t), M_k(t)) = \chi_{kt,2}(z_k^1 + z_k^2 l_k) \quad (2.26c)$$

$$s_k(t^+) - s_k(t^{++}) = \chi_{kt,1} G_k\big(\omega_{kt}, s_k(t^+)\big) + \chi_{kt,2}(s_k(t^+) - s_k^{new}), \forall t \quad (2.26d)$$

$$G_k\big(\omega_{kt}, s_k(t^+)\big) = \frac{g_k^1 s_k(t^+)}{g_k^2 s_k(t^+) + g_k^3} \omega_{kt} \quad (2.26e)$$

$$0 \le \omega_{kt} \le R_{kt} = \left(\frac{g_k^2}{g_k^1} + \frac{g_k^3}{g_k^1 s_k(t^+)}\right) \max(0, \min\{s_k(t^+) - s_k^*, g_k^1 s_k(t^+)\}), \forall t \quad (2.26f)$$

$$\chi_{kt,1} + \chi_{kt,2} \le 1, \forall t \quad (2.26g)$$

$$h_k(t^{++}) = h_k(t^+)\big(1 - \chi_{kt,2}\big), \forall t \quad (2.26h)$$

$$s_k^{new} \le s_k(t) \le s_k^{max}, \forall t \quad (2.26i)$$

$$T_k^{min} \chi_{kt,2} \le h_k(t) \chi_{kt,2} \le T_k^{max} \chi_{kt,2}, \forall t \quad (2.26j)$$

$$q_k(0) = (s_k(0), h_k(0)) \quad (2.26k)$$

The models for user cost $C_k^U\big(q_k(t)\big)$, rehabilitation cost $C_{k,1}(q_k(t), M_k(t))$ and reconstruction cost $C_{k,2}(q_k(t), M_k(t))$ in period t for segment k are described

in (2.26a-c) respectively, where c_k^1, c_k^2, m_k^1, m_k^2, z_k^1 and z_k^2 are (non-negative) cost coefficients. Constraints (2.26d) indicate the roughness reduction caused by a rehabilitation or reconstruction activity, where G_k represents the rehabilitation effectiveness defined in (2.26e); $s_k(t^+)$ and $s_k(t^{++})$ denote the roughness indices right before and after the MR&R activity, respectively; s_k^{new} is the roughness index immediately after a reconstruction; and g_k^1, g_k^2, and g_k^3 are coefficients. Constraints (2.26f) stipulate the upper bound, R_{kt}, for ω_{kt}. Constraints (2.26g) ensure that at most one activity is executed per period. Constraints (2.26h) ensure that the pavement age is reset to 0 after reconstruction. Constraints (2.26i-j) specify the upper and lower bounds of the roughness level and the pavement's lifecycle length (i.e., the duration between two consecutive reconstruction activities). Constraint (2.26k) defines the initial pavement condition state.

Some of the parameter values used in our numerical cases are summarized in Table 2.3, while others are same as those summarized in Table 2.1. To account for heterogeneity among segments, we specify that the initial pavement states, traffic loading, structure number and some cost coefficients follow certain distributions, where $U[a, b]$ and $\Gamma[a, b]$ represent continuous and discrete uniform distributions, both bounded by a and b.

Table 2.3. Parameter values

Parameter	Value	Unit	Parameter	Value	Unit
c_k^1	U[38500, 42500]	\$/IRI/km/lane/ million ESAL	cm_k	U[34000, 38000]	\$/km/lane
α	0.9524	-	$\mathbb{t}$	3	year
SN_k	Γ[7,11]	SN	T	20	year

h_{k0}	$\Gamma[1,30]$	year			

2.2.3.2 Validation of the segment-level solution algorithm

At each time t, we have 9 available actions (do nothing, do rehabilitation or do reconstruction with $\Delta t = 1, 2$ or 3). Then given the Lagrange multiplier λ, we approximate the value function by a complete set of ordinary polynomials of total degree 4 of $s_k(t)$ and $e^{\Theta_t^1 h_{kt}}$ as follows:

$$\tilde{Q}_k^t(q_k(t), M_k(t), \Delta t|\lambda) = \sum_{i=1}^{\ell_n(t)} \rho_k^i(t|\lambda) \cdot \tilde{\xi}_k^i(q_k(t), M_k(t), \Delta t)$$

$$= \sum_{action=1}^{9} \sum_{i=1}^{\ell_n(t)} \rho_{k,action}^i(t|\lambda) \cdot \tilde{\xi}_{k,action}^i\big(q_k(t)\big)$$

$$= \sum_{action=1}^{9} \sum_{i=1}^{\ell_n(t)} \rho_{k,action}^i(t|\lambda) \cdot \tilde{\xi}_{k,action}^i(s_k(t), h_{kt})$$

$$= \sum_{action=1}^{9} \sum_{0 \le k_1+k_2 \le 4} \rho_{k,action}^{k_1,k_2}(t|\lambda) \cdot s_k(t)^{k_1} e^{k_2 \Theta_t^1 h_{kt}} \quad (2.27)$$

where k_1 and k_2 are non-negative integers; i.e.,

$(k_1, k_2) \in$

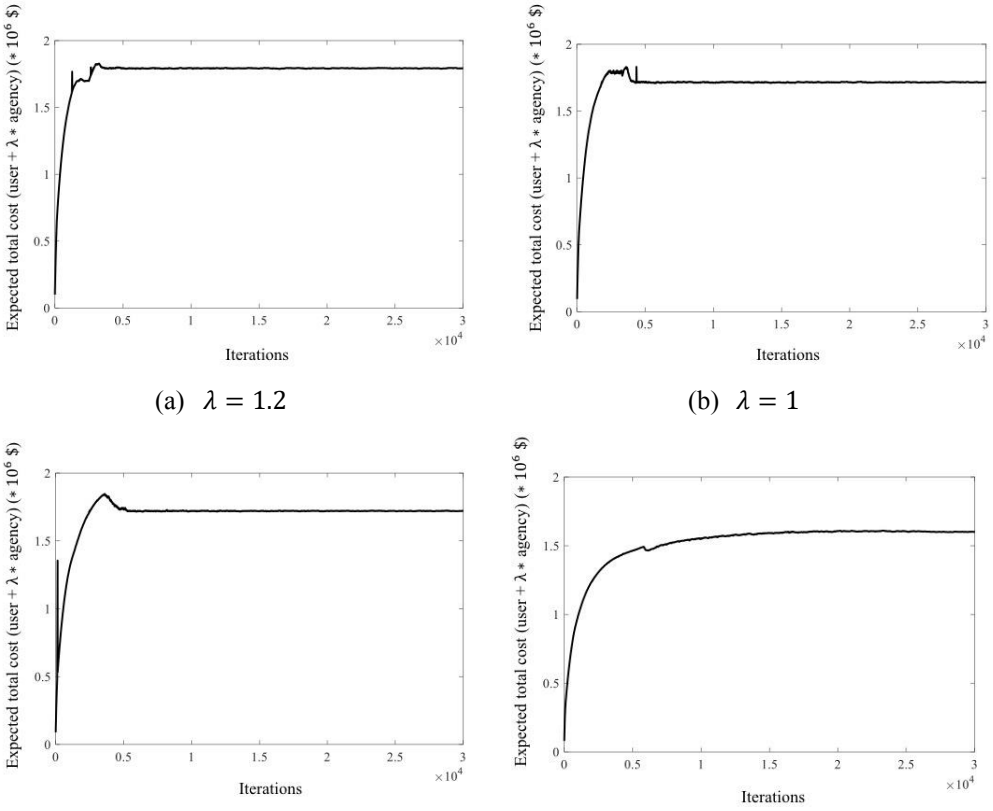
$\{(0,1), (1,0), (0,2), (1,1), (2,0), (0,3), (1,2), (3,0), (0,4), (2,2), (3,1), (4,0)\}$.

Given the *Q*-factors (defined in Appendix H), the corresponding total number of parameters for segment k at time t is $\ell_k(t) = 13 \times 9 = 135$. The step-size parameter used in this section is presented as follows (Powell, 2007):

$$\gamma_{\bar{n}} = \gamma_0 \frac{\left(\frac{b}{\bar{n}} + a\right)}{\left(\frac{b}{\bar{n}} + a + \bar{n}^{\beta}\right)} \quad (2.28)$$

where $\gamma_0 = 10^{-4}, b = 500, a = 300$ and $\beta = 0.6$.

To verify the convergence of the segment-level approximate dynamic-programming method, we present the numerical cases with an initial condition state $q_k(0) = (1.5,10)$ and various values of $\lambda \in \{0.6,0.8,1,1.2\}$ in Figure 2.11, which shows that the costs predicted by ADP converge after 5,000-7,000 iterations.



(a) $\lambda = 1.2$ (b) $\lambda = 1$

(c) $\lambda = 0.8$ (d) $\lambda = 0.6$

Figure 2.11. Convergence of the segment-level approximate dynamic programming method.

2.2.3.3 System-level numerical cases under combined budget

To better represent the heterogeneity among segments, we first generate a 50-sgement pavement system randomly with $s_k(0) \sim U[1,3]$, $SN_k \sim \Gamma[7,11]$, $l_k \sim U[0.4,0.9]$ and $h_{k0} \sim \Gamma[1,30]$, as illustrated in Figure 2.12. Each circle in the figure represents a pavement segment. The horizontal and vertical positions of a circle indicate the segment's initial age and initial roughness, respectively. The size of each circle is proportional to the segment's structural number. The circle's color represents the segment's traffic loading.

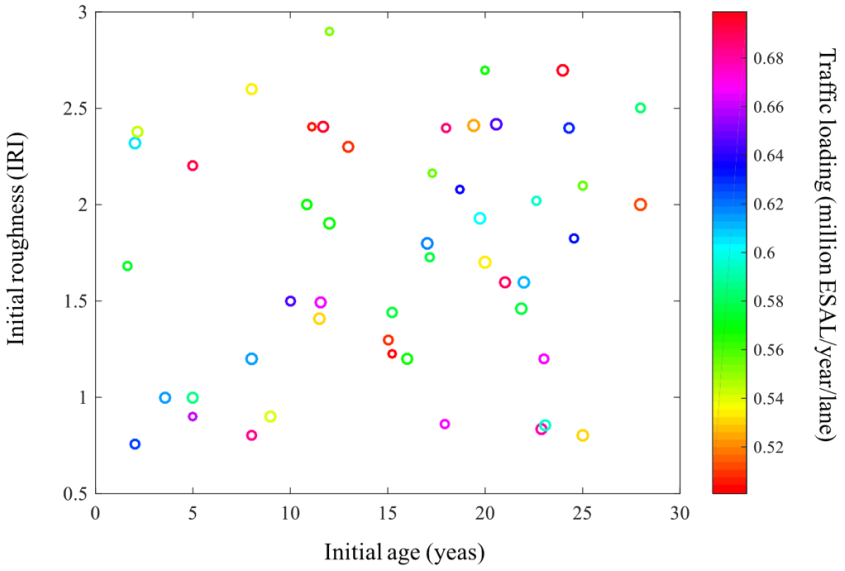


Figure 2.12. Initial conditions of the tested pavement system (50 segments).

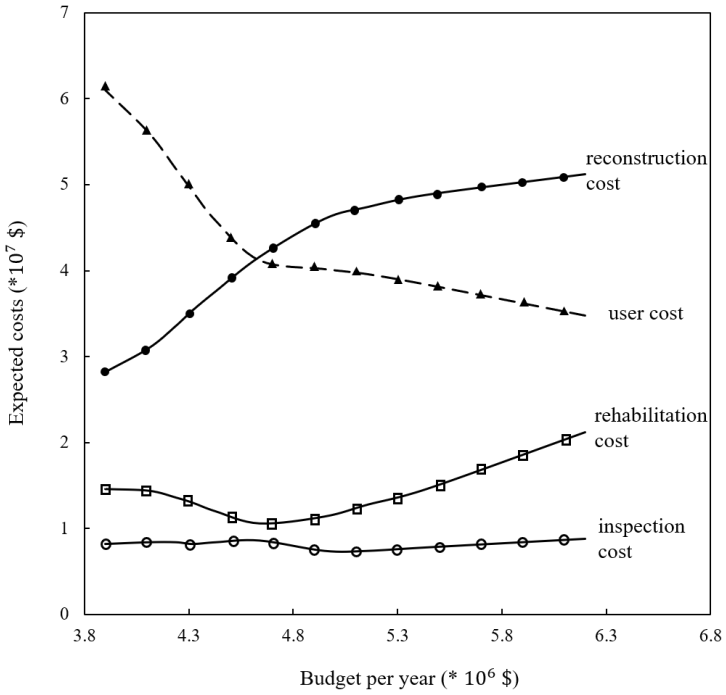


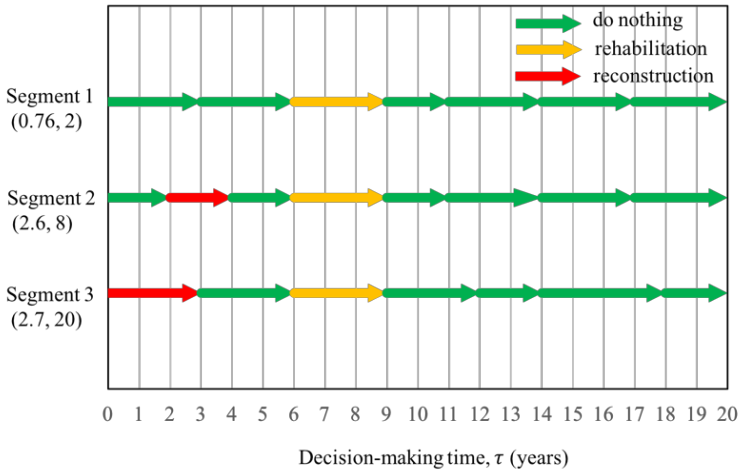
Figure 2.13. Effects of the combined agency budget on the system-level (50 segments) optimal costs.

The optimal expected user and agency costs are plotted against the combined annual budget: $B \in [3.9 \times 10^6, 6.2 \times 10^6]$ \$/year in Figure 2.13. The figure shows that the optimal expected user cost (the dashed curve with triangle markers) decreases as B grows, which is as expected. Meanwhile, the reconstruction cost (the dot-marked solid curve) increases with B. This means with a larger budget

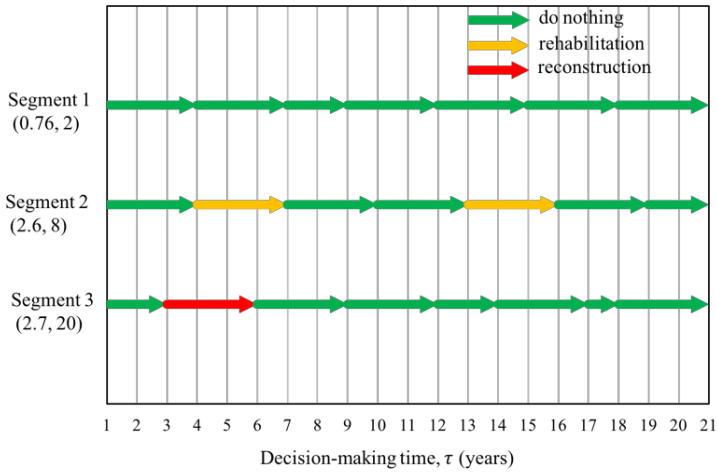
available, the agency will apply more reconstruction to improve the condition of the pavement system and reduce user cost. However, we note that the rehabilitation cost (the square-marked solid curve) diminishes at first and then increases as B rises. This is because when the budget is highly limited, the agency tends to perform more rehabilitation activities to postpone reconstruction activities, which are more expensive than rehabilitation activities. This finding is consistent with the results reported under the deterministic scenario in Section 2.1. As the budget increases, the budget becomes comparatively sufficient, and the agency will normally implement more rehabilitation and reconstruction activities to reduce user cost. We also find that the inspection cost (the circle-marked solid curve) is insensitive to B which indicates that a certain number of inspections are always required. When the budget is limited, inspections are needed to make more cost-effective rehabilitation and reconstruction decisions. On the other hand, when the budget is adequate, more rehabilitation and reconstruction activities are conducted, which also require more inspections since inspections are needed before every MR&R activity.

To examine how the budget affects the optimal rehabilitation and reconstruction plans for each segment at the system level, we compare realizations of optimal pavement management policy for three typical segments selected from the system tested under two budget values (Figure 2.14): (1) $B = 5.16 \times 10^6$ \$/year (adequate budget) and (2) $B = 4.32 \times 10^6$ \$/year (limited budget). The three selected segments are $(0.76, 2)$, $(2.8, 8)$ and $(2.7, 20)$, which represent segments with good, moderate and poor initial conditions, respectively. In each figure, the arrow length represents the interval between two consecutive inspections. Each arrow starts at a decision-making moment, and its color (green,

yellow or red) indicates the MR&R activity (do nothing, rehabilitation or reconstruction, respectively). As expected, fewer reconstructions and more 'do-nothing' are performed when the budget is limited, which is consistent with the findings in Figure 2.13. A comparison of optimal management policies among these three segments reveals that the optimal rehabilitation and reconstruction plans are significantly affected by the initial pavement conditions.



(a)



(b)

Figure 2.14. A realization of optimal pavement management policy under different budget constraints: (a) the annual budget is 5.16×10^6 \$/year; (b) the annual budget is 4.32×10^6 \$/year.

To examine the updating process of the deterioration model, the updating trajectories of Θ^1 and Θ^2 are plotted in Figure 2.15a and b for two identical systems with the same annual budget ($B = 4.8 \times 10^6$ \$/year), but with different historical data sets[6]. Figure 2.15a is plotted when the number of historical data points is insufficient, while Figure 2.15b is plotted with sufficient historical data points, which are collected before $\tau = \tau_0$, where τ_0 stands for the initial decision-making year. The accurate values of Θ^1 and Θ^2 in each figure are set to 0.04 and 930, respectively; see the red dot. The initial points presented in the two figures (i.e., $(0.02, 0.07)$ and $(0.03, 0.081)$) are the initial value of Θ^1 and $\Theta^2 \times 10^{-4}$ at $\tau = \tau_0$, which are obtained by the maximum likelihood estimates from the historical data. As expected, the limited historical data result in inaccurate initial parameter values. The figures reveal in both cases that the accuracy of deterioration model improves as more and more inspection results become available. To examine how the updating process contributes to pavement management, we also compare the total user costs against those without the updating process in the same simulation environment. We find that the user cost is reduced by 11.9% (i.e., from 1.18×10^7 \$ to 1.04×10^7 \$) for the case of Figure 2.15a and by 7.5% (i.e., reduced from 1.07×10^7 \$ to 0.99×10^7 \$) for the case of Figure 2.15b. A comparison between the two figures reveals that the updating process's effect tends to become more significant when the historical data is insufficient (i.e., when the deterioration model is very inaccurate initially).

[6] Here we assume Θ^3 and σ^2 is known without error for simplicity ($\Theta^3 = -4; \sigma^2 = 0.25$).

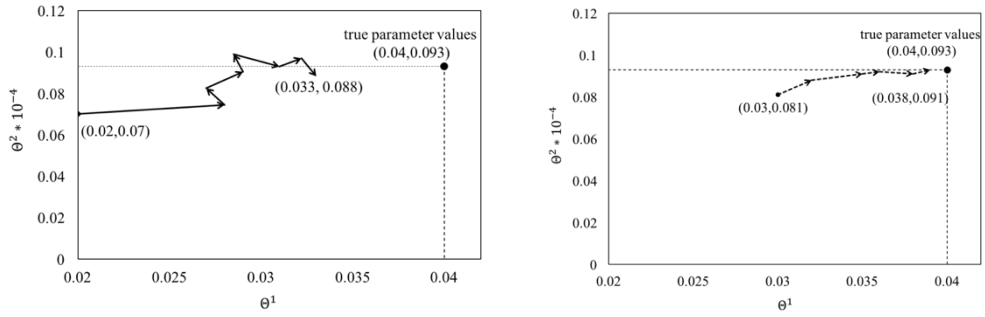


Figure 2.15. Convergence of the estimated parameters with different historical data size: (a) insufficient historical data points; (b) sufficient historical data points.

2.3 Summary of findings

This chapter formulated general mathematical models for the joint optimization of MR&R planning for a system of heterogeneous pavement segments under both deterministic and stochastic scenarios. Our numerical case studies reveal a number of useful findings and managerial insights.

For the work in the deterministic scenario, the key findings can be summarized as follows:

i) Compared with the conventional practice of using separate budgets for each treatment, optimally allocating a combined agency budget among treatments can reduce the total cost by up to 4%.

ii) Incorporating maintenance in the optimal MR&R planning will result in a total cost saving of over 6%. More importantly, it can significantly lower the minimum budget required to keep pavement system workable (by over 9% in our numerical cases).

iii) The agency should perform fewer reconstructions and more rehabilitations when the budget is more limited.

iv) The pavements' initial condition has a significant effect on the minimum budget required and the transient periods of the optimal MR&R plan, but has almost no effect on the steady-state periods of the MR&R plan.

The key findings from our work in the stochastic scenario can be summarized as follows:

(i) The agency should perform fewer reconstructions when the budget is limited; note this is consistent with the finding under the deterministic scenario.

(ii) Fewer rehabilitations are performed as the budget increases from a very-limited situation. But this trend reverses as the budget becomes abundant.

(iii) The optimal inspection schedules are insensitive to the budget.

(iv) Updating of the deterioration model in a rolling horizon is crucial for making efficient inspection and MR&R decisions, especially when the historical data is insufficient.

All these insights are helpful for agencies to plan future pavement management activities.

Chapter 3. Joint optimization of utilization and replacement for truck fleets

In this chapter, we develop an efficient solution procedure for a general truck fleet management model. The chapter is organized as follows. Section 3.1 presents a general discrete-time problem formulation. Section 3.2 proposes a bi-level solution approach, in which the lower level optimizes each truck's utilization by using continuous approximation (CA) and the Lagrange multiplier method; and the upper level optimizes the purchase and retirement schedules of the trucks by tabu search. The computation time and solution quality of our approach are tested in Section 3.3. Section 3.4 provides numerical case studies. The findings are summarized in Section 3.5.

3.1 A general formulation

In this section, we propose a general discrete-time formulation for the truck fleet management problem, which does not depend on the exact mathematical forms of cost functions.

The objective of the problem is to minimize the total discounted cost incurred by truck purchase, operation, maintenance and retirement over a given finite planning horizon T, as shown in (3.1a) below. The time is discretized by a prespecified time unit, which can be a year, a quarter, a month, a week, or even a day. If a smaller time unit is selected, a more accurate and detailed fleet utilization and replacement plan can be obtained, but the number of variables will also increase in polynomial order, which will result in a much higher computational cost. The decision

variables include the quantity of trucks purchased at time t, denoted by P_t ($1 \leq t \leq T$); the type of trucks purchased at time t, denoted by γ_t ($1 \leq t \leq T$) (note that we assume only one type of trucks is purchased at a time); the mileage served at time τ by a truck purchased at time t, denoted by $u_{\tau,t}$ ($1 \leq t \leq \tau \leq T$); and the time when the trucks in cohort t are retired, S_t ($1 \leq t \leq T$), where the cohort t includes all the trucks that are purchased at time t. We assume that the trucks belonging to one cohort have the same utilization and retirement plans. This assumption is consistent with the so-called "no-splitting property" reported in the literature of parallel replacement problems (Jones et al., 1991; Hartman, 2004; Guerrero et al, 2013). It is based upon the intuition that uneven distribution of workload among assets of the same type, age and cumulative utilization may lead to higher total cost, yet proof of this property was furnished under idealized assumptions only. The general formulation of the problem is presented as follows:

$$\min J = \sum_{t=1}^{T} A(\gamma_t) P_t e^{-rt} + \sum_{t=1}^{T} \sum_{\tau=t}^{S_t} P_t u_{\tau,t} M(y_{\tau,t}, \gamma_t) e^{-r\tau} - \sum_{t=1}^{T} P_t F\left(y_{S_t,t}, \gamma_t\right) e^{-rS_t} \quad (3.1a)$$

subject to:

$$\sum_{t:\, 1 \leq t \leq \tau \leq S_t} P_t u_{\tau,t} = D_\tau, \ 1 \leq \tau \leq T \quad (3.1b)$$

$$y_{\tau,t} = \sum_{s=t}^{\tau} u_{s,t}, \text{ and } y_{t-1,t} = 0, \ 1 \leq t \leq \tau \leq S_t \leq T \quad (3.1c)$$

$$\gamma_t \in H, \ 1 \leq t \leq T \quad (3.1d)$$

S_t is an integer satisfying $1 \leq t \leq S_t \leq T$ (3.1e)

P_t is an integer satisfying $P_t \geq 0, \ 1 \leq t \leq T$ (3.1f)

$$0 \leq u_{\tau,t} \leq U, 1 \leq t \leq \tau \leq S_t \leq T \quad (3.1g)$$

$$y_{S(t),t} \leq \bar{y}, \ 1 \leq t \leq S_t \leq T \quad (3.1h)$$

The first term on the right-hand-side of (3.1a) is the total discounted truck purchase cost, where $A(\gamma_t)$ denotes the unit cost for purchasing a truck of type γ_t. The second term is the total discounted operating and maintenance (O&M) cost of the truck fleet, where $M(y_{\tau,t}, \gamma_t)$ denotes the unit O&M cost per mile, which is a function of the odometer reading (i.e. cumulative mileage), $y_{\tau,t}$, of a truck in

cohort t at the end of time τ (see the definition of $y_{\tau,t}$ in constraint (3.1c)). We assume $M(y_{\tau,t}, \gamma_t) > 0$ and $\frac{\partial M}{\partial y_{\tau,t}} > 0$; the latter is reasonable because the maintenance cost increases, while the fuel efficiency decreases as vehicle mileage increases (CARB, 2008). The last term in the objective function denotes the total discounted salvage value, where $F(y_{S_t,t}, \gamma_t)$ indicates the salvage value of a truck in cohort t that retires at S_t. We assume $F(y_{S_t,t}, \gamma_t) \geq 0$ and $\frac{\partial F}{\partial y_{S_t,t}} < 0$; i.e., the salvage value is a decreasing function of the odometer reading. We further assume that when $\tau = S_t$, i.e., when cohort t is about to retire at the present time, $\frac{\partial M}{\partial y_{S_t,t}} - \frac{\partial^2 F}{\partial y_{S_t,t}^2} > 0$ for all $y_{S_t,t}$. This assumption is also reasonable. Note that $-\frac{\partial F}{\partial y_{S_t,t}}$ is the marginal salvage value loss as $y_{S_t,t}$ increases at retirement time S_t, and thus $M\left(y_{S_t,t}, \gamma_t\right) - \frac{\partial F}{\partial y_{S_t,t}}$ is the unit utilization cost per mile for a truck at retirement time. This assumption means that the unit utilization cost per mile for a truck at retirement time increases with the terminal mileage $y_{S_t,t}$.[7] The r denotes the discount rate.

The constraints (3.1b) are demand constraints, where D_τ is the given demand at time τ (measured in miles). For simplicity, we assume this demand is infinitely divisible. The types of trucks are defined by a finite set, H, as shown in constraint (3.1d). Constraints (3.1e-h) specify the bounds for S_t, P_t, $u_{\tau,t}$ and $y_{\tau,t}$,

[7] This assumption is needed when we develop the demand allocation rule (see Section 3.3). However, we understand that this assumption is not always true; in fact, its validity depends largely on the salvage value function F, which is determined by the market of used trucks. If this assumption is not satisfied, we can still develop an approach that significantly reduces the size of the solution space. But this approach is not as efficient as the one we present later in this chapter. Refer to Section 3.3 for details.

respectively; S_t and P_t are integers, where U is the maximum mileage a truck can serve per unit time, and $\bar{y}$ is the maximum allowable cumulative mileage.

3.2 Solution approach

Formulation (3.1a-h) is a non-convex mixed-integer nonlinear program with $\frac{T(T+7)}{2}$ decision variables P_t, S_t, γ_t and $u_{\tau,t}$ (or $y_{\tau,t}$)[8] $(1 \le t \le \tau \le S_t \le T)$. The optimal solution is very difficult to obtain using previous methods, like the branch-and-bound algorithm and dynamic programming, especially when T is large. To solve this problem, we first converted the original program (3.1a-h) to a continuous-time approximation model, where the decision variables and unit costs are written as functions of continuous time, and the summations are replaced by integrals (Section 3.2.1). From the first-order conditions of the CA model, we derive analytical conditions that an optimal solution (of the CA model) must satisfy (Section 3.2.2). With a certain assumption, these optimality conditions are then built upon to develop a near-optimal rule for allocating demand to the truck fleet at any time, given that the trucks' purchase and retirement schedules are fixed (Section 3.2.3). This demand allocation rule solves the lower-level problem, i.e., the optimization of truck utilization plans. This can reduce the original program with $\frac{T(T+7)}{2}$ decision variables to a new (upper-level) problem with as few as $3T + 1$ decision variables, including P_t, γ_t, S_t $(1 \le t \le T)$, and a Lagrange multiplier variable. (To see how this would greatly reduce the size of the solution space, note that when $T = 50$, the original program (3.1a-h) would have 1,425

[8] From constraints (3.1c), we see that $u_{\tau,t}$ $(1 \le t \le \tau \le S_t \le T)$ can also be written as a function of $y_{\tau,t}$ $(1 \le t \le \tau \le S_t \le T)$. Hence, the decision variables of the original problem (3.1a-h) can be defined as P_t, S_t, γ_t and $y_{\tau,t}$ $(1 \le t \le \tau \le S_t \le T)$.

decision variables, while the upper-level problem has only 151 variables.) The upper-level problem is then solved by a metaheuristic search algorithm (in this chapter, the tabu search algorithm), which is proposed in Section 3.2.4. The details are furnished as follows.

3.2.1 The CA model

In the CA model presented in (3.2a-h), the discrete-time decision variables P_t, γ_t, S_t, $y_{\tau,t}$ and $u_{\tau,t}$ are replaced by the continuous decision functions $P(t)$, $\gamma(t)$, $S(t)$, $u(\tau,t)$ and $y(\tau,t)$ ($1 \le t \le \tau \le S_t \le T$); the discrete-time demand function D_τ is replaced by the continuous function $D(\tau)$ ($0 \le \tau \le T$); the relation between the cumulative mileage and the mileage served at time τ is now written in the form of partial differential equation (3.2c); the summations in (3.1a-h) are replaced by integrals; and the integer constraints for $S(t)$ and $P(t)$ are removed. The rest of the formulation is the same as in (3.1a-h), including the time variables t and τ, although they are now continuous variables.

$$\min \tilde{J} = \int_{t=0}^{T} A(\gamma(t))P(t)e^{-rt}dt +$$

$$\int_{t=0}^{T}\int_{\tau=t}^{S(t)} P(t)u(\tau,t)M\big(y(\tau,t),\gamma(t)\big)\,e^{-r\tau}d\tau dt -$$

$$\int_{t=0}^{T} P(t)F(y(S(t),t),\gamma(t))e^{-rS(t)}\,dt \qquad (3.2a)$$

subject to:

$$\int_{t:0\le t\le\tau\le S(t)} P(t)u(\tau,t)\,dt = D(\tau), \quad \text{for } \tau \in [0,T] \qquad (3.2b)$$

$$\frac{\partial y(\tau,t)}{\partial \tau} = u(\tau,t), \text{ for } t \in [0,T], \tau \in [t,S(t)] \qquad (3.2c)$$

$$\gamma(t) \in H, \text{ for } t \in [0,T] \qquad (3.2d)$$

$$t \le S(t) \le T \text{ for } t \in [0,T] \qquad (3.2e)$$

$$P(t) \ge 0, \text{ for } t \in [0,T] \qquad (3.2f)$$

$$0 \le u(\tau,t) \le U, \text{ for } t \in [0,T], \tau \in [t,S(t)] \qquad (3.2g)$$

$$y(S(t),t) \le \bar{y}, \quad \text{for } t \in [0,T] \tag{3.2h}$$

Note that the above CA model also has a physical meaning: it can be considered as a "limiting" model of the original discrete-time program (3.1a-h) when the time interval for decisions approaches zero (i.e., when a decision can be made at any time); and the new decision functions $P(t)$ and $u(\tau,t)$ can be regarded as purchase and utilization rates per unit of time. Hence, the optimal solution of (3.2a-h) and the corresponding minimum cost $\tilde{J}^*$ should by nature be close to the optimal solution of the discrete program (3.1a-h) and the minimum cost J^*, respectively, especially when the time unit is small.

3.2.2 The optimality conditions of the CA model

We now fix $\gamma(t)$ and $S(t)$ $(t \in [0,T])$ and introduce Lagrange multipliers $\lambda(\tau)$ $(\tau \in [0,T])$, $\mu(\tau,t)$, $\varphi_1(\tau,t)$, $\varphi_2(\tau,t)$ and $\omega(t)$ $(t \in [0,T], \tau \in [t,S(t)])$ to relax constraints (3.2b), (3.2c), (3.2g) and (3.2h), respectively, (where $\varphi_1(\tau,t)$ and $\varphi_2(\tau,t)$ are used to relax the right- and left-hand sides of (3.2g), respectively). The corresponding Lagrangian function is presented as follows:

$$\tilde{L} = \tilde{J} + \int_{\tau=0}^{T} \lambda(\tau) D(\tau) e^{-r\tau}\, d\tau - \int_{t=0}^{T} \int_{\tau=t}^{S(t)} \lambda(\tau) P(t) u(\tau,t) e^{-r\tau} d\tau dt + \int_{t=0}^{T} \int_{\tau=t}^{S(t)} \mu(\tau,t) \left(u(\tau,t) - \frac{\partial y(\tau,t)}{\partial \tau} \right) e^{-r\tau} d\tau dt + \int_{t=0}^{T} \int_{\tau=t}^{S(t)} \varphi_1(\tau,t)\,(u(\tau,t) - U) e^{-r\tau} d\tau dt - \int_{t=0}^{T} \int_{\tau=t}^{S(t)} \varphi_2(\tau,t)\, u(\tau,t) e^{-r\tau} d\tau dt + \int_{t=0}^{T} \omega(t)(y(S(t),t) - \bar{y})\, e^{-rS(t)} dt \tag{3.3}$$

By examining the Karush-Kuhn-Tucker (KKT) necessary conditions for optimality of (3.3), we develop the following conditions, which should hold at optimality. The details of the derivation are shown in Appendix I for simplicity.

For any $t \in [0,T]$, either:

$$P(t) = 0 \tag{3.4}$$

or: for any $\tau \in [t, S(t))$, one of the following three conditions holds:

$$u(\tau, t) = 0 \tag{3.5a}$$

$$u(\tau, t) = U \tag{3.5b}$$

$$\tilde{z}(y(\tau, t)|\tau, t) \equiv M\big(y(\tau, t), \gamma(t)\big) = \lambda(\tau) - \frac{1}{r}\frac{d\lambda(\tau)}{d\tau} \tag{3.5c}$$

and for $\tau = S(t)$, one of the following four conditions holds:

$$u(S(t), t) = 0 \tag{3.6a}$$

$$u(S(t), t) = U \tag{3.6b}$$

$$y(S(t), t) = \bar{y} \tag{3.6c}$$

$$\tilde{z}(y(S(t), t)|S(t), t) \equiv M\big(y(S(t), t), \gamma(t)\big) - \frac{\partial F}{\partial y(S(t), t)} = \lambda(\tau) \tag{3.6d}$$

The above conditions can be explained as follows: for any truck in an existing cohort t, before its retirement time, either its utilization rate is 0 or the maximum rate U, or its cumulative mileage satisfies (3.5c); at its retirement time, either its utilization rate is 0 or the maximum rate U, or its cumulative mileage satisfies (3.6c) or (3.6d). Note that $\tilde{z}(y(\tau, t)|\tau, t)$ $(0 \le t \le \tau \le S(t) \le T)$ represents the unit utilization cost per mile at time τ per truck of cohort t. For simplicity, we term this $\tilde{z}(y(\tau, t)|\tau, t)$ as the "z-score" of cohort t at time τ. The right-hand sides of (3.5c) and (3.6d) reveal that this z-score is by and large independent of the purchase time t of the trucks. Specifically, for two different cohorts $t_1 \neq t_2$, we have:

$\tilde{z}(y(\tau, t_1)|\tau, t_1) = \tilde{z}(y(\tau, t_2)|\tau, t_2)$, if $\max\{t_1, t_2\} \le \tau < \min\{S(t_1), S(t_2)\}$, or if $\tau = S(t_1) = S(t_2)$ (3.7)

Equation (3.7) implies that at the same time τ, all the non-retiring cohorts (i.e., those with $S(t) > \tau$) should have the same unit utilization cost per mile; and the same is true for all the retiring cohorts (i.e., those with $S(t) = \tau$). This is intuitive: the less-expensive trucks in terms of utilization cost will be used more. In particular, the non-retiring cohorts of the same type should have the same cumulative mileage ; see equation (3.5c). And the same is true for the retiring cohorts; see (3.6d). However, a non-retiring cohort and a retiring cohort may not have the same z-score at the same time. These claims hold true only when the mileage bounds (i.e. equations (3.5a-b) and (3.6a-c)) are not attained.

Conditions (3.4-3.7) also imply that if $\lambda(\tau)$ and $P(t)$ are known for all $\tau, t \in [0, T]$, there exists a way to develop the optimal utilization plan, $u(\tau, t)$, for all cohorts. We next return to the original discrete-time program (3.1a-h) to develop a near-optimal demand allocation rule, which solves the lower level problem.

3.2.3 The lower-level solution approach

The discrete-time analogue of the optimality conditions (3.4-3.6d) is presented as follows:

For any $t = 1,2, \dots, T$, either:

$$P_t = 0 \tag{3.8}$$

or: for any $\tau = t, t+1, \dots, S_t - 1$, one of the following three conditions holds:

$$u_{\tau,t} = 0 \tag{3.9a}$$

$$u_{\tau,t} = U \tag{3.9b}$$

$$z_{\tau,t}(y_{\tau,t}) \equiv M(y_{\tau,t}, \gamma_t) = \lambda_\tau - \frac{1}{r}(\lambda_{\tau+1} - \lambda_\tau) \tag{3.9c}$$

and for $\tau = S_t$, one of the following three conditions holds:

$$u_{S_t,t} = U \tag{3.10a}$$

$$y_{S_t,t} = \bar{y} \tag{3.10b}$$

$$z_{S_t,t}\left(y_{S_t,t}\right) \equiv M\left(y_{S_t,t}, \gamma_t\right) - \frac{\partial F}{\partial y_{S_t,t}} = \lambda_\tau \tag{3.10c}$$

Note that under the discrete-time case, $u_{S_t,t} = 0$ should never happen; otherwise, the cohort t should retire at $S_t - 1$, but not S_t.

Built upon the above conditions (which do not guarantee global optimality because of the continuous approximation), we can determine $y_{\tau,t}$ when λ_t, S_t and γ_t are given for $1 \le t \le \tau \le S_t \le T$. We first define:

$$\zeta_{\tau,t} = \begin{cases} -\frac{1}{r}\lambda_{\tau+1} + \left(1 + \frac{1}{r}\right)\lambda_\tau & \text{if } 1 \le t \le \tau < S_t \le T \\ \lambda_\tau & \text{if } 1 \le t \le \tau = S_t \le T \end{cases} \tag{3.11}$$

and

$$\hat{y}_{\tau,t} \equiv z_{\tau,t}^{-1}(\zeta_{\tau,t}) \tag{3.12}$$

which is the inverse function of $z_{\tau,t}\left(y_{\tau,t}\right)$ so that $z_{\tau,t}\left(\hat{y}_{\tau,t}\right) = \zeta_{\tau,t}$.

For a cohort t, the $y_{\tau,t}$ is derived recursively as follows. For any $\tau = t, t+1, \ldots, S_t - 1$, let

$$y_{\tau,t} = \text{mid}\left\{y_{\tau-1,t}, \hat{y}_{\tau,t}, y_{\tau-1,t} + U\right\} \tag{3.13}$$

where the operator "mid" gives the middle value of the three arguments. Equation (3.13) is true because the z-score function $z_{\tau,t}\left(y_{\tau,t}\right) \equiv M\left(y_{\tau,t}, \gamma_t\right)$ monotonically increases with respect to $y_{\tau,t}$; therefore, $z_{\tau,t}^{-1}(\zeta_{\tau,t})$ is single-valued.

For $\tau = S_t$, under the assumption that $\frac{\partial M}{\partial y_{S_t,t}} - \frac{\partial^2 F}{\partial y_{S_t,t}^2} > 0$ for all $y_{S_t,t}$, $z_{\tau,t}\left(y_{\tau,t}\right)$ still monotonically increases with respect to $y_{S_t,t}$ (see (3.10c)); so

$$y_{S_t,t} = \text{mid}\left\{y_{S_t-1,t}, \hat{y}_{S_t,t}, \min\left\{y_{S_t-1,t} + U, \bar{y}\right\}\right\} \tag{3.14}$$

Equations (3.13) and (3.14) indicate that for any cohort t at time τ, the cumulative mileage of each truck tends to reach a *desired* level $\hat{y}_{\tau,t}$. This means when a truck's previous cumulative mileage $y_{\tau-1,t}$ exceeds $\hat{y}_{\tau,t}$, its present utilization will be set as $u_{\tau,t} = 0$. However, if a truck's cumulative mileage cannot reach $\hat{y}_{\tau,t}$ at the present time τ (i.e., when $y_{\tau-1,t} + U < \hat{y}_{\tau,t}$ if $\tau < S_t$; and when $\min\{y_{\tau-1,t} + U, \bar{y}\} < \hat{y}_{\tau,t}$ if $\tau = S_t$), its present utilization will be set as $u_{\tau,t} = U$ if it will not be retired at τ, and as $u_{\tau,t} = \min\{U, \bar{y} - y_{\tau-1,t}\}$ if it will be retired at τ.

Given the $y_{\tau,t}$ for all τ, t such that $1 \le t \le \tau \le S_t \le T$, the $u_{\tau,t}$ can be derived simply by:

$$u_{\tau,t} = y_{\tau,t} - y_{\tau-1,t} \tag{3.15}$$

Note that $y_{t-1,t} = 0$ for all $t = 1,2,\dots,T$. Finally, the P_τ can be calculated recursively using the demand constraints (3.1b):

$$P_\tau = \frac{1}{u_{\tau,\tau}}\left(D_\tau - \sum_{t:t<\tau\le S_t} P_t u_{\tau,t}\right) \tag{3.16}$$

Equations (3.11-3.16) can be used to find near-optimal plans for truck purchase and utilization when λ_t, γ_t and S_t $(1 \le t \le T)$ are given. This process can reduce the number of decision variables from $\frac{T(T+7)}{2}$ to $3T$. The remaining optimization problem with $3T$ variables is thus much easier to solve in a reasonable computation time. However, equation (3.16) may produce fractional values for P_τ, which does not satisfy the integer constraint (3.1f). Hence, we need to take one more step to address this issue. For example, we can search in the neighborhood of the fractional values for P_τ given by (3.16) for an optimal integer solution. This additional step will inevitably increase the computation cost, since

every time P_τ is changed, all the $u_{s,t}$ and $y_{s,t}$ for $s \geq \tau$ need to be adjusted to ensure that the demand constraints (3.1b) are satisfied. Hence, in what follows we present a more efficient solution procedure for the optimization of the truck utilization plan (i.e. $u_{\tau,t}$ or $y_{\tau,t}$), where instead of fixing λ_t, γ_t and S_t, we fix P_t, γ_t, S_t $(1 \leq t \leq T)$ and an additional variable λ_1 (thus, the number of variables to be solved at the upper level is $3T + 1$). We start by presenting the following rules for demand allocation.

Proposition 3.1. At any time $\tau \in [0, T]$, the existing cohorts are divided into two batches: the non-retiring cohorts with $S_t > \tau$ and the retiring cohorts with $S_t = \tau$. Given P_t, γ_t and S_t for all the $t \in [0, T]$, demand D_τ should be allocated to all the available trucks so that the z-scores of all the trucks in the non-retiring batch are as close to each other as possible, and that the z-scores of all the trucks in the retiring batch are also as close to each other as possible. Specifically, the following rules are used for demand allocation:

i) In each batch, demand will be continuously and evenly allocated to the cohort(s) of trucks with the lowest z-score. Note that the z-score of a truck will increase as more mileage is assigned to it. When the z-score of those trucks that are being assigned mileage increases to be equal to the z-score of some other cohort(s), the remaining demand will be evenly allocated to all these cohorts of trucks that have an equal (lowest) z-score. In other words, the demand allocation process intends to make the z-scores of all trucks in the batch equal. This is like pouring water into a set of communicating vessels, where the z-score of each cohort of trucks is like the water level in each vessel. This process will continue until there is no more demand to allocate.

ii) If a truck's mileage at time τ reaches the maximum limit U, or if a retiring truck's cumulative mileage reaches $\bar{y}$, no more demand will be allocated to this truck.

With Proposition 3.1 and given P_t, γ_t, S_t $(1 \le t \le T)$ and λ_1, we propose the following recursive procedure for calculating $y_{\tau,t}$ $(1 \le t \le \tau \le S_t \le T)$.

Algorithm 3.1:

Step 1. At time $\tau = 1$, set $y_{1,1} = u_{1,1} = \frac{D_1}{P_1}$. We have $u_{1,1} \in [0, U]$, or there will be no feasible solution with the given P_1.
Step 2. Assuming $S_1 > 1$, from (3.9c) we have $\lambda_2 = (r+1)\lambda_1 - rM(y_{1,1}, \gamma_1)$. (If $S_1 = 1$, the time $\tau = 1$ is considered a "break point". This special case will be addressed momentarily.)
Step 3. Set $\tau = 2$.
Step 4. At time τ, for each retiring cohort t (if any), calculate $\hat{y}_{\tau,t} = z_{\tau,t}^{-1}(\lambda_\tau)$, where $z_{\tau,t}$ is defined by (3.10c). Then calculate $y_{\tau,t}$ using equation (3.14). The remaining demand (if any) is $D_\tau - \sum_{w:S_w=\tau} P_w u_{\tau,w}$. Allocate this remaining demand to all the non-retiring cohorts according to Proposition 3.1.
Step 5. Now consider all the non-retiring cohorts whose mileage at time τ is *non-zero*. Set z_τ to be the highest z-score among these non-retiring cohorts and let $\lambda_{\tau+1} = (r+1)\lambda_\tau - rz_\tau$.
Step 6. If $\tau = T$, end. Otherwise, set $\tau \leftarrow \tau + 1$ and return to *Step 4*.

An exception arises in the above procedure when all the existing cohorts retire at the same time τ (i.e. there is no non-retiring cohort), or when the non-retiring cohorts at time τ all have zero mileage. In this case, we call the time τ a "break point". At a break point, there is no way to calculate $\lambda_{\tau+1}$ using the above procedure. Thus, we also need to specify the value of $\lambda_{\tau+1}$ for the above procedure to continue. The variable $\lambda_{\tau+1}$ will also be searched as part of the

upper-level solution procedure; i.e., the number of variables to be solved at the upper level will be $3T + 1 + K$, where K is the number of break points through the planning horizon ($K \leq T - 1$). Fortunately, we find in our numerical case studies that these kinds of exceptions are rare.

The procedure can also be modified to apply to a case where the assumption $\frac{\partial M}{\partial y_{S_t,t}} - \frac{\partial^2 F}{\partial y_{S_t,t}^2} > 0$ is not true for some t, i.e., when for some $y_{S_t,t}$, $\frac{\partial M}{\partial y_{S_t,t}} - \frac{\partial^2 F}{\partial y_{S_t,t}^2} \leq 0$. In this case, function $z_{S_t,t}(y_{S_t,t})$ for some t is not monotonic, and thus $z_{S_t,t}^{-1}(\zeta_{S_t,t})$ may be multi-valued. In fact, $y_{S_t,t}$ can be any value in the set $\{y_{S_t-1}, \min\{y_{S_t-1,t} + U, \bar{y}\}, \hat{y}_{S_t,t}^N\}$, where $\hat{y}_{S_t,t}^N$ represents all the possible roots of equation (3.10c) which are contained in the interval $[y_{S_t-1,t}, \min\{y_{S_t-1,t} + U, \bar{y}\}]$. Then in Step 4 of the above procedure (if there are retiring trucks at the present time), each possible value of $y_{S_t,t}$ (for each retiring cohort) will create a "branch" for the following recursive steps. Note here that Proposition 3.1 can still be applied to the non-retiring trucks, because for non-retiring trucks the function $z_{\tau,t}(y_{\tau,t}) \equiv M(y_{\tau,t}, \gamma_t)$ is still a monotonic function of $y_{\tau,t}$. Hence, the recursive procedure can still be used after making the above changes (i.e., adding the branches). At the final time T, there will be a number of branches, with each branch associated with a total discounted cost. The branch with the lowest total discounted cost is the solution to the lower-level problem.

This procedure can reduce the original program (3.1a-h) to minimizing $J(\lambda_1, P_t, \gamma_t, S_t: t = 1,2, \dots, T)$, subject to constraints (3.1d-f). This reduced program is not differentiable. In the next section, we present a tabu search algorithm (Glover, 1986) to solve this upper-level problem.

3.2.4 The upper-level solution procedure

The upper-level solution procedure also consists of two sublevels: the top sublevel searches for the optimal solution of the continuous variable λ_1, while the bottom sublevel optimizes the discrete variables P_t, γ_t and S_t $(t = 1,2,\dots,T)$ using tabu search. We first discuss the solution method of the bottom sublevel problem.

To start the tabu search process, we need a feasible initial solution, denoted by $\boldsymbol{x}^0 \equiv \{P_t^0, \gamma_t^0, S_t^0 : t = 1,2,\dots,T\}$. The quality of this initial solution may significantly affect the performance of the tabu search. In this section, we first propose a greedy heuristic algorithm to generate a fairly good initial solution $\boldsymbol{x}^0$ at a relatively low computation cost. The algorithm relies on a simplified demand allocation rule, in which the demand is always allocated to the truck(s) with the lowest z-score, regardless of whether it is being retired at that time or not; i.e., we ignore the difference between the two batches of trucks in the demand allocation rule. We determine P_t and γ_t recursively as t progresses from 1 to T: at each time t, P_t is determined as the minimum number of new trucks needed to serve demand at time t; and the best γ_t is selected from the set of truck types H. Note that the selection of γ_t is made by comparing the costs associated with all truck types at time t, but not by comparing the costs associated with all the possible combinations of γ_t's over the entire planning horizon $[1, T]$ since the latter procedure has a much higher computation cost. At the beginning, S_t is set to T for all cohorts, and later the algorithm will check for each cohort if an earlier retirement would yield a lower total cost. The details of the algorithm are furnished as follows.

Algorithm 3.2:

Step 1. Set $i = 1$.

Step 2. Set $j = 0,\ bestCost = \infty$.

Step 3. If $j \geq 1,\ P_j > 0$ and $S_j > i - 1$, set $\tilde{S}_j = S_j$ and $S_j = j - 1$.

Step 4. For each truck type $\gamma \in H$, find the least-cost solution through the following steps:

Step 4.1. Set $\tau = i$. Keep the purchase, utilization and retirement schedules before time i.

Step 4.2. Purchase the minimum number of new trucks given by $P_\tau = \left\lceil \frac{D_\tau - V_\tau}{U} \right\rceil$, where the ceiling function $\lceil a \rceil$ returns the minimum integer that is greater than or equal to a; and V_τ denotes the maximum total mileage that all the existing trucks can serve at τ, i.e.,

$$V_\tau \equiv \begin{cases} \max\left\{\sum_{1 \leq t < \tau \leq S_t} u_{\tau,t} \,|\, u_{\tau,t} \leq U, y_{\tau,t} \leq \bar{y}\right\}, & \text{if } \tau > 1 \\ 0, & \text{if } \tau = 1 \end{cases}.$$

Set $S_\tau = T$.

Step 4.3. Continuously and evenly allocate demand to the trucks with the smallest z-score, subject to boundary constraints (3.1g) and (3.1h), until there is no more demand to allocate. If for some $t\ (1 \leq t \leq \tau)$ we have $y_{\tau,t} = \bar{y}$, set $S_t = \tau$.

Step 4.4. If $\tau < T$, set $\tau \leftarrow \tau + 1$ and return to *Step 4.2*; otherwise, go to *Step 4.5*.

Step 4.5. Evaluate the total discounted cost from time 1 to T, which is denoted by $totalCost$. If $bestCost > totalCost$, then record the utilization plan of all the cohorts at time i, the purchase and retirement schedule of cohort i, and all the associated changes made to the retirement schedules of the cohorts before time i (if any).

Step 5. If $j \geq 1,\ S_j = i - 1$, and $bestCost$ is NOT reduced in the previous *Step 4*, set $S_j = \tilde{S}_j$.

Step 6. If $j < i - 1$, set $j \leftarrow j + 1$ and return to *Step 3*. Otherwise, go to *Step 7*.

Step 7. If $i < T$, set $i \leftarrow i + 1$ and go to *Step 2*. Otherwise, end.

We next present the tabu search algorithm to improve the solution. We define a *move* as a change from solution $\boldsymbol{x}$ to a new feasible solution, where the change is one of the following: i) add or reduce one purchased truck at a time; or ii) change the types of trucks purchased at one time; or iii) postpone or advance the retirement

time of a cohort by one time unit. The *neighborhood* of $\boldsymbol{x}$, denoted by $\mathcal{N}(\boldsymbol{x})$, is a set of solutions that can be obtained by making one move from $\boldsymbol{x}$. In each iteration, we make a move using the following rules:

i) If a move in $\mathcal{N}(\boldsymbol{x}) \cap TL$ produces a total cost that is lower than the best solution so far, set the current move as the one in $\mathcal{N}(\boldsymbol{x}) \cap TL$ that produces the lowest total cost. (This rule is termed the *aspiration level criterion* in the literature of tabu search.)

ii) If there is no move in $\mathcal{N}(\boldsymbol{x}) \cap TL$ that can produce a lower total cost, set the current move as the one in $\mathcal{N}(\boldsymbol{x}) \backslash TL$ that produces the lowest total cost. Note that a move will still be made even if the new total cost produced by the selected move is higher than the best solution so far.

Here TL denotes the *tabu list*, which records the *reverse* moves of the latest consecutive moves up to a size defined by $tabu\_size$. The tabu list is updated after every iteration. Rule ii) requires it to find the best neighboring solution that is not associated with any move in the tabu list, where the tabu list is used to prevent the algorithm from returning to a solution attained in a previous iteration step. Rule i) specifies that if a move in the tabu list can yield a better solution than the best one so far, we select that move and the associated solution. Rule i) overrides Rule ii). The tabu search ends when no better solution is found after max_num1 consecutive iteration steps. The algorithm is presented as follows.

Algorithm 3.3:

Step 1. Set $\boldsymbol{x} = \boldsymbol{x}^0$, where $\boldsymbol{x}^0$ is a feasible initial solution, possibly generated by the greedy heuristic algorithm. Set $iter = iter_{best} = 0$, and $TL = \emptyset$. Set the best solution so far $\boldsymbol{x}^* = \boldsymbol{x}$.

Step 2. Examine all the possible moves of $\boldsymbol{x}$ and perform the move selected via the rules specified above (i.e. the best tabu move that passes the aspiration level criterion, or the best non-tabu move). Denote the resulting solution as $\tilde{\boldsymbol{x}}$. Set $\boldsymbol{x} = \tilde{\boldsymbol{x}}$. Set $iter \leftarrow iter + 1$.

Step 3. Update the tabu list TL and the best solution reached so far: if $J(\boldsymbol{x}) < J(\boldsymbol{x}^*)$, set $\boldsymbol{x}^* = \boldsymbol{x}$ and $iter_{best} = iter$.

Step 4. If $iter - iter_{best} < max\_num1$, return to *Step 2*; otherwise, end.

This algorithm will generate a heuristic solution of $\boldsymbol{x}^* = \{P_t^*, \gamma_t^*, S_t^* : t = 1,2, \dots, T\}$ for a given λ_1. Finally, we present the following algorithm to find the optimal Lagrange multiplier λ_1 at the top sublevel.

Algorithm 3.4:

Step 1. Randomly select the values of $\lambda_1^{(0)}$ and $\lambda_1^{(1)}$ from a predefined range, Ω. For each of $\lambda_1^{(0)}$ and $\lambda_1^{(1)}$, find the heuristic solutions, $\boldsymbol{x}^*\left(\lambda_1^{(0)}\right)$ and $\boldsymbol{x}^*\left(\lambda_1^{(1)}\right)$, using the tabu search algorithm, where the initial solutions are generated from the greedy heuristic algorithm. The total discounted costs are denoted as $J^*\left(\lambda_1^{(0)}\right)$ and $J^*\left(\lambda_1^{(1)}\right)$, respectively.

Step 2. Set $k = 2, k_{best} = 1$. Denote the best solution so far as λ_1^*, set $\lambda_1^* = \underset{\lambda_1 \in \left\{\lambda_1^{(0)}, \lambda_1^{(1)}\right\}}{\operatorname{argmin}} \{J^*(\lambda_1)\}$.

Step 3. Set $\lambda_1^{(k)} = \lambda_1^{(k-1)} - \alpha_{k-1} \frac{J^*\left(\lambda_1^{(k-1)}\right) - J^*\left(\lambda_1^{(k-2)}\right)}{\lambda_1^{(k-1)} - \lambda_1^{(k-2)}}$, where α_{k-1} is a positive step size. Calculate $J^*\left(\lambda_1^{(k)}\right)$ using the tabu search algorithm, where $\boldsymbol{x}^*\left(\lambda_1^{(k-1)}\right)$ is taken as the initial solution in the tabu search.

Step 4. If $J^*(\lambda_1^*) > J^*\left(\lambda_1^{(k)}\right)$, set $\lambda_1^* = \lambda_1^{(k)}$ and $k_{best} = k$.

Step 5. If $k - k_{best} < max\_num2$, set $k = k + 1$ and return to *Step 3*; otherwise, end.

3.3 Performance of our solution approach

Section 3.3.1 presents the specific cost functions and parameter values that we use in Sections 3.3 and 3.4 of this thesis. Section 3.3.2 validates the solution quality of our lower-level CA approach against two commonly used commercial solvers using a battery of numerical instances and compares their computation costs. Section 3.3.3 compares the solution quality and computation cost of our bi-level approach against previous methods used in the literature. All the numerical instances in Sections 3.3 and 3.4 of this paper are carried out via Matlab R2015b on a HP 3.20GHz personal computer with 4GB RAM.

3.3.1 Cost functions and parameter values

The cost functions used in our numerical instances are borrowed from Guerrero et al. (2013). They were originally derived from CARB (2008). These cost functions are presented as follows:

$$A(\gamma_t) = A_p + \frac{k_1 \gamma_t^2}{k_2 - \gamma_t} \tag{3.17a}$$

$$M(y_{\tau,t}, \gamma_t) = O(\gamma_t) + W(y_{\tau,t}, \gamma_t) \tag{3.17b}$$

$$O(\gamma_t) = \theta_M + k_0 + (\theta_F + p_F)(1 - \gamma_t) f \tag{3.17c}$$

$$W(y_{\tau,t}, \gamma_t) = (k_{m0} + \beta \gamma_t) y_{\tau,t} \tag{3.17d}$$

$$F(y_{S_t,t}, \gamma_t) = A(\gamma_t) k_d (1 - k_x y_{S_t,t}) \tag{3.17e}$$

$$\bar{y} = 1/k_x \tag{3.17f}$$

where $O(\gamma_t)$ and $W(y_{\tau,t}, \gamma_t)$ denote the unit operation and maintenance cost per mile, respectively.

The $A_p, k_1, k_2, \theta_M, k_0, \theta_F, p_F, f, k_{m0}, \beta, k_d$ and k_x are constant parameters. Their definitions and values are summarized in Table 3.1. Most of these parameter

values are borrowed from Guerrero et al. (2013) and CARB (2008). Note here that γ_t represents the fuel saving efficiency of the truck type; i.e., $\gamma_t = 0.3$ indicates that the truck type will use 30% less fuel for every mile travelled than the benchmark (i.e. the type with $\gamma_t = 0$). Also note that under these parameter values, our assumptions specified in section 2, i.e., $M > 0$, $\frac{\partial M}{\partial y_{\tau,t}} > 0$, $F \geq 0$, $\frac{\partial F}{\partial y_{S_t,t}} < 0$, and $\frac{\partial M}{\partial y_{S_t,t}} - \frac{\partial^2 F}{\partial y_{S_t,t}^2} > 0$, are all satisfied. Hence, the solution procedure described above can be directly applied. The unit of time is a year.

Table 3.1 Parameter values

Parameter	**Notation**	**Value**	**Unit**
Fixed truck purchase cost	A_p	1.3E5	\$/truck
Coefficient for variable truck purchase cost	k_1	3.8E5	\$/truck
Coefficient for variable truck purchase cost	k_2	0.6	-
Baseline toll	θ_M	0	\$/mile
Fixed operating cost	k_0	0.647	\$/mile
Baseline fuel tax	θ_F	0	\$/gallon
Fuel price	p_F	4	\$/gallon
Base-line fuel efficiency	f	0.169	Gallons/mile
Fixed maintenance cost coefficient	k_{m0}	1.85E-7	\$/odometer-mile
Variable maintenance cost coefficient	β	2.57E-7	\$/odometer-mile
Instantaneous depreciation for the salvage value	k_d	0.75	-
Mileage depreciation for the salvage value	k_x	9.77E-7	odometer-mile^{-1}
Maximum mileage served per truck per unit of time	U	1E5	mile
Discount factor	r	0.07	-
Set of truck types	H	$\{0, 0.3\}$	-
Planning horizon	T	5-50	year

The instance-specific demand parameters D_τ $(1 \leq \tau \leq T)$ are furnished in the following sections.

3.3.2 Validation of the lower-level CA approach

To verify the performance of our lower-level CA approach, we tested 5 batches of numerical instances, each with $T = 10, 20, 30, 40, 50$, respectively. For each

batch of instances, we tested 20 numerical instances. In each instance the demand D_t in each year $t \in \{1,2,\dots,T\}$ was generated from a uniform distribution over the support $[2.0E6, 2.8E6]$. The P_t, γ_t and S_t $(1 \le t \le T)$ were also randomly selected from certain uniform distributions (but we ensure that feasible solutions can be developed under these values).

In our CA approach, the λ_1 is searched using an algorithm similar to subgradient search Algorithm 3.4, except that the solutions under each given λ_1 are produced by the demand allocation algorithm (Algorithm 3.1) instead of the tabu search method (Algorithm 3.3). The initial range of λ_1, Ω, is set to be $[5,20]$. This range is determined by trial and error in the numerical tests: if λ_1 is too small, in the recursive demand allocation process (Algorithm 3.1), a λ_t for some $t > 1$ may be negative; on the other hand, if λ_1 is too large, the $\hat{y}_{\tau,t}$ calculated from the z-score at τ, t will always be greater than $y_{\tau-1,t} + U$, which means the values of λ_t will have no effect on the optimal utilization plan.[9] The α_k in the subgradient search algorithm is set to a constant value 2×10^{-6} for $k = 1,2,\dots$. This step size value is selected so that the search of λ_1 spans the range Ω. The max_num2 is set to be 10.

When P_t, γ_t and S_t $(1 \le t \le T)$ are given, the original program (3.1a-h) is convex, given the exact cost functions presented in (3.17a-f); thus, the global optimal solution can be obtained via some commercial solvers for convex programming (without knowing λ_1): e.g. the *CVX* solver in Matlab (Boyd and Vandenberghe, 2004). Therefore, in this section we use the CVX solution as a

[9] In fact, our extensive numerical tests show that the value of λ_1 has only a modest effect on the demand allocation solution.

benchmark and evaluate our solution quality and runtime against CVX. Note that in general, the lower-level model may be not convex; consequently, we also compare the performance of our algorithm against a commonly used constrained nonlinear program solver, the *fmincon* tool in Matlab (in our comparison, the sequential quadratic programming (SQP) algorithm is selected in the fmincon tool). Specifically, we define the following cost gaps:

$$\text{Gap1} = \frac{\text{minimum total cost of CA approach} - \text{minimum cost of CVX}}{\text{minimum cost of CVX}}$$

$$\text{Gap2} = \frac{\text{minimum total cost of fmincon} - \text{minimum cost of CVX}}{\text{minimum cost of CVX}}$$

$$\text{Gap3} = \frac{\text{minimum M}-\text{S cost of CA approach} - \text{minimum M}-\text{S cost of CVX}}{|\text{minimum M}-\text{S cost of CVX}|}$$

$$\text{Gap4} = \frac{\text{minimum M}-\text{S cost of fmincon} - \text{minimum M}-\text{S cost of CVX}}{|\text{minimum M}-\text{S cost of CVX}|}$$

Table 3.2. Cost gaps and runtimes for our CA approach and the commercial solvers

T	**Gap1 (%)**		**Gap2 (%)**		**Gap3 (%)**		**Gap4 (%)**		**Average runtime (second)**		
	Avg	Max	Avg	Max	Avg	Max	Avg	Max	CA approach	CVX	fmincon
10	0.05	0.06	0.06	0.14	1.89	2.24	2.03	4.63	0.19	0.82	1.3
20	0.07	0.09	0.18	0.24	0.18	0.24	5.45	6.90	0.26	1.13	5.4
30	0.11	0.14	0.36	0.41	2.87	3.76	12.12	14.12	0.33	1.42	15.7
40	0.13	0.22	0.65	0.75	2.80	3.27	13.89	20.49	0.42	1.72	33.3
50	0.12	0.14	0.52	0.65	2.92	3.79	15.24	18.94	0.47	2.03	102

When P_t, γ_t and S_t $(1 \leq t \leq T)$ are given, the terms regarding purchase cost and operation cost are fixed. Therefore, we also compare the 'maintenance cost − salvage value' gap in this section, denoted as 'M-S'. Note that the term 'maintenance cost − salvage value' is negative, and consequently, the denominators used in Gap 3 and Gap 4 are the absolute values. The cost gaps and runtimes of our approach and the two solvers are summarized in Table 3.2.

The tabulated values show that the solutions of our CA approach are very close to the global optima; note that for all the batches of numerical instances examined here, the average Gap1 and Gap3 never exceed 0.2% and 3%, respectively, and the maximum Gap1 and Gap 3 never exceed 0.25% and 4%, respectively. Our approach has much shorter computation times than the CVX solver, however, and its computational advantage grows as the problem size increases; when $T = 50$, our approach reduces the computation cost 76% as compared to that of the CVX solver. This means the computational complexity of our CA approach grows much slower than the prevailing commercial solver. The table also shows that the fmincon solver produces much worse solutions than the CA approach, and the runtimes are even longer. We next examine how the proposed bi-level solution approach performs in terms of solution quality and computational efficiency.

3.3.3 Performance of the bi-level approach

We tested 7 batches of numerical instances, each with $T = 5, 6, 10, 20, 30, 40, 50$, respectively. Each batch includes 10 instances, with the annual demand randomly generated from the uniform distribution over $[2.0E6, 2.8E6]$. For tabu search Algorithm 3.3, the parameter $tabu\_size$ is carefully selected: if $tabu\_size$ is too small, the iterative search may be trapped in a loop that contains a local minimum, and this will prevent the search from exploring the remaining majority of the solution space; on the other hand, if $tabu\_size$ is too large, the very long tabu list may also prevent the algorithm from finding a better solution. Intuitively, the suitable $tabu\_size$ is an increasing function of the problem size, and the same parameter value can be used for problems of similar sizes. We found the values of $tabu\_size$ for the 7 batches of numerical instances by trial and error; they are presented in Table 3.. The max_num1 is set to be 50.

We compared the solution and runtime of our bi-level approach against a benchmark approach. The benchmark approach employs CVX to solve the lower-level problem (demand allocation) to the global optimum. For the upper-level problem, we used exhaustive search to obtain the global optimum for $T = 5$ and 6, and the same tabu search method (Algorithm 3.3) for $T \geq 10$ (note here that the exhaustive search will fail due to the huge computation cost). Note for $T \geq 10$ that the benchmark solutions are not the global optima of the original program (3.1a-h). In fact the global optima are very difficult to obtain because of the curse of dimensionality. However, they are almost the best solutions we can obtain for these numerical instances. We also compared our bi-level approach with the heuristic algorithm used by Guerrero et al. (2013). Specifically, the following cost gaps are defined:

$$\text{Gap5} = \frac{\text{minimum cost of bi-level approach} - \text{minimum cost of benchmark approach}}{\text{minimum cost of benchmark approach}}$$

$$\text{Gap6} = \frac{\text{minimum cost of Guerrero's approach} - \text{minimum cost of benchmark approach}}{\text{minimum cost of benchmark approach}}$$

The averages and the maxima of the above cost gaps for each batch of numerical instances are furnished in Table 3.. The table also shows the average runtimes for the three solution approaches.

The table shows that our bi-level approach produces solutions that are very close to the benchmark solutions; note that the average and maximum of Gap1 never exceed 1%. Moreover, the runtimes in our approach are much shorter than those of the benchmark approach, and the former's computation advantage increases as the problem size grows. In addition, the optimal costs obtained by our approach are on average 13-21% lower than the costs resulting from the previous method

used by Guerrero et al. (2013). All of these factors verify that our bi-level approach can produce very good solutions with reasonable runtimes.

Table 3.3. Cost gaps and runtimes for our bi-level approach, Guerrero's approach, and the benchmark approach

T	Tabu_size	Gap5		Gap6		Average runtime (second)		
		Average	Maximum	Average	Maximum	Our approach	CVX + exhaustive search or tabu	Guerrero's heuristic
5	8	0.31%	0.82%	13.89%	18.23%	1.91	357.13	4.31
6	10	0.42%	0.79%	20.54%	24.06%	2.43	2590.79	5.15
10	20	0.21%	0.41%	17.19%	27.76%	41.24	189.58	22.35
20	25	0.26%	0.32%	18.24%	23.12%	171.17	895.12	88.23
30	60	0.23%	0.34%	19.75%	22.03%	298.74	1275.82	98.93
40	125	0.27%	0.39%	17.31%	23.38%	520.26	2206.32	104.08
50	210	0.29%	0.43%	18.75%	25.65%	942.84	4082.65	149.51

3.4 Numerical case studies

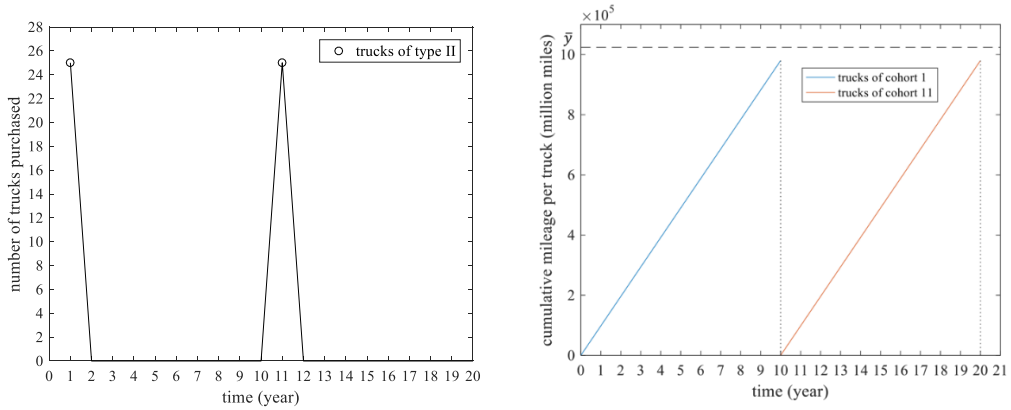
In this section, we examine a medium-size problem with $T = 20$ years and various demand patterns. All the other parameter values are the same as in section 3.3. Specifically, we examine three demand patterns: a constant demand for all years (section 3.4.1), a linearly increasing demand (section 3.4.2) and an exponentially increasing demand (section 3.4.3).

3.4.1 Constant demand pattern

We first assume $D_\tau = D = 2.45\text{E}6$ miles for $\tau \in \{1, \dots, T\}$. In Figure 3.1, the optimal purchase plan and utilization schedules are plotted against time. The minimal total discounted cost found is $J = 4.01 \times 10^7$ \$, with 50 trucks purchased in two batches in year 1 and 11. As shown in Figure 3.1a, the trucks are purchased at equal intervals (every 10 years), and new trucks are purchased only after the trucks in the previous cohort are retired. The evenly distributed truck

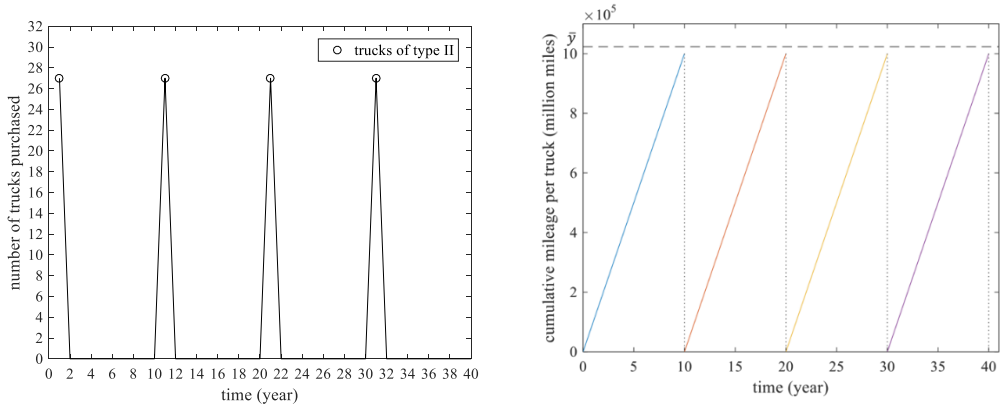
purchase plan is the consequence of a constant demand pattern. The vertical dashed lines in Figure 3.1b represent the time the trucks in each cohort are retired. In Figure 3.1b, the slope of each cumulative mileage curve is 1E5, the maximum rate per truck to serve, which means each truck is fully utilized in each year before it is retired. Note that the maximum allowable cumulative mileage ($\bar{y}$) is 1.02E6 miles as shown by the horizontal dashed line in Figure 3.1b. In the scenario of constant demand, trucks in each cohort are not be retired until they exceed $\bar{y}$ if they are continuously used to meet demand. This is because the unit purchase cost of a type II truck is higher than that of a type I truck, so the operators tend to use type II trucks to the maximum so that the operating and maintenance cost savings of type II trucks outweigh its higher purchase price. In other words, the optimal plan for an individual truck of type II is to use it for 10 years and to use it to the limit (1E5) every year. Under this constant demand case, the optimal utilization plan is close to the optimal plan of an individual truck without the demand constraints.

To further verify this finding, we then examine the optimal purchase plan utilization schedules for longer planning horizons ($T = 40$ years and 45 years) and with a different demand rate ($D_\tau = D = 2.7\text{E}6$). These cases present a similar trend, i.e., trucks are purchased at roughly equal intervals and are fully utilized in each year before being retired, as shown in Figure 3.2 and Figure 3.3. When the length of the planning horizon is not an integer multiple of 10 years, like 45 years, the operators tend to use the trucks purchased in earlier years less to gain more salvage value because of the discount factor.



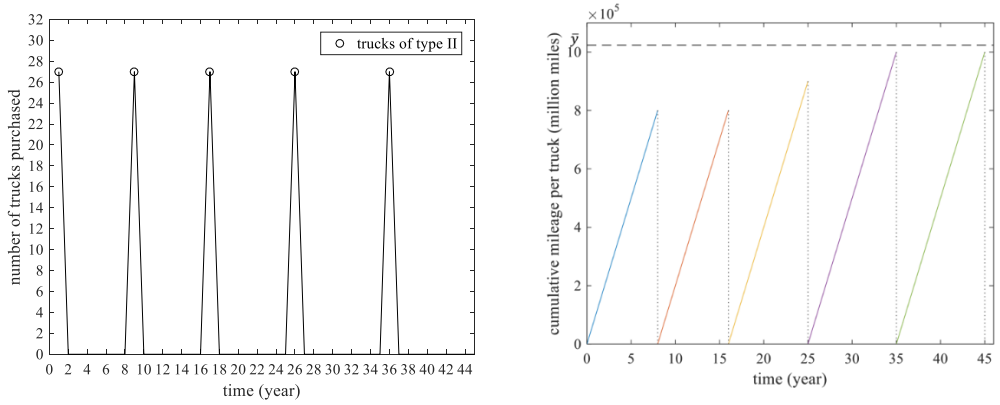
(a) (b)

Figure 3.1. Optimal utilization and replacement schedules for constant demand (D_τ=D=2.45E6 miles): (a) quantity and types of trucks purchased in each year; (b) utilization trajectory of trucks in each cohort.



(a) (b)

Figure 3.2. Optimal utilization and replacement schedules for constant demand (D_τ =D=2.7E6 miles) when T=40: (a) quantity and types of trucks purchased in each year; (b) utilization trajectory of trucks in each cohort.



(a) (b)

Figure 3.3. Optimal utilization and replacement schedules for constant demand (D_τ=D=2.7E6 miles) when T=45: (a) quantity and types of trucks purchased in each year; (b) utilization trajectory of trucks in each cohort.

3.4.2 Linearly increasing demand pattern

A linearly increasing demand pattern is shown in Figure 3.4 with $D_\tau = 1.5 + 0.1(\tau - 1)(1 \le \tau \le 20)$, where the aggregate demand over 20 years is equal to that in the constant-demand case ($D_\tau = D = 2.45E6$). In the results shown in Figure 3.5a, periodic purchase plans do not exist any more; instead, small quantities of trucks are purchased in most years to satisfy the added demand. Totally 58 trucks are purchased, which is more than in the constant demand case (50 trucks). The largest number of trucks are purchased in year 1 and year 11. Trucks of type I are purchased only near the end of planning horizon (years 16-20). Compared to the constant demand scenario, these trucks have much smaller cumulative mileage before retirement as shown in Figure 3.5b, so type I trucks are purchased due to their low purchase price. We also find that some trucks serve the maximum cumulative mileage (i.e., trucks in cohort 1, 4, 5, 6, 7). In other words, the demand allocation among trucks in a linearly increasing-demand pattern is quite uneven. The total discounted cost in this scenario is $J = 3.64 \times 10^7$ \$, which is lower than the constant-demand case ($J = 4.01 \times 10^7$ \$). This is because

the purchase and utilization in the later years have smaller discounting coefficients, and demand in those late years are low (see Figure 3.4). By examining the cumulative mileage trajectory of the trucks, we verified that the demand allocation rule (i.e., less-used trucks will be used first) is always satisfied; note that the two trajectories of non-retiring trucks of same type never cross each other. Also, the terminal cohorts are treated separately from the non-retiring cohorts (i.e. the less-used trucks are not always used first), see evidently years 17 and 18 in Figure 3.5b.

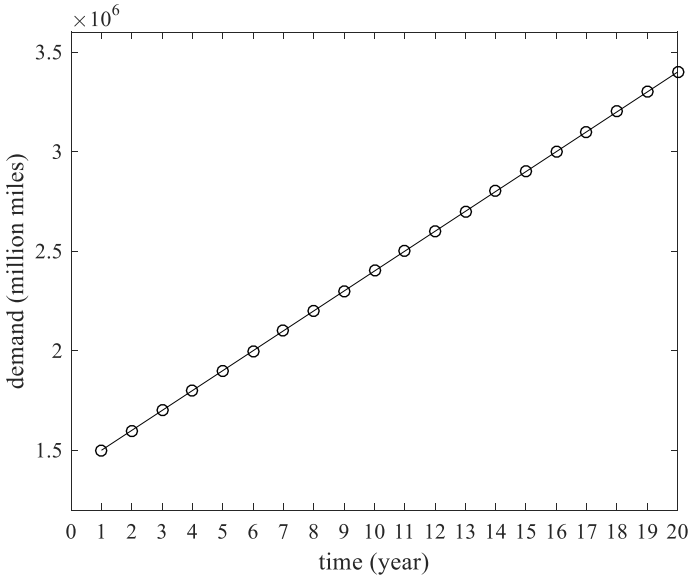
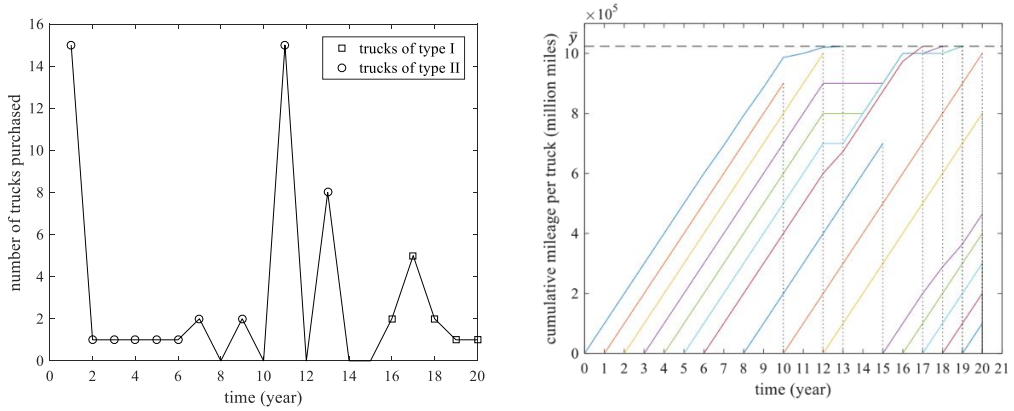


Figure 3.4. Linearly increasing demand.

We also find that older trucks may be retired later than younger trucks. Trucks in cohort 1 are retired in year 13, while trucks in cohorts 2 and 3 are retired in years 10 and 12, respectively, which violates the 'older cluster replacement' property of Jones et al. (1991). The property states that the older assets should be replaced before the new one. It is another intuitive property that has been widely believed to hold true at the optimality. In our case, the violation of 'older cluster replacement' is a consequence of the 'no splitting' property assumed earlier in this chapter, and the solution based on this assumption is suboptimal. Note that if cohort 1 can be split, and only one truck in cohort 1 is retired, then the total cost will be further

reduced. This is a limitation of our model. However, our problem can be viewed as an exception where the no-splitting property and older cluster replacement property cannot co-exist in the optimal solution. This is different from the finding in Jones et al. (1991), which claimed (for their specific problem) that both intuitive properties hold true at the optimality.



(a) (b)

Figure 3.5. Optimal utilization and replacement schedules for linearly increasing demand: (a) quantity and types of trucks purchased in each year; (b) utilization trajectories of trucks in each cohort.

3.4.3 Exponentially increasing demand pattern

We generated an exponentially increasing demand pattern, i.e., $D_\tau = 1.2 \times (1.06964)^{\tau-1}$ $(1 \le \tau \le 20)$, as shown in Figure 3.6, where the sum of total demand in 20 years is still equal to that in the previous two scenarios. The best total discounted cost found so far is $J = 3.44 \times 10^7$ \$, with 60 trucks purchased, most of them are purchased in years 1 and 11 (Figure 3.7a). Figure 3.7b shows that the operators need to buy trucks almost every year. We note that some trucks are not retired even though they make no contribution to demand during some periods (e.g., trucks in cohort 5 for years 14-19, trucks in cohort 6 for years 14-18, trucks

in cohort 7 for years 15-18, and trucks in cohort 8 for years 16-18). They are kept and used in later years to meet the demand surge.

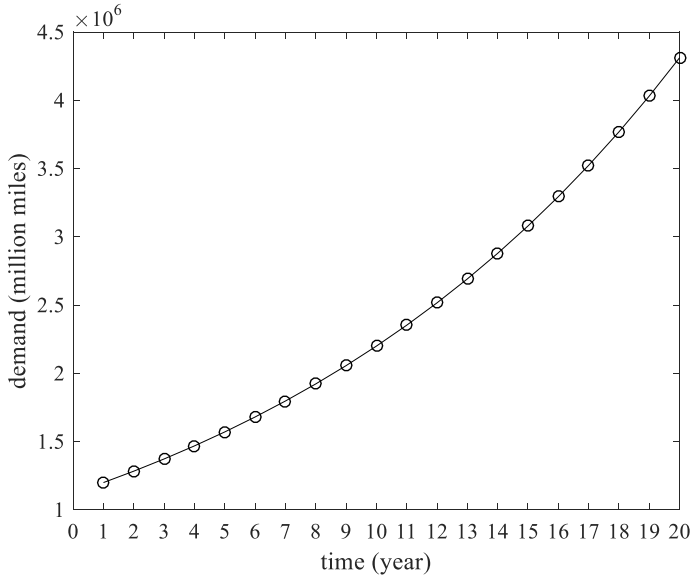
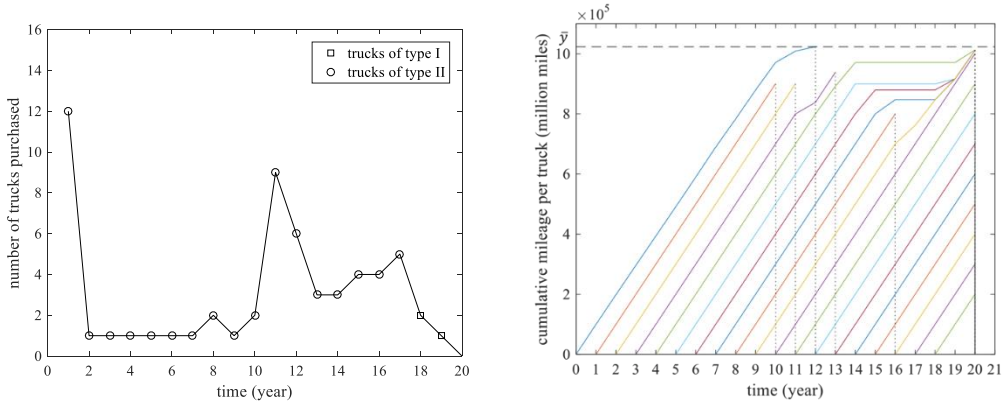


Figure 3.6. Exponentially increasing demand.



(a) (b)

Figure 3.7. Optimal utilization and replacement schedules for exponentially increasing demand: (a) quantity and types of trucks purchased in each year; (b) utilization trajectories of trucks in each cohort.

3.5 Summary of findings

In this chapter, a general discrete-time optimization model was developed for trucking companies to manage fleet utilization and replacement schedules on a finite planning horizon. Following is a summary of the findings:

(i) The demand allocation rule indicates that at optimality, demand D_τ should be allocated to all available trucks, so that the z-scores of all trucks in the non-retiring batch are as close to each other as possible, and the z-scores of all trucks in the retiring batch are also as close to each other as possible.

(ii) The older cluster replacement property and the no-splitting property cannot both be true.

(iii) Under the constant demand case, the optimal utilization plan is close to the optimal plan of an individual truck without the demand constraints.

(iv) Under linearly increasing and exponentially increasing demand patterns, trucking companies should buy trucks of type I near the end of the planning horizon. At other times, they should choose type II trucks, which are more utilized due to their lower utilization cost.

Chapter 4. Conclusions and future work

The contributions of the thesis are summarized in Section 4.1. Directions for future work are discussed in Section 4.2.

4.1 Conclusions

In this thesis, we examined the optimal planning of two common types of transportation asset systems: highway pavement systems belonging to the category of transportation infrastructure systems, and truck fleet systems belonging to the category of vehicle fleet systems. Although these systems have long been studied in the literature, better models have been developed in this thesis that are: i) more comprehensive, for they incorporate more treatment types, inspection planning, model parameter updating, and utilization planning of the assets; ii) more general, as the formulations and solution approaches can be applied with various specific forms of cost, deterioration, and treatment effectiveness models; and iii) more efficient, since our new solution approaches can furnish higher-quality solutions (sometimes global optima subject to certain tolerance level) within much shorter runtimes, and their computational complexity grows much slower than previous methods used in this realm. Lagrange multiplier method plays a key role in our methodology, relaxing the coupling constraints across heterogeneous assets and linking up various mathematical techniques (e.g. approximate dynamic programming, continuous approximation, subgradient search, metaheuristics and greedy algorithms) to formulate efficient solution approaches.

Contributions of the three research topics addressed in this thesis are summarized as follows.

For deterministic pavement system management, a general mathematical model is formulated for the joint optimization of MR&R planning for a system of heterogeneous pavement segments. A Lagrange multiplier approach is proposed in combination with derivative-free quasi-Newton methods to solve the system-level problem. By relaxing the budget constraints, our approach decomposes the system-level problem into multiple segment-level subproblems, which can take different forms and the solutions can be more easily derived. To the best of our knowledge, this is the first formulation of, and solution to, the system-level MR&R optimization problem that incorporates preventive maintenance activities, modelled by a more realistic formulation to fit the real data. Despite the added complexity, the optimization problem can be solved in moderate runtimes, thanks in part to the derivative-free quasi-Newton methods and the efficient segment-level heuristic. More importantly, the runtime increases linearly with the number of segments in a system, which ensures the applicability of our solution approach to large-scale systems. The computational efficiency is achieved without compromising solution quality. Particularly for problems under combined budget constraints, our approach guarantees global optimality or near-optimality at the system level as long as the segment-level subproblems are solved at or near the optima. Note that high-quality solutions are always preferred because cost savings of even 1% may result in savings of millions of dollars.

For stochastic pavement system management, a stochastic model for the joint optimization of inspection and MR&R policies is formulated for heterogeneous pavement systems under model uncertainty. Built upon a Markov Decision Process framework, the model finds the optimal condition-based management scheme

(including the inspection schedules) on a rolling-horizon basis. The model is solved via a Lagrange multiplier approach combined with approximate dynamic programming. The solution method is also independent of the specific forms of segment-level models. A statistical learning approach is also employed to update the deterioration prediction after every inspection, and as a result the model uncertainty diminishes over the planning horizon. We showed the effectiveness of our model and the computational efficiency of the solution procedure. Numerical analysis unveils a number of useful insights; please refer to Section 2.3 for more details.

For truck fleet management, a general discrete-time optimization model is formulated for truck operators to manage the fleet utilization and replacement schedules on a finite planning horizon. The model is solved by first converting it to a continuous approximation formulation and solving the CA model via a Lagrange multiplier approach. The resulting optimality conditions are converted back to the discrete time, based upon which some rules governing the optimal utilization planning are created. The process can largely reduce the size of solution space, and the remaining problem can be efficiently solved by metaheuristic methods. We show that this solution procedure can furnish high-quality solutions in much shorter runtimes as compared to commercial solvers and previous methods proposed in the literature.

4.2 Future work

A number of future research opportunities can be built upon this thesis. Some of these opportunities are summarized below:

For infrastructure systems management:

1 *For the pavement management problem, the segments are essentially independent, the only linkage is via the joint budget constraint. In reality, adjacent segments might influence each other. For example, the deterioration of adjacent segments might be correlated. When rehabilitation and especially reconstruction is performed, there might be cost savings (due to economies of scale, e.g. fixed cost of equipment rental, etc.) if adjacent segments undergo the same treatment at the same time, even when the treatment is not (yet) warranted based on the standalone evaluation for a single segment. I am not asking you to do this, but it would be good to include a discussion on how the models might need to be changed to accommodate this realistic aspect. This would be a most interesting extension of the thesis.*

(i) We can extend our deterministic and stochastic models by considering varying annual budgets that are not transferable between the years since this is how government agencies manage budgets at present.

(ii) Another direction is to incorporate other objectives than costs (e.g. greenhouse gas emissions) and more realism for pavement systems (e.g. network connectivity constraints and time-varying traffic loadings determined by demand forecasting and traffic assignment over the pavement network). Meanwhile, when MR&R activities are applied on adjacent segments simultaneously, there might be cost savings due to economies of scale. To capture this, we could use a matrix to record the connectivity among segments and multiplied the corresponding agency cost by a weighting factor if two adjacent segments are maintained

simultaneously. The details to formulate such model is under way, however, similar method can still be used to solve the problem.

(iii) Some assumptions regarding the deterioration model can be relaxed to better utilize historical data collected from other pavement networks via appropriate statistical approaches, such as Full Bayesian (FB) or Empirical Bayesian (EB) methods. These methods can adjust biases resulting from heterogeneity among the networks.

(iv) We can also consider to model the uncertainty in inspection measurements since pavement inspection technologies are imperfect.

For vehicle fleet management:

(i) Our work can be built upon to model uncertainties in demand (Hartman, 2004; List et al., 2003), operation conditions (List et al., 2003), and occasional vehicle breakdowns (Stasko and Gao, 2012; Childress and Durango-Cohen, 2005).

(ii) It is more realistic to express demand in terms of freight tonnage, truckloads, or number of containers, and consider the OD pairs instead of mileage. While the assumption of infinitely divisible demand is relaxed (e.g., a job may not be divided into two trucks and two small jobs cannot always be combined to one truck), similar demand allocation rules as those specified in Proposition 3.1 can still be used to formulate good heuristic truck utilization plans. We can further consider multiple truck types with different freight-carrying capacities in the model.

(iii) We will also consider to relax the 'no splitting' assumption to develop better, more realistic truck fleet management plans.

Appendices

Appendix A. Glossary of symbols

k – the index of segment and $k \in \{1,2,\cdots,K\}$

C_k^U – total discounted user cost for segment k over a given planning horizon

$C_k^U(t)$ – user cost for segment k during period t

C_{kp} – total discounted agency cost for treatment p on segment k over a given planning horizon

$C_{kp}(t)$ – agency cost for treatment p on segment k during period t

$q_k(t)$ – condition states of segment k at the beginning of period t, including roughness level $s_k(t)$ and pavement age h_{kt}

$q_k(0)$ – initial state of segment k

p – the index of treatment (i.e., 1 for rehabilitation, 2 for reconstruction and 3 for preventive maintenance)

$\boldsymbol{\Phi_k}$ – segment-specific equality constraints for segment k

$\boldsymbol{\Psi_k}$ – segment-specific inequality constraints for segment k

Z_k – the sum of discounted user and agency costs for segment k over a given planning horizon

Z – the sum of discounted user and agency costs for pavement system over a given planning horizon

B – annual combined budget

B_p – annual separate budget for treatment p

$x_{kt,p}$ – a binary variable; it is equal to 1 if a rehabilitation (corresponding to p=1), reconstruction ($p = 2$) or preventive maintenance ($p = 3$) activity is executed in year t for segment k, respectively, and 0 otherwise

v_{kt} – maintenance intensity for segment k during period t

$\bar{v}_k$ – the upper bound of maintenance intensity for segment k

ω_{kt} – rehabilitation intensity for segment k during period t

R_{kt} – the upper bound of rehabilitation intensity for segment k during period t

b_k – deterioration rate for segment k

b_k^L – the lower bound of deterioration rate for segment k

l_k – traffic loading on segment k

r – discounted rate

F_k – history-dependent deterioration process (deterministic scenario) for segment k

$\tilde{\mathcal{F}}_\tau$ – the best available deterioration model at the decision-making time point τ (stochastic scenario)

s_k^L – the best possible roughness level after a rehabilitation

s_k^{new} – the roughness level immediately after a reconstruction for segment k

s_k^{max} – the upper bound of roughness level for segment k

T_k^{min} – the lower bound of lifecycle duration for segment k

T_k^{max} – the upper bound of lifecycle duration for segment k

$I_k(t^-)$ – a binary variable; it specifies whether to perform an inspection at t (i.e., $I_k(t^-) = 1$) or nor ($I_k(t^-) = 0$) for segment k

$M_k(t^+)$ – MR&R decisions applied at time t for segment k

$\mathcal{L}$ – set of all possible inspection policies

$\mathcal{L}_k$ – set of all possible inspection policies for segment k, $\mathcal{L}_\mathrm{k} \in \{0,1\}$

$\mathcal{M}$ – set of all possible MR&R policies

$\mathcal{M}_\mathrm{k}$ – set of all possible MR&R policies for segment k

t^- – the start time of period t, i.e., the time immediately before inspection

t^+ – the time immediately after inspection and before MR&R activities in period t

t^{++} – the time after the application of MR&R activities in period t

$\mathbb{t}$ – the maximum duration between two consecutive inspections

P_t – quantity of trucks purchased at time t

γ_t – type of trucks purchased at time t

$u_{\tau,t}$ – the mileage served at time τ by a truck purchased at time t

$y_{\tau,t}$ – the cumulative mileage served at time τ by a truck purchased at time t

S_t – the time when trucks in cohort t retired

$A(\gamma_t)$ – the unit cost for purchasing a truck of type γ_t

$M(y_{\tau,t}, \gamma_t)$ – the unit O&M cost per mile

$O(\gamma_t)$ – the unit operation cost per mile for truck of type γ_t

$W(y_{\tau,t}, \gamma_t)$ – the unit maintenance cost per mile for truck of type γ_t

$F(y_{S_t,t}, \gamma_t)$ – the salvage value of a truck in cohort t that retired at S_t

U – the maximum mileage a truck can serve per unit time

$\bar{y}$ – the maximum allowable cumulative mileage

D_t – the given demand at time t

Appendix B. Proof sketch of Lemma 2.1

First, note that the case of $\lambda^* = 0$ is trivial. In the following proof, we assume that $\lambda^* \neq 0$ and $\lambda^H \neq 0$ (note it is unlikely that $\lambda^* \neq 0$ and $\lambda^H = 0$ when δ_1 and δ_2 are both small). So from (2.6a and b), we have:

$$\sum_{k=1}^{K} C_k\left(\boldsymbol{x}_k^*(\lambda^*)\right) = \sum_{k=1}^{K} C_k\left(\boldsymbol{x}_k^H(\lambda^H)\right) = B \quad \text{(B1)}$$

Then,

$$\left|\sum_{k=1}^{K}\left(C_k\left(\boldsymbol{x}_k^*(\lambda^H)\right) - C_k\left(\boldsymbol{x}_k^*(\lambda^*)\right)\right)\right| = \left|\sum_{k=1}^{K}\left(C_k\left(\boldsymbol{x}_k^*(\lambda^H)\right) - C_k\left(\boldsymbol{x}_k^H(\lambda^H)\right)\right)\right| \leq$$

$$\sum_{k=1}^{K}\left|C_k\left(\boldsymbol{x}_k^*(\lambda^H)\right) - C_k\left(\boldsymbol{x}_k^H(\lambda^H)\right)\right| \leq K \cdot \delta_1 \quad \text{(B2)}$$

On the other hand, since $\boldsymbol{x}_k^*(\lambda)$ minimizes $Z_k(\boldsymbol{x}_k) + \lambda C_k(\boldsymbol{x}_k)$, we have:

$$Z_k\left(\boldsymbol{x}_k^*(\lambda^*)\right) + \lambda^* \cdot C_k\left(\boldsymbol{x}_k^*(\lambda^*)\right) \le Z_k\left(\boldsymbol{x}_k^*(\lambda^H)\right) + \lambda^* \cdot C_k\left(\boldsymbol{x}_k^*(\lambda^H)\right) \quad \text{(B3)}$$

and,

$$Z_k\left(\boldsymbol{x}_k^*(\lambda^H)\right) + \lambda^H \cdot C_k\left(\boldsymbol{x}_k^*(\lambda^H)\right) \le Z_k\left(\boldsymbol{x}_k^*(\lambda^*)\right) + \lambda^H \cdot C_k\left(\boldsymbol{x}_k^*(\lambda^*)\right) \quad \text{(B4)}$$

The (B3) and (B4) can be combined into:

$$\lambda^* \cdot \left(C_k\left(\boldsymbol{x}_k^*(\lambda^*)\right) - C_k\left(\boldsymbol{x}_k^*(\lambda^H)\right)\right) \le Z_k\left(\boldsymbol{x}_k^*(\lambda^H)\right) - Z_k\left(\boldsymbol{x}_k^*(\lambda^*)\right) \le \lambda^H \cdot$$

$$\left(C_k\left(\boldsymbol{x}_k^*(\lambda^*)\right) - C_k\left(\boldsymbol{x}_k^*(\lambda^H)\right)\right) \quad \text{(B5)}$$

Hence,

$$\left|\sum_{k=1}^K \left(Z_k\left(\boldsymbol{x}_k^*(\lambda^*)\right) - Z_k\left(\boldsymbol{x}_k^*(\lambda^H)\right)\right)\right| \le \max\{\lambda^*, \lambda^H\} \cdot \left|\sum_{k=1}^K \left(C_k\left(\boldsymbol{x}_k^*(\lambda^*)\right) -\right.\right.$$

$$\left.\left.C_k\left(\boldsymbol{x}_k^*(\lambda^H)\right)\right)\right| \le \max\{\lambda^*, \lambda^H\} \cdot K\delta_1 \quad \text{(B6)}$$

Now we have:

$$\left|\sum_{k=1}^K Z_k\left(\boldsymbol{x}_k^*(\lambda^*)\right) - \sum_{k=1}^K Z_k\left(\boldsymbol{x}_k^H(\lambda^H)\right)\right| \le \left|\sum_{k=1}^K Z_k\left(\boldsymbol{x}_k^*(\lambda^*)\right) -\right.$$

$$\left.\sum_{k=1}^K Z_k\left(\boldsymbol{x}_k^*(\lambda^H)\right)\right| + \left|\sum_{k=1}^K Z_k\left(\boldsymbol{x}_k^*(\lambda^H)\right) - \sum_{k=1}^K Z_k\left(\boldsymbol{x}_k^H(\lambda^H)\right)\right| \le \max\{\lambda^*, \lambda^H\} \cdot$$

$$K\delta_1 + \sum_{k=1}^K \left|Z_k\left(\boldsymbol{x}_k^*(\lambda^H)\right) - Z_k\left(\boldsymbol{x}_k^H(\lambda^H)\right)\right| \le K \cdot (\max\{\lambda^*, \lambda^H\}\,\delta_1 + \delta_2) \quad \text{(B7)}$$

Note that in a real pavement system, $C_k\left(\boldsymbol{x}_k^*(\lambda^H)\right) - C_k\left(\boldsymbol{x}_k^*(\lambda^*)\right)$ can be either positive or negative for any k; and the positive and negative components of the sum $\sum_{k=1}^K \left(C_k\left(\boldsymbol{x}_k^*(\lambda^H)\right) - C_k\left(\boldsymbol{x}_k^*(\lambda^*)\right)\right)$ will cancel each other out. Thus, inequality (B2) may be very weak, and as a result, (B7) may be weak too.

Appendix C. Proof of Lemma 2.2

We prove Lemma 2.2 by contradiction. Suppose there exists $\lambda_1 > \lambda_2 \geq 0$, such that $V(\lambda_1) \geq V(\lambda_2)$. We denote $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$ as the solutions associated with λ_1 and λ_2, respectively; i.e., for each $k = 1,2,\ldots,K$, $\boldsymbol{x}_k^1$ is the unique minimizer of $H_k(\boldsymbol{x}_k, \lambda_1) \equiv C_k^U(\boldsymbol{x}_k) + (1+\lambda_1)C_k(\boldsymbol{x}_k)$, and $\boldsymbol{x}_k^2$ is the unique minimizer of $H_k(\boldsymbol{x}_k, \lambda_2) \equiv C_k^U(\boldsymbol{x}_k) + (1+\lambda_2)C_k(\boldsymbol{x}_k)$.

We then have:

$0 > \sum_{k=1}^K [H_k(\boldsymbol{x}_k^1, \lambda_1) - H_k(\boldsymbol{x}_k^2, \lambda_1)]$

$= \sum_{k=1}^K \{[C_k^U(\boldsymbol{x}_k^1) + (1+\lambda_1)C_k(\boldsymbol{x}_k^1)] - [C_k^U(\boldsymbol{x}_k^2) + (1+\lambda_1)C_k(x_k^2)]\}$

$= \sum_{k=1}^K [C_k^U(\boldsymbol{x}_k^1) - C_k^U(\boldsymbol{x}_k^2)] + (1+\lambda_1)\big(V(\lambda_1) - V(\lambda_2)\big)$

$\geq \sum_{k=1}^K [C_k^U(\boldsymbol{x}_k^1) - C_k^U(\boldsymbol{x}_k^2)] + (1+\lambda_2)\big(V(\lambda_1) - V(\lambda_2)\big)$

$= \sum_{k=1}^K [H_k(\boldsymbol{x}_k^1, \lambda_2) - H_k(\boldsymbol{x}_k^2, \lambda_2)] > 0$

Contradiction! ■

Note if $\boldsymbol{x}_k^1$ and $\boldsymbol{x}_k^2$ are not unique minimizers, then the above equation may hold without creating contradiction, but only when $H_k(\boldsymbol{x}_k^1, \lambda_1) = H_k(\boldsymbol{x}_k^2, \lambda_1)$ and $H_k(\boldsymbol{x}_k^1, \lambda_2) = H_k(\boldsymbol{x}_k^2, \lambda_2)$ for all k (i.e., $\boldsymbol{x}_k^1$ and $\boldsymbol{x}_k^2$ are dual optima under both λ_1 and λ_2), and $V(\lambda_1) = V(\lambda_2)$. Further note that the optimal $C_k(\boldsymbol{x}_k)$ should be a non-increasing function of λ, we have $C_k(\boldsymbol{x}_k^1) = C_k(\boldsymbol{x}_k^2)$ and $C_k^U(\boldsymbol{x}_k^1) = C_k^U(\boldsymbol{x}_k^2)$ for all k. The above conditions are too strict to be satisfied by realistic segment-level models. Thus, we reckon Lemma 2.2 as a general result.

Appendix D. Proof sketch of the convergence of Algorithm 2.1

We assume that $V(\lambda)$ is continuously differentiable everywhere and the unique root of $V(\lambda)$ is λ^* (since $V(\lambda)$ is a decreasing function of λ). We prove the convergence of Algorithm 2.1 by contradiction. Suppose the stop criterion cannot

be attained as n increases. Initially, we have $\lambda^0 < \lambda^1$ and $V(\lambda^0) > 0$. According to Lemma 2.2, we have $V(\lambda^1) < V(\lambda^0)$. One of the following two cases will occur.

Case 1: $V(\lambda^n) > 0$ for all $n \geq 1$

In this case $\lambda^{n+1} = \lambda^n - V(\lambda^n)\frac{\lambda^n - \lambda^{n-1}}{V(\lambda^n) - V(\lambda^{n-1})} > \lambda^n$ for all the n; i.e., the sequence $\{\lambda^n\}$ is strictly increasing. Thus $\{\lambda^n\}$ should be bounded above, because otherwise $V(\lambda^n)$ would be 0 or negative for sufficiently large n. That means $\{\lambda^n\}$ has a supremum: $\tilde{\lambda} = \sup\{\lambda^n\}$. Let $V(\tilde{\lambda}) = \kappa$ as shown in Figure. D1a. We have:

$$\lim_{n\to\infty} V(\lambda^n) = \kappa > 0 \tag{D1}$$

$$\lim_{n\to\infty} \lambda^n = \tilde{\lambda} \tag{D2}$$

However,

$$\tilde{\lambda} = \lim_{n\to\infty}(\lambda^n - V(\lambda^n)\frac{\lambda^n - \lambda^{n-1}}{V(\lambda^n) - V(\lambda^{n-1})}) = \tilde{\lambda} - \kappa \cdot V'(\tilde{\lambda}) > \tilde{\lambda} \tag{D3}$$

Contradiction!

Case 2: $V(\lambda^n) < 0$ for some $n \geq 1$.

According to Algorithm 2.1, we have $V(\lambda^{n-1}) \cdot V(\lambda^n) < 0$ for all $n \geq n'$, where $n' = \min\{n | V(\lambda^n) < 0\}$; i.e., $\{\lambda^n\}$ oscillates on both sides of λ^*. Then there must be infinite number of λ^n's on at least one side of λ^*. There are two subcases:

(1) Infinite number of λ^n's occur only on one side of λ^*. The contradiction can be shown using the same method presented in Case 1.

(2) Infinite numbers of λ^n's occur on both sides of λ^*. We denote $\{\lambda_L^n\}$ as the λ^n's on the left side of λ^* and $\{\lambda_R^n\}$ as those on the right side. We define

$\tilde{\lambda}_L = \sup\{\lambda_L^n\}$ and $\tilde{\lambda}_R = \inf\{\lambda_R^n\}$ as shown in Figure. D1b, where $V(\tilde{\lambda}_L) = \kappa_L$ and $\mathrm{V}(\tilde{\lambda}_\mathrm{R}) = \kappa_\mathrm{R}$. We have:

$$\lim_{n\to\infty} V(\lambda_L^n) = \kappa_L > 0 \tag{D4}$$

$$\lim_{n\to\infty} \lambda_L^n = \tilde{\lambda}_L \tag{D5}$$

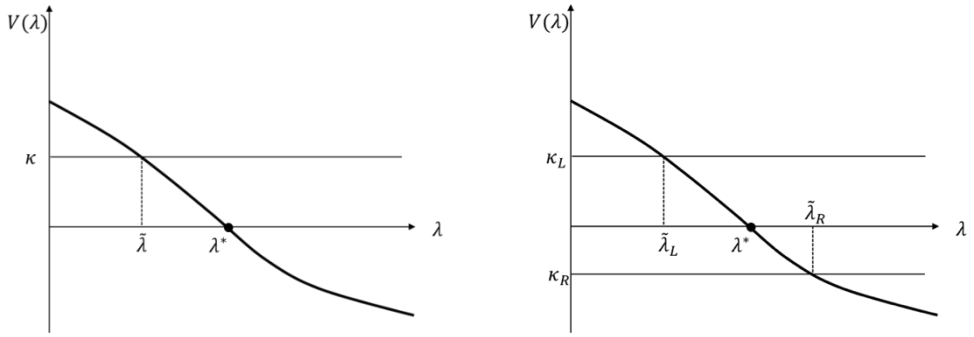
$$\lim_{n\to\infty} V(\lambda_R^n) = \kappa_R < 0 \tag{D6}$$

$$\lim_{n\to\infty} \lambda_R^n = \tilde{\lambda}_R \tag{D7}$$

And,

$$\lim_{n\to\infty} \lambda^{n+1} = \lim_{n\to\infty} (\lambda^n - V(\lambda^n)\frac{\lambda^n - \lambda^{n-1}}{V(\lambda^n) - V(\lambda^{n-1})}) \in (\tilde{\lambda}_L, \tilde{\lambda}_R) \tag{D8}$$

Contradiction! ■



(a) (b)

Figure. D1. Illustrations of the two cases of contradiction: (a) case 1; (b) case 2.

Appendix E. The dynamic programming approach to the deterministic segment-level problem

We first reproduced the dynamic programming method used by Lee and Madanat (2014a, 2015) with only minor modifications to solve subproblem 1. To this end, we assume that the maintenance intensity $v_{k\tau}$ and the pavement roughness level $s_k(\tau)$ take values form predefined discrete sets, i.e., $v_{k\tau} \in$

$\left\{0, \frac{1}{d}D_{k\tau}, \frac{2}{d}D_{k\tau}, \cdots, D_{k\tau}\right\}$ for $\tau \in \{1, \cdots, T_k\}$, and $s_k(\tau) \in M'_{k\tau} = \left\{s_k^{new}, s_k^{new} + \frac{\bar{s}_k(\tau) - s_k^{new}}{N}, s_k^{new} + 2\frac{\bar{s}_k(\tau) - s_k^{new}}{N}, \cdots, \bar{s}_k(\tau)\right\}$ for $\tau \in \{0, \cdots, T_k\}$, where $\bar{s}_k(\tau)$ is the maximum allowed roughness for segment k in year τ; i.e., if $s_k(\tau) > \bar{s}_k(\tau)$, the roughness level would exceed s_k^{max} in the following year $\tau + 1$. There are $d + 2$ decision options in each year $\tau \in \{1, \cdots, T_k\}$: do nothing; rehabilitation only; and maintenance only with intensity $\frac{1}{d}D_{k\tau}, \frac{2}{d}D_{k\tau}, \cdots, D_{k\tau}$, respectively. Let $Y_k(q_{k\tau})$ denote the cost-to-go in year τ (i.e. the minimum total discounted cost from year τ to T_k), the algorithm is described as follows:

Step 1. For each $T_k \in \left\{T_k^{min}, \cdots, T_k^{max}\right\}$, set the boundary condition as $Y_k(q_k(T_k)) = 0, \ \forall\, q_k(T_k)$. For each year $\tau = T_k - 1, T_k - 2, \cdots, 0$, $s_k(\tau) \in M'_{k\tau}$ and $h_{k\tau} = h_{k0} + \tau$, we generate $Y_k(q_k(\tau))$ in the backward direction by the Bellman equation:

$$Y_k(q_k(\tau)) = \min_{x_{k\tau,3}, x_{k\tau,1}, v_{k\tau}} \left\{\int_\tau^{\tau+1} l_k(c_k^1 s_k(u) + c_k^2)\, e^{-ru} du + x_{k\tau,3}(\gamma_k^1 v_{k\tau} + \gamma_k^2) e^{-r\tau} + x_{k\tau,1}(m_k^1 R_{k\tau} + m_k^2) e^{-r\tau} + Y_k(q_k(\tau + 1))\right\} \quad \text{(E1)}$$

where

$$q_k(\tau + 1) = \{s_k(\tau + 1), h_{k,\tau+1}\} = \left\{F_k\left(s_k(\tau) - x_{k\tau,1} G_k\big(R_{k\tau}, s_k(\tau)\big), 1, h_{k\tau}, \bar{b}_k - x_{k\tau,3} E_k(v_{k\tau}, s_k(\tau))\right), h_{k\tau} + 1\right\} \quad \text{(E2)}$$

$$Y_k(q_k(\tau + 1)) = \frac{s_k - s_k(\tau+1)}{s_k - s'_k} Y_k\big(s'_k, h_{k,\tau+1}\big) + \frac{s_k(\tau+1) - s'_k}{s_k - s'_k} Y_k\big(s_k, h_{k,\tau+1}\big) \quad \text{(E3)}$$

s'_k and s_k are the two consecutive roughness indices in $M'_{k,\tau+1}$ that satisfy $s'_k \le s_k(\tau + 1) \le s_k$.

Step 2. For each year $\tau = 0,1,\cdots,T_k - 1$, record the optimal decision in the forward direction:

$$\left(x^*_{k\tau,3}, x^*_{k\tau,1}, v^*_{k\tau}\right) = \underset{x_{k\tau,3},x_{k\tau,1},v_{k\tau}}{\operatorname{argmin}} \left\{\int_\tau^{\tau+1} l_k(c_k^1 s_k(u) + c_k^2)\, e^{-ru} du + x_{k\tau,3}(\gamma_k^1 v_{k\tau} + \gamma_k^2) e^{-r\tau} + x_{k\tau,1}(m_k^1 R_{k\tau} + m_k^2) e^{-r\tau} + Y_k(q_k(\tau+1))\right\} \quad \text{(E4)}$$

where $q_k(0) = \{s_k(0), h_{k0}\}$.

Step 3. Find the T_k that minimizes $\frac{Z_k^S}{1-e^{-rT_k}}$.

In the first step, we apply the Bellman equation (E1) recursively to generate $Y_k(q_k(\tau))$ for all $\tau \in \{0,\cdots,T_k-1\}, s_k(\tau) \in M'_{k\tau}$ and $h_{k\tau} = h_{k0} + \tau$. The pavement state in year $\tau+1$ is calculated by equation (E2). The cost-to-go $Y_k(q_k(\tau+1))$ is approximately by linear interpolation between $Y_k(s'_k, h_{k,\tau+1})$ and $Y_k(s_k, h_{k,\tau+1})$; see equation (E3). The optimal decision $(x^*_{k\tau,3}, x^*_{k\tau,1}, v^*_{k\tau})$ that minimizes $Y_k(q_k(\tau))$ in each year τ is obtained and recorded in step 2 with the initial state $q_k(0)$.

To solve subproblem 2, we make the following changes to the above algorithm: i) in year 0 there are $d+2$ decisions as in other years; and ii) T_k is replaced by t_k^T, whose range is $\left[T_k^{min\prime}, T_k^{max\prime}\right]$, where $T_k^{min\prime} = \max\{0, T_k^{min} - h_{k0}\}$ and $T_k^{max\prime} = \max\{0, T_k^{max} - h_{k0}\}$. Finally, we choose the t_k^T that minimizes equation (2.12).

Appendix F. General patterns for the solution regions as defined in Figure 2.7a-d

Figure. F1 presents all the seven possible patterns of the solution regions. The result in Lee and Madanat (2015) belongs to the pattern shown by Figure. F1a, where RC_{min} and RC_{max} denote the minimum and maximum reconstruction costs that are required when the lifecycle duration is T_k^{max} and T_k^{min}, respectively. (Note that this is the only pattern described in Lee and Madanat, 2015.) This pattern occurs if: i) a feasible solution exists when no maintenance or rehabilitation is applied, and only the minimum reconstruction is executed; and ii) the maximum reconstruction budget RC_{max} will be binding when no maintenance or rehabilitation is applied. If only condition ii) is false, i.e., RC_{max} is unbinding even if no maintenance or rehabilitation is applied, then the upper boundary of region ***D*** would hit the horizontal axis before crossing the vertical line at RC_{max}. This will render the pattern shown by Figure. F1b.

On the other hand, if the above condition i) is false, then the lower boundary of region ***D*** will decline as the reconstruction budget increases. This oblique lower boundary may end by: I) hitting the horizontal axis (before reaching RC_{max}); II) crossing the upper boundary of ***D*** (before reaching RC_{max}); and III) crossing the vertical line at RC_{max}. Case I) can be further divided into two patterns: when the upper boundary of ***D*** ends at the vertical line at RC_{max} (Figure. F1c), and when that boundary also ends at the horizontal axis (Figure. F1d). In case II), the two boundaries of region ***D*** merge to a single line which is decreasing as reconstruction budget increases. This line will cross the horizontal axis (Figure. F1e) or the vertical line at RC_{max} (Figure. F1f). Finally, case III) will render the patterns described by Figure. F1g. Note any interface that appears on the right of RC_{max}

has to be horizontal. The results shown in Figure 2.7a-d belong to the pattern in Figure. F1e.

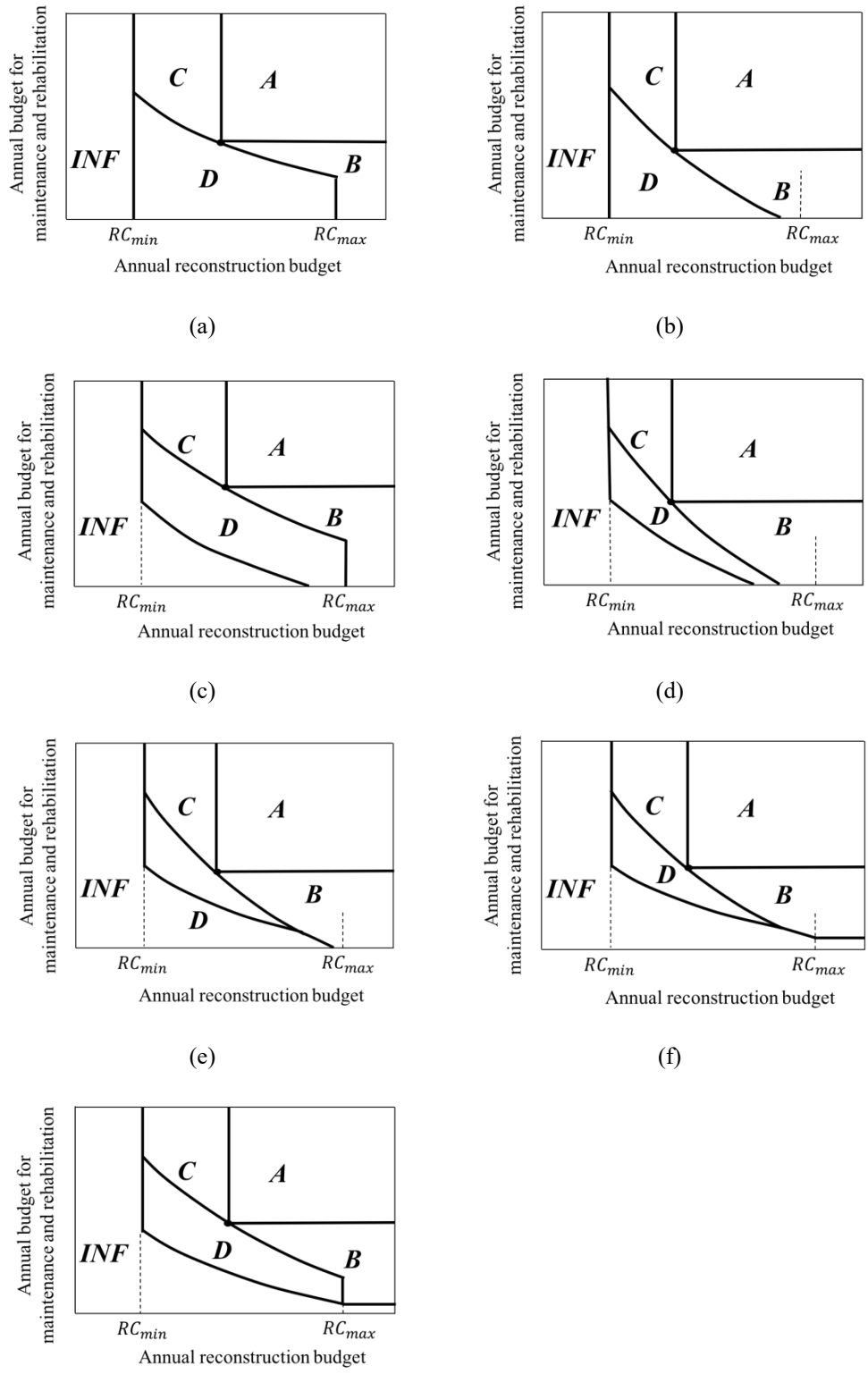


Figure F1. Patterns for the solution regions ***A***, ***B***, ***C***, ***D***, and ***INF***

Appendix G. Updating the deterioration model

We assume that the error term $\mathcal{E}_\tau$ follows the normal distribution, denoted as $\mathcal{E}_\tau \sim N(0, \sigma_\tau^2)$. All the inspection data are given in terms of $(s_{before}^{\hat{d},k}, s_{after}^{\hat{d},k}, h_{before}^{\hat{d},k}, h_{after}^{\hat{d},k})$, where $s_{before}^{\hat{d},k}$ and $s_{after}^{\hat{d},k}$ are two consecutive inspection results on segment k; $h_{before}^{\hat{d},k}$ and $h_{after}^{\hat{d},k}$ denote the segment age when the two consecutive inspections are applied; and $\hat{d}$ is the data index. The size of data used for calibration increases with more inspections results, which will improve the accuracy of our deterioration model. We assume that at time τ, $\widehat{D}$ groups of data are available, including all the historical data obtained previously and the newest inspection data acquired at time τ. As the error terms during each time interval are i.i.d. random variables, we can derive the following:

$$(s_{after}^{\hat{d},k} - s_{before}^{\hat{d},k} e^{\Theta_\tau^1\left(h_{after}^{\hat{d},k} - h_{before}^{\hat{d},k}\right)} - \Theta_\tau^2 l_k \cdot (h_{after}^{\hat{d},k} - h_{before}^{\hat{d},k}) \cdot (1 + SN_k)^{\Theta_\tau^3} e^{\Theta_\tau^1 (h_{after}^{\hat{d},k} - h_{before}^{\hat{d},k} + 1)}) \sim N(0, (h_{after}^{\hat{d},k} - h_{before}^{\hat{d},k})\sigma_\tau^2) \quad \text{(G1)}$$

$$\frac{1}{\sqrt{h_{after}^{\hat{d},k} - h_{before}^{\hat{d},k}}} \left(s_{after}^{\hat{d},k} - s_{before}^{\hat{d},k} e^{\Theta_\tau^1 (h_{after}^{\hat{d},k} - h_{before}^{\hat{d},k})} - \Theta_\tau^2 l_k \cdot \left(h_{after}^{\hat{d},k} - h_{before}^{\hat{d},k}\right) \cdot (1 + SN_k)^{\Theta_\tau^3} e^{\Theta_\tau^1 \left(h_{after}^{\hat{d},k} - h_{before}^{\hat{d},k} + 1\right)} \right) \sim N(0, \sigma_\tau^2) \quad \text{(G2)}$$

, i.e.,

We define:

$$\bar{G}\left(s_{before}^{\hat{d},k}, s_{after}^{\hat{d},k}, h_{before}^{\hat{d},k}, h_{after}^{\hat{d},k}\right) = \frac{1}{\sqrt{h_{after}^{\hat{d},k} - h_{before}^{\hat{d},k}}} \left(s_{after}^{\hat{d},k} - s_{before}^{\hat{d},k} e^{\Theta_\tau^1\left(h_{after}^{\hat{d},k} - h_{before}^{\hat{d},k}\right)} - \Theta_\tau^2 l_k \cdot (h_{after}^{\hat{d},k} - h_{before}^{\hat{d},k}) \cdot (1 + SN_k)^{\Theta_\tau^3} e^{\Theta_\tau^1 (h_{after}^{\hat{d},k} - h_{before}^{\hat{d},k} + 1)} \right) \quad \text{(G3)}$$

Then the likelihood function is:

$$\bar{\bar{L}}(\Theta_\tau^1, \Theta_\tau^2, \Theta_\tau^3, \sigma_\tau^2) = \prod_{\hat{d}=1}^{\widehat{D}} \hat{f}\left(\bar{G}\left(s_{before}^{\hat{d},k}, s_{after}^{\hat{d},k}, h_{before}^{\hat{d},k}, h_{after}^{\hat{d},k}\right)\right) \quad \text{(G4)}$$

where $\hat{f}(\cdot)$ denotes the probability density function of normal distribution with mean 0 and variance σ_τ^2.

We define:

$$\hat{l}(\Theta_\tau^1, \Theta_\tau^2, \Theta_\tau^3, \sigma_\tau^2) = log\bar{\bar{L}}(\Theta_\tau^1, \Theta_\tau^2, \Theta_\tau^3, \sigma_\tau^2)$$

$$= -\frac{\hat{D}}{2} log(2\pi) - \frac{\hat{D}}{2} log\, \sigma_\tau^2 - \frac{\sum_{\hat{d}=1}^{\hat{D}} \bar{G}^2\left(s_{before}^{\hat{d},k}, s_{after}^{\hat{d},k}, h_{before}^{\hat{d},k}, h_{after}^{\hat{d},k}\right)}{2\sigma_\tau^2} \quad \text{(G5)}$$

Then the newest value of parameters $\Theta_\tau^1, \Theta_\tau^2, \Theta_\tau^3, \sigma_\tau^2$ is updated by solving the following equations:

$$\frac{\partial \hat{l}}{\partial \Theta_\tau^i} = -\frac{1}{\sigma_\tau^2} \sum_{\hat{d}=1}^{\hat{D}} \bar{G}\left(s_{before}^{\hat{d},k}, s_{after}^{\hat{d},k}, h_{before}^{\hat{d},k}, h_{after}^{\hat{d},k}\right)$$

$$\cdot \frac{\partial \bar{G}\left(s_{before}^{\hat{d},k}, s_{after}^{\hat{d},k}, h_{before}^{\hat{d},k}, h_{after}^{\hat{d},k}\right)}{\partial \Theta_\tau^i} = 0, \forall i = 1, 2, 3 \quad \text{(G6)}$$

$$\frac{\partial \hat{l}}{\partial \sigma_\tau^2} = -\frac{\hat{D}}{2\sigma_\tau^2} + \frac{\sum_{\hat{d}=1}^{\hat{D}} \bar{G}^2\left(s_{before}^{\hat{d},k}, s_{after}^{\hat{d},k}, h_{before}^{\hat{d},k}, h_{after}^{\hat{d},k}\right)}{2\sigma_\tau^4} = 0 \quad \text{(G7)}$$

Appendix H. Approximate dynamic programming for segment-level problem

To better exploit the trade-off between solution optimality and computational tractability, we solve the MDP problem (2.24a-c) by using approximate dynamic programming, given the Lagrange multiplier λ. The details are outlined in the following subsections: first, we give an approximation of the value function; then we discuss the learning process by updating the parameters of the approximated value functions based on the simulated sample path; finally, we specify the trade-off between exploration and exploitation.

H.1 Value function approximation

In this thesis, we approximate the value function as a linear combination of basic functions, as shown in (H1).

$$\tilde{Q}_k^t(q_k(t^+), M_k(t^+), \Delta t|\lambda) = \sum_{i=1}^{\ell_k(t)} \rho_k^i(t|\lambda) \cdot \tilde{\xi}_k^i(q_k(t^+), M_k(t^+), \Delta t) \quad \text{(H1)}$$

where

$\tilde{Q}_k^t(q_k(t^+), M_k(t^+), \Delta t|\lambda)$: the approximated value of the expected cost-to-go from period $t^+ + \Delta t$ to the end of the planning horizon T, when action $M_k(t^+)$ and inspection $I_k(t + \Delta t^-)$ are applied to segment k with state $q_k(t^+)$, termed the *Q*-function;

$\tilde{\xi}_k^i(q_k(t^+), M_k(t^+), \Delta t)$: a basis function, which captures important attributes of segment conditions, available MR&R and inspection options, termed the *Q*-factor;

$\rho_k^i(t|\lambda)$: a weighting factor associated with the basis function $\tilde{\xi}_k^i(q_k(t^+), M_k(t^+), \Delta t)$, which is refined within the ADP framework;

$\ell_k(t)$: total number of basis functions for segment k during period t.

Given the *Q*-function, $\tilde{Q}_k^t(q_k(t^+), M_k(t^+), \Delta t|\lambda)$, the optimal segment-level activity option at period t can be obtained as follows:

$$\min_{M_k(t^+) \in \mathcal{M}_{k,\Delta t}} \left\{ R_k(q_k(t^+), M_k(t^+), \Delta t) + \tilde{Q}_k^t(q_k(t^+), M_k(t^+), \Delta t|\lambda) \right\} \quad \text{(H2)}$$

where

$$R_k(q_k(t^+), M_k(t^+), \Delta t) = \lambda \sum_{p=1}^{P} C_{kp}(q_k(t^+), M_k(t^+)) +$$

$$\sum_{\mu=0}^{\Delta t-1} \alpha^{\mu} \underset{s_k(t+\mu^-)|q_k(t^+), M_k(t^+)}{E} \left[C_k^U\left(q_k(t+\mu^-)\right) \right] + \alpha^{\Delta t} \cdot \lambda \cdot cm_n \quad \text{(H3)}$$

is termed the step reward. (H2) shows that once we know the estimated *Q*-function, the optimal management option can be obtained directly, where the step reward $R_k(q_k(t^+), M_k(t^+), \Delta t)$ can be obtained by the Monte Carlo simulation. This can considerably reduce the complexity caused by computing the expectation of future cost, especially when the outcome space is large.

H.2 Updating the value function approximation

For a given ADP iteration $\bar{n}$, we assume that the current condition state of segment k is $q_{k,\bar{n}}(t^+)$, and the optimal action $\{\overline{M}_{k,\bar{n}}^t, \overline{\Delta t}_{k,\bar{n}}^t\}$ is selected according to Equation (H4):

$$\{\overline{M}_{k,\bar{n}}^t, \overline{\Delta t}_{k,\bar{n}}^t\} = \underset{M_k(t)\in\mathcal{M}_k,\Delta t}{\operatorname{argmin}} \{R_k(q_k(t^+), M_k(t^+), \Delta t) + \tilde{Q}_{k,\bar{n}-1}^t(q_k(t^+), M_k(t^+), \Delta t|\lambda)\} \quad \text{(H4)}$$

where the Q-function $\tilde{Q}_{k,\bar{n}-1}^t(q_k(t^+), M_k(t^+), \Delta t|\lambda)$ is obtained during the previous iteration $\bar{n}-1$. Once action $\overline{M}_{k,\bar{n}}^t$ has been performed and Δt is determined, a sample realization of the future state $q_{k,\bar{n}}(t+\Delta t^+)$ is generated by a Monte Carlo simulation procedure, following the stochastic deterioration process, $\tilde{\mathcal{F}}_\tau(\cdot)$. The procedure can be symbolically described as:

$$q_{k,\bar{n}}(t+\Delta t^+) = \tilde{\mathcal{F}}_\tau(q_{k,\bar{n}}(t^+), \overline{M}_{k,\bar{n}}^t, \overline{\Delta t}_{k,\bar{n}}^t) \quad \text{(H5)}$$

Once a sample state of period $t+\Delta t$ is ascertained, the optimal action for period $t+\Delta t$ can be selected by repeatedly using (H4). The procedure is subsequently repeated until the end of the planning horizon. Then we get a sequence of state-action pairs:

$$\left[\left(q_{k,\bar{n}}(t_1^+), \overline{M}_{k,\bar{n}}^{t_1}, \overline{\Delta t}_{k,\bar{n}}^{t_1}\right), \left(q_{k,\bar{n}}(t_2^+), \overline{M}_{k,\bar{n}}^{t_2}, \overline{\Delta t}_{k,\bar{n}}^{t_2}\right), \left(q_{k,\bar{n}}(t_3^+), \overline{M}_{k,\bar{n}}^{t_3}, \overline{\Delta t}_{k,\bar{n}}^{t_3}\right)\cdots, \left(q_{k,\bar{n}}(t_z^+), \overline{M}_{k,\bar{n}}^{t_z}, \overline{\Delta t}_{k,\bar{n}}^{t_z}\right), q_{k,\bar{n}}(T)\right] \quad \text{(H6)}$$

where

$t_{i+1} - t_i = \overline{\Delta t}_{k,\bar{n}}^{t_i}$ with $t_1 = 0,\ T - t_z \le \mathbb{t}, 0 \le i \le z-1.$

The state-action sequence (H5) is termed a sample path. We repeat the process for $\bar{N}$ iterations, and at each iteration, we assume that the current Q-function is the best estimate of future cost-to-go.

We next use the $\mathrm{TD}(\ddot{\lambda})$ method to update the parameters of the approximated value function backward recursively, based on the realized sample path. The error-term is defined as:

$$\Delta_{k,\bar{n}}(t) = \sum_{\tau'=t}^{T-1}(\alpha\ddot{\lambda})^{\tau'-t}(R_k(q_{k,\bar{n}}(t_{i+1}^+), \overline{M}_{k,\bar{n}}^{t_{i+1}}, \overline{\Delta t}_{k,\bar{n}}^{t_i+1})$$

$$+\alpha\tilde{Q}_{k,\bar{n}-1}^{t_{i+1}}(q_{k,\bar{n}}(t_{i+1}^+), \overline{M}_{k,\bar{n}}^{t_{i+1}}, \overline{\Delta t}_{k,\bar{n}}^{t_i+1}|\lambda) - \tilde{Q}_{k,\bar{n}-1}^{t_i}(q_{k,\bar{n}}(t_i^+), \overline{M}_{k,\bar{n}}^{t_i}, \overline{\Delta t}_{k,\bar{n}}^{t_i}|\lambda)) \quad \text{(H7)}$$

Once the temporal difference error $\Delta_{k,\bar{n}}(t)$ is calculated for a given period t in the ADP iteration $\bar{n}$, the weighting factors can be updated using a stochastic gradient algorithm:

$$\rho_{k,\bar{n}}^i(t) = \rho_{k,\bar{n}}^i(t) + \gamma_{\bar{n}}\frac{\partial\hat{\vartheta}_{k,\bar{n}}(t)}{\partial\rho_k^i}\cdot\Delta_{k,\bar{n}}(t) \quad \text{(H8)}$$

$$\hat{\vartheta}_{k,\bar{n}}(t) = \min_{M_k(t^+)\in\mathcal{M}_{k,\Delta t}} R_k(q_{k,\bar{n}}(t^+), M_k(t^+), \Delta t) +$$

$$\tilde{Q}_{k,\bar{n}-1}^t(q_{k,\bar{n}}(t^+), M_k(t^+), \Delta t|\lambda) \quad \text{(H9)}$$

Note that to guarantee the convergence of the parameters being updated, three basic conditions should be satisfied for the step-size parameter $\gamma_{\bar{n}}$:

$$\gamma_{\bar{n}} \geq 0, \forall\bar{n} = 1,2,\cdots \quad \text{(H10a)}$$

$$\sum_{\bar{n}=1}^{\infty}\gamma_{\bar{n}} = \infty \quad \text{(H10b)}$$

$$\sum_{\bar{n}=1}^{\infty}\gamma_{\bar{n}}^2 < \infty \quad \text{(H10c)}$$

More details about the selection of step-size parameter $\gamma_{\bar{n}}$ are provided in Section 2.2.3. Given the details above, the proposed ADP framework for solving a segment-level optimization problem on a finite time horizon can be summarized as follows:

Algorithm H1:

Step 1. Initialization:

Step 1.1. Initialize $\tilde{Q}_{k,0}^{t}(q_k(t^+), M_k(t^+), \Delta t|\lambda)$, $\forall M_k(t^+) \in \mathcal{M}_k, \forall t = 0,1,\cdots T-1$.

Step 1.2. Choose an initial state $q_{k,0}(0^+)$.

Step 2: Do for $\bar{n} = 1,2,\cdots \bar{N}$ (record the index of iterations in our ADP framework):

Step 2.1: For $t = 0,1,\cdots T-1$:

Step 2.1a: Solve $\hat{\vartheta}_{k,\bar{n}}(t) = \min_{M_k(t^+)\in\mathcal{M}_k} R_k\big(q_{k,\bar{n}}(t^+), M_k(t^+), \Delta t\big) + \tilde{Q}_{k,\bar{n}-1}^{t}\Big(\big(q_{k,\bar{n}}(t^+), M_k(t^+), \Delta t|\lambda\big)\Big)$, and denote $\bar{M}_{k,\bar{n}}^{t}$ as the corresponding optimal action.

Step 2.1b: Compute $q_{k,\bar{n}}\big(t + \overline{\Delta t}_{k,\bar{n}}^{t}\big) \leftarrow \tilde{\mathcal{F}}_{\tau}(q_{k,\bar{n}}(t^+), \bar{M}_{k,\bar{n}}^{t}, \overline{\Delta t}_{k,\bar{n}}^{t})$.

Step 2.2: Initialize $\Delta_{k,\bar{n}}(T) = 0$, $\hat{\vartheta}_{k,\bar{n}}(T) \leftarrow H_k^*(q_k(T)|\lambda)$.

Step 2.3: For $t = T-1,\cdots 1,0$, do the following:

Step 2.3a: $\varsigma_k(t) \leftarrow \hat{\vartheta}_{k,\bar{n}}(t+1) - \tilde{Q}_{k,\bar{n}-1}^{t}\Big(\big(q_{k,\bar{n}}(t^+), \bar{M}_{k,\bar{n}}^{t}, \overline{\Delta t}_{k,\bar{n}}^{t}|\lambda\big)\Big)$

Step 2.3b: $\Delta_{k,\bar{n}}(t) \leftarrow \Delta_{k,\bar{n}}(t) + \varsigma_k(t)$

Step 2.3c: $\rho_{k,\bar{n}}^{i}(t) \leftarrow \rho_{k,\bar{n}-1}^{i}(t) + \gamma_{\bar{n}} \frac{\partial \hat{\vartheta}_{k,\bar{n}}(t)}{\partial \rho_k^i} \cdot \Delta_{k,\bar{n}}(t)$

Step 2.3d: $\Delta_{k,\bar{n}}(t) \leftarrow \alpha\ddot{\lambda}\Delta_{k,\bar{n}}(t-1)$

H.3 Exploration vs. exploitation

One of the challenges in approximate dynamic programming is the trade-off between exploration and exploitation. The decision-maker has to exploit what is already known to obtain the lowest cost, but also has to explore to get more information. In our paper, we employ an ε-greedy policy, where with small probability ε, the maintenance option is selected at random from amongst all the actions with equal probability, independent of the action-value estimates, as shown below:

For $k = 1,2,\cdots K, t = 1,2,\cdots T$

$$\text{Prob}\left(M_{k,\bar{n}}(t^+)\middle|q_{k,\bar{n}}(t)\right)=\begin{cases}1-\varepsilon+\frac{\varepsilon}{\mathbb{A}}, & \text{if } M_{k,\bar{n}}(t^+)=\bar{M}_{k,\bar{n}}^t, \Delta t=\overline{\Delta t}_{k,\bar{n}}^t \\ \frac{\varepsilon}{\mathbb{A}}, & \text{otherwise}\end{cases}$$

$$\forall\, M_{k,\bar{n}}(t^+)\in\mathcal{M}_k, 1\le\Delta t\le\mathbb{t} \tag{H11}$$

where $\mathbb{A}$ denotes the size of the available action set.

Appendix I. Derivation of (3.4a-b)

If we plug (3.2a) into (3.3), we have:

$$\tilde{L}=\int_{t=0}^{T}A(\gamma(t))P(t)e^{-rt}dt+$$

$$\int_{t=0}^{T}\int_{\tau=t}^{S(t)}P(t)u(\tau,t)M(y(\tau,t),\gamma(t))\,e^{-r\tau}d\tau dt-$$

$$\int_{t=0}^{T}P(t)F(y(S(t),t),\gamma(t))e^{-rS(t)}\,dt+\int_{\tau=0}^{T}\lambda(\tau)D(\tau)e^{-r\tau}\,d\tau-$$

$$\int_{t=0}^{T}\int_{\tau=t}^{S(t)}\lambda(\tau)P(t)u(\tau,t)e^{-r\tau}d\tau dt+\int_{t=0}^{T}\int_{\tau=t}^{S(t)}\Big(\mu(\tau,t)u(\tau,t)+$$

$$y(\tau,t)\left(\frac{\partial\mu(\tau,t)}{\partial\tau}-r\mu(\tau,t)\right)\Big)e^{-r\tau}d\tau dt-\int_{t=0}^{T}\mu(S(t),t)y(S(t),t)e^{-rS(t)}dt+$$

$$\int_{t=0}^{T}\int_{\tau=t}^{S(t)}\varphi_1(\tau,t)\,(u(\tau,t)-U)e^{-r\tau}d\tau dt-\int_{t=0}^{T}\int_{\tau=t}^{S(t)}\varphi_2(\tau,t)\,u(\tau,t)e^{-r\tau}d\tau dt+$$

$$\int_{t=0}^{T}\omega(t)(y(S(t),t)-\bar{y})\,e^{-rS(t)}dt \tag{I1}$$

We take the partial derivatives of (I1) with respect to $u(\tau,t)$ and $y(\tau,t)$, and present part of the Karush-Kuhn-Tucker (KKT) necessary conditions for optimality as follows:

i) Stationarity: $\frac{\partial\tilde{L}}{\partial u(\tau,t)}=0,\ \frac{\partial\tilde{L}}{\partial y(\tau,t)}=0$ (I2)

ii) Complementary slackness:

$$\varphi_1(\tau,t)(u(\tau,t)-U)=0,\ \text{ for } t\in[0,T],\tau\in[t,S(t)] \tag{I3a}$$

$$\varphi_2(\tau,t)u(\tau,t)=0,\ \text{ for } t\in[0,T],\tau\in[t,S(t)] \tag{I3b}$$

$$\omega(t)(y(S(t),t)-\bar{y})=0,\quad \text{for } t\in[0,T] \tag{I3c}$$

and iii) Dual feasibility:

$$\varphi_1(\tau,t),\varphi_2(\tau,t),\omega(t)\geq 0,\ \text{for } t\in[0,T],\tau\in[t,S(t)] \tag{I4}$$

Note that not all the KKT conditions are presented here because some of them are not used in the following derivation of the optimality conditions.

The first stationarity condition in (I2) gives the following (I5a); the second stationarity condition in (I2) leads to the following (I5b) when $\tau < S(t)$, and (I5c) when $\tau = S(t)$:

$$P(t)M\big(y(\tau,t),\gamma(t)\big) - P(t)\lambda(\tau) + \mu(\tau,t) + \varphi_1(\tau,t) - \varphi_2(\tau,t) = 0 \tag{I5a}$$

$$P(t)u(\tau,t)\frac{\partial M}{\partial y(\tau,t)} + \frac{\partial \mu(\tau,t)}{\partial \tau} - r\mu(\tau,t) = 0 \tag{I5b}$$

$$P(t)\frac{\partial F}{\partial y(S(t),t)} + \mu(S(t),t) - \omega(t) = 0 \tag{I5c}$$

Take the partial derivative of both sides of (I5a) with respect to τ:

$$P(t)\frac{\partial M}{\partial y(\tau,t)}\frac{\partial y(\tau,t)}{\partial \tau} - P(t)\frac{d\lambda(\tau)}{d\tau} + \frac{\partial \mu(\tau,t)}{\partial \tau} + \frac{\partial \varphi_1(\tau,t)}{\partial \tau} - \frac{\partial \varphi_2(\tau,t)}{\partial \tau}$$

$$= P(t)u(\tau,t)\frac{\partial M}{\partial y(\tau,t)} - P(t)\frac{d\lambda(\tau)}{d\tau} + \frac{\partial \mu(\tau,t)}{\partial \tau} + \frac{\partial \varphi_1(\tau,t)}{\partial \tau} - \frac{\partial \varphi_2(\tau,t)}{\partial \tau} = 0 \tag{I6}$$

When we subtract (I5b) from (I6), we have:

$$\mu(\tau,t) = \frac{1}{r}\left(P(t)\frac{d\lambda(\tau)}{d\tau} - \frac{\partial \varphi_1(\tau,t)}{\partial \tau} + \frac{\partial \varphi_2(\tau,t)}{\partial \tau}\right) \tag{I7}$$

Then we plug (I7) into (I5a):

$$P(t)M\big(y(\tau,t),\gamma(t)\big) = \lambda(\tau)P(t) - \frac{1}{r}\left(P(t)\frac{d\lambda(\tau)}{d\tau} - \frac{\partial \varphi_1(\tau,t)}{\partial \tau} + \frac{\partial \varphi_2(\tau,t)}{\partial \tau}\right) -$$

$$\varphi_1(\tau,t) + \varphi_2(\tau,t) \tag{I8a}$$

On the other hand, we subtract (I5c) from (I5a) for $\tau = S(t)$:

$$P(t)M\big(y(S(t),t),\gamma(t)\big) - P(t)\frac{\partial F}{\partial y(S(t),t)} = P(t)\lambda\big(S(t)\big) - \varphi_1(S(t),t) +$$

$$\varphi_2(S(t),t) - \omega(t) \qquad \text{(I8b)}$$

Equations (I8a) and (I8b) apply when $\tau < S(t)$ and when $\tau = S(t)$, respectively. The (I8a) can be further simplified by examining the values of $\varphi_1(\tau,t)$ and $\varphi_2(\tau,t)$ for any given τ and t. One of the following three cases will arise:

i) $\varphi_1(\tau,t) = \varphi_2(\tau,t) = 0$: This means the constraint (3.2g) is unbinding for the given τ and t; i.e., $0 < u(\tau,t) < U$. Note further that $\varphi_1(\tau,t), \varphi_2(\tau,t) \geq 0$ due to the dual feasibility condition (I4). Thus, $\varphi_1(\tau,t) = \varphi_2(\tau,t) = 0$ implies that $\frac{\partial \varphi_1(\tau,t)}{\partial \tau} = \frac{\partial \varphi_2(\tau,t)}{\partial \tau} = 0$ (this relies on the implicit assumption that $\varphi_1(\tau,t)$ and $\varphi_2(\tau,t)$ are continuous and differentiable with respect to τ, which has been used in similar studies: e.g., Jin and Kite-Powell, 2000). Hence, (I8a) can be re-arranged as:

$$P(t) \cdot \left[M\big(y(\tau,t),\gamma(t)\big) - \left(\lambda(\tau) - \frac{1}{r}\frac{d\lambda(\tau)}{d\tau}\right)\right] = 0 \qquad \text{(I9)}$$

ii) $\varphi_1(\tau,t) = 0$ but $\varphi_2(\tau,t) \neq 0$: The left part of (3.2g) is binding for the given τ and t; i.e., $u(\tau,t) = 0$.

iii) $\varphi_1(\tau,t) \neq 0$ but $\varphi_2(\tau,t) = 0$: The right part of (3.2g) is binding for the given τ and t; i.e., $u(\tau,t) = U$.

Note that $\varphi_1(\tau,t)$ and $\varphi_2(\tau,t)$ cannot be both positive because the left and right parts of (3.2g) cannot be binding at the same time.

Similar reasoning applies to (I8b). Specifically, one of the following four cases will arise:

i) $\varphi_1(S(t),t) = \varphi_2(S(t),t) = \omega(t) = 0$: Both constraints (3.2g) and (3.2h) are unbinding; i.e., $0 < u(\tau,t) < U$ and $y(S(t),t) < \bar{y}$. We have:

$$P(t) \cdot \left[M\big(y(S(t),t),\gamma(t)\big) - \frac{\partial F}{\partial y(S(t),t)} - \lambda\big(S(t)\big)\right] = 0 \qquad \text{(I10)}$$

ii) $\omega(t) \neq 0$ (and no further constraints on $\varphi_1(\tau,t)$ and $\varphi_2(\tau,t)$): The constraint (3.2h) is binding for the given τ and t; i.e., $y(S(t),t) = \bar{y}$.

iii) $\omega(t) = \varphi_1(\tau,t) = 0$ but $\varphi_2(\tau,t) \neq 0$: The left part of (3.2g) is binding for the given τ and t; i.e., $u(\tau,t) = 0$.

iv) $\omega(t) = \varphi_2(\tau,t) = 0$ but $\varphi_1(\tau,t) \neq 0$: The right part of (3.2g) is binding for the given τ and t; i.e., $u(\tau,t) = U$.

The above results are summarized in (3.4), (3.5a-c) and (3.6a-d).

References

Adelman, D., Mersereau, A., 2008. Relaxations of weakly coupled stochastic dynamic programs. Operations Research, 56(3), 712-727.

Ahmed, S.B., 1973. Optimal equipment replacement policy: an empirical study. Journal of Transport Economics and Policy, 7 (1), 71-79.

ASCE, 2017. 2017 infrastructure report card: A comprehensive assessment of America's Infrastructure. (accessed on May 2, 2017). American Society of Civil Engineers, Reston, Virginia. http://www.infrastructurereportcard.org/wp-content/uploads/2016/10/2017-Infrastructure-Report-Card.pdf

Bai, Y., Gungor, O.E., Hernandez-Urrea, J.A., Ouyang, Y., Al-Qadi, I.L., 2015. Optimal pavement design and rehabilitation planning using a mechanistic-empirical approach. EURO Journal on Transportation and Logistics, 4(1), 57-73.

Bean, J. C., Lohmann, J. R., Smith, R. L., 1984. A dynamic infinite horizon replacement economy decision model. The Engineering Economist, 30(2), 99-120.

Bellman, R.A., 1957. Markovian Decision Process. Rand Corp., Santa Monica, CA No. P-1066

Bertsimas, D., Mersereau, A., 2007. A learning approach for interactive marketing to a customer segment. Operations Research, 55(6), 1120-1135.

Beutler, F., Ross, K., 1985. Optimal policies for controlled Markov chains with a constraint. Operations Research, 55(6), 1120.

Blum, C., Roli, A., 2003. Metaheuristics in combinatorial optimization: Overview and conceptual comparison. ACM Computing Surveys, 35(3), 268-308.

Bretthauer, K.M. and Shetty, B., 1995. The nonlinear resource allocation problem. Operations research, 43(4), 670-683.

CARB , 2008a. Technical Support Document: Proposed Regulation for In-use On-road Diesel Vehicles: Appendix J, Mobile Sources Control Division, Heavy-Duty Diesel In-Use Strategies Branch. California Air Resources Board. October 2008.

Carnahan, J.V., Davis, W.J., Shahin, M.Y., Kean, P.L., Wu, M.I., 1987. Optimal maintenance decisions for pavement management. ASCE Journal of Transportation Engineering, 113 (5), 554-572.

CBO, 2016. Approaches to make federal highway spending more productive (accessed on May 2, 2017). Congressional Budget Office of the United States, Washington, DC. www.cbo.gov/sites/default/files/114th-congress-2015-2016/reports/50150-Federal_Highway_Spending.pdf.

Chan, W., Fwa, T., Tan, C., 1994. Road maintenance planning using genetic algorithms. I: Formulation. Journal of Transportation Engineering, 120(5), 693-709.

Chang, P. T., 2005. Fuzzy strategic replacement analysis. European Journal of Operational Research, 160(2), 532-559.

Childress, S., Durango-Cohen, P., 2005. On parallel machine replacement problems with general replacement cost functions and stochastic deterioration. Naval Research Logistics, 52(5), 409-419.

Chisholm, A. H., 1974. Effects of tax depreciation policy and investment incentives on optimal equipment replacement decisions. American Journal of Agricultural Economics, 56(4), 776-783.

Chu, J., Chen, Y., 2012. Optimal threshold-based network-level transportation infrastructure life-cycle management with heterogeneous maintenance actions. Transportation Research Part B: Methodological, 46(9), 1123-1143.

Chu, J., Huang, K., 2018. Mathematical programming framework for modeling and comparing network-level pavement maintenance strategies. Transportation Research Part B: Methodological, 109, 1-25.

Cooper, M.W., 1980. The use of dynamic programming methodology for the solution of a class of nonlinear programming problems. Naval Research Logistics Quarterly, 27(1), 89-95.

Dantzig, G., 2016. Linear programming and extensions. Princeton university press.

Drinkwater, R. W., Hastings, N. A., 1967. An economic replacement model. Operational Research Quarterly, 18(2), 121-138.

Durango-Cohen, P., 2004. Maintenance and repair decision making for infrastructure facilities without a deterioration model. Journal of Infrastructure Systems, 10(1), 1-8.

Durango-Cohen, P., Madanat, S., 2002. Optimal maintenance and repair policies in infrastructure management under uncertain facility deterioration rates: an adaptive control approach. Transportation Research Part A: Policy and Practice, 36(9), 763-778.

Durango-Cohen, P., Sarutipand, P., 2007. Capturing interdependencies and heterogeneity in the management of multifacility transportation infrastructure system. Journal of Infrastructure Systems, 13(2), 115-123.

Evans, J. J., 1989. Replacement, obsolescence and modifications of ships. Maritime Policy & Management, 16(3), 223-231.

Feinberg, E., Shwartz, A., 1996. Constrained discounted dynamic programming. Mathematics of Operations Research, 21(4), 922-945.

Feinberg, E., Shwartz, A., 1999. Constrained dynamic programming with two discount factors: Applications and an algorithm. IEEE Transactions on Automatic Control, 44(3), 628-631.

Fernandez, J., Friesz, T., 1981. Influence of demand-quality interrelationships on optimal policies for stage construction of transportation facilities. Transportation Science, 15(1), 16-31.

Friesz, T., Fernandez, J., 1979. A model of optimal transport maintenance with demand responsiveness. Transportation Research Part B: Methodological, 13(4), 317-339.

Fwa, T., Chan, W., Tan, C., 1996. Genetic-algorithm programming of road maintenance and rehabilitation. Journal of Transportation Engineering, 122(3), 246-253.

Fwa, T., Tan, C., Chan, W., 1994. Road maintenance planning using genetic algorithms. I: Analysis. Journal of Transportation Engineering, 120(5), 710-722.

Ghellinck, G. T., Eppen, G. D., 1967. Linear programming solutions for separable Markovian decision problems. Management Science, 13(5), 371-394.

Glover, F., 1986. Future paths for integer programming and links to artificial intelligence. Computers and Operation Research,13(5), 533-549.

Golabi, K., Kulkarni, R., Way, G., 1982. A statewide pavement management system. Interfaces, 12 (6), 5-21.

Gray Plant Moody, 2008. Investigative report to joint committee to investigate the I-35W bridge collapse. Minneapolis, MN.

Guerrero, S. E., Madanat, S. M., Leachman, R. C., 2013. The Trucking Sector Optimization Model: A tool for predicting carrier and shipper responses to

policies aiming to reduce GHG emissions. Transportation Research Part E: Logistics and Transportation Review, 59, 85-107.

Gupta, O.K. and Ravindran, A., 1985. Branch and bound experiments in convex nonlinear integer programming. Management science, 31(12), 1533-1546.

Gu, W., Ouyang, Y., Madanat, S., 2012. Joint optimization of pavement maintenance and resurfacing planning. Transportation Research Part B: Methodological, 46 (4), 511-519.

Hajibabai, L., Bai, Y., Ouyang, Y., 2014. Joint optimization of freight facility location and pavement infrastructure rehabilitation under network traffic equilibrium. Transportation Research Part B: Methodological, 63, 38-52.

Hartman, J. C., 1999. A general procedure for incorporating asset utilization decisions into replacement analysis. The Engineering Economist, 44(3), 217-238.

Hartman, J. C., 2004. Multiple asset replacement analysis under variable utilization and stochastic demand. European Journal of Operational Research, 159(1), 145-165.

Hawkins, J., 2003. A Langrangian decomposition approach to weakly coupled dynamic optimization problems and its applications (Doctoral dissertation, Massachusetts Institute of Technology).

Hopp, W. J., Jones, P. C., Zydiak, J. L., 1993. A further note on parallel machine replacement. Naval Research Logistics, 40(4), 575-579.

Jin, D., Kite-Powell, H. L., 2000. Optimal fleet utilization and replacement. Transportation Research Part E: Logistics and Transportation Review, 36(1), 3-20.

Jones, P. C., Zydiak, J. L., Hopp, W. J., 1991. Parallel machine replacement. Naval Research Logistics, 38(3), 351-365.

Jorge, N., Stephen, J.W., 2006. Numerical Optimization, 286-290.

Karabakal, N., Bean, J. C., Lohmann, J. R., 2000. Solving large replacement problems with budget constraints. The engineering economist, 45(4), 290-308.

Karabakal, N., Lohmann, J. R., Bean, J. C., 1994. Parallel replacement under capital rationing constraints. Management Science, 40(3), 305-319.

Keles, P., Hartman, J. C., 2004. Case study: bus fleet replacement. The Engineering Economist, 49(3), 253-278.

Koopman, B.O., 1953. The optimum distribution of effort. Journal of the Operations Research Society of America, 1(2), 52-63.

Kuhn, K., Madanat, S., 2005. Model uncertainty and the management of a system of infrastructure facilities. Transportation Research Part C: Emerging Technologies, 13 (5), 391-404.

Labi, S., Sinha, K.C., 2003. The effectiveness of maintenance and its impact on capital expenditures. Indiana Department of Transportation.

Lasdon, L.S., 2002. Optimization theory for large systems. Courier Corporation.

Lee, J., Madanat, S., 2014a. Jointly optimal policies for pavement maintenance, resurfacing and reconstruction. EURO Journal on Transportation and Logistics, 4(1), 75-95.

Lee, J., Madanat, S., 2014b. Joint optimization of pavement design, resurfacing and maintenance strategies with history-dependent deterioration models. Transportation Research Part B: Methodological, 68, 141-153.

Lee, J., Madanat, S., 2015. A joint bottom-up solution methodology for system-level pavement rehabilitation and reconstruction. Transportation Research Part B: Methodological, 78, 106-122.

Lee, J., Madanat, S., Reger, D., 2016. Pavement systems reconstruction and resurfacing policies for minimization of life-cycle costs under greenhouse gas

emissions constraints. Transportation Research Part B: Methodological, 93, 618-630.

Lee, J., Madanat, S., 2017. Optimal policies for greenhouse gas emission minimization under multiple agency budget constraints in pavement management. Transportation Research Part D: Transport and Environment, 55, 39-50.

List, G. F., Wood, B., Nozick, L. K., Turnquist, M. A., Jones, D. A., Kjeldgaard, E. A., Lawton, C. R., 2003. Robust optimization for fleet planning under uncertainty. Transportation Research Part E: Logistics and Transportation Review, 39(3), 209-227.

Li, Y., Madanat, S., 2002. A steady-state solution for the optimal pavement resurfacing problem. Transportation Research Part A: Policy and Practice, 36(6), 525-535.

Lomnicki, Z.A., 1965. A “branch-and-bound” algorithm for the exact solution of the three-machine scheduling problem. Journal of the Operational Research Society, 16(1), 89-100.

Luss, H., 2012. Equitable Resource Allocation: Models, Algorithms and Applications (Vol. 101). John Wiley & Sons.

Madanat, S., 1993. Incorporating inspection decisions in pavement management. Transportation Research Part B: Methodological, 27(6), 425-438.

Madanat, S., Ben-Akiva, M., 1994. Optimal inspection and repair policies for infrastructure facilities. Transportation Science, 28(1), 55-62.

Madanat, S., Park, S., Kuhn, K., 2006. Adaptive optimization and systematic probing of infrastructure system maintenance policies under model uncertainty. Journal of infrastructure systems, 12(3), 192-198.

Mamlouk, M.S., Dosa, M., 2014. Verification of effectiveness of chip seal as a pavement preventive maintenance treatment using LTPP data. International Journal of Pavement Engineering 15(10), 879-888.

Markow, M., Balta, W., 1985. Optimal rehabilitation frequencies for highway pavements. Transportation Research Record, 1035, 31-43.

Marsten, R.E. and Morin, T.L., 1978. A hybrid approach to discrete mathematical programming. Mathematical programming, 14(1), 21-40.

Medury, A., Madanat, S., 2013. Incorporating network considerations into pavement management systems: A case for approximate dynamic programming. Transportation Research Part C: Emerging Technologies, 33, 134-150.

Miyamoto A., Kawamura K., Nakamura H., 2000. Bridge management system and maintenance optimization for existing bridges. Computer-Aided Civil and Infrastructure Engineering, 15, 45-55.

McClurg, T., Chand, S., 2002. A parallel machine replacement model. Naval Research Logistics, 49(3), 275-287.

Morin, T.L. and Marsten, R.E., 1976a. Branch-and-bound strategies for dynamic programming. Operations Research, 24(4), 611-627.

Morin, T.L. and Marsten, R.E., 1976b. An algorithm for nonlinear knapsack problems. Management Science, 22(10), 1147-1158.

Nicholson, T.A.J. and Pullen, R.D., 1971. Dynamic programming applied to ship fleet management. Journal of the Operational Research Society, 22(3), 211-220.

Oakford, R. V., Lohmann, J. R., Salazar, A., 1984. A dynamic replacement economy decision model. IIE transactions, 16(1), 65-72.

Ohlmann, J., Bean, J., 2009. Resource-constrained management of heterogeneous assets with stochastic deterioration. European Journal of Operational Research, 199(1), 198-208.

Ouyang, Y., 2007. Pavement resurfacing planning on highway networks: A parametric policy iteration approach. Journal of Infrastructure Systems (ASCE) 13(1), 65-71.

Ouyang, Y., Madanat, S., 2004. Optimal scheduling of rehabilitation activities for multiple pavement facilities: exact and approximate solutions. Transportation Research Part A: Policy and Practice, 38(5), 347-365.

Ouyang, Y., Madanat, S., 2006. An analytical solution for the finite-horizon pavement resurfacing planning problem. Transportation Research Part B: Methodological, 40 (9), 767-778.

Parthanadee, P., Buddhakulsomsiri, J., Charnsethikul, P., 2012. A study of replacement rules for a parallel fleet replacement problem based on user preference utilization pattern and alternative fuel considerations. Computers & Industrial Engineering, 63(1), 46-57.

Patriksson, M., 2008. A survey on the continuous nonlinear resource allocation problem. European Journal of Operational Research, 185(1), 1-46.

Powell, W.B., Shapiro, J.A. and Simão, H.P., 2002. An adaptive dynamic programming algorithm for the heterogeneous resource allocation problem. Transportation Science, 36(2), 231-249.

Rashid, M.M., Tsunokawa K., 2012. Trend curve optimal control model for optimizing pavement maintenance strategies consisting of various treatments. Computer-Aided Civil and Infrastructure Engineering, 27(3), 155-169.

Redmer, A., 2009. Optimisation of the exploitation period of individual vehicles in freight transportation companies. Transportation Research Part E: Logistics and Transportation Review, 45, 978-987.

Reid, D. W., Bradford, G. L., 1983. On optimal replacement of farm tractors. American Journal of Agricultural Economics, 65(2), 326-331.

Ross, K., Varadarajan, R., 1989. Markov decision processes with sample path constraints: the communicating case. Operations Research, 37(5), 780-790.

Ross, K., Varadarajan, R., 1991. Multichain Markov decision processes with a sample path constraint: A decomposition approach. Mathematics of Operations Research, 16(1), 195-207.

Sathaye, N., Madanat, S., 2011. A bottom-up solution for the multi-facility optimal pavement resurfacing problem. Transportation Research Part B: Methodological, 45 (7), 1004-1017.

Sathaye, N., Madanat, S., 2012. A bottom-up optimal pavement resurfacing solution approach for large-scale networks. Transportation Research Part B: Methodologcial, 46 (4), 520-528.

Sethi, S. and Sorger, G., 1991. A theory of rolling horizon decision making. Annals of Operations Research, 29(1), 387-415.

Scarf, P., Dwight, R., McCusker, A., Chan, A., 2007. Asset replacement for an urban railway using a modified two-cycle replacement model. Journal of the Operational Research Society, 58(9), 1123-1137.

Simms, B. W., Lamarre, B. G., Jardine, A. K. S., Boudreau, A., 1984. Optimal buy, operate and sell policies for fleets of vehicles. European Journal of Operational Research, 15(2), 183-195.

Sivrikaya-Şerifoğlu, F., Ulusoy, G., 1999. Parallel machine scheduling with earliness and tardiness penalties. Computers & Operations Research, 26(8), 773-787.

Smilowitz, K., Madanat, S., 2000. Optimal inspection and maintenance policies for infrastructure networks. Computer-Aided Civil and Infrastructure Engineering, 15(1), 5-13.

Stasko, T. H., Gao, H. O., 2012. Developing green fleet management strategies: Repair/retrofit/replacement decisions under environmental regulation. Transportation Research Part A: Policy and Practice, 46(8), 1216-1226.

Tan, C. H., Hartman, J. C., 2010. Equipment replacement analysis with an uncertain finite horizon. IIE Transactions, 42(5), 342-353.

Tang, J., Tang, K., 1993. A note on parallel machine replacement. Naval Research Logistics, 40(4), 569-573.

Thomas, D., 1990. Intra-household resource allocation: An inferential approach. Journal of human resources, 635-664.

Transportation Officials, 2011. AASHTO transportation asset management guide: A focus on implementation. AASHTO.

Tsunokawa, K., Hiep, D.V., Ul-Isalm, R., 2006. True optimization of pavement maintenance options with what-if models. Computer-Aided Civil and Infrastructure Engineering, 21, 193-204.

Tsunokawa, K., Schofer, J., 1994. Trend curve optimal control model for highway pavement maintenance: case study and evaluation. Transportation Research Part A: Policy and Practice, 28 (2), 151-166.

Usher, J. S., Whitfield, G. M., 1993. Evaluation of used-system life cycle costs using fuzzy set theory. IIE transactions, 25(6), 84-88.

Vander Veen, D. J.,1985. Parallel replacement under nonstationary deterministic demand. Ph. D. dissertation, Department of Industrial and Operations Engineering, The University of Michigan, Ann Arbor.

Vander Veen, D. J., Jordan, W. C., 1989. Analyzing trade-offs between machine investment and utilization. Management Science, 35(10), 1215-1226.

Whittle, P., 1988. Restless bandits: Activity allocation in a changing world. Journal of applied probability, 25(A), 287-298.

Wijsmuller, M. A., Beumee, J. G. B., 1979. Investment and replacement analysis in shipping. International Shipbuilding Progress, 26(294).

Williamson, J.G., 1971. Optimal replacement of capital goods: the early New England and British textile firm. Journal of Political Economy, 79, 1320-1334.

Yeo, H., Yoon, Y., Madanat, S., 2013. Algorithms for bottom-up maintenance optimization for heterogeneous infrastructure systems. Structure and Infrastructure Engineering, 9(4), 317-328.

Zhang, L., Fu, L., Gu, W., Ouyang, Y., Hu, Y., 2017. A general iterative approach for the system-level joint optimization of pavement maintenance, rehabilitation, and reconstruction planning. Transportation Research Part B: Methodological, 105, 378-400.