https://doi.org/10.1016/j.tre.2021.102268

A two-stage approach for fleet management optimization under time-varying demand

Le Zhang
E-mail address: le.zhang@njust.edu.cn
School of Economics and Management, Nanjing University of Science and Technology, China

Weihua Gu [*]
E-mail address: weihua.gu@polyu.edu.hk
Department of Electrical Engineering, The Hong Kong Polytechnic University, Hong Kong SAR

Liangliang Fu
E-mail address: fuliangliang1984@hotmail.com
Department of Electrical Engineering, The Hong Kong Polytechnic University, Hong Kong SAR

Yu Mei
E-mail address: yu.mei@connect.polyu.hk
Department of Electrical Engineering, The Hong Kong Polytechnic University, Hong Kong SAR

Yaohua Hu
E-mail address: hyh19840428@163.com
College of Mathematics and Statistics, Shenzhen University, China

Abstract

An efficient two-stage heuristic approach is developed for solving the fleet management problem under time-varying demand. Stage 1 of the approach optimizes the vehicles' utilization schedule. Continuous-time approximation is employed to yield a set of near-optimality conditions that can greatly reduce the solution space of this stage. Stage 2 then optimizes the vehicle purchase and retirement schedules. Numerical experiments showed that our approach outperformed a number of previous methods and commercial solvers by large margins in terms of solution quality, computational efficiency, or both.

Keywords: vehicle fleet management; two-stage optimization; continuous-time approximation; first-order condition; time-varying demand

1. Introduction

Freight and passenger transport service providers operate vehicle fleets (e.g., trucks, buses, ships and aircrafts) of variable sizes to serve time-varying demands. Optimal decisions on fleet management are crucial for those service providers to minimize the overall purchase, operation, maintenance, and retirement costs of their fleets. These decisions pertain to: (i) when to purchase new vehicles and retire old ones; and (ii) how to utilize the fleet to meet forecasted demands.

The fleet management problem under demand constraint belongs to the realm of “parallel replacement problems” in the literature (Vander Veen, 1985; Leung and Tanchoco, 1990; Jones et al., 1991; Karabakal et al., 1994). This class of problems aim to find the optimal replacement schedules (or more generally, the purchase and retirement schedules if the number of assets is not fixed) of assets (in our case, the vehicles) that minimize the total cost over a given planning horizon. The assets considered in these problems are interdependent due to budget constraints (Karabakal et al., 1994; Lee and Madanat, 2015; Lee et al., 2016; Zhang et al., 2017), economies of scale (Jones et al., 1991; Büyüktahtakın et al., 2014), demand constraints (Wu et al., 2003; Wu et al., 2005; Guerrero et al., 2013; Guerrero, 2014; Seif et al., 2019; Shields et al., 2019), or combinations of the above (Büyüktahtakın and Hartman, 2016; Des-Bordes and Büyüktahtakın, 2017).

The parallel replacement problems are known to be difficult to solve due to the large solution space (Vander Veen, 1985). As a result, heuristic approaches were often used instead of exact methods (e.g., Karabakal et al., 2000). Simplifying assumptions were also made to reduce the solution complexity. Specifically, many works assumed that an asset’s unit operation and maintenance (O&M) cost per period or per utilization unit (e.g., mile) was a constant (Li et al., 2018), or a function of the asset age (Wu et al., 2003, 2005; Redmer, 2009; Parthanadee et al., 2012; Yatsenko and Hritonenko, 2015; Abdi and Taghipour, 2018; Islam and Lownes, 2019) or maintenance type (Ngo et al., 2018). However, for vehicle assets, empirical studies have shown that their unit O&M costs depend rather on their cumulative mileages than on the above factors (CARB, 2008; Hartman and Tan 2014). Hence, in the fleet management problem, the vehicles’ utilization in terms of mileage (which is a continuous variable), or the mileage-based demand assignment to the vehicles, must be jointly optimized with the fleet purchase and retirement schedules. This joint optimization problem is nonlinear and has a much greater solution space. It thus becomes rather difficult to develop an efficient solution method for this problem. Although some previous studies have also jointly optimized assets’ purchase and retirement plan together with their utilization schedules, most of those works did not account for the dependency of unit O&M cost on an asset’s cumulative utilization (e.g., Wu et al., 2003, 2005; Büyüktahtakın and Hartman, 2016; Des-Bordes and Büyüktahtakın, 2017). Only a handful of those joint optimization studies considered the impacts of cumulative utilization on the unit O&M cost. Regrettably, some of them assumed simple, binary utilization variables (Seif et al, 2019; Shields et al., 2019). Others relied on either a linear modeling approach associated with even larger solution spaces (Hartman, 1999), or overly-simplified heuristic approaches that may result in poor solution quality (Jin and Kite-Powell, 2000; Guerrero et al., 2013). In short, an efficient approach to solving the fleet management problem is still lacking.

Of note, some works in the literature also developed useful analytical insights that have practical implications or can assist in the development of efficient solution methods. For example, Jones et al. (1991) showed for a replacement problem of single-type assets that two properties, namely the “no-splitting” property and the “older cluster replacement” property, should hold simultaneously at optimality. The former means assets of the same age must be replaced at the same time; and the latter means old assets should be replaced before new ones. However, these two seemingly intuitive properties were only proved for cases where the number of assets is fixed (i.e., always a new asset replacing an old one) and where demand or utilization is not concerned (Tang and Tang, 1993; Hopp et al., 1993; McClurg and Chand, 2002; Childress and Durango-Cohen, 2005). In the fleet management problem, however, the optimal fleet size naturally varies in response to the fluctuating demand. We show by numerical examples that the widely-cited "no-splitting" and "older cluster replacement" properties cannot both hold at optimality in this case. Thus, those earlier insights also cannot be applied to solve our joint optimization problem with time-varying demand.

In light of the above, this paper develops an efficient heuristic approach for solving a general fleet management optimization model. The model is a generalization of the truck fleet optimization model proposed by Guerrero et al. (2013), which jointly optimized the truck mileages assigned and the purchase and retirement schedule of multiple types of trucks subject to a time-varying demand constraint. Our approach solves the problem in two stages. Stage 1 optimizes the vehicle mileage assignment problem given the vehicle purchase and retirement schedules. Solution at this stage utilizes an analytical property developed from a continuous-time approximation of the original, discrete-time nonlinear model. This property indicates that, at the optimality, mileage should be allocated to those vehicles with the lowest marginal utilization cost. Built upon this property, we propose a Stage-1 solution approach that can greatly reduce the solution space without notably compromising the solution quality. Stage 2 employs a tabu search algorithm (Glover and Laguna, 1998) to optimize the fleet purchase and retirement schedule. The benefits of our two-stage approach are demonstrated through extensive numerical experiments. For cases where the Stage-1 problem is convex (the simpler case), our approach produced solutions within 0-2% of those developed by a commercial solver (i.e., CVX in Matlab) using only 0.3-13% of the latter's runtimes. Even greater advantages were observed for more general cases with a non-convex Stage-1 problem, where our approach outperformed previous methods in both solution quality and computational efficiency.

The rest of the paper is organized as follows. Section 2 presents the general problem formulation and an equivalent two-stage formulation. Section 3 proposes the heuristic approach. The computation time and solution quality of our approach are tested in Section 4. Numerical case studies are furnished in Section 5. Section 6 demonstrates the robustness of numerical solutions when some parameter values contain errors and uncertainties, and when actual vehicle utilizations deviate from the optimal schedule. Insights and potential extensions are discussed in Section 7.

2. Problem formulations

Section 2.1 presents a general formulation. Section 2.2 presents an equivalent two-stage formulation. Notations used in this paper are summarized in Appendix A.

2.1. A general formulation

The problem is formulated as [P1] below, where the decision variables are: the number of vehicles purchased at time t (those vehicles are termed *cohort* t from now on), denoted by P_t; the type of vehicles in cohort t, γ_t; the mileage served at time τ by a vehicle in cohort t, $u_{\tau,t}$; and the time when the vehicles in cohort t are retired, S_t. The subscripts in the above notations satisfy $1 \le t \le \tau \le T$, where T denotes the planning horizon. The unit of time can be a year, a month, or even a day. Here we assume that all the vehicles in a specific cohort are of the same type, have the same utilization plan over their service lives, and retire at the same time. This assumption is consistent with the "no-splitting" property specified by Jones et al. (1991), and with those commonly assumed in the literature (e.g., Parthanadee et al., 2012; Guerrero et al., 2013; Laksuwong et al., 2014).

[P1]

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

subject to:

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

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

$$\gamma_t \in H,\ 1\le t\le T \quad \text{(1d)}$$

$$S_t\ \text{is an integer},\ 1\le t\le S_t\le T \quad \text{(1e)}$$

$$P_t\ \text{is an integer},\ P_t\ge 0,\ 1\le t\le T \quad \text{(1f)}$$

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

$$y_{S_t,t}\le\bar{y},\ 1\le t\le S_t\le T \quad \text{(1h)}$$

In the RHS of the objective function (1a), the first term is the total discounted vehicle purchase cost, where $A(\gamma_t)$ denotes the cost for purchasing a type-γ_t vehicle, and r the discount rate; the second term is the total discounted O&M cost, where $M(y_{\tau,t},\gamma_t)$ denotes the unit O&M cost per vehicle per mile, and $y_{\tau,t}$ a cohort-t vehicle's cumulative mileage at τ; and the last term is the total discounted salvage value, where $F(y_{S_t,t},\gamma_t)$ indicates the salvage value of a cohort-t vehicle that retires at S_t. The following three assumptions are made for functions $M(\cdot)$ and $F(\cdot)$:

(i) $M\left(y_{\tau,t},\gamma_t\right) > 0$ and $\frac{\partial M}{\partial y_{\tau,t}} > 0$, meaning that the unit O&M cost increases with $y_{\tau,t}$ (CARB, 2008);

(ii) $F\left(y_{S_t,t},\gamma_t\right) \ge 0$ and $\frac{\partial F}{\partial y_{S_t,t}} < 0$, meaning that the salvage value decreases with $y_{S_t,t}$; and

(iii) $\frac{\partial}{\partial y_{S_t,t}}\left(M - \frac{\partial F}{\partial y_{S_t,t}}\right) > 0$, meaning that the utilization cost per mile at a vehicle's retirement time, $M - \frac{\partial F}{\partial y_{S_t,t}}$, increases with its final mileage $y_{S_t,t}$.[1]

Constraint (1b) specifies that a given demand at each time τ, denoted by D_τ (measured by miles), has to be satisfied. For simplicity, the demand is assumed to be infinitely divisible between vehicles. Constraint (1c) defines $y_{\tau,t}$ $(1\le t\le\tau\le S_t)$ as the cumulative mileage of a cohort-t vehicle at τ. Constraint (1d) specifies the set of vehicle types, denoted by H. Constraints (1e-h) are the boundary and integer constraints for S_t, P_t, $u_{\tau,t}$ and $y_{\tau,t}$, respectively, where U is the maximum mileage a vehicle can serve per unit time, and $\bar{y}$ the maximum allowable cumulative mileage.

Program [P1] is a mixed-integer nonlinear program with $\frac{T(T+7)}{2}$ decision variables. The nonlinearity is due to the demand constraint (1b) and the O&M cost term in the objective function. It is also nonconvex in general. Thus, its exact solution is very difficult to obtain when T is large. We next reformulate it as a two-stage problem, for which a heuristic approach will be developed in Section 3.

2.2. The equivalent two-stage formulation

We propose the following two-stage formulation. The Stage-1 problem [P2] optimizes the vehicle utilization plan, i.e., $u_{\tau,t}$ $(1\le t\le\tau\le S_t)$, for a given set of P_t, γ_t, and S_t $(1\le t\le T)$. The

---

[1] Assumption (iii) simplifies our solution approach. However, a similar but moderately more complicated solution approach can still be developed if this assumption is relaxed. See Section 3.1.3 for more details.

Stage-2 problem [P3] optimizes P_t, γ_t, and S_t $(1 \le t \le T)$ given that the optimal $u_{\tau,t}$ $(1 \le t \le \tau \le S_t)$ is expressed as a function of P_t, γ_t, and S_t $(1 \le t \le T)$.

[P2]

$$\min_{u_{\tau,t}} J' = \sum_{t=1}^{T}\sum_{\tau=t}^{S_t} P_t u_{\tau,t} M(y_{\tau,t},\gamma_t) e^{-r\tau} - \sum_{t=1}^{T} P_t F\left(y_{S_t,t},\gamma_t\right) e^{-rS_t} \qquad (2)$$

subject to: (1b), (1c), (1g), and (1h)

[P3]

$$\min_{P_t,\gamma_t,S_t} J = \sum_{t=1}^{T} A(\gamma_t) P_t e^{-rt} + \sum_{t=1}^{T}\sum_{\tau=t}^{S_t} P_t u_{\tau,t} M(y_{\tau,t},\gamma_t) e^{-r\tau} - \sum_{t=1}^{T} P_t F\left(y_{S_t,t},\gamma_t\right) e^{-rS_t} \qquad (3a)$$

subject to: (1c)-(1f), and

$$u_{\tau,t} = g_{\tau,t}^{u}(\{P_t,\gamma_t,S_t, 1 \le t \le T\}), 1 \le t \le \tau \le S_t \le T, \qquad (3b)$$

where $g_{\tau,t}^{u}(\cdot)$ denotes the optimal solution of $u_{\tau,t}$ expressed as a function of given P_t, γ_t, and S_t $(1 \le t \le T)$, which is found by solving [P2]. An optimal solution to [P3] must also be optimal to the original program [P1] and vice versa. In other words, [P1] an [P3] are equivalent.

We next present the heuristic approach for solving the two-stage formulation.

3. The solution approach

The key element of our approach is a near-optimal solution to the Stage-1 problem [P2], as described in Section 3.1. Section 3.2 presents the tabu search algorithm for solving the Stage-2 problem [P3].

3.1. A heuristic solution to [P2]

We first convert the discrete-time formulation [P2] to a continuous-time approximation model [P4], as presented in Section 3.1.1. An optimality property is developed analytically for [P4] in Section 3.1.2. Built upon this property, a heuristic solution to [P2] is presented in Section 3.1.3.

3.1.1. The continuous-time approximation model

Continuous-time approximation, or more generally, the continuous approximation technique, was often used in the literature of pavement management optimizations (Rashid and Tsunokawa, 2012), supply chain and logistics system optimizations (Tsao and Lu, 2012), and public transportation network optimizations (Chen et al., 2015; Chen and Nie, 2018; Mei et al., 2020). The technique approximates numerous discrete variables and parameters by a few continuous functions. The resulting program becomes parsimonious and can often be tackled using calculus of variations.

Specifically, we approximate [P2] by the following program [P4], where the discrete-time parameters P_t, γ_t, S_t, and D_τ $(0 < t,\tau \le T)$ are replaced by the continuous-time functions $P(t)$, $\gamma(t)$, $S(t)$, and $D(\tau)$ $(0 < t,\tau \le T)$, and the variables $u_{\tau,t}$ and $y_{\tau,t}$ $(0 < t \le \tau \le S_t)$ by $u(\tau,t)$ and $y(\tau,t)$ $(0 < t \le \tau \le S(t))$, respectively. Note that $P(t)$ and $u(\tau,t)$ denote the vehicle purchase rate at t and the utilization rate at τ per vehicle of cohort t, respectively. For simplicity, other notations are kept unchanged. The relation between $y_{\tau,t}$ and $u_{\tau,t}$, (1c), is now written as a partial differential equation (4c). The summations in [P2] are replaced by the integrals in [P4].

[P4]

$$\min J' = \int_{t=0}^{T}\int_{\tau=t}^{S(t)} P(t) u(\tau,t) M\big(y(\tau,t),\gamma(t)\big) e^{-r\tau} d\tau dt - \int_{t=0}^{T} P(t) F(y(S(t),t),\gamma(t)) e^{-rS(t)}\, dt \qquad (4a)$$

subject to:

$$\int_{t:0\leq t\leq\tau\leq S(t)} P(t)u(\tau,t)\,dt = D(\tau), \text{ for } \tau\in(0,T] \tag{4b}$$

$$\frac{\partial y(\tau,t)}{\partial\tau} = u(\tau,t), \text{ for } t\in(0,T], \tau\in[t,S(t)] \tag{4c}$$

$$0\leq u(\tau,t)\leq U, \text{ for } t\in(0,T], \tau\in[t,S(t)] \tag{4d}$$

$$y(S(t),t)\leq\bar{y}, \text{ for } t\in(0,T] \tag{4e}$$

[P2] asymptotically converges to [P4] when the time interval for decisions approaches zero (i.e., when the decisions can be made with infinitesimal intervals). Hence, the optimal solution to [P4] should be close to the optimal solution to [P2], especially when the time interval is small.

3.1.2. An optimality property of the continuous-time model

First, define the *z-score* of cohort t at time τ, $z(y(\tau,t),\tau,t)$ $(0<t\leq\tau\leq S(t))$, as follows:

$$z(y(\tau,t),\tau,t)\equiv M\big(y(\tau,t),\gamma(t)\big) \text{ for } \tau\in\big[t,S(t)\big),\ t\in(0,T] \tag{5a}$$

$$z(y(S(t),t),S(t),t)\equiv M\big(y(S(t),t),\gamma(t)\big)-\frac{\partial F}{\partial y(S(t),t)} \text{ for } t\in(0,T]. \tag{5b}$$

The z-score can be interpreted as the cost for a cohort-t vehicle to cover an additional mile at τ: for a non-retiring vehicle at τ (i.e. a vehicle with $S(t)>\tau$), the z-score is equal to the unit O&M cost; while for a retiring vehicle (i.e. one with $S(t)=\tau$), it is the unit O&M cost minus the marginal salvage value. In other words, the z-score essentially represents a vehicle's marginal utilization cost, accounting for the differences between vehicle types and between non-retiring and retiring vehicles.

We now present the following proposition:

Proposition 1. At the optimality of [P4], if $P(t)\neq 0$ for a $t\in(0,T]$, then for any $\tau\in\big[t,S(t)\big)$, one of the following three conditions holds:

$$u(\tau,t)=0, \tag{6a}$$

$$u(\tau,t)=U, \text{ or} \tag{6b}$$

$$z(y(\tau,t),\tau,t)=\lambda(\tau)-\frac{1}{r}\frac{d\lambda(\tau)}{d\tau}; \tag{6c}$$

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

$$u(S(t),t)=0, \tag{7a}$$

$$u(S(t),t)=U, \tag{7b}$$

$$y(S(t),t)=\bar{y}, \text{ or} \tag{7c}$$

$$z(y(S(t),t),S(t),t)=\lambda(\tau) \tag{7d}$$

where $\lambda(\tau)$ $(\tau\in(0,T])$ is the Lagrange multiplier for relaxing constraint (4b). Proof of Proposition 1 employs the first-order necessary conditions of [P4]. The details are relegated to Appendix B.

The first half of Proposition 1 means that, at a given τ, the z-scores of all the non-retiring vehicles, regardless of their cohorts, should be equal (note that the RHS of (6c) is only a function of τ but not of the cohort index t), if their utilization is neither zero nor U. This is intuitive from the economic point of view. Recall that the z-score is the marginal utilization cost. If two non-retiring vehicles with different z-scores are used at the same time, shifting some demand from the vehicle with a higher z-score to the other vehicle will reduce the total cost. This kind of demand shift can be carried on within the fleet until some vehicles have no demand to shift out (i.e., $u(\tau,t)=0$), others have reached the maximum utilization ($u(\tau,t)=U$), and the remaining vehicles all have the same z-score. A similar note can be made for retiring vehicles, except that a retiring vehicle's cumulative mileage is capped by $\bar{y}$. Note that a non-retiring vehicle and a retiring vehicle at the same τ may not have equal z-scores.

Proposition 1 implies that the optimal solution to [P4] can be derived if $\lambda(\tau)$ $(\tau \in (0,T])$ is known. Inspired by this, the discrete-time program [P2] can be solved using a discrete-time analog of Proposition 1, which is presented next.

3.1.3. A heuristic approach for solving [P2]

The approach is built upon a discrete-time analog of Proposition 1, which is presented below:
Proposition 2. A near-optimal solution to [P2] can be developed to satisfy the following conditions: if $P_t \neq 0$ for a $t \in \{1,2,\ldots,T\}$, then for any $\tau \in \{t, t+1, \ldots, S_t - 1\}$, one of the following three conditions holds:

$$u_{\tau,t} = 0 \tag{8a}$$

$$u_{\tau,t} = U \tag{8b}$$

$$z_{\tau,t}(y_{\tau,t}) \equiv M(y_{\tau,t}, \gamma_t) = \lambda_\tau - \frac{1}{r}(\lambda_{\tau+1} - \lambda_\tau) \tag{8c}$$

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

$$u_{S_t,t} = U \tag{9a}$$

$$y_{S_t,t} = \bar{y} \tag{9b}$$

$$z_{S_t,t}(y_{S_t,t}) \equiv M(y_{S_t,t}, \gamma_t) - \frac{\partial F}{\partial y_{S_t,t}} = \lambda_\tau \tag{9c}$$

where $z_{\tau,t}(y_{\tau,t})$ is the z-score at τ for a cohort-t vehicle, $1 \le t \le \tau \le S_t$; and λ_τ $(\tau \in \{1,2,\ldots,T\})$ the Lagrange multiplier for relaxing (1b). Note that (7a) in Proposition 1 is dropped in the discrete-time case because, if $u_{S_t,t} = 0$, then cohort t should retire at $S_t - 1$ instead of S_t.

Proposition 2 does not guarantee global optimality[2]. However, since Proposition 2 and [P2] are discrete-time analogs of Proposition 1 and [P4], respectively, and Proposition 1 states the optimality conditions of [P4], we believe a solution developed using Proposition 2 would be near-optimal. We next show how such a solution can be developed.

The solution will be derived in an iterative fashion. First, when $\tau = 1$, we have $u_{1,1} = \frac{D_1}{P_1}$ (without loss of generality, we assume $D_1 > 0$ and thus $P_1 > 0$). Now suppose cohort 1 does not retire at $\tau = 1$. Then (8c) holds at $\tau = 1$, i.e., $z_{1,1}(y_{1,1}) = M(y_{1,1}, \gamma_1) = \lambda_1 - \frac{1}{r}(\lambda_2 - \lambda_1)$. If λ_1 is given exogenously, then λ_2 can be derived from the above equation.

Now suppose λ_τ is already known, allocate the demand D_τ among the existing fleet as follows:

(i) For a retiring cohort t (i.e., $\tau = S_t$), calculate $\hat{y}_{S_t,t} = z^{-1}_{S_t,t}(\lambda_\tau)$ from (9c), where $z^{-1}_{S_t,t}(\cdot)$ is the inverse function of $z_{S_t,t}(\cdot)$. Note that assumption (iii) in Section 2.1 means $\frac{dz_{S_t,t}}{dy_{S_t,t}} > 0$, and this results in a single-valued $\hat{y}_{S_t,t}$. The $y_{S_t,t}$ is then calculated as $y_{S_t,t} = \min\{\hat{y}_{S_t,t}, y_{S_t-1,t} + U, \bar{y}\}$. This means that, if a retiring cohort's cumulative mileage cannot reach $\hat{y}_{S_t,t}$, it must be equal to $y_{S_t-1,t} + U$ or $\bar{y}$, whichever is lower. One can easily verify that the above $y_{S_t,t}$ satisfies (9a-c). The $u_{S_t,t}$ can be calculated as $y_{S_t,t} - y_{S_t-1,t}$.

(ii) After allocating the demand to all the retiring cohorts, calculate the remaining demand. The remaining demand will be first allocated to the non-retiring cohort(s) with the lowest z-score. When that lowest z-score increases and catches up with a previously higher z-score, the demand

---

[2] The optimality property of [P2] that are similar to Proposition 1 cannot be developed because the first-order conditions of [P2] are more complicated and cannot be simplified in a way similar to Appendix B. In other words, equal z-score (i.e., (8c) and (9c)) is not an optimality property for the discrete-time model.

will also be allocated to the cohorts that have that previously higher z-score (this is like flooding a staircase step by step with water). If a cohort's mileage per vehicle reaches U, no more demand will be fed to this cohort. The process ends when no more demand is left. Then calculate $u_{\tau,t}$ for all the non-retiring cohorts.

(iii) Calculate the highest z-score of all the non-retiring cohorts that have received demand in step (ii). Use that z-score and (8c) to calculate $\lambda_{\tau+1}$. (The highest z-score is associated with the last non-retiring cohort(s) that receives demand before the process in step (ii) ends.)

Pseudo code of the above approach is summarized in Appendix C.1. Note, however, that the above process can be iterated only if there exists at least one non-retiring cohort that receives some demand at each time τ. If at a certain τ there is no non-retiring cohort, steps (ii-iii) cannot be executed and $\lambda_{\tau+1}$ cannot be derived. In this case, $\lambda_{\tau+1}$ needs to be given exogenously so that the iteration process can resume. We term the time i $(1 \leq i \leq T)$ when a new λ_i needs to be specified exogenously as a "breakpoint". (The first breakpoint is the start time, $i = 1$.) The λ_i's associated with breakpoints can be optimized using some derivative-free gradient or subgradient search methods (see, e.g., Rios and Sahinidis, 2013).[3] Appendix C.2 furnishes a derivative-free approximate gradient algorithm for optimizing these λ_i's.

Of a related note, if assumption (iii) in Section 2.1 is relaxed, then $\hat{y}_{S_t,t} = z_{S_t,t}^{-1}(\lambda_\tau)$ may be multi-valued in the above step (i). If $z_{S_t,t}^{-1}(\lambda_\tau)$ returns a small finite set of values (which is usually the case), then the Stage-1 problem can still be solved by a modified approach in which all possible values of $\hat{y}_{S_t,t}$ are enumerated. However, this modified approach would exhibit a greater computational complexity.

3.2. A tabu-search method for solving [P3]

The first step of the tabu search method is to obtain a feasible initial solution to [P3]. This solution, denoted by $\boldsymbol{x}^0 \equiv \{P_t^0, \gamma_t^0, S_t^0 : t = 1,2,\ldots,T\}$, is generated by a greedy heuristic algorithm. Pseudo code of this greedy heuristic algorithm is provided in Appendix C.3.

We now describe the tabu search algorithm. The description is kept short in the interest of brevity because the algorithm is only a standard practice of the tabu search method. For more details on the theory of tabu search, please refer to Glover and Laguna (1998).

Define a *move* as a change from a feasible solution $\boldsymbol{x}$ to a new feasible solution, where the change can be one of the following: (i) $P_t \rightarrow P_t + 1$ or $P_t - 1$ (if $P_t > 0$) for a certain t; (ii) γ_t switches to another value in H for a certain t; and (iii) $S_t \rightarrow S_t + 1$ (if $S_t < T$) or $S_t - 1$ (if $S_t > t$) for a certain t. At each move, the heuristic approach presented in Section 3.1.3 is executed to find the vehicle utilization schedule, and the discounted total cost J is calculated. If no feasible utilization schedule is obtained, J is set to infinity. Define the *neighborhood* of $\boldsymbol{x}$, $\mathcal{N}(\boldsymbol{x})$, as the set of feasible solutions that can be obtained by making one move from $\boldsymbol{x}$. Further define the *tabu list*, TL, as the list of inverse moves of those most recent moves performed. The maximum length of tabu list is denoted as $tabu\_size$. In each iteration, a move is made according to one of the following two rules:

(i) If no move in $\mathcal{N}(\boldsymbol{x})$ can produce a lower total cost as compared to the best solution so far, set the current move to the one in $\mathcal{N}(\boldsymbol{x}) \setminus TL$ that produces the lowest total cost. Following this rule, a move is made even if it produces a higher cost than the best solution so far.

---

[3] The number of breakpoints is generally small. For most numerical instances in this paper, λ_1 is the only Lagrange multiplier that needs to be optimized via search methods.

(ii) If a move in $\mathcal{N}(\boldsymbol{x}) \cap TL$ produces a lower total cost than the best solution so far, set the current move to the lowest-cost move in $\mathcal{N}(\boldsymbol{x})$.

The tabu list TL is updated after each iteration. It is used to prevent the algorithm from returning to a solution attained in a previous iteration. Rule (i) finds the best neighboring solution that is generated not from any move in the tabu list. However, if a move in the tabu list can yield a better solution than the best one so far, that move is still selected according to rule (ii). The algorithm ends when no better solution is found after max_num_tb consecutive iterations. The pseudo code of this algorithm is provided in Appendix C.4.

4. Performance of the two-stage approach

Section 4.1 presents the cost functions and parameter values used in numerical experiments. Section 4.2 evaluates the solution quality and computational efficiency of our approach. All the numerical instances were carried out via Matlab R2016b on an HP 3.20GHz personal computer with 4GB RAM.

4.1. Cost functions and parameter values

We first consider a special case with cost functions borrowed from Guerrero et al. (2013) for a truck fleet management problem. They are presented as follows:

$$A(\gamma_t) = A_p + \frac{k_1 \gamma_t^2}{k_2 - \gamma_t} \tag{10a}$$

$$M(y_{\tau,t}, \gamma_t) = \theta_M + k_0 + (\theta_F + p_F)(1 - \gamma_t) f + (k_{m0} + \beta\gamma_t) y_{\tau,t} \tag{10b}$$

$$F(y_{S_t,t}, \gamma_t) = A(\gamma_t) k_d (1 - k_x y_{S_t,t}) \tag{10c}$$

$$\bar{y} = 1/k_x \tag{10d}$$

where $A_p, k_1, k_2, \theta_M, k_0, \theta_F, p_F, f, k_{m0}, \beta, k_d$ and k_x are constant parameters, whose definitions and values are summarized in Table 1. Those values were also borrowed from Guerrero et al. (2013)[4]. Here γ_t represents the fuel-saving efficiency of cohort-t trucks. A larger γ_t renders a lower unit O&M cost, but a higher purchase cost. Note that assumptions (i-iii) specified in Section 2.1 are all satisfied here. Values of D_τ $(1 \leq \tau \leq T)$ are specified for each numerical instance separately, as described in the following sections.

Table 1. Parameter definitions and values

Parameter	**Notation**	**Value**	**Unit**
Fixed truck purchase cost	A_p	1.3E5	\$/truck
Coefficient for the variable truck purchase cost	k_1	3.8E5	\$/truck
Coefficient for the 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
Baseline fuel efficiency	f	0.169	gallons/mile
Fixed maintenance cost coefficient	k_{m0}	1.85E-7	\$/mile
Variable maintenance cost coefficient	β	2.57E-7	\$/mile
Instantaneous depreciation for the salvage value	k_d	0.75	-
Mileage depreciation for the salvage value	k_x	9.77E-7	mile^{-1}
Maximum mileage served per truck per unit time	U	1E5	mile

---

[4] The only exception is that the value of k_1 is different. If the original value was used, type-II trucks would be too advantageous over type-I trucks, and would be the only truck type selected in a solution.

Discount rate (when the time unit is one year)	r	0.07 if the time unit is a year; 0.07/12 if that is a month	-
Set of truck types	H	$\{0, 0.3\}$	-
Planning horizon	T	5-50	year

Note under this special case that the Stage-1 problem [P2] happens to be convex. Hence, its optimal solution can be obtained via gradient search methods or commercial solvers such as the CVX solver (Boyd and Vandenberghe, 2004), which will be used as a benchmark method for comparison against our approach.

To examine the performance of our approach for the more general non-convex Stage-1 problems, we also conduct numerical tests using a second set of cost models, where (10b) is replaced by:

$$M(y_{\tau,t}, \gamma_t) = \theta_M + k_0 + (\theta_F + p_F)(1 - \gamma_t)f + (k_{m0} + \beta\gamma_t)y_{\tau,t}^2. \tag{11}$$

This renders a non-convex Stage-1 problem. All the other cost models and parameter values are the same as in the convex cost models.

4.2 Performance of our heuristic approach

We tested totally 9 batches of numerical instances. For the first 7 batches, we set $T = 5, 6, 10, 20, 30, 40, 50$ years, respectively; and for the last 2 batches, $T = 60, 120$ months, respectively, to reflect finer planning time intervals. Each batch includes 10 instances with D_τ ($\tau \in \{1,2, \dots, T\}$) randomly generated from a uniform distribution: over the support $[2.0\text{E}6, 2.8\text{E}6]$ miles for the first 7 batches, and $\left[\frac{2.0\text{E}6}{12}, \frac{2.8\text{E}6}{12}\right]$ miles for the last 2 batches.

We first use the convex cost functions given by (10a-d). For the tabu search algorithm for [P3], different values of $tabu\_size$ were used for problems of different sizes. This is because a too small $tabu\_size$ will render the search process easily trapped around a local minimum, while a too large $tabu\_size$ may prevent the algorithm from finding a better solution (Glover and Laguna, 1998). The 2nd column of Table 2 shows the $tabu\_size$ found by trial and error for the 9 batches of numerical instances (the same $tabu\_size$ can often be used for problems of similar sizes). The parameter max_num_tb was set to 15.

Solutions and computation times of our approach are compared against three benchmark approaches. The first one is the heuristic approach proposed in Guerrero et al. (2013), where the trucks' utilization plan and retirement schedules were optimized separately using a simplified time-invariant model. The second benchmark approach is borrowed from Hartman (1999), where the original non-linear model [P1] is linearized by discretizing the vehicle mileage using an interval $\mathbb{w}$. The resulting mixed integer linear program (MILP) is then solved by CPLEX. The details of this approach and the MILP model are furnished in Appendix D. In the third benchmark approach, CVX is employed to solve [P2] to global optimality; exhaustive search (for smaller instances with $T = 5$ and 6) and the tabu search method described in Section 3.2 (for larger-scale instances with $T \geq 10$) are used to solve [P3]. Note that exhaustive search would fail for larger-scale instances due to the curse of dimensionality. Global optima are thus obtained only for smaller instances.

We calculate the following three relative errors between the solutions produced by our approach and the three benchmark approaches:

$$\varepsilon_{Guerrero} = \frac{[\text{minimum cost of Guerrero's approach}] - [\text{minimum cost of our approach}]}{[\text{minimum cost of our approach}]};$$

$$\varepsilon_{Hartman} = \frac{[\text{minimum cost of Hartman's approach}] - [\text{minimum cost of our approach}]}{[\text{minimum cost of our approach}]};$$

$$\varepsilon_{CVX}=\frac{[\text{minimum cost of CVX}-\text{based approach}]-[\text{minimum cost of our approach}]}{[\text{minimum cost of our approach}]};$$

The means of $\varepsilon_{Guerrero}$, $\varepsilon_{Hartman}$, and ε_{CVX} for each of the 9 batches of instances are presented in columns 3-6 of Table 2. The mean of $\varepsilon_{Hartman}$ is presented for two different values of $\mathbb{u}$: 5×10^4 and 5×10^3 miles. A positive error indicates that our solution is better than the corresponding benchmark. We also present the minima of ε_{CVX} errors in the 7th column of the table, which indicates the maximum gaps between our solutions and the CVX-based ones (which are better). We further show the mean runtimes for the four solution approaches in the last five columns of the table.

Table 2. Relative cost errors and runtimes for the four solution approaches when [P2] is convex

T	**Tabu size**	**Mean $\varepsilon_{Guerrero}$**	**Mean $\varepsilon_{Hartman}$**		**ε_{CVX}**		**Mean runtime (sec)**				
			$\mathbb{u}$ =5E4	$\mathbb{u}$ =5E3	**Mean**	**Min**	**Our approach**	**Guerrero's approach**	**Hartman's approach** $\mathbb{u}$ =5E4	$\mathbb{u}$ =5E3	**CVX-based approach**
5	8	12.13%	11.48%	0.79%	-0.32%	-0.60%	3.68	4.31	29.51	17879.01	357.13
6	10	17.00%	12.30%	0.75%	-0.43%	-0.54%	8.03	5.15	40.52	24018.04	2590.79
10	20	16.33%	12.45%	0.78%	-0.35%	-0.55%	20.36	22.35	126.21	81754.78	179.23
20	25	15.36%	12.63%	4.97%	-0.37%	-0.62%	46.20	88.23	582.79	86400*	405.12
30	60	16.48%	11.48%	6.12%	-0.31%	-0.43%	167.12	98.93	2150.77	86400*	1335.15
40	125	14.63%	13.09%	7.09%	-0.27%	-0.36%	274.78	104.08	7981.26	86400*	2229.23
50	210	15.65%	11.67%	7.37%	-0.31%	-0.45%	503.25	149.51	48931.07	86400*	4827.69
60	300	13.77%	13.98%	8.45%	-0.47%	-0.58%	1365.79	237.81	86400*	86400*	10729.87
120	500	15.52%	14.22%	8.90%	-1.13%	-1.57%	3198.34	634.09	86400*	86400*	27802.14

[*] For these instances, Hartman's approach did not converge after 24 hours (86400 seconds). Thus, only the best solutions recorded in 24 hours were used here.

Comparison against each benchmark approach unveils distinct results. First, column 3 of the table shows that our approach produced costs that are on average 12-17% lower than Guerrero's approach, showing the advantage of our approach over Guerrero's despite the lower runtimes of the latter approach (see columns 8 and 9). This is because the overly-simplified utilization optimization model in Guerrero's approach significantly undermined the solution quality.

Columns 4 and 5 show that our approach also outperformed Hartman's linear modeling approach by a large margin in terms of solution quality, especially when $T\geq20$. Although Harman's approach can attain the global optimum when the discretization interval $\mathbb{u}$ approaches zero, a large $\mathbb{u}$ such as those used in the above tests can render considerable errors. This is why it loses to our heuristic approach even in terms of solution quality. On the other hand, further decreasing $\mathbb{u}$ does not improve the solution quality of Harman's approach, since the runtime increases exponentially with T and soon becomes prohibitively high (e.g., over 24 hours); see columns 10-11 of the table.

Finally, comparison against the CVX-based approach unveils that our approach produced costs that are very close to the latter approach, with a gap less than 1% for most cases; see columns 6 and 7. On the other hand, our average runtime is only 0.3-13% of the CVX-based approach; see the last column. (Closer investigation unveils that for each instance of $T\geq10$, the numbers of tabu search iterations executed in Stage 2 are similar between our approach and the CVX-based one, meaning that the runtime saving is mainly attributed to our heuristic method for solving the Stage-1 problem [P2].) In short, results in Table 2 indicate that our approach performed very good in both solution quality and computational efficiency.

Note that the benefits of our approach are limited when [P2] is convex, because a convex [P2] can be efficiently solved to global optimality. However, such a convexity is not guaranteed for the general case. We next show that our approach would perform even better when a non-convex [P2] is used (i.e., when (10b) is replaced by (11)). For the non-convex case, we employ two commonly used solvers as benchmark approaches for solving [P2]: the “fmincon” solver in Matlab using the sequential quadratic programming algorithm (Osorio and Bierlaire, 2013), and the SCIP solver (Wei et al., 2014). Tabu search is still used in both benchmark approaches to solve [P3]. Hartman’s linear modeling approach is still used as the third benchmark. In addition to $\varepsilon_{Hartman}$, the following error terms are calculated:

$$\varepsilon_{fmincon} = \frac{\text{[minimum cost of fmincon−based approach]−[minimum cost of our approach]}}{\text{[minimum cost of our approach]}}$$

$$\varepsilon_{SCIP} = \frac{\text{[minimum cost of scip−based approach]−[minimum cost of our approach]}}{\text{[minimum cost of our approach]}}$$

Means of these error terms are presented in columns 3-6 of Table 3, and the runtimes of the four approaches are presented in columns 7-11 of that table. These values show that, for every value of T examined, our approach always outperformed all the three benchmark methods in terms of both solution quality and computational cost. The advantage increased with the problem size. When $T = 120$, the cost reductions as compared to the benchmark approaches are 6-14%. Also note for $T \geq 30$ that the benchmark approaches often failed to attain convergence within 24 hours, while our approach still found solutions within 1 hour. We believe these results have compellingly demonstrated the benefits of our solution approach.

Table 3. Relative cost errors and runtimes for the four solution approaches when [P2] is non-convex

T	**Tabu size**	**Mean $\varepsilon_{fmincon}$**	**Mean ε_{SCIP}**	**Mean $\varepsilon_{Hartman}$**		**Mean runtime (sec)**				
				$\mathbb{u}$ =5E4	$\mathbb{u}$ =5E3	**Our approach**	**fmincon-based approach**	**SCIP-based approach**	**Hartman’s approach** $\mathbb{u}$ =5E4	**Hartman’s approach** $\mathbb{u}$ =5E3
10	20	1.81%	0.87%	11.30%	0.91%	25.03	402.09	1944.09	157.21	8027.78
20	25	2.26%	0.65%	11.69%	5.13%	50.87	2960.32	14691.22	607.91	86400*
30	60	1.70%	1.06%	12.57%	5.92%	145.87	20604.11	86400*	1848.60	86400*
40	125	4.04%	2.39%	12.70%	7.38%	259.04	86400*	86400*	8110.33	86400*
50	210	4.84%	3.44%	13.31%	7.95%	483.91	86400*	86400*	49902.08	86400*
60	300	5.67%	3.56%	13.88%	8.61%	1507.05	86400*	86400*	86400*	86400*
120	500	6.30%	6.17%	13.94%	9.27%	3409.61	86400*	86400*	86400*	86400*

* For these instances, the corresponding approaches did not converge after 24 hours. Thus, only the best solutions recorded in 24 hours were used for each instance.

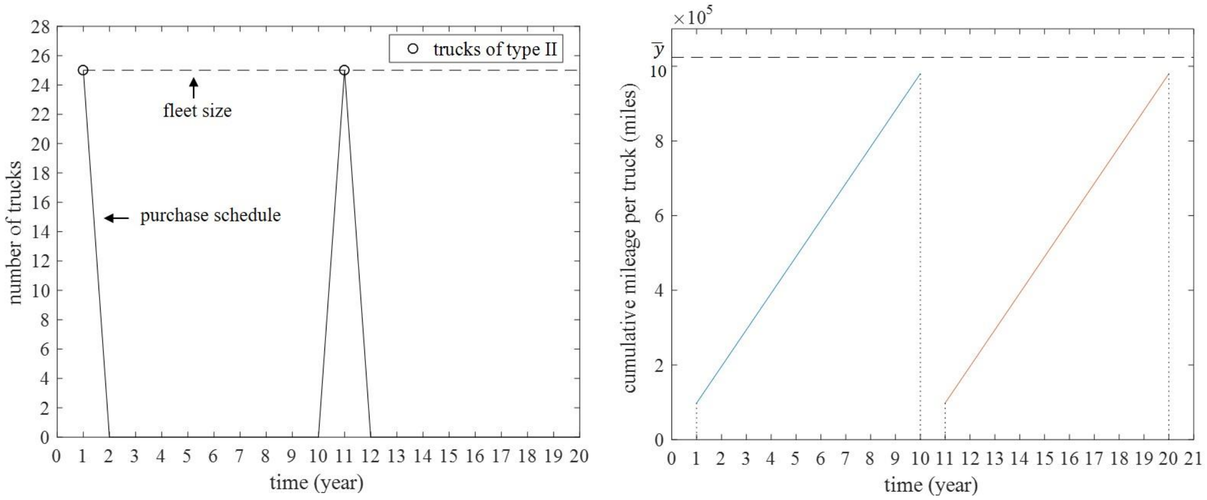
5. Numerical case studies

To examine the optimal fleet management plans, in this section we present solutions of numerical instances with $T = 20$ years under three demand patterns: a constant demand (Section 5.1), a linearly increasing demand (Section 5.2) and a demand pattern with a demand drop in middle years (Section 5.3). The convex cost models and parameter values in Section 4.1 are used.

5.1. Constant demand pattern

First assume $D_\tau = 2.45\text{E}6$ miles, $\forall \tau \in \{1, \ldots, T\}$. The optimal truck purchase plan and fleet size over the planning horizon are plotted as the solid and dashed curves, respectively, in Figure 1a. The figure shows that two equal-sized cohorts are purchased in years 1 and 11, and each cohort contains 25 type-II trucks (i.e., $\gamma_t = 0.3$) with 10-year service lives. Figure 1b plots the cumulative mileage trajectories for the two cohorts as solid curves. These linear trajectories reveal that each truck in the two cohorts serves a fixed annual mileage (0.98E5 miles), which is only slightly below $U = 1\text{E}5$ miles. This indicates that only the minimum number of trucks required (i.e., $\left\lceil \frac{D_\tau}{U} \right\rceil = 25$) are purchased for each cohort, and that each truck is almost fully utilized every year until its cumulative mileage is close to the limit $\bar{y}$ (as marked by the dashed horizontal line in Figure 1b). This periodic truck purchase and utilization plan is a natural result of the constant demand. Only type-II trucks are used in this plan because, when a truck is nearly fully utilized, a type-II truck's cost per mile served is lower despite its higher purchase cost. This periodic solution pattern was consistently observed when the constant demand D_τ took other values, and when T was an integer multiple of 10 years. Note that 10 years is the maximum service life of a fully-utilized truck before its cumulative mileage reaches $\bar{y}$.

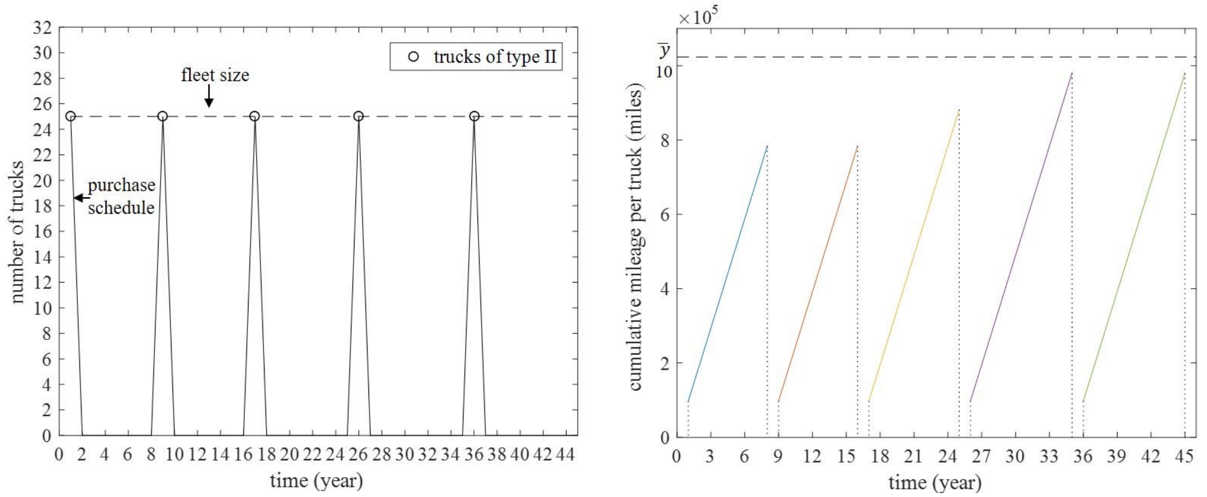
Results are a little different when T is not an integer multiple of 10 years. Figures 2a and b show the optimal truck purchase plan and cumulative mileage trajectories, respectively, for an instance with $T = 45$ years and the same constant demand $D_\tau = 2.45\text{E}6$ miles, $\forall \tau \in \{1, \dots, T\}$. Five equal-sized cohorts, each containing 25 type-II trucks, are purchased at year 1, 9, 17, 26, and 36. Note that the service lives of the three early cohorts are less than 10 years. It is more economical to shorten the lives of earlier cohorts since their salvage values are less discounted.



(a) truck purchase plan and fleet size

(b) trucks' cumulative mileage trajectories

Figure 1. Optimal truck management plan for $D_\tau = 2.45\text{E}6$ miles, $\tau \in \{1, \dots, 20\}$.



(a) truck purchase plan and fleet size

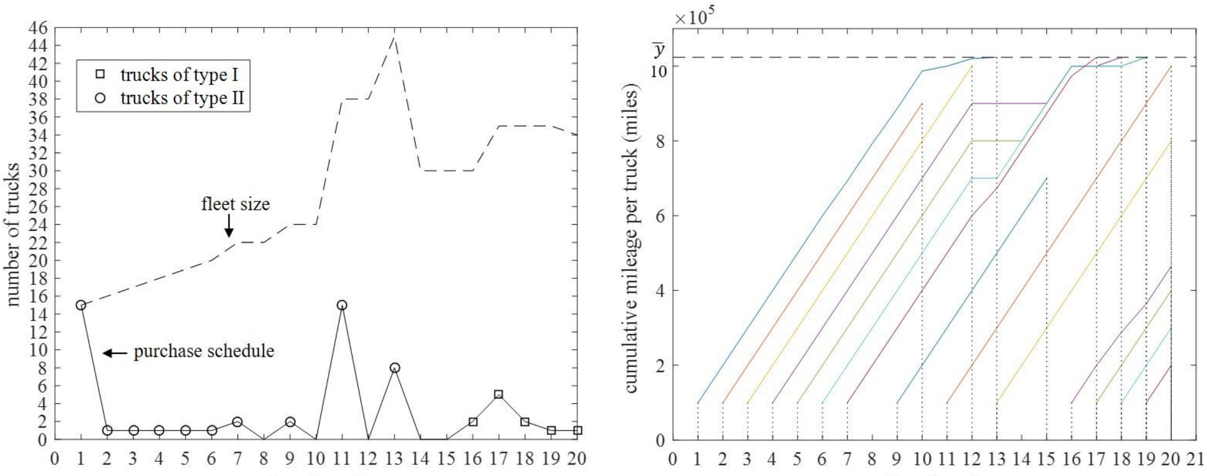
(b) trucks' cumulative mileage trajectories

Figure 2. Optimal truck management plan for $D_\tau = 2.45\text{E}6$ miles, $\tau \in \{1, \dots, 45\}$

5.2. Linearly increasing demand pattern

Now assume a linearly increasing demand as described by $D_\tau = (1.4 + 0.1\tau)\text{E}6$ miles ($1 \leq \tau \leq T$). The optimal truck purchase plan and cumulative mileage trajectories are plotted in Figures 3a and b, respectively. The figures unveil a number of findings regarding the optimal fleet management plan.

Note first that the truck purchase plan is no longer periodic under this time-varying demand. In fact, cohorts of different sizes and types are purchased in 15 of the 20 years. The largest two cohorts still appear in years 1 and 11, each consists of 15 type-II trucks. The other 13 cohorts are much smaller: they collectively consist of 28 trucks. This is intuitive: 15 trucks are needed to meet D_1, and they also serve the majority of demand in years 2-10; small cohorts of 1-2 trucks are purchased over those years to serve the demand increments. In year 11, cohort 1 is near $\bar{y}$ and thus replaced by cohort 11. Smaller cohorts are again added over the following years to serve the incremental demand. The fleet size curve in Figure 3a shows that although demand increases over time, the optimal fleet size is not always increasing. In addition, type-I trucks are purchased in the last 5 years, rendering a mixed fleet. This is because trucks purchased near the end of planning horizon will serve less mileage in their short service lives, and thus cheaper type-I trucks are preferred. Furthermore, some cohorts (i.e., cohorts 2, 9, 13, 16-20) are retired far before reaching their mileage limit to save the cost. This is again due to the time-varying demand. Finally, this solution violates the "older cluster replacement" property of Jones et al. (1991); see that cohort 1 is retired in year 13 while cohorts 2 and 3 are retired in years 10 and 12, respectively. This occurs mainly because cohorts are not equal-sized due to the time-varying demand, and thus the retirement decision is also affected by cohort sizes, in addition to each cohort's cumulative mileage (and age).



(a) truck purchase plan and fleet size (b) trucks' cumulative mileage trajectories

Figure 3. Optimal truck management plan for $D_\tau = (1.4 + 0.1\tau)\text{E}6$ miles, $\tau \in \{1, \ldots, 20\}$

We further examined more instances under linear demands with different annual increments. The demands are denoted by $D_\tau = (1.4 + \beta\tau)\text{E}6$ miles ($1 \leq \tau \leq T$) for $\beta \in [0.05, 0.3]$. Figure 4 shows how the optimal cost J (the solid line with circular markers) and the total number of trucks (the dashed line with diamond markers) vary with β. It unveils that the total cost increases linearly with β, and the total number of trucks increases faster than the cost. The latter is also expected: when the demand becomes more uneven, more trucks will retire before being fully utilized, and thus more trucks are needed to serve the demand.

5.3. A demand pattern with a drop in middle years

For the last numerical instance, a demand pattern as shown in Figure 5 is used. This demand pattern contains a sharp drop in year 5 (e.g., due to an economic recession or the appearance of a business competitor); the demand then stays low for years 5-9 and recovers gradually from year 10 on. We examine this instance to learn how the optimal fleet management plan, especially the purchase and retirement plan, varies in response to an expected demand drop. The optimal truck purchase plan, fleet size, and the cumulative mileage trajectories are plotted against time in Figures 6a and b, respectively. The figures show that totally 12 cohorts of trucks are used, with the largest cohorts being purchased in years 1 and 12. Compared to the previous instances, the solution of this instance features a "more mixed" fleet of different truck types. In particular, the earlier cohorts are of type-I, probably because they are expected to retire earlier due to the forecasted demand drop. The optimal fleet size stays roughly invariant over the demand "valley", since a later demand recovery is also expected. Cohorts 2 and 4 retire earlier than cohort 1, indicating again a violation of the "older cluster replacement" property. This is because cohort 1 is much larger and is better retained for serving the recovered demand after year 9.

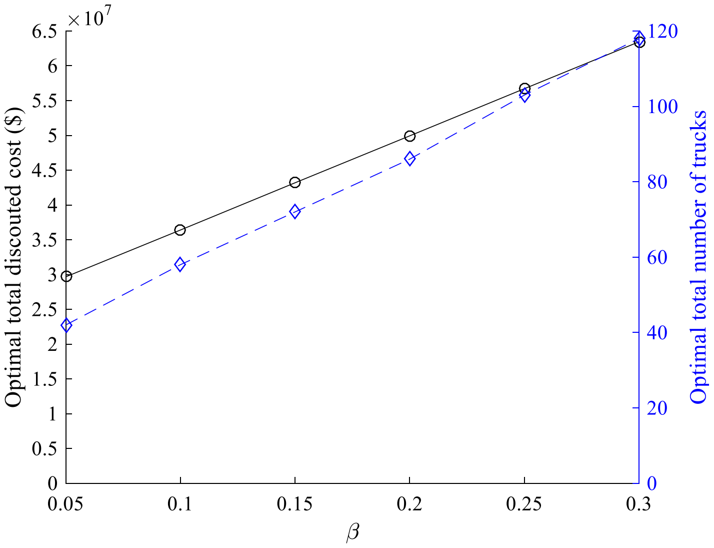


Figure 4. Optimal cost and total number of trucks versus β for $D_\tau = (1.4 + \beta\tau)\text{E6}$ miles, $\tau \in \{1, \ldots, 20\}$

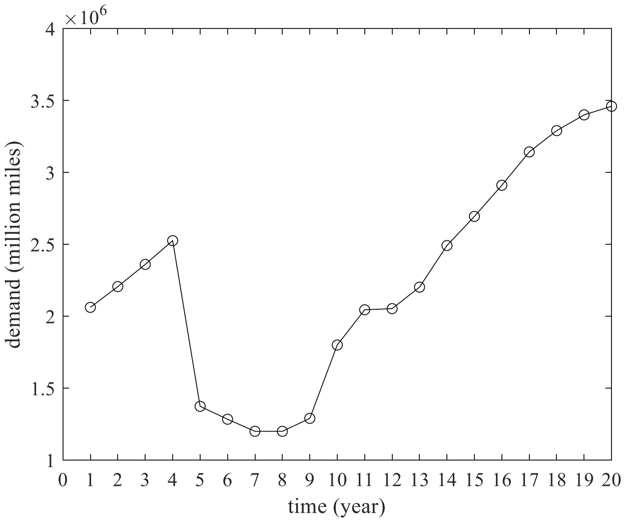
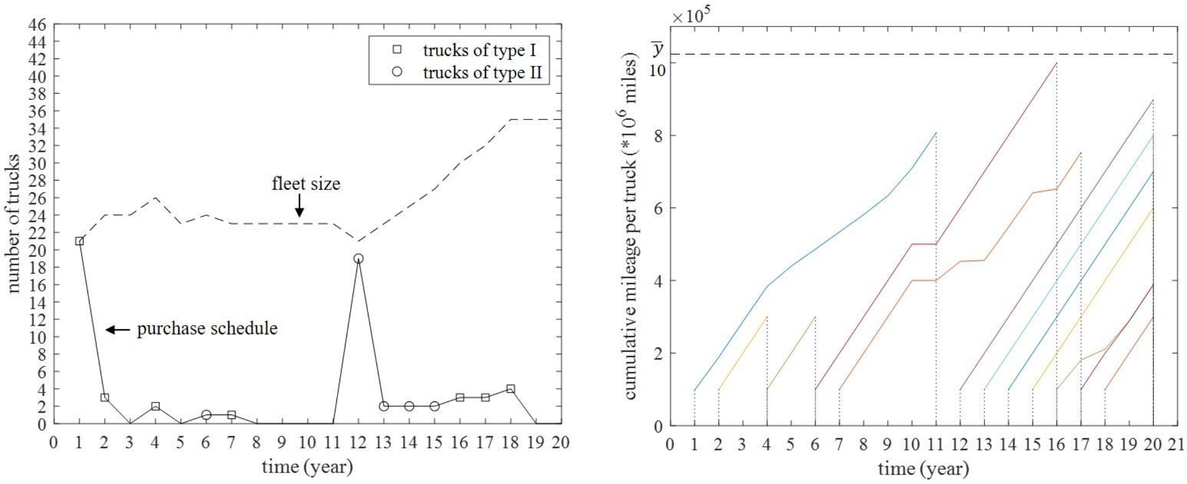


Figure 5. A demand pattern with a demand drop in middle years



(a) truck purchase plan and fleet size (b) trucks' cumulative mileage trajectories

Figure 6. Optimal truck management plan for the demand with a drop in middle years

6. Robustness of the optimal solutions

In real practice, many operating parameter values are subject to estimation errors and uncertainties. In addition, actual vehicle utilizations can also deviate from the optimal plan. This section shows that the optimal fleet management plan is robust to these errors and deviations.

In our first batch of robustness tests, we study how an "optimal" plan developed using inaccurate parameter estimates performs in the true environment. To this end, we first examine a scenario where the discount rate estimate contains an error. We assume the estimated discount rate is $r(1+\varepsilon)$, where r is the true value and ε is the relative estimation error. We use randomly generated demand patterns for $T = 20$ years, the convex Stage-1 formulation, and the parameter values given in Table 1. We evaluate: (i) the true total cost, $\hat{J}$, if the "optimal" plan developed by using the inaccurate estimate $r(1+\varepsilon)$ is implemented in the true environment; and (ii) the optimal total cost, J^*, for the optimal plan developed by using the true parameter r. We find that the difference between $\hat{J}$ and J^* (averaged across 10 numerical instances) is consistently below 0.2% for any given ε satisfying $|\varepsilon| \leq 15\%$. This indicates that the estimation error in discount rate would not significantly undermine the performance of our solution. Similar results were also found for other model parameters, including the O&M cost parameters and the salvage value function parameters.

In addition, we consider a scenario where future demand estimates are inaccurate, and the accurate demands are known when they are realized. (A similar scenario is where some vehicles' utilization trajectories unexpectedly deviate from an optimal plan, and the deviations are known when they occur.) Thus, we can re-optimize the fleet management plan when the accurate information is known. To see how this re-optimization approach performs, we examine a 20-year instance where the estimated demand in year 5 contains an error. This demand is represented by $D_5(1+\varepsilon)$, where D_5 is the true value and ε is the estimation error. In year 1, the fleet management plan is optimized using the estimated demand for years 1-20 (the same parameter values as the last batch of tests are used). Then in year 5, after knowing the true demand D_5, we re-optimize the plan for years 5-20.[5] Thus, the original plan was implemented in years 1-4 and the updated one in the remaining years. The total cost

---

[5] The re-optimization problem involves an initial fleet consisting of cohorts that were purchased (and not retired) by year 4. Although [P1] did not consider any initial fleet, it can be easily modified to model one. Our solution approach, including the demand allocation rule and the tabu search algorithm can be readily applied.

is calculated and compared against the optimal cost developed by assuming that the accurate demand D_5 was known in the beginning of planning horizon. We find for $|\varepsilon| \leq 30\%$ that the error between the two cost values never exceeded 1.5%.

The above results revealed that moderately inaccurate parameter values would not undermine the quality of our solution. They verified the practicality of our model and solution approach.

7. Conclusions

A two-stage approach is proposed for solving the discrete-time fleet management problem under time-varying demand. By exploiting a set of near-optimal conditions developed from a continuous-time approximation of the original formulation, the number of decision variables is reduced from $\frac{T(T+7)}{2}$ to $3T + n$, where n is the number of breakpoints and is small in most cases (see Section 3.1.3). Numerical experiments showed that our approach outperformed existing solution approaches in terms of solution quality or computational efficiency, and oftentimes both, by significant margins. The advantage is greater for problems with a non-convex Stage-1 formulation, and for problems of larger sizes. The results manifested that our approach is an important improvement over the existing ones despite its heuristic nature, since exact solutions to the fleet management problem are unavailable for large-scale instances.

Thanks to the above advantages, the proposed approach can be used to solve larger-scale problems with longer planning horizons or more vehicle types, and problems with a finer decision-making time scale (e.g., a month or a week instead of a year). The generality of our problem formulation also allows it to be applied to the management of various fleet types, including coach buses and aircrafts.

Our work also demonstrated the potential of using continuous-time approximation for efficiently solving asset management problems with large numbers of variables. The key insight unveiled by this method, i.e., that the marginal utilization costs of distinct assets at a given time tend to be equal, is consistent with economic intuition. This insight and the resulting demand allocation rule (see again Section 3.1.3) can be potentially extended to solve more realistic problems such as: (i) problems with indivisible demands, e.g., containerized cargo with multiple origins and destinations; and (ii) problems with stochastic demand and operating conditions[6]. Works in the above directions are under investigation now.

Our numerical results also show that the widely-cited “older cluster replacement” does not hold in an optimal fleet management plan. This, however, could possibly be a consequence of our assumption of the “no-splitting” property, meaning that the two seemingly intuitive properties cannot both hold at the optimality. In the future work we also plan to explore more realistic scenarios where the “no-splitting” assumption is relaxed, i.e., where vehicles in the same cohort can have different utilizations and retirement times.

Acknowledgements

This study is supported by General Research Funds (Project No. 15280116 and 15224818) provided by the Research Grants Council of Hong Kong and a start-up grant provided by the Hong Kong Polytechnic University (Project ID: P0001008).

---

[6] See, e.g., List et al. (2003), Hartman (2004), Childress and Durango-Cohen (2005), Stasko and Gao (2012), and Zheng and Chen (2016) for studies that assumed stochastic demand or operating conditions.

Appendix A. List of notations

Table A1. List of notations

Notation	Description	Notation	Description
Decision variables			
P_t	Number of vehicles purchased at time t	$P(t)$	Continuous-time form of P_t
γ_t	Type of vehicles in cohort t	$\gamma(t)$	Continuous-time form of γ_t
$u_{\tau,t}$	Mileage served at τ by a cohort-t vehicle	$u(\tau,t)$	Continuous-time form of $u_{\tau,t}$
$y_{\tau,t}$	Cumulative mileage at τ of a cohort-t vehicle	$y(\tau,t)$	Continuous-time form of $y_{\tau,t}$
S_t	Time when cohort-t vehicles are retired	$S(t)$	Continuous-time form of S_t
$P_{t,\gamma}$	Number of type-γ vehicles purchased at t	$Q_{t,\gamma}$	Equals 1 if type-γ vehicles are purchased at time t, and 0 otherwise
$X(u_l)_{y,w,t,\gamma}$	Number of type-γ vehicles in use at time t with utilization u_l, age w, and cumulative utilization y	$Z(u_l)_{y,w,t,\gamma}$	Equals 1 if type-γ vehicles with age w and cumulative utilization y are used at level u_l at time t, and 0 otherwise
$S_{y,w,t,\gamma}$	Number of type-γ vehicles retired at time t, with age w and cumulative utilization y	$W_{y,w,t,\gamma}$	Equals 1 if type-γ vehicles with age w and cumulative utilization y are retired at time t, and 0 otherwise
Parameters and other variables			
D_τ	Demand at τ	$D(\tau)$	Continuous-time form of D_τ
U	Maximum mileage per vehicle in a unit time	H	Set of vehicle types
$\bar{y}$	Maximum allowable cumulative mileage	$A(\cdot)$	Unit purchase cost per vehicle
$M(\cdot)$	Unit operating and maintenance cost per mile	$F(\cdot)$	Salvage value of a vehicle
T	Planning horizon	r	Discount rate
$\lambda(\tau)$	Lagrange multiplier for relaxing constraint (4b)	$z(\cdot)$	z-score
λ_τ	Lagrange multiplier for relaxing constraint (1b)	$z_{\tau,t}(\cdot)$	Discrete-time form of $z(\cdot)$
$F_{y,w,t,\gamma}$	Salvage value of a cohort-t vehicle of type γ, age w with cumulative mileage y	$g^u_{\tau,t}(\cdot)$	Optimal solution for $u_{\tau,t}$ under a given vehicle purchase and retirement plan
$M(u_l)_{y,w,t,\gamma}$	O&M cost of a cohort-t vehicle of type γ, age w with cumulative utilization y and current utilization level u_l	$\bar{M}$	A sufficiently large number

Appendix B. Proof of Proposition 1

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 the constraints (4b)-(4e) of [P4], respectively (where $\varphi_1(\tau,t)$ and $\varphi_2(\tau,t)$ are used to relax the right and left inequalities of (4d), respectively). The Lagrange function is presented as:

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

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

$$\int_{\tau=0}^{T} \lambda(\tau)D(\tau)e^{-r\tau}\,d\tau - \int_{t=0}^{T}\int_{\tau=t}^{S(t)} \lambda(\tau)P(t)u(\tau,t)e^{-r\tau}d\tau dt + \int_{t=0}^{T}\int_{\tau=t}^{S(t)} \left(\mu(\tau,t)u(\tau,t) + y(\tau,t)\left(\frac{\partial\mu(\tau,t)}{\partial\tau} - r\mu(\tau,t)\right)\right)e^{-r\tau}d\tau dt - \int_{t=0}^{T}\mu(S(t),t)y(S(t),t)e^{-rS(t)}dt + \int_{t=0}^{T}\int_{\tau=t}^{S(t)} \varphi_1(\tau,t)\,(u(\tau,t)-U)e^{-r\tau}d\tau dt - \int_{t=0}^{T}\int_{\tau=t}^{S(t)} \varphi_2(\tau,t)\,u(\tau,t)e^{-r\tau}d\tau dt + \int_{t=0}^{T}\omega(t)(y(S(t),t)-\bar{y})\,e^{-rS(t)}dt \tag{B1}$$

For the second equality above,

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

results from integration by parts.

Take the partial derivatives of (B1) with respect to $u(\tau,t)$ and $y(\tau,t)$, part of the first-order conditions for optimality are:

(i) Stationarity: $\frac{\partial L}{\partial u(\tau,t)} = 0,\ \frac{\partial L}{\partial y(\tau,t)} = 0$ (B2)

(ii) Complementary slackness:

$\varphi_1(\tau,t)(u(\tau,t)-U) = 0,$ for $t \in (0,T], \tau \in [t,S(t)]$ (B3a)

$\varphi_2(\tau,t)u(\tau,t) = 0,$ for $t \in (0,T], \tau \in [t,S(t)]$ (B3b)

$\omega(t)(y(S(t),t)-\bar{y}) = 0,$ for $t \in (0,T]$ (B3c)

and (iii) Dual feasibility:

$\varphi_1(\tau,t), \varphi_2(\tau,t), \omega(t) \geq 0,$ for $t \in (0,T], \tau \in [t,S(t)]$ (B4)

Note that not all the first-order conditions are presented here because some of them will not be used in the following derivation. Nevertheless, (B2)-(B4) are still necessary conditions of the optimality.

Equations (B2) lead to the following (B5a-c):

$$P(t)M\big(y(\tau,t),\gamma(t)\big) - P(t)\lambda(\tau) + \mu(\tau,t) + \varphi_1(\tau,t) - \varphi_2(\tau,t) = 0 \tag{B5a}$$

$$P(t)u(\tau,t)\frac{\partial M}{\partial y(\tau,t)} + \frac{\partial\mu(\tau,t)}{\partial\tau} - r\mu(\tau,t) = 0 \text{ for } \tau < S(t) \tag{B5b}$$

$$P(t)\frac{\partial F}{\partial y(S(t),t)} + \mu(S(t),t) - \omega(t) = 0 \tag{B5c}$$

Take the partial derivative of both sides of (B5a) with respect to τ:

$$P(t)\frac{\partial M}{\partial y(\tau,t)}\frac{\partial y(\tau,t)}{\partial\tau} - P(t)\frac{d\lambda(\tau)}{d\tau} + \frac{\partial\mu(\tau,t)}{\partial\tau} + \frac{\partial\varphi_1(\tau,t)}{\partial\tau} - \frac{\partial\varphi_2(\tau,t)}{\partial\tau}$$
$$= P(t)u(\tau,t)\frac{\partial M}{\partial y(\tau,t)} - P(t)\frac{d\lambda(\tau)}{d\tau} + \frac{\partial\mu(\tau,t)}{\partial\tau} + \frac{\partial\varphi_1(\tau,t)}{\partial\tau} - \frac{\partial\varphi_2(\tau,t)}{\partial\tau} = 0 \tag{B6}$$

Subtract (B5b) from (B6):

$$\mu(\tau,t) = \frac{1}{r}\left(P(t)\frac{d\lambda(\tau)}{d\tau} - \frac{\partial\varphi_1(\tau,t)}{\partial\tau} + \frac{\partial\varphi_2(\tau,t)}{\partial\tau}\right) \tag{B7}$$

Then plug (B7) into (B5a):

$$P(t)M\big(y(\tau,t),\gamma(t)\big) = \lambda(\tau)P(t) - \frac{1}{r}\left(P(t)\frac{d\lambda(\tau)}{d\tau} - \frac{\partial\varphi_1(\tau,t)}{\partial\tau} + \frac{\partial\varphi_2(\tau,t)}{\partial\tau}\right) - \varphi_1(\tau,t) + \varphi_2(\tau,t) \tag{B8a}$$

On the other hand, subtract (B5c) from (B5a) for $\tau = S(t)$:

$$P(t)M\big(y(S(t),t),\gamma(t)\big) - P(t)\frac{\partial F}{\partial y(S(t),t)} = P(t)\lambda\big(S(t)\big) - \varphi_1(S(t),t) + \varphi_2(S(t),t) - \omega(t) \tag{B8b}$$

Equations (B8a) and (B8b) apply to the cases of $\tau < S(t)$ and $\tau = S(t)$, respectively. In the former case, by examining (B8a) and the values of $\varphi_1(\tau,t)$ and $\varphi_2(\tau,t)$ for any given τ and t, we find that one of the following three cases will arise:

(i) When $\varphi_1(\tau,t) = \varphi_2(\tau,t) = 0$, constraint (4d) is unbinding; i.e., $0 < u(\tau,t) < U$. Since $\varphi_1(\tau,t), \varphi_2(\tau,t) \geq 0$ (see (B4)), we have $\frac{\partial\varphi_1(\tau,t)}{\partial\tau} = \frac{\partial\varphi_2(\tau,t)}{\partial\tau} = 0$. (Note that this relies on an 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 other similar studies, e.g., Jin and Kite-Powell, 2000). Hence, (B8a) can be re-arranged as:

$$P(t) \cdot \left[M\big(y(\tau,t), \gamma(t)\big) - \left(\lambda(\tau) - \frac{1}{r}\frac{d\lambda(\tau)}{d\tau}\right)\right] = 0 \qquad \text{(B9)}$$

(ii) When $\varphi_1(\tau,t) = 0$ but $\varphi_2(\tau,t) \neq 0$, we have $u(\tau,t) = 0$.

(iii) Lastly, when $\varphi_1(\tau,t) \neq 0$ but $\varphi_2(\tau,t) = 0$, we have $u(\tau,t) = U$.

Note that at least one of $\varphi_1(\tau,t)$ and $\varphi_2(\tau,t)$ must be zero, because the left and right inequalities of (4d) cannot be binding simultaneously.

A similar reasoning applies to (B8b). Specifically, one of the following four cases will arise:

(i) When $\varphi_1(S(t),t) = \varphi_2(S(t),t) = \omega(t) = 0$, both constraints (4d) and (4e) are unbinding; i.e., $0 < u(\tau,t) < U$ and $y(S(t),t) < \bar{y}$. Then we have:

$$P(t) \cdot \left[M\big(y(S(t),t), \gamma(t)\big) - \frac{\partial F}{\partial y(S(t),t)} - \lambda\big(S(t)\big)\right] = 0 \qquad \text{(B10)}$$

(ii) When $\omega(t) \neq 0$, we have $y(S(t),t) = \bar{y}$.

(iii) When $\omega(t) = \varphi_1(S(t),t) = 0$ but $\varphi_2(S(t),t) \neq 0$, we have $u(S(t),t) = 0$.

(iv) Lastly, when $\omega(t) = \varphi_2(S(t),t) = 0$ but $\varphi_1(S(t),t) \neq 0$, $u(S(t),t) = U$.

By rearranging the above results, we have Proposition 1. ■

Appendix C. Solution algorithms

C.1 The solution algorithm for solving [P2]

Algorithm 1: Finding optimal $u_{\tau,t}$ for $1 \leq t \leq \tau \leq S_t$, given P_t, γ_t, S_t $(1 \leq t \leq T)$, and λ_i's at all breakpoints $i \in \{1, 2, \ldots, T\}$

```
Initialize $u_{\tau,t} = y_{\tau,t} = 0$ for $1 \leq t \leq \tau \leq S_t$.
Find the first cohort, $\tilde{t}$, whose service life is longer than one time unit. Since $\tilde{t}$ is a breakpoint, $\lambda_{\tilde{t}}$ is given by the condition of the algorithm.
For all $\tau \in \{1, \cdots, \tilde{t}\}$: if $P_\tau > 0$, set $u_{\tau,\tau} = \frac{D_\tau}{P_\tau}$, $y_{\tau,\tau} = u_{\tau,\tau}$.
Set $\lambda_{\tilde{t}+1} = (1+r)\lambda_{\tilde{t}} - r z_{\tilde{t},\tilde{t}}(y_{\tilde{t},\tilde{t}})$, where $z_{\tilde{t},\tilde{t}}(y_{\tilde{t},\tilde{t}})$ is calculated by (8c).
For $\tau = \tilde{t}+1, \ldots, T$:
    For each retiring cohort $t$ at $\tau$:
        Set $y_{\tau,t} = \min\{z_{\tau,t}^{-1}(\lambda_\tau), y_{\tau-1,t} + U, \bar{y}\}$, where $z_{\tau,t}^{-1}(\lambda_\tau)$ is the inverse function of (9c);
        and $u_{\tau,t} = y_{\tau,t} - y_{\tau-1,t}$.
    End For
    If there exists at least one non-retiring cohort at $\tau$ and $D_\tau - \sum_{t:t\leq\tau\leq S_t} P_t u_{\tau,t} > 0$:
        Do:
            Allocate the remaining demand to the non-retiring cohort $t$ with the lowest z-score unless $u_{\tau,t}$ reaches $U$; keep $z_{\tau,t}$ and $u_{\tau,t}$ updated.
        Until $D_\tau - \sum_{t:t\leq\tau\leq S_t} P_t u_{\tau,t} = 0$ (i.e., all the demand has been allocated)
```

```
        Set $\lambda_{\tau+1} = (1+r)\lambda_\tau - r \cdot \begin{bmatrix}\text{maximum z-score among all the non-}\\ \text{retiring cohorts receiving demand at }\tau\end{bmatrix}$.
    Else:
        $\tau + 1$ is a breakpoint, and thus $\lambda_{\tau+1}$ is given by the condition of the algorithm.
    End If
End For
```

Output $u_{\tau,t}$ for $1 \le t \le \tau \le S_t$ and J' calculated using (2).

C.2 The derivative-free approximate gradient algorithm for optimizing λ_i's at breakpoints

The following pseudo code optimizes λ_1 only, assuming that it is the only breakpoint. If there are more breakpoints, they will be optimized with embedded iteration loops.

Algorithm 2: Finding optimal λ_1, given P_t, γ_t, and S_t $(1 \le t \le T)$

```
Randomly initialize $\lambda_1^{(0)}$ and $\lambda_1^{(1)}$ using a predefined range $\Omega$; calculate the optimal total cost of [P2]
using **Algorithm 1**, i.e., $J'(\lambda_1^{(0)})$ and $J'(\lambda_1^{(1)})$.
Define $\lambda_1^*$ as the value of $\lambda_1$ that attains the lowest $J'$ so far.
Do:
    Let $\lambda_1^{(k)} = \lambda_1^{(k-1)} - \alpha_{k-1} \frac{J'\left(\lambda_1^{(k-1)}\right) - J'\left(\lambda_1^{(k-2)}\right)}{\lambda_1^{(k-1)} - \lambda_1^{(k-2)}}$, where $\alpha_{k-1}$ is a positive step size.
    Calculate $J'\left(\lambda_1^{(k)}\right)$ and update $\lambda_1^*$.
    Set $k \leftarrow k + 1$.
Until $\lambda_1^*$ has not been changed for $max\_num1$ steps
```

Output λ_1^* and $J'(\lambda_1^*)$.

In our numerical case studies presented in Sections 4 and 5, we set $\Omega = [5,10]$, $max\_num1 = 10$, and $\alpha_k = 2 \times 10^{-6}, \forall k$.

C.3 The greedy heuristic algorithm for developing an initial solution to [P3]

At each present time i (i progresses from 1 to T), the greedy heuristic determines P_i and γ_i as follows:

(i) For the present time i and all the future times, purchase the minimum number of vehicles required to meet the demand, assuming that all these vehicles retire at T and have the same type $\gamma \in H$; and find the γ that minimizes the cost.

(ii) Examine if retiring an existing cohort at the present time i will reduce the cost.

The algorithm is detailed as follows.

Algorithm 3: Finding an initial [P3] solution x^0

```
For $i = 1, \cdots, T$:                //$i$ represents the present time
    For $j = 0, \cdots, i-1$:          //$j$ is used to examine if retiring an existing cohort before the present
                                       time $i$ can reduce cost
        //Examine the case where cohort $j$ retires right before the present time.
        If $j \ge 1, P_j > 0$ and $S_j > i - 1$: set $\tilde{S}_j = S_j$, $S_j = i - 1$.
        For each $\gamma \in H$:
            For all the future times $\tau = i, \cdots, T$:
```

```
                Set $P_\tau$ to the minimum number of vehicles required to satisfy the demand constraint; set $S_\tau = T$, $\gamma_\tau = \gamma$.
                Continuously allocate $D_\tau$ to the vehicles with the lowest z-score, while satisfying boundary constraints (1g-h).
            End For
            Calculate cost $J$ using (3a); record the lowest-cost solution so far as $\{P_t, \gamma_t, S_t: t = 1,2,\ldots,T\}$.
        End For
        If $S_j = i - 1$: set $S_j = \tilde{S}_j$.        //Revert $S_j$.
    End For
End For
Output $\boldsymbol{x^0} = \{\boldsymbol{P_t}, \boldsymbol{\gamma_t}, \boldsymbol{S_t}: \boldsymbol{t} = \mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{T}\}$.
```

C.4 The tabu search algorithm for solving [P3]

Algorithm 4: Finding a heuristic solution $\boldsymbol{x}^* \equiv \{\boldsymbol{P}_t^*, \boldsymbol{\gamma}_t^*, \boldsymbol{S}_t^*: \boldsymbol{t} = \mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{T}\}$

```
Initialize $\boldsymbol{x} = \boldsymbol{x}^0$ using Algorithm 3, $TL = \emptyset$, and $\boldsymbol{x}^* = \boldsymbol{x}$;
Do:
    Find the best move in $\mathcal{N}(\boldsymbol{x}) \cap TL$ that yields the lowest cost $J$; denote the solution as $\widetilde{\boldsymbol{x}}$.
    If $J(\widetilde{\boldsymbol{x}}) < J(\boldsymbol{x}^*)$:
        Set $\boldsymbol{x}^* = \boldsymbol{x} = \widetilde{\boldsymbol{x}}$;
        Update $TL$.
    Else:
        Find the best move in $\mathcal{N}(\boldsymbol{x}) \backslash TL$ that yields the lowest $J$; denote the solution as $\widetilde{\boldsymbol{x}}$.
        Set $\boldsymbol{x} = \widetilde{\boldsymbol{x}}$;
        Update $TL$.
    End If
Until $\boldsymbol{x}^*$ has not been changed for $max\_num\_tb$ steps
Output $\boldsymbol{x}^*$.
```

Appendix D. Formulation using Hartman's linear modeling approach

To convert [P1] to a linear program following Hartman's approach (1999), we discretize the demand and utilization values using an interval $\mathbb{u} > 0$. Thus, vehicle utilization levels in a period can only take values from a finite set, i.e., $u_l \in \{0, \mathbb{u}, 2\mathbb{u}, \ldots, u_{max}\}$. Decision variables of the linearized problem are defined as follows:

$P_{t,\gamma}$: number of type-γ vehicles purchased at time $t, 1 \le t \le T, \gamma \in H$;

$Q_{t,\gamma}$: binary variable that equals 1 if type-γ vehicles are purchased at time t, and 0 otherwise, $1 \le t \le T, \gamma \in H$;

$X(u_l)_{y,w,t,\gamma}$: number of type-γ vehicles in use at time t with utilization u_l, age w, and cumulative utilization y, $0 \le u_l \le u_{max}, 0 \le y \le \bar{y}, 1 \le w, t \le T$, $\gamma \in H$;

$Z(u_l)_{y,w,t,\gamma}$: binary variable that equals 1 if type-γ vehicles with age w and cumulative utilization y are used at level u_l at time t, and 0 otherwise, $0 \le u_l \le u_{max}, 0 \le y \le \bar{y}, 1 \le w, t \le T$;

$S_{y,w,t,\gamma}$: number of type-γ vehicles retired at time t, with age w and cumulative utilization y, $0 \le y \le \bar{y}, 1 \le w, t \le T$;

$W_{y,w,t,\gamma}$: binary variable that equals 1 if type-γ vehicles with age w and cumulative utilization y are retired at time t, and 0 otherwise, $0 < y \leq \bar{y}, 1 \leq w, t \leq T$.

[P1] is then reformulated as:

$$\min \sum_{t=1}^{T}\sum_{\gamma} A(\gamma)P_{t,\gamma}e^{-rt} + \sum_{t=1}^{T}\sum_{w=1}^{T}\sum_{y=0}^{\bar{y}}\sum_{u_l=0}^{u_{max}}\sum_{\gamma} M(u_l)_{y,w,t,\gamma}X(u_l)_{y,w,t,\gamma}e^{-rt} - \sum_{t=1}^{T}\sum_{w=1}^{T}\sum_{y=0}^{\bar{y}}\sum_{\gamma} F_{y,w,t,\gamma}S_{y,w,t,\gamma}e^{-rt} \tag{D1}$$

subject to:

$$\sum_{w=1}^{T}\sum_{y=0}^{\bar{y}}\sum_{u_l=0}^{u_{max}}\sum_{\gamma} X(u_l)_{y,w,t,\gamma}u_l \geq D_t \quad \forall\, 1 \leq t \leq T \tag{D2}$$

$$P_{t,\gamma} - \sum_{u_l=0}^{u_{max}} X(u_l)_{0,1,t,\gamma} = 0, \ \forall\, 1 \leq t \leq T,\ \gamma \in H \tag{D3}$$

$$P_{t,\gamma} \geq Q_{t,\gamma}, \ \forall\, 1 \leq t \leq T, \gamma \in H \tag{D4}$$

$$P_{t,\gamma} \leq \bar{M} Q_{t,\gamma}, \ \forall\, 1 \leq t \leq T, \gamma \in H \tag{D5}$$

$$\sum_{\gamma} Q_{t,\gamma} \leq 1, \ \forall\, 1 \leq t \leq T \tag{D6}$$

$$\sum_{u_l=0}^{u_{max}} X(u_l)_{y,w,1,\gamma} = 0, \ \forall 0 \leq y \leq \bar{y}, 2 \leq w \leq T, \gamma \in H \text{ and } \forall 0 < y \leq \bar{y}, 1 \leq w \leq T, \gamma \in H \tag{D7}$$

$$\sum_{u_l=0}^{u_{max}} X(u_l)_{y-u_l,w-1,t-1,\gamma} - S_{y,w-1,t-1,\gamma} - \sum_{u_l=0}^{u_{max}} X(u_l)_{y,w,t,\gamma} = 0, \ \forall 0 < y \leq \bar{y}, 2 \leq w \leq t \leq T, \gamma \in H \tag{D8}$$

$$\sum_{u_l=0}^{u_{max}} X(u_l)_{y-u_l,w,T,\gamma} - S_{y,w,T,\gamma} = 0, \ \forall 0 < y \leq \bar{y}, 1 \leq w \leq T,\ \gamma \in H \tag{D9}$$

$$\sum_{u_l=0}^{u_{max}} Z(u_l)_{y,w,t,\gamma} + W_{y,w-1,t-1,\gamma} \leq 1, \ \forall 0 < y \leq \bar{y}, 2 \leq w \leq t \leq T,\ \gamma \in H \tag{D10}$$

$$\sum_{u_l=0}^{u_{max}}\sum_{\gamma} Z(u_l)_{0,1,1,\gamma} = 1 \tag{D11}$$

$$X(u_l)_{y,w,t,\gamma} \geq Z(u_l)_{y,w,t,\gamma}, \ \forall 0 \leq u_l \leq u_{max}, 0 \leq y \leq \bar{y}, 1 \leq w \leq t \leq T,\ \gamma \in H \tag{D12}$$

$$X(u_l)_{y,w,t,\gamma} \leq \bar{M} Z(u_l)_{y,w,t,\gamma}, \ \forall 0 \leq u_l \leq u_{max}, 0 \leq y \leq \bar{y}, 1 \leq w \leq t \leq T,\ \gamma \in H \tag{D13}$$

$$S_{y,w,t,\gamma} \geq W_{y,w,t,\gamma}, \ \forall 0 < y \leq \bar{y}, 1 \leq w \leq t < T,\ \gamma \in H \tag{D14}$$

$$S_{y,w,t,\gamma} \leq \bar{M} W_{y,w,t,\gamma}, \ \forall 0 < y \leq \bar{y}, 1 \leq w \leq t < T,\ \gamma \in H \tag{D15}$$

$$Q_{t,\gamma} \in \{0,1\}, \ \forall 1 \leq t \leq T, \gamma \in H \tag{D16}$$

$$W_{y,w,t,\gamma} \in \{0,1\}, \ \forall 0 < y \leq \bar{y}, 1 \leq w \leq t < T,\ \gamma \in H \tag{D17}$$

$$Z(u_l)_{y,w,t,\gamma} \in \{0,1\}, \ \forall 0 \leq u_l \leq u_{max}, 0 \leq y \leq \bar{y}, 1 \leq w \leq t \leq T, \gamma \in H \tag{D18}$$

$$P_{t,\gamma} \in \mathbb{Z}, \ \forall\, 1 \leq t \leq T, \gamma \in H \tag{D19}$$

The objective function (D1) consists of the vehicle purchase cost, O&M cost, and salvage value, where $A(\gamma)$ is the purchase cost of a type-γ vehicle; $M(u_l)_{y,w,t,\gamma}$ the O&M cost of a type-γ vehicle with age w, cumulative utilization y, and present utilization level u_l; and $F_{y,w,t,\gamma}$ the salvage value of a type-γ vehicle with age w and cumulative mileage y. Constraint (D2) specifies that all the demand must be met. (D3-15) are flow conservation constraints, where $\bar{M}$ is a sufficiently large number. Constraints (D16-18) define $Q_{t,\gamma}$, $W_{y,w,t,\gamma}$ and $Z(u_l)_{y,w,t,\gamma}$ as binary variables. Constraint (D19) stipulates that $P_{t,\gamma}$ is integer-valued. The definitions of all the other parameters, including H and r, are given in Table 1. Discrete values of $M(u_l)_{y,w,t,\gamma}$ and $F_{y,w,t,\gamma}$ can be calculated using cost models presented in Section 4.1.

References

Abdi, A., Taghipour, S., 2018. An optimization model for fleet management with economic and environmental considerations, under a cap-and-trade market. Journal of Cleaner Production, 204, 130-143.

Büyüktahtakın, İ. E., Hartman, J. C., 2016. A mixed-integer programming approach to the parallel replacement problem under technological change. International journal of production research, 54(3), 680-695.

Büyüktahtakın, İ. E., Smith, J. C., Hartman, J. C., Luo, S., 2014. Parallel asset replacement problem under economies of scale with multiple challengers. The Engineering Economist, 59(4), 237-258.

Boyd, S., Vandenberghe, L., 2004. Convex optimization. Cambridge university press.

CARB, 2008. 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.

Chen, H., Gu, W., Cassidy, M.J., Daganzo, C.F., 2015. Optimal transit service atop ring-radial and grid street networks: a continuum approximation design method and comparisons. Transportation Research Part B: Methodological, 81, 755-774.

Chen, P.W., Nie, Y.M., 2018. Optimal design of demand adaptive paired-line hybrid transit: Case of radial route structure. Transportation Research Part E: Logistics and Transportation Review, 110, 71-89.

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.

Des-Bordes, E., Büyüktahtakın, İ. E., 2017. Optimizing capital investments under technological change and deterioration: A case study on MRI machine replacement. The Engineering Economist, 62(2), 105-131.

Glover, F., Laguna, M., 1998. Tabu search. Handbook of Combinatorial Optimization, Springer, Boston, MA, 2093-2229.

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.

Guerrero, S.E., 2014. Modeling fuel saving investments and fleet management in the trucking industry: the impact of shipment performance on GHG emissions. Transportation Research Part E: Logistics and Transportation Review, 68, 178-196.

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.

Hartman, J. C., Tan, C. H., 2014. Equipment replacement analysis: a literature review and directions for future research. The Engineering Economist, 59(2), 136-153.

Hopp, W.J., Jones, P.C., Zydiak, J.L., 1993. A further note on parallel machine replacement. Naval Research Logistics, 40(4), 575-579.

Islam, A., Lownes, N., 2019. When to go electric? A parallel bus fleet replacement study. Transportation Research Part D: Transport and Environment, 72, 299-311.

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.

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.

Laksuwong, E., Pannakkong, W., Parthanadee, P., Buddhakulsomsiri, J., 2014. A study of a single-period model for the parallel fleet replacement problem. In Applied Mechanics and Materials, 619, 364-370. Trans Tech Publications.

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.

Leung, L.C., Tanchoco, J.M.A., 1990. Multiple machine replacement analysis. Engineering Costs and Production Economics, 20(3), 265-275.

Li, L., Lo, H.K., Xiao, F., Cen, X., 2018. Mixed bus fleet management strategy for minimizing overall and emissions external costs. Transportation Research Part D: Transport and Environment, 60, 104-118.

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.

McClurg, T., Chand, S., 2002. A parallel machine replacement model. Naval Research Logistics, 49(3), 275-287.

Mei, Y., Gu, W., Cassidy, M.J., Fan, W., 2020. Planning a form of skip-stop transit service under heterogeneous demands. Under review.

Ngo, H.H., Shah, R., Mishra, S., 2018. Optimal asset management strategies for mixed transit fleet. Transportation Research Part A: Policy and Practice, 117, 103-116.

Osorio, C., Bierlaire, M., 2013. A simulation-based optimization framework for urban transportation problems. Operations Research, 61(6), 1333-1345.

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.

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.

Rios, L.M., Sahinidis, N.V., 2013. Derivative-free optimization: a review of algorithms and comparison of software implementations. Journal of Global Optimization, 56(3), 1247-1293.

Seif, J., Shields, B. A., Yu, A. J., 2019. Parallel machine replacement under horizon uncerainty. The Engineering Economist, 64(1), 1-23.

Shields, B. A., Seif, J., Yu, A. J., 2019. Parallel machine replacement with shipping decisions. International Journal of Production Economics, 218, 62-71.

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.

Tang, J., Tang, K., 1993. A note on parallel machine replacement. Naval Research Logistics, 40(4), 569-573.

Tsao, Y.C., Lu, J.C., 2012. A supply chain network design considering transportation cost discounts. Transportation Research Part E: Logistics and Transportation Review, 48(2), 401-414.

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.

Wei, K., Li, X., Lin, S., Yue, C., Li, S., 2014. A simulated annealing based heuristic for the multi-source single-path multi-commodity network flow problem. In Proceedings of the 11th International Conference on Service Systems and Service Management (ICSSSM), 1-6.

Wu, P., Hartman, J. C., Wilson, G. R., 2003. A demand-shifting feasibility algorithm for Benders decomposition. European Journal of Operational Research, 148(3), 570-583.

Wu, P., Hartman, J. C., Wilson, G. R., 2005. An integrated model and solution approach for fleet sizing with heterogeneous assets. Transportation Science, 39(1), 87-103.

Yatsenko, Y., Hritonenko, N., 2015. Two-cycle optimization in replacement models with non-exponential technological improvement. IMA Journal of Management Mathematics, 28(3), 359-372.

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.

Zheng, S., Chen, S., 2018. Fleet replacement decisions under demand and fuel price uncertainties. Transportation Research Part D: Transport and Environment, 60, 153-173.