A joint routing and speed optimization problem

Ricardo Fukasawa,* Qie He† Fernando Santos‡ Yongjia Song§

Abstract

Vehicle speed plays a significant role on fuel consumption and greenhouse gas emissions in freight transportation. We study a general joint routing and speed optimization problem with the goal of minimizing the total fuel, emission, and labor costs; the fuel consumption and emission rate is allowed to be an arbitrary strictly convex function of vehicle speed. We propose a novel set partitioning formulation for this problem, in which each column represents a combination of a route and a subset of customers on the route whose time-window constraints are active, in contrast to the set partitioning formulation of the vehicle routing problem in which each column represents a route. This formulation facilitates the development of an efficient labeling algorithm for the pricing problem in a branch-cut-and-price framework. We test our formulation and branch-cut-and-price algorithm on a variety of instances on United Kingdom cities and modified Solomon instances. Our algorithm is capable of solving all the 25-customer instances including the most different UK-A family and several 50-customer instances; only 10-customer UK-A instances are solved to optimality to date by other published research.

1 Introduction

With the growing public concern of sustainability and climate change, there is an urgent need in current transportation practice to improve energy efficiency and reduce greenhouse gas emissions. The energy consumption and emission of gasoline- and diesel-powered vehicles depend on many factors (Demir et al. 2011), among which the travel distance, carrying load, and vehicle speed are the significant ones that can be affected by operational decisions such as routing and speed control. In this paper, we propose a transportation model that jointly optimizes the routing and speed decisions in order to save fuel and reduce emissions.

While many recent vehicle routing problems (VRPs) have considered minimizing fuel consumption and emissions as the new objective (Demir et al. 2014b, Eglese and Bektaş 2014, Lin et al. 2014), few of them treat vehicle speed as a decision separate from routing. It has been shown that vehicle speed plays a significant role on fuel consumption and emissions, e.g., the fuel consumption rate of a vehicle driving at a very high or a very low speed can be twice as much of that of a vehicle at a medium speed (Barth et al. 2000). Thus when there is operational flexibility on speed control, such as in maritime and railway transportation, significant savings on fuel consumption and reduction on emissions can be expected by considering the routing and speed decisions together. For example, Bektaş and Laporte (2011) recently extend the classical vehicle routing problem with

---

*Department of Combinatorics and Optimization, Faculty of Mathematics, University of Waterloo. Email: rfukasawa@uwaterloo.ca.
†Department of Industrial and Systems Engineering, University of Minnesota. Email: qhe@umn.edu.
‡Department of Combinatorics and Optimization, Faculty of Mathematics, University of Waterloo. Email: fasantos@uwaterloo.ca.
§Department of Statistical Sciences and Operations Research, Virginia Commonwealth University. Email: ysong3@vcu.edu.
time windows (VRPTW) by incorporating speed decisions: the average speed over each arc of the network is cast as a decision variable in addition to the routing decision, and the goal is to minimize the total fuel, emission, and labor costs. In their model, the emission is assumed to be proportional to the fuel consumption, and the fuel consumption is a function of travel distance, carrying load, and average vehicle speed, based on an empirical model for one particular type of heavy-duty diesel trucks (Barth et al. 2004). They call the problem the pollution routing problem (PRP). Computational results based on instances generated from United Kingdom cities show that this model yields an up to 10% reduction in CO$_2$ emissions, compared to the traditional VRP model whose objective is to minimize total distance. In this paper, we propose the joint routing and speed optimization problem (JRSP) to improve energy efficiency and reduce emission. The JRSP extends the PRP in the following manner: we do not restrict the fuel consumption rate function to any particular form, as long as it is a strictly convex function of the vehicle speed. Many fuel consumption and emission models in the literature, whether developed based on regression models or physics rules, satisfy the strictly convexity assumption, including MEET (the methodology for calculating transport emissions and energy consumption) developed by the European Commission (Hickman et al. 1999), COPERT (the computer program to calculate emissions from road transportation) (Kouridis et al. 2010), and CMEM (the comprehensive model emission model, the model employed in the PRP) (Barth et al. 2000).

The combination of routing and speed decisions makes the JRSP significantly more difficult to solve than the VRPs. The most successful exact algorithm in practice for the VRPs is the branch-and-price (BP) algorithm (Barnhart et al. 1998, Desrochers et al. 1992, Lübbeke and Desrochers 2005), in which a set covering/partitioning formulation is solved by branch and bound, the linear programming relaxation at each node of the branch-and-bound tree is solved by column generation, and the problem of generating a column of the linear program is called pricing. Cutting planes can also be added at each node of the branch-and-bound tree to further strengthen the relaxation, leading to the branch-cut-and-price (BCP) algorithm. One of the most critical components that affect the computational efficiency of these algorithms is how the pricing problem is solved. The pricing problem of the VRP is typically formulated as an elementary shortest path problem with resource constraints (ESPPRC) (Irnich and Desaulniers 2005), known to be strongly NP-hard (Dror 1994), and solved by dynamic programming or heuristics. For the PRP and the JRSP, however, performing the BP algorithm using the ESPPRC becomes quite challenging. The major obstacle is that speed over each arc is no longer a resource that gets accumulated throughout the route. In fact it is a decision variable that affects both feasibility and optimality of the routes. In the general framework of the ESPPRC proposed in Irnich and Desaulniers (2005), if we consider time (instead of speed) as a resource, then the resource extension function will be neither separable nor linear, and the set of resource vectors becomes an infinite set. Typically, as seen in Irnich and Desaulniers (2005), at least one of these conditions needs to be met to make the ESPPRC more tractable. The only BP algorithm that we are aware of to solve models with both routing and speed decisions is given by Dabia et al. (2015). The authors study a variant of the PRP in which the vehicle speed must be kept as a constant for each selected route and the departure time at the depot is allowed to vary. They solve this variant with a BP algorithm based on the set-partitioning formulation for the VRPTW, and develop novel and complicated dominance rules for the pricing algorithm to manage the combinatorial explosion in the number of routes that need to be checked. However, the authors also point out the assumption that the vehicle speed is constant for each selected route is crucial for their algorithm; otherwise the pricing algorithm must keep track of every possible speed over every possible arc, leading to an enormous computational effort.

In this paper, we propose a new set partitioning formulation for the general JRSP. In contrast to Dabia et al. (2015), we do not assume that the vehicle speed is kept as a constant for each
selected route. In our proposed formulation, each column represents a combination of a route and a subset of customers on that route whose time-window constraints are active (i.e., service time occurs in one of the bounds of the time window). Notice that each column in a “typical” set partitioning formulation of the VRP just represents a route. We solve the set partitioning model in a BCP framework. The pricing problem—a joint shortest path and speed optimization problem—is solved by a new labeling algorithm. The novelty of the labeling algorithm is that, for every path extension, we generate three (instead of one) labels for each customer to be visited in the next step: two labels in which the last customer is served at the beginning or end of her time window and one label in which the vehicle speeds are kept constant on the last few arcs of the path. The rationale behind this is based on the following observation: during the dynamic generation of a route, if the service starting time at the last visited customer is known, the optimal speeds over arcs from the depot to that customer can be calculated and fixed without considering the rest of the customers on the route; otherwise the optimal speeds have to be constant over arcs between the last customer whose time-window constraint is active and the last visited customer, due to the strict convexity of the fuel consumption function. Note that the constant speed property is exactly the critical assumption needed for the BP algorithm in Dabia et al. (2015). With the new formulation, we do not need to keep track of every possible speed combination when generating a route, and are able to derive easy-to-check dominance rules that keep the pricing problem manageable. We test our formulation and algorithm on a variety of instances for the PRP. Our algorithm is capable of solving all the 25-customer instances and several 50-customer instances among the most difficult UK-A instances with provable optimality; only 10-customer instances were solved to optimality to date by other published research. The contributions of this paper can be summarized as follows.

- We propose a general joint routing and speed optimization model to improve fuel efficiency and reduce emissions. The model is able to accommodate any fuel consumption and emission model as long as it is a strictly convex function of the vehicle speed.
- We propose a new set partitioning formulation for the JRSP, where each column represents a combination of a route and a set of customers whose time-windows constraints are active over that route. This formulation enables a practical BCP algorithm with easy-to-check dominance rules for the pricing algorithm.
- We significantly advance the state-of-the-art in the exact solution of the PRP. Our BCP algorithm is capable of solving all the 25-customer and several 50-customer PRP instances in the most difficult UK-A instances with provable optimality. The largest size of UK-A instances solved to optimality reported up to date in the literature is 10.

The rest of the paper is organized as follows. Section 2 reviews models and algorithms related to the JRSP. Section 3 describes the PRP and introduces the new set partitioning formulation. Sections 4 and 5 elaborate the labeling algorithm for the pricing problem as well as the dominance rules used in the labeling algorithms. Extensive computational results are provided in Section 6. We conclude in Section 7 with some discussions on future research directions.

2 Literature review

2.1 Green vehicle routing

Green vehicle routing is a broad term referring to vehicle routing problems that explicitly take into account the negative impact on the environment of using vehicles, in particular the greenhouse
gas emissions. We focus our review on routing models aiming at saving fuel consumption and reducing emissions, and divide them into two main categories: models that do not explicitly consider speeds as decisions and models with speeds as decisions. Most research falls into the first category, extending the classical VRP to energy and emission oriented objectives. Kara et al. (2007) propose the energy minimizing capacitated VRP, in which the energy cost over each arc is proportional to the product of the arc length and vehicle gross weight; the problem is first formulated as an integer linear program with the standard flow-arc formulation in Kara et al. (2007) and later solved by a branch-cut-and-price algorithm in Fukasawa et al. (2014). Xiao et al. (2012) propose the fuel consumption VRP, in which the fuel consumption over an arc is modeled as a linear function of the vehicle carrying load over that arc, based on data from the Ministry of Land, Infrastructure, Transport and Tourism of Japan. They solve the model by simulated annealing. Figliozzi (2010) extends the time-dependent VRPTW to minimize total CO$_2$ emissions, models emissions over each arc as a function of the departure time at the starting vertex of the arc, and solves the model by a heuristic. To the best of our knowledge, the first vehicle routing model considering the speed decisions explicitly is Jabali et al. (2012). The authors assume that the whole planning horizon is divided into two periods: a peak period with fixed low vehicle speed and an off-peak period with high free-flow speed, and the goal is to find a set of routes and a uniform upper bound on the free-flow speed over all arcs to minimize the total fuel, emission, and labor costs. The model is solved by a tabu search heuristic. Another important class of routing models that explicitly consider speed decisions is the pollution routing problem, which we review later. Since this is an emerging area, our review is by no means exhaustive; we refer interested readers to several recent surveys (Demir et al. 2014b, Eglese and Bektaş 2014, Lin et al. 2014, Bektaş et al. 2016).

### 2.2 The speed optimization problem

Given a fixed route, the problem of finding the optimal speed over each arc to minimize the total fuel consumption while respecting time-window constraints is called the speed optimization problem (SOP). The SOP is first considered by Fagerholt et al. (2010) in the context of maritime transportation. When the fuel consumption function is convex in the average speed over each arc, the SOP is a convex minimization problem. Norstad et al. (2011) develop an iterative algorithm to solve the SOP. The algorithm runs in time quadratic in the number of vertices on the route, and its correctness is proven in Hvattum et al. (2013) assuming the fuel consumption function is convex and non-decreasing in the average speed. Given a fixed route, the PRP is reduced to the SOP with additional labor cost, which is solved in Demir et al. (2012) by modifying the algorithm in Norstad et al. (2011) and used as a subroutine in their adaptive large neighborhood search heuristic for the PRP. Kramer et al. (2014) further modify the algorithm in Norstad et al. (2011) to solve the SOP with labor cost and varying departure time at the depot. The correctness of these adapted algorithms in Demir et al. (2012), Kramer et al. (2014) is proved recently in Kramer et al. (2015).

### 2.3 The pollution routing problem

The pollution routing problem, proposed by Bektaş and Laporte (2011), can be seen as a combination of the VRPTW and the SOP. It aims to find a set of routes as well as speeds over each arc of the routes to minimize the total fuel, emission, and labor costs, while respecting constraints on time and vehicle capacities. The emission function used in the PRP is based on the comprehensive modal emission model developed by Barth et al. (2004) for one type of heavy-duty diesel trucks. The PRP generalizes the energy minimizing VRP (Kara et al. 2007) and the fuel consumption VRP (Xiao et al. 2012), and is equivalent to the classical VRPTW under some condition with fixed
speeds (Fukasawa et al. 2014). Several extensions of the PRP have been studied. Franceschetti et al. (2013) consider the impact of congestion on the vehicle speed and emissions in the PRP, using the time-dependence setting proposed in Jabali et al. (2012). Demir et al. (2014a) study the bi-objective PRP with the goals of minimizing the fuel and emission cost and travel duration simultaneously. Koç et al. (2014) show additional reduction on emissions can be achieved with heterogeneous vehicles in the PRP.

Most of the methods for the PRP and its extensions are approximation or heuristic methods. Under the assumption that the speed over each arc can take a set of discrete values, the PRP and time-dependent PRP can be formulated as mixed-integer linear programs (Bektas and Laporte 2011, Franceschetti et al. 2013). Demir et al. (2012) use an adaptive large neighborhood search heuristic to find promising routes in the PRP, and apply a modified iterative algorithm to solve the SOP for each route. Kramer et al. (2014) develop a multi-start iterative local search framework for the PRP with varying departure time; the framework integrates a set partitioning formulation for the PRP and the modified iterative algorithm for the SOP. The only exact methods for the PRP that we are aware of are the recent work by Dabia et al. (2015) and Fukasawa et al. (2015). Dabia et al. (2015) study a variant of the PRP in which the vehicle speed must be kept as a constant for each selected route and the departure time at the depot is allowed to vary. They solve this variant with a BP algorithm based on the set-partitioning formulation for the VRPTW, and develop novel and complicated dominance rules for the pricing algorithm to manage the combinatorial explosion in the number of routes that need to be checked. Fukasawa et al. (2015) consider the PRP with departure time at the depot as additional decision variables, and propose the first set of exact formulations for the PRP as mixed-integer convex programs, based on disjunctive programming techniques. They solve the mixed-integer convex programs by branch-and-cut algorithms. However, the instance size that can be solved to optimality reported in both papers is limited. In particular, the largest size that can be solved in both papers for the most difficult UK-A instances is 10.

2.4 The branch-and-price algorithm for the VRP

The BP and BCP algorithms have consistently been the most successful exact methods for a large variety of vehicle routing and crew scheduling problems in practice (Barnhart et al. 1998, Desrochers et al. 1992, Pecin et al. 2014), due to the modeling power of the set partitioning/covering formulation in handling complex costs and side constraints (Desaulniers et al. 1998), the tight relaxation bound compared to other formulations, and the dynamic generation of columns in solving the relaxation (Desaulniers et al. 2006, Lübbecke and Desrosiers 2005), among others. For the VRP, the pricing problem of these algorithms is usually formulated as an elementary shortest path problem with resource constraints. This problem is strongly NP-hard (DrOR 1994), and to solve it efficiently in practice to generate a column, problem-specific dominance rules are needed to avoid enumerating too many path extensions. In addition, the requirement of an elementary route is usually relaxed, allowing routes with cycles (Desrochers et al. 1992) or ng-path (Baldacci et al. 2011), to speed up the column generation procedure.

3 Problem description and formulation

3.1 Problem description

Throughout the paper, we work on the delivery version of the JRSP with asymmetric distance matrix and homogeneous vehicles. The pickup version can be handled in a similar way. Let $N = (V, A)$ be a complete directed graph and $V = \{0, 1, 2, \ldots, n\}$, where vertex 0 denotes the
depot and $V_0 := \{1, 2, \ldots, n\}$ denotes the set of customers. There are $K$ identical vehicles, each with capacity $Q$, available at the depot. Each customer $i \in V_0$ has a demand $q_i$, a service time window $[a_i, b_i]$, within which the customer can be served, and a service duration $\tau_i$. For each $(i, j) \in A$, $d_{ij}$ is the distance from vertex $i$ to vertex $j$, and $l_{ij}$ and $u_{ij}$ are the lower and upper speed limits over arc $(i, j)$, respectively.

**Definition 1.** A route is a walk $(i_0, i_1, \ldots, i_h, i_{h+1})$, where $i_0 = i_{h+1} = 0$, $i_j \in V_0$ for $j = 1, \ldots, h$. An elementary route is a route in which no customer is visited more than once. A $q$-route is a route over which the total demands do not exceed the vehicle capacity.

The goal of the JRSP is to select a set of elementary routes for all the vehicles and speeds over each arc of the elementary routes to minimize the total cost and respect the following constraints:

1. Each customer is served within its time window (the vehicle will wait if it arrives early);
2. The total demands on each route do not exceed the vehicle capacity $Q$;
3. The speed over each arc does not violate the speed limits over that arc, i.e., the speed lies within $[l_{ij}, u_{ij}]$.

The cost over a route consists of fuel cost and time-dependent cost. First, the fuel consumption rate (liter per meter) is modeled as a function of the traveling speed $v$ and carrying load $f$ of the vehicle; in particular, the fuel consumption rate

$$c(v, f) = G(v) + \lambda f + \eta$$

is separable in $v$ and $f$. The speed-dependent part $G(v)$ only depends on the vehicle speed and is the same for all arcs, and the load-dependent part $\lambda f + \eta$ has aggregated parameters $\lambda$ and $\eta$ that are allowed to vary over different arcs. Given a route $r = (i_0 = 0, i_1, \ldots, i_h, i_{h+1} = 0)$, let $v = (v_{i_k-1,i_k})_{k=1}^{h+1}$ be the speed vector over $r$, $f_{i_k-1,i_k}$ be the carrying load of the vehicle when it traverses arc $(i_{k-1}, i_k)$ for $k = 1, \ldots, h + 1$, and $t_r$ be the total travel time over $r$. Then the cost over route $r$ is defined as follows:

$$c_r(v) = \sum_{k=1}^{h+1} d_{i_{k-1},i_k} c_{i_{k-1},i_k} (v_{i_{k-1},i_k}, f_{i_{k-1},i_k}) + \underbrace{p t_r}_{\text{time-dependent cost}}$$
$$= \sum_{k=1}^{h+1} \left[ d_{i_{k-1},i_k} G(v_{i_{k-1},i_k}) + d_{i_{k-1},i_k} (\lambda f_{i_{k-1},i_k} + \eta_{i_{k-1},i_k}) \right] + pt_r,$$

where $p$ is a constant and the term $pt_r$ represents the cost associated with the total travel time. In the PRP, $p$ is the hourly salary of a driver, and $pt_r$ represents the labor cost over route $r$. The cost over a route defined in (2) is quite general, since it includes several VRP models in the literature as special cases.

1. In the VRPTW, $G(v) = p = 0$, $\lambda_{ij} = 0$ and $\eta_{ij} = 1$ for any $(i, j) \in A$.
2. In the energy minimization VRP, $G(v) = p = 0$, $\lambda_{ij} = 1$ and $\eta_{ij} = 0$ for any $(i, j) \in A$.
3. In the fuel consumption VRP, $G(v) = p = 0$, $\lambda_{ij} = \lambda$ and $\eta_{ij} = \eta$ for any $(i, j) \in A$.
4. In the PRP, $G(v) = \frac{\pi_1}{v} + \pi_2 v^2$ for some parameters $\pi_1$ and $\pi_2$, $\lambda_{ij} = \lambda$ and $\eta_{ij} = \mu$ for any $(i, j) \in A$. 

6
3.2 A new set partitioning formulation for the JRSP

In the set partitioning formulation for many vehicle routing models, a column typically corresponds to a candidate route that respects some operational constraints, and the coefficient in the objective function for that column is the cost of that candidate route. In order to apply this formulation to the JRSP, we first need to define the cost of a route, which depends not only on the route itself but also on the speed vector over that route. One choice would be to define the cost of a route as the minimum cost of the route over all feasible speed vectors over that route. But this formulation gives rise to a challenging pricing problem, since it is already nontrivial to compute the optimal speed vector over a given route, and will be more difficult to derive any dominance rule for the labeling algorithm.

To circumvent this difficulty of the pricing problem, we propose a novel set partitioning formulation, in which each column is defined as a combination of a route and a set of active customers along the route with their service start times. We first introduce the concept of an active customer.

**Definition 2.** Given a route with a speed vector on the route, a customer $i_j$ is active, if it is served at its earliest or latest available time, i.e., $a_{i_j}$ or $b_{i_j}$, and we call $j$ an active index on the route.

In our new set partitioning formulation, each column no longer represents a candidate route; instead, it represents a triple $(r, I, s)$, where $r = \{i_0 = 0, i_1, \ldots, i_h, i_{h+1} = 0\}$ is a candidate route, $I$ is the set of active indices on route $r$, and $s = (s_j)_{j \in I}$ is a vector of service start times at active customers with $s_j \in \{a_{i_j}, b_{i_j}\}$ for any $j \in I$. The cost $c_{r,I,s}$ of a column in the formulation is defined as the minimum cost of route $r$ over any speed vector that guarantees that customer $i_j$ is served at $s_j$ for all $j \in I$. In particular,

$$c_{r,I,s} = \min c_r(v)$$

s.t. \[ t_{i_l} \geq t_{i_{l-1}} + \tau_{i_{l-1}} + \frac{d_{i_{l-1},i_l}}{v_{i_{l-1},i_l}}, \quad l = 1, \ldots, h, \]

\[ t_r \geq t_{i_h} + \tau_{i_h} + \frac{d_{i_h,0}}{v_{i_h,0}}, \]

\[ a_{i_l} \leq t_{i_l} \leq b_{i_l}, \quad l = 1, \ldots, h, \]

\[ t_{i_j} = s_j, \quad j \in I, \]

where the calculation of $c_r(v)$ is given in (2). We set $c_{r,I,s}$ to be $+\infty$ if the cumulative demands on the route exceed the vehicle capacity or the optimization problem in (3) is infeasible. Let $R$ be the set of all the routes in the graph. Let $\Omega$ be the set of all such triples $(r, I, s)$, i.e., $\Omega = \{(r, I, s) | I \text{ is a subset of indices of route } r, r \in R\}$. Note that $\Omega$ is a finite set. Let $a_{ir}$ indicate the number of times customer $i$ appears on route $r$. Our set partitioning formulation of the JRSP is stated as follows.

$$\min \sum_{r \in \Omega} c_{r,I,s} z_{r,I,s}$$

s.t. \[ \sum_{(r,I,s) \in \Omega} a_{ir} z_{r,I,s} = 1, \quad i \in V_0, \]

\[ \sum_{(r,I,s) \in \Omega} z_{r,I,s} = K, \]

\[ z_{r,I,s} \in \mathbb{Z}_+, \quad (r, I, s) \in \Omega. \]
It is clear that (4) gives a valid formulation for the JRSP, since all possible sets of active customers are considered on any given route and only elementary routes give integer feasible solutions. This new formulation contains more columns than the traditional formulation, but its pricing problem is much easier to solve. This is due to the following two facts. First, computing \( c_{r,l,s} \) is more efficient than computing the minimum cost of a route \( r \) over all feasible speeds. Second, this enables simple dominance rules that we derive in Section 5, which speed up the labeling algorithm for the pricing problem. In addition, we note that we may replace routes by elementary routes, \( ng \)-routes, \( k \)-cycle free \( q \)-routes, or any subset of routes that includes elementary routes. All these formulations are valid.

Formulation (4) can be solved by a branch-and-bound algorithm, in which the linear programming relaxation at each node of the branch-and-bound tree is solved by column generation. At each iteration of the column generation, a restricted linear program is solved with a subset of columns in \( \Omega \). Let the dual variables corresponding to constraints (4b) and (4c) be \( \mu_i \) for \( i \in V_0 \) and \( \nu \), respectively. Then the pricing problem to generate a new column is formulated as follows.

\[
\min_{(r,l,s) \in \Omega} \bar{c}_{r,l,s} := c_{r,l,s} - \sum_{i \in V_0} \mu_i a_{ir} - \nu,
\]

(5)

We are able to solve this pricing problem efficiently for large-scale instances under certain assumptions. Before introducing the labeling algorithm to solve (5) in Section 4, we state some assumptions needed for the algorithm.

**Assumption 1.** Each customer is served at its earliest available time and leaves for the next vertex on the route as soon as service is finished.

**Assumption 2.** The speed-dependent part \( G(v) \) in the fuel consumption rate function (1) is strictly convex and differentiable. Furthermore, \( G(v) \) is non-decreasing within \([l_{ij}, u_{ij}]\).

**Assumption 3.** The speed limits are the same across all the arcs, i.e., \( u_{ij} = u \) and \( l_{ij} = l \) for each \((i, j) \in A\).

We note that none of these assumptions is needed for the validity of formulation (4). Assumption 1 can be made without loss of generality. If a customer is not served immediately when available, then the vehicle must wait at the customer. We can instead serve the customer first and wait, without changing the speed over each arc and the overall cost on that route. The same argument can be made if there is any wait after service ends before the vehicle leaves for the next customer. Under Assumption 1, the arrival and service start times at each customer on a route are uniquely determined by the speed vector on that route, so is the cost over that route. Both Assumptions 2 and 3 are needed for the pricing problem (5) to be practically tractable. Without the convexity assumption in Assumption 2, the speed optimization problem itself would be difficult to solve in general. The convexity and differentiability assumption is satisfied by most fuel consumption models we are aware of in the routing literature, such as the fuel and emission model in the PRP (Bektaş and Laporte 2011) and maritime transportation problem (Fagerholt et al. 2010), MEET (Hickman et al. 1999), COPERT (Kouridis et al. 2010), and the fuel flow model for aircraft cruise control (Aktürk et al. 2014). The assumption that \( G \) is nondecreasing within \([l_{ij}, u_{ij}]\) is without loss of generality. Any speed that lies within a nonincreasing interval of \( G \) will not be optimal, since the vehicle can always travel at the highest possible speed in that interval, incurring no more fuel cost and less time-dependent cost. Thus we can preprocess the interval \([l_{ij}, u_{ij}]\) to satisfy the nondecreasing assumption, without eliminating any optimal speed. Assumption 3 is more restrictive and is needed for our labeling algorithm in Section 4. We should mention though that this assumption is satisfied by all the PRP benchmark instances in the literature.
The computation of $c_{r,I,s}$ and the labeling algorithm are motivated by the following observation.

**Proposition 1.** Any optimal solution for the speed optimization problem over a given route satisfies the following property: for each non-active customer $i$, the optimal speed over the arc into $i$ is equal to the optimal speed over the arc out of $i$.

**Proof.** See the appendix.

**Remark 1.** Proposition 1 implies that the optimal speed should be the same over each arc between two consecutive active customers. On the other hand, there will be no waiting at any non-active customer due to Assumption 1. Thus given a route, two consecutive active customers on the route, and their service start times, finding out the optimal speed in between becomes a univariate convex optimization problem. Then computing the cost $c_{r,I,s}$ is reduced to $|I|+1$ univariate convex optimization problems, each of which can be solved dynamically and separately without knowing the whole route. This nice property motivates the new set partitioning formulation (4), and forms the foundation of our labeling algorithm in Section 4 for the pricing problem.

**Remark 2.** With Assumption 1, Proposition 1 states that a non-active customer $i$ in an optimal speed vector on a route satisfies the following two conditions:

1. The speed over the arc into $i$ is equal to the speed over the arc out of $i$;
2. There is no waiting at customer $i$.

The converse is not true though, i.e., a customer satisfying the above two conditions may happen to be served at the boundary of its time window under the optimal speeds. This important observation also affects the design of our labeling algorithm, which we illustrate below.

## 4 The labeling algorithm

The key ingredient to solve (4) is to find the triple $(r,I,s)$ that corresponds to the smallest reduced cost. This means that we need to find, not only the route, but also the actual speeds under which that route is traveled. We achieve this goal by dynamically generating the route $r$ as well as the set of active indices $I$ and the vector $s$ in a labeling algorithm. We first need the following crucial concept for our labeling algorithm.

**Definition 3.** Given a route with a speed vector on the route, a customer $j$ is *seamless*, if the following two conditions hold:

1. The speed over the arc into $j$ is equal to the speed over the arc out of $j$;
2. There is no waiting at customer $j$.

A seamless customer is just a customer satisfying the two conditions in Remark 2, the conditions satisfied by a non-active customer given any optimal speed vector. Given a route with an optimal speed vector, a customer on the route is either active or seamless (and it could be both). The reason we make this distinction between non-active customer and seamless customers is purely technical, since if a customer is required to be non-active, then the interval of feasible speeds could be open and the optimal solution of an optimization problem over an open set may not exist.

The general idea of our labeling algorithm is as follows. When we extend a walk to a new customer $j$, we deem $j$ as either active, so that its service start time is $a_j$ or $b_j$, or seamless, so
that the speeds into and out of \( j \) are the same and there is no waiting at \( j \). Let \( i_w \) be the last known active customer on the walk. The optimal speeds and cost on the walk before \( i_w \) have been computed and stored while the walk is constructed, and the speed over arcs between \( i_w \) and the last customer \( j \) on the walk is constant. Dabia et al. (2015) have successfully developed a BP algorithm for the PRP when speed is constant over the entire route. Motivated by that and with the new formulation, we propose an efficient labeling algorithm to solve the pricing problem (5). We formalize the idea below.

### 4.1 Label definition

We first define some notations related to a walk. Given a walk \( P = (i_0, i_1, \ldots, i_h) \), let \( q(P) = \sum_{k=1}^{h} q_{i_{k-1}i_k} \) be the cumulative load on \( P \), \( \tau_{i_l,i_m}(P) = \sum_{k=l}^{m} \tau_{i_k} \) be the total service time spent on customers between \( i_l \) and \( i_m \) on \( P \), \( D_{i_l,i_m}(P) = \sum_{k=l}^{m} d_{i_{k-1}i_k} \), and \( D_{i_l,i_m}(P) = \sum_{k=l+1}^{m} \lambda_{i_{k-1}i_k} d_{i_{k-1}i_k} \), for \( 0 \leq l \leq m \leq h \). We write \( \tau_{i_l,i_m}, D_{i_l,i_m}, \text{ and } D_{i_l,i_m} \), instead of \( \tau_{i_l,i_m}(P), D_{i_l,i_m}(P), \text{ and } D_{i_l,i_m}(P) \) when the context is clear.

We develop a forward labeling algorithm to solve the pricing problem (5). A label \( L = (P, w, s) \) is associated with: (i) a walk \( P = (i_0, i_1, \ldots, i_h) \) with \( i_0 = 0 \) and \( q(P) \leq Q \); (ii) the index \( w \) of the last known active customer on \( P \); (iii) the service start time \( s \) of customer \( i_w \), which is either \( a_{i_w} \) or \( b_{i_w} \). Note that we only need to store the last known active index instead of all active indices, since the walk \( P \) is generated dynamically, and the optimal speed-dependent cost before \( i_w \) is stored and updated along the walk. A label \( L = (P, w, s) \) contains the following attributes:

- \( M \), the set of forbidden vertices, i.e., vertices that cannot be visited directly after vertex \( i_h \) along \( P \). For example, if we consider only elementary routes in our set-partitioning formulation, then \( M \) is the set of vertices in \( P \); if we consider q-routes, then \( M = \{i_h\} \).

- \( S \), a set of feasible speeds \( v \) such that the time-window constraint of each customer between \( i_w \) and vertex \( i_h \) along \( P \) is satisfied, by traveling at constant speed \( v \) between \( i_w \) and \( i_h \) without any waiting in between. In particular,

\[
S = \left\{ v \in [l, u] \mid s + \tau_{i_w,i_j} + \frac{D_{i_w,i_j}}{v} \in [a_{i_j}, b_{i_j}], \forall j = w + 1, \ldots, h \right\},
\]

- \( \Gamma \), the total service time spent on customers from \( i_w \) to \( i_h \) along \( P \), i.e., \( \Gamma = \tau_{i_w,i_h} \).

- \( D \), the total distance between vertex \( i_w \) and vertex \( i_h \) along \( P \), i.e., \( D = D_{i_w,i_h} \).

- \( F_{load} \), the load-dependent fuel cost along \( P \). In particular,

\[
F_{load} = \sum_{j=1}^{h} \bar{D}_{0,i_j}q_{i_j} + \sum_{j=1}^{h} d_{i_{j-1}i_j} \eta_{i_{j-1}i_j}.
\]

- \( F_{speed} \), the optimal speed-dependent fuel cost up to customer \( i_w \) along \( P \). This value is updated when the last known active customer along \( P \) is updated. The initialization and update of \( F_{speed} \) is explained in Section 4.2.

These attributes can be seen as resources as in the resource constrained shortest path problem. However, the proposed new definition of a label is a departure from the standard labeling algorithm, in which a label represents a walk. A new dominance rule will be developed in Section 5 based on this new definition of labels.
4.2 Label initialization and extension

From this point on, we will refer to labels and their attributes with superscripts whenever there is a need to differentiate between two labels. For example, given a label \( L^f = (P^f, w^f, s^f) \) with superscript \( f \), we will refer to its attributes as \( M^f, S^f \), and so on. Given two walks \( P \) and \( P' \) with \( P' \) starts at the ending vertex of \( P \), define \( P \oplus P' \) to be the walk obtained by concatenating \( P' \) to \( P \), that is, \( P \oplus P' = (i_1, \ldots, i_l) \) if \( P = (i_1, \ldots, i_k) \) and \( P' = (i_k, \ldots, i_l) \).

The initial label \( L^0 = (P^0, w^0, s^0) \) is created by setting \( P^0 = (0) \), \( w^0 = 0 \), and \( s^0 = 0 \). Table 1 illustrates the initial values of attributes associated with \( L^0 \). When a label \( L = (P, w, s) \) with \( P = (i_0, \ldots, i_h) \) is extended to a customer \( j \), we set \( j = i_{h+1} \) and create three new labels depending on whether \( j \) is active or seamless: \( L^1 = (P \oplus (i_h, j), h + 1, a_j) \) with customer \( j \) being active and served at \( a_j \), \( L^2 = (P \oplus (i_h, j), h + 1, b_j) \) with customer \( j \) being active and served at \( b_j \), and \( L^3 = (P \oplus (i_h, j), w, s) \) with customer \( j \) being seamless and the last known active customer remains to be \( i_w \). Table 1 illustrates how the attributes of the three labels are updated according to the attributes of \( L \). We discard the newly generated label, whenever the set \( S \) becomes empty or the optimization problem for computing attribute \( F_{\text{speed}} \) during the label extension is infeasible.

<table>
<thead>
<tr>
<th>( L^0 )</th>
<th>( L^1 )</th>
<th>( L^2 )</th>
<th>( L^3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M )</td>
<td>( \emptyset )</td>
<td>( M \cup {j} ) for elementary routes, ( {j} ) for q-routes</td>
<td></td>
</tr>
<tr>
<td>( S )</td>
<td>([l, u])</td>
<td>([l, u])</td>
<td>( S \cap {v \mid a_j \leq s + \tau_{i_w, i_h} + \frac{D_{i_w,j}}{v} \leq b_j} )</td>
</tr>
<tr>
<td>( \Gamma )</td>
<td>0</td>
<td>( \Gamma + \tau_j )</td>
<td></td>
</tr>
<tr>
<td>( D )</td>
<td>0</td>
<td>( D + d_{i_h,j} )</td>
<td></td>
</tr>
<tr>
<td>( F_{\text{load}} )</td>
<td>0</td>
<td>( F_{\text{load}} + D_{0,j} q_j + d_{i_h,j} h_{i_h,j} )</td>
<td></td>
</tr>
<tr>
<td>( F_{\text{speed}} )</td>
<td>0</td>
<td>( F_{\text{speed}} + F^1 )</td>
<td>( F_{\text{speed}} + F^2 )</td>
</tr>
</tbody>
</table>

Table 1: Label initialization and extension from a label \( L \).

The entry \( F^1 \) (or \( F^2 \)) in Table 1 denotes the optimal speed-dependent fuel costs when the vehicle leaves customer \( i_w \) at time \( s \), travels at some constant speed in \( S \) to customer \( j \) without any waiting in between, and serves customer \( j \) at time \( a_j \) (or \( b_j \)). The value of \( F^1 \) is calculated through solving a univariate convex optimization problem, as illustrated below.

\[
F^1 = \min D_{i_w,j} G(v) \quad \text{s.t. } v \in S, \quad \text{subject to} \quad s + \tau_{i_w, i_h} + \frac{D_{i_w,j}}{v} \leq a_j. \quad (6a)
\]

The calculation of \( F^2 \) is similar to that of \( F^1 \), as illustrated below.

\[
F^2 = \min D_{i_w,j} G(v) \quad \text{s.t. } v \in S, \quad \text{subject to} \quad s + \tau_{i_w, i_h} + \frac{D_{i_w,j}}{v} = b_j. \quad (7a)
\]

4.3 Label termination

Label termination refers to the extension of a label \( L \) to the depot, in which the vehicle leaves customer \( i_w \) at time \( s \) and travels at a constant speed \( v \in S \) along \( P \) back to the depot. The
total travel time along the resulting route \( r \) is
\[
t_r = s + \tau_{i_w,i_h} + D_{i_w,0}/v,
\]
so the time-dependent cost is \( pt_r \). Let the minimum sum of the speed-dependent cost between \( i_w \) and the depot and the time-dependent cost \( pt_r \) be \( F_{\text{speed},\text{time}} \). It can be computed through solving the univariate convex optimization below.

\[
F_{\text{speed},\text{time}} = \min \, D_{i_w,0} G(v) + p(s + \tau_{i_w,i_h} + \frac{D_{i_w,0}}{v})
\]
\[
\text{s.t. } v \in S,
\]
\[
s + \tau_{i_w,i_h} + \frac{D_{i_w,0}}{v} \leq b_0.
\]

The route \( r \) is discarded if (8) is infeasible. The total cost of the terminated label \( L \) is

\[
c_L := F_{\text{load}} + d_{i_h,0} \eta_{i_h,0} + F_{\text{speed}} + F_{\text{speed},\text{time}}.
\]

The correctness of the labeling algorithm is shown by Proposition 2 below.

**Proposition 2.** Suppose when a label \( L \) terminates, it recursively defines a route \( r \) and a set \( I \) of active indices over \( r \) with the corresponding service start times \( s \) at the active customers. Then the cost \( c_{r,I,s} \) is given by \( c_L \).

**Proof.** It is clear that the load-dependent fuel cost in \( c_L \) is equal to the load-dependent fuel cost in \( c_{r,I,s} \). Now consider the speed-dependent and time-dependent costs. According to Proposition 1, given a route and an optimal speed vector over the route, a customer is either active or seamless. The optimal speed at which the vehicle travels from active customer \( i \) to active customer \( j \), through a sequence of seamless customers, can be computed independent of the speed decisions before customer \( i \) and after customer \( j \). The way each label is extended and terminated guarantees the optimal speed over each arc is computed correctly, so we have \( c_{r,I,s} = c_L \). \( \square \)

## 5 Dominance rules

A crucial part of any labeling algorithm is the development of dominance rules, which allows one to discard a significant number of (potentially exponentially many) labels. Without such dominance rules, the implementation of most pricing algorithms would be impractical. We devote this section to the development of dominance rules that can be applied to the labels developed in Section 4. Intuitively, a label is dominated if any possible extension of the label cannot lead to a triple \((r, I, s)\) with a better reduced cost. We formalize below what we mean by “possible extension” and “lead to.”

We first introduce two new notations for label \( L^I = (P^I, w^I, s^I) \). Let \( i^I \) denote the last vertex of \( P^I \). Let \( v^I \) denote the speed vector over arcs from the depot to \( i^I \) along \( P^I \), obtained during the update of \( F_{\text{speed}} \). That is, the component of \( v^I \) is computed by recursively solving the optimization problem (6) or (7), during the generation of label \( L^I \).

**Definition 4.** Let \( P \) be a walk and \( v_P \) be a speed vector over arcs of \( P \). Given a label \( L^I \), the pair \((P, v_P)\) is a feasible extension of \( L^I \), if the following conditions hold:

- The walk \( P \) starts at vertex \( i^I \), \( P^I \oplus P \) is a route, and the total demands on \( P^I \oplus P \) do not exceed the vehicle capacity.
• There exists \( v \in S^I \) such that \((\bar{v}^I, v, v_P)\) is a feasible speed vector over the route \( P^I \oplus P \). That is, a vehicle can travel from the depot to vertex \( i_{w^I} \) with \( \bar{v}^I \), travel from \( i_{w^I} \) to \( i^I \) at constant speed \( v \), travel from \( i^I \) to the depot along \( P \) with \( v_P \), and the time-window constraints of all the customers on \( P^I \oplus P \) are satisfied.

Note that if \((P, v_P)\) is a feasible extension of the label \( L^I \), then the set of active customers on route \( P^I \oplus P \) with the speed vector \((\bar{v}^I, v, v_P)\) for some \( v \in S^I \) will be consistent with those in label \( L^I \).

Let \( E(L^I) \) denote the set of all possible feasible extensions of \( L^I \).

**Definition 5.** A label \( L^2 = (P^2, w^2, s^2) \) is a dominated label (so it can be discarded), if there exists a label \( L^1 = (P^1, w^1, s^1) \) such that the following conditions hold:

1. \( E(L^2) \subseteq E(L^1) \).

2. For any \((P, v_P)\) \( \in E(L^2) \) and \( v \in S^2 \) such that \((\bar{v}^2, v, v_P)\) is a feasible speed vector over \( P^2 \oplus P \), let \( T^2 \) be the set of active customers on \( P^2 \oplus P \) with this speed vector, and let \( s^2 \) be the vector of the corresponding service start times at these customers. There exists a set \( I^3 \) of active customers on \( P^1 \oplus P \) with the corresponding service start times \( s^3 \) such that \( \bar{c}_{P^1 \oplus P, P^2, s^3} < \bar{c}_{P^2 \oplus P, P^2, s^2} \).

Definition 5 does not state that \( L^2 \) is dominated by \( L^1 \), rather it states that by comparing \( L^1 \) and \( L^2 \), \( L^2 \) is dominated by some label \( L^3 \) with \( P^3 = P^1 \). Indeed the conditions in Definition 5 cannot guarantee that extending \( L^1 \) with speed vector \( \bar{v}^1 \) always leads to a column with a better reduced cost than any extension of \( L^2 \). This is another feature that is different from the usual domination conditions for regular labels that are defined just by routes.

The existence of such a label \( L^1 \) in Definition 5, however, is in general hard to assert unless we have already generated such a label \( L^1 \) and can check \( L^2 \) against it. We will replace the conditions on Definition 5 by a set of sufficient conditions that are easier to check. We first introduce two functions needed for the sufficient conditions. Given a label \( L \), let \( T : S \rightarrow \mathbb{R} \) be the function that computes the service finish time at the last vertex \( i_h \) when the vehicle travels at speed \( v \) between vertex \( i_w \) and vertex \( i_h \) along \( P \). Specifically,

\[
T(v) = s + \Gamma + \frac{D}{v}. \tag{10}
\]

Let \( C : S \rightarrow \mathbb{R} \) be the function that computes the total fuel cost up to vertex \( i_h \) along \( P \) with a constant speed \( v \) between vertex \( i_w \) and vertex \( i_h \). Specifically,

\[
C(v) = F_{\text{load}} + F_{\text{speed}} + D \cdot G(v). \tag{11}
\]

Let \( \mu(P) = \sum_{i=1}^{h} \mu_{i_{i-1,i}} + \sum_{j=1}^{h} \mu_{i_j} \) be the sum of all dual variables along \( P \), where \( \mu_{i,j} \) and \( \mu_i \) are the dual variables over arc \((i, j)\) and vertex \( i \), respectively.

We propose the following set of sufficient conditions:

**Proposition 3.** The label \( L^2 = (P^2, w^2, s^2) \) is a dominated label, if there exists a label \( L^1 = (P^1, w^1, s^1) \) such that the following conditions hold:

1. \( i^1 = i^2 \);

2. \( M^1 \subseteq M^2 \);

3. \( q(P^2) \leq q(P^1) \);
4. For any \( v_2 \in S^2 \), there exists \( v_1 \in S^1 \) such that

\[
T^1(v_1) \leq T^2(v_2)
\]  

(4-a) and

\[
C^1(v_1) - \mu(P^1) + (Q - q(P^2)) \max\{0, \bar{D}^1_{0,j} - \bar{D}^2_{0,j}\} < C^2(v_2) - \mu(P^2).
\]  

(4-b)

Proof. See the appendix.

We discuss in detail how to check the dominance rule in Proposition 3. Given two labels \( L^1 \) and \( L^2 \) that end at the same vertex, conditions 2–3 are easy to check. To check if condition 4 holds efficiently, we divide it into four cases. To simplify notation, let the last vertex in both labels be \( j \), the index of vertex \( j \) on \( P^1 \) be \( h^1 \), and the index of vertex \( j \) on \( P^2 \) be \( h^2 \).

5.1 Case 1: \( w^1 = h^1 \) and \( w^2 = h^2 \)

In this case, vertex \( j \) is the last active vertex on both \( P^1 \) and \( P^2 \). Therefore, \( T^1(v) = s^1 + \tau_j \), \( T^2(v) = s^2 + \tau_j \), \( C^1(v) = F^1_{\text{load}} + F^1_{\text{speed}} \), and \( C^2(v) = F^2_{\text{load}} + F^2_{\text{speed}} \). Condition 4 is satisfied if and only if \( s^1 \leq s^2 \) and \( F^1_{\text{load}} + F^1_{\text{speed}} < F^2_{\text{load}} + F^2_{\text{speed}} \).

5.2 Case 2: \( w^1 < h^1 \) and \( w^2 = h^2 \)

The service start time \( s^2 \) at \( j \) in label \( L^2 \) can only be \( a_j \) or \( b_j \), and \( C^2(v) = F^2_{\text{load}} + F^2_{\text{speed}} \). If \( s^2 = a_j \), consider condition (4-a), we must have \( T^1(v_1) = a_j + \tau_j \), which uniquely determines a value \( v_1 \). If \( c_1 \in S^1 \), we can check if (4-b) holds by comparing two numbers, otherwise we cannot satisfy (4-a).

If \( s^2 = b_j \), then any \( v_1 \in S^1 \) satisfies that \( T^1(v_1) \leq T^2(v_2) = b_j + \tau_j \). We then check if there exists \( v_1 \in S^1 \) such that (4-b) is satisfied, which is equivalent to check if the following inequality holds:

\[
\min_{v_1 \in S^1} C^1(v_1) - \mu(P^1) + (Q - q(P^2)) \max\{0, \bar{D}^1_{0,j} - \bar{D}^2_{0,j}\} < F^2_{\text{load}} + F^2_{\text{speed}} - \mu(P^2).
\]

Since \( C^1(v_1) \) is a univariate convex function and \( S^1 \) is a closed interval, \( \min_{v_1 \in S^1} C^1(v_1) \) can be computed efficiently.

5.3 Case 3: \( w^1 = h^1 \) and \( w^2 < h^2 \)

The service start time \( s^1 \) at \( j \) in label \( L^1 \) can only be \( a_j \) or \( b_j \), and \( C^1(v) = F^1_{\text{load}} + F^1_{\text{speed}} \). If \( s^1 = b_j \), then \( T^1(v_1) = b_j + \tau_j \geq T^2(v_2) \) for any \( v_2 \in S^2 \). Condition (4-a) holds only if \( T^1(v_1) = b_j + \tau_j = T^2(v_2) \) for all \( v_2 \in S^2 \). It implies that \( S^2 \) must contain only a single element. Thus it is easy to check if (4-b) holds.

If \( s^1 = a_j \), then \( T^1(v_1) = a_j + \tau_j \leq T^2(v_2) \) for any \( v_2 \in S^2 \). We then check if for all \( v_2 \in S^2 \) condition (4-b) is satisfied, which is equivalent to check if the following inequality holds:

\[
F^1_{\text{load}} + F^1_{\text{speed}} - \mu(P^1) + (Q - q(P^2)) \max\{0, \bar{D}^1_{0,j} - \bar{D}^2_{0,j}\} < \min_{v_2 \in S^2} C^2(v_2) - \mu(P^2).
\]

Since \( C^2(v_2) \) is a univariate convex function and \( S^2 \) is a closed interval, \( \min_{v_2 \in S^2} C^2(v_2) \) can be computed efficiently.
5.4 Case 4: \( w^1 < h^1 \) and \( w^2 < h^2 \)

This is the most tricky case. We first assume \( S^1 = [v^1_{\min}, v^1_{\max}] \) and \( S^2 = [v^2_{\min}, v^2_{\max}] \).

**Proposition 4.** Condition (4-a) in Proposition 3 holds if and only if \( T^1(v^1_{\max}) \leq T^2(v^2_{\max}) \).

**Proof.** Note that \( T^1 \) and \( T^2 \) are both monotonically decreasing. Suppose for any \( v_2 \in S^2 \), there exists \( v_1 \in S^1 \) such that \( T^1(v_1) \leq T^2(v_2) \). Then for \( v^2_{\max} \), there exists \( v_1 \in S^1 \) such that \( T^1(v_1) \leq T^2(v^2_{\max}) \). Thus \( T^1(v^1_{\max}) \leq T^1(v_1) \leq T^2(v^2_{\max}) \). On the other hand, if \( T^1(v^1_{\max}) \leq T^2(v^2_{\max}) \), then for any \( v_2 \in S^2 \), \( v^1_{\max} \in S^1 \) satisfies that \( T^1(v^1_{\max}) \leq T^2(v^2_{\max}) \leq T^2(v_2) \).

To efficiently check condition (4-b) in Proposition 3, we first need to define two functions that will be used throughout the rest of this section. Let

\[
\beta(v) = \frac{D^1}{T^2(v) - s^1 - \Gamma^1} = \frac{D^1}{D^2/v + s^2 + \Gamma^2 - s^1 - \Gamma^1} = \frac{D^1}{D^2/v + \delta},
\]

where \( \delta = s^2 + \Gamma^2 - s^1 - \Gamma^1 \) is a constant. It is not difficult to show the following properties of function \( \beta \) hold.

**Proposition 5.**

- For any \( v \in S^2 \) and \( \beta(v) \in S^1 \), \( T^1(\beta(v)) = T^2(v) \).
- Function \( \beta \) is monotonically increasing.
- \( T^1(v_1) \leq T^2(v_2) \Leftrightarrow \beta(v_2) \leq v_1 \), and in particular, \( T^1(v^1_{\max}) \leq T^2(v^2_{\max}) \Leftrightarrow \beta(v^2_{\max}) \leq v^1_{\max} \).

A vehicle is able to travel from \( i_{w^1} \) along \( P^1 \) at speed \( \beta(v) \), and arrive at the last vertex \( j \) at the same time as traveling from \( i_{w^2} \) along \( P^2 \) at speed \( v \). The function \( \beta \) can also be seen as the composition of \( (T^1)^{-1} \) and \( T^2 \).

Let \( H(v) = C^1(\beta(v)) - C^2(v) \), i.e., the cost difference between the two labels if they arrive at the same time. It will be seen later that checking whether Condition (4-b) holds is closely related to find a maximum of \( H(v) \) over a closed interval.

**Lemma 1.** When \( v \in [l, u] \), the derivative of \( H(v) \) is 0 if and only if \( \beta(v) = v \).

**Proof.** See the appendix.

The condition \( \beta(v) = v \) implies \( \delta v = D^1 - D^2 \). If \( \delta = 0 \), then \( \beta(v) = v \) if and only if \( D^1 = D^2 \). When \( \delta \neq 0 \), define \( v^* = (D^1 - D^2)/\delta \). Observe that \( T^1(v^*) = T^2(v^*) \). If the speed \( v^* \) is in \( S^1 \cap S^2 \), then it is the speed at which a vehicle can travel from either \( i_{w^1} \) along \( P^1 \) or \( i_{w^2} \) along \( P^2 \) and arrive at \( j \) at the same time.

**Lemma 2.** Suppose \( \delta = 0 \). Then the maximum of \( H(v) \) over an interval \([v_{\min}, v_{\max}]\) with \( v_{\min} > 0 \) is attained at \( v_{\max} \) if \( D^1 > D^2 \), at \( v_{\min} \) if \( D^1 < D^2 \), and at any point in \([v_{\min}, v_{\max}]\) if \( D^1 = D^2 \).

**Proof.** When \( D^1 = D^2 \), \( \beta(v) = v \) for any \( v \) and \( H(v) \) is constant. When \( D^1 > D^2 \), \( \beta(v) > v \) and \( H'(v) > 0 \) for any \( v \in [v_{\min}, v_{\max}] \), based on the proof of Lemma 1. Thus \( H(v) \) is strictly increasing and its maximum is attained at \( v_{\max} \). The case of \( D^1 < D^2 \) can be proved in a similar way.

Condition (4-b) in Proposition 3 can be checked by simply comparing numbers, based on the result below.
Proposition 6. Suppose that condition (4-a) holds, i.e., $T^1(v_1^{\text{min}}) \leq T^2(v_2^{\text{max}})$. Condition (4-b) in Proposition 5 holds if and only if the following inequality holds
\[ z^* - \mu(P^1) + (Q - q(P^2)) \max\{0, D_{0,1}^1 - D_{0,2}^2 \} + \mu(P^2) < 0, \]
where the value of $z^*$ is calculated as follows.

1. If $T^1(v_1^{\text{min}}) \leq T^2(v_2^{\text{max}})$, then $z^* = C^1(v_1^{\text{min}}) - C^2(v_2^{\text{min}})$.

2. If $T^2(v_2^{\text{max}}) < T^1(v_1^{\text{min}}) \leq T^2(v_2^{\text{min}})$, then $z^*$ equals to the maximum of $C^1(v_1^{\text{min}}) - C^2(v_2^{\text{min}})$, $H(v_2^{\text{max}})$, and $H(v^*)$ if $v^*$ is well-defined ($\delta \neq 0$) and $v^* \in [D^2v_2^{\text{min}}/(D^1 - \delta_{v_1}^\text{min}), v_2^{\text{max}}]$.

3. If $T^2(v_2^{\text{min}}) < T^1(v_1^{\text{min}})$, then $z^*$ equals to the maximum of $H(v_2^{\text{min}})$, $H(v_2^{\text{max}})$, and $H(v^*)$ if $v^*$ is well-defined ($\delta \neq 0$) and $v^* \in S^2$.

Proof. To check if Condition (4-b) holds is equivalent to solve a univariate constrained optimization problem in $v_2$. To see this, fix $v_2 \in S^2$ and define
\[ \phi(v_2) = \min C^1(v_1) \quad \text{(12a)} \]
subject to
\[ T^1(v_1) = s^1 + \Gamma^1 + \frac{D_1^1}{v_1} \leq s^2 + \Gamma^2 + \frac{D_2^2}{v_2} = T^2(v_2) \quad \text{(12b)} \]
\[ v_1^{\text{min}} \leq v_1 \leq v_1^{\text{max}}. \quad \text{(12c)} \]
Constraint (12b) is equivalent to $v_1 \geq \beta(v_2)$. Based on the assumption that condition 4-a holds, and Proposition 5, we have $\beta(v_2^{\text{max}}) \leq v_2^{\text{max}}$, so $\beta(v_2) \leq v_2^{\text{max}}$, $\forall v_2 \in S^2$. Therefore, $\phi(v_2) = C^1(v_1^{\text{min}})$ if $\beta(v_2) \leq v_1^{\text{min}}$ and $\phi(v_2) = C^1(\beta(v_2))$ otherwise. Define
\[ \psi(v_2) = \phi(v_2) - C^2(v_2) \quad \text{(13a)} \]
\[ z^* = \max\{\psi(v_2) \mid v_2 \in S_2\} \quad \text{(13b)} \]
Then (4-b) is satisfied if and only if
\[ z^* - \mu(P^1) + (Q - q(P^2)) \max\{0, D_{0,1}^1 - D_{0,2}^2 \} + \mu(P^2) < 0. \]
We consider the following three cases to compute $z^*$.

1. $T^1(v_1^{\text{min}}) \leq T^2(v_2^{\text{max}})$. Then for any $v_2 \in S^2$, $v_1^{\text{min}} \geq \beta(v)$ and $\phi(v) = C^1(v_1^{\text{min}})$.

2. $T^2(v_2^{\text{max}}) < T^1(v_1^{\text{min}}) \leq T^2(v_2^{\text{min}})$. Then $\beta(v_2^{\text{min}}) \leq v_1^{\text{min}} < \beta(v_2^{\text{max}})$. Since $\beta$ is monotonically increasing, there exists $\tilde{v} \in S^2$ such that $\beta(\tilde{v}) = v_1^{\text{min}}$. In particular, $\tilde{v} = D_2v_2^{\text{min}}/(D_1 - \delta_{v_1}^{\text{min}})$. Then $\phi(v) = C^1(v_1^{\text{min}})$ for $v \in [v_2^{\text{min}}, \tilde{v}]$, and $\phi(v) = C^1(\beta(v))$ for $v \in [\tilde{v}, v_2^{\text{max}}]$.

3. $T^2(v_2^{\text{min}}) < T^1(v_1^{\text{min}})$. Then for any $v_2 \in S^2$, $\beta(v) > v_1^{\text{min}}$ and $\phi(v) = C^1(\beta(v))$. Thus $\psi(v) = C^1(\beta(v)) - C^2(v_2)$ for any $v \in S^2$. Therefore from Lemmas 1 and 2, the maximum of $H(v)$ is attained at $v_2^{\text{min}}$, $v_2^{\text{max}}$, or $v^*$ if $v^*$ is well-defined and $v^* \in S^2$. 

□
6 Computational Experiments

This section describes extensive computational experiments conducted on the PRP instances to assess the performance of our algorithm. We also compare our computational results with two existing exact algorithms in the literature related to the PRP, a branch-and-cut algorithm proposed by Fukasawa et al. (2015) and a BP algorithm proposed by Dabia et al. (2015).

6.1 Computational performance of six variants of the proposed algorithms on the PRP instances

6.1.1 Implementation details

We consider three variants of the set-partitioning formulation introduction in Section 3. These variants differ on the what \( r \) represents in each triple \((r, I, s)\): an elementary route, a q-route, or a 2-cycle-free q-route (a q-route that forbids cycles of type \( i - j - i \)). The linear programming relaxation with elementary routes provides the tightest lower bound, and each pricing problem takes the longest time to solve; on the other hand, the linear programming relaxation with q-routes gives the worst lower bound, and the pricing problem takes the least time to solve. We also consider whether or not to add valid inequalities to strengthen each formulation. We consider the rounded capacity inequalities from the standard VRP in the following form

\[
\sum_{i \in S} \sum_{j \notin S} x_{ij} \geq \xi(S),
\]

where \( S \) is a subset of all the customers and \( \xi(S) \) is the minimum number of vehicles needed to visit customers in set \( S \) (Toth and Vigo 2014). These inequalities are separated via a heuristic from Lysgaard (2003). In total, six algorithms (three choices of columns and whether or not to add valid inequalities) are implemented and their performances are compared.

We discuss how the generic labeling algorithm described in Section 4 and Section 5 specializes to the PRP. In the PRP, the speed-dependent fuel cost rate function \( G(v) \) in (1) equals \( \frac{\pi_1}{v} + \pi_2 v^2 \) for some parameters \( \pi_1 \) and \( \pi_2 \), \( \lambda_{ij} = \lambda \) and \( \eta_{ij} = \eta \) for any \((i, j) \in A\). All these parameters are given in Demir et al. (2012). Function \( G(v) \) attains its minimum at \( v_F = (\frac{\pi_1}{2\pi_2})^{\frac{1}{3}} \), so we set \( l_{ij} = \max\{v_F, l_{ij}\} \) for any \((i, j) \in A\). For label termination, the objective function in the optimization problem (8) is (ignoring the constant) \( G(v) + \frac{p}{v} = \frac{\pi_1 + p}{v} + \pi_2 v^2 \), which is also convex in \( v \). Define \( v_{FD} := (\frac{\pi_1 + p}{2\pi_2})^{\frac{3}{2}} \). Assume the feasible region of the problem (8) is \( S = [v_{min}, v_{max}] \), then the optimal objective of (8) is attained at \( v_{FD} \) if \( v_{FD} \in S \), \( v_{max} \) if \( v_{FD} > v_{max} \), and \( v_{min} \) if \( v_{FD} < v_{min} \).

6.1.2 Test instances

Our computational experiments are conducted on two sets of test instances. The first set consists of instances generated using the geographical locations of United Kingdom cities. This set includes three groups of instances, UK-A, UK-B, and UK-C, and they differ on the time-window length for each customer. The UK-A instances were proposed in Bektaş and Laporte (2011), have the widest time windows (around 20,000 seconds for each customer), and can be downloaded at

http://www.apollo.management.soton.ac.uk/prplib.htm.

UK-B and UK-C instances were proposed in Kramer et al. (2014), have narrower time windows than UK-A instances (between 2,000 and 5,000 seconds for each customer in the UK-B instances.
and between 2,000 and 15,000 seconds for each customer in the UK-C instances), and are available at


The second set of test instances were proposed by Solomon (1987) for the VRPTW and adapted by Dabia et al. (2015) for the PRP. Solomon’s instances are organized into six groups according to the geographical locations of the customers: Random (R), Clustered (C), and Randomly Clustered (RC), and the time window ranges: tighter (1) or wider (2). Dabia et al. (2015) modified these instances for the PRP by setting the planning horizon to 24 hours and adjusting other parameters accordingly; see Dabia et al. (2015) for more details.

We test the six algorithms on instances with up to 50 customers. All the test instances are pre-processed using the procedure described in Fukasawa et al. (2015). Although larger instances are available on both test sets, they are out of reach for our current exact approach. Hereinafter, we refer to each instances with two additional parameters: the number of customers \( n \) and the index \( i \) of the instance in its group. Specifically, we use label UK\( n \)-\( G \)-\( i \) for the UK instances, in which \( G \in \{ A, B, C \} \) represents the group of each instance, and GT\( n \)-\( G \)-\( T \)-\( i \) for the Solomon instances, in which \( G \in \{ R, C, RC \} \) and \( T \in \{ 1, 2 \} \) represent the group of each instance. For example, UK25A-1 represents the first instance in the UK-A group with 25 customers, and R103-10 represents the third instance in group R1 (Random, tighter time windows) with 10 customers of the Solomon instances.

### 6.1.3 Computational results

All the algorithms discussed in this paper are coded in C++. Computational experiments are conducted with an AMD machine running at 2.3 GHz with 264 GBytes of RAM memory, under Linux Operating System. SCIP Optimization Suite 3.2.0 is used for the BCP framework. Following previous PRP results in the literature, a time limit of one hour is imposed to the algorithm execution for the UK instances and two hours for Solomon instances.

We compare the performance of the six different variants of the proposed algorithm, using performance profiles (Dolan and More 2002) in Figure 1. For each instance, we record the computational time that each variant of the algorithm takes to obtain the optimal solution, and we set the time to infinity if the time limit is reached. The performance profile is then plotted as follows. For a given algorithm, a point \((x, y)\) in the performance profile indicates that \(y\) percentage of instances are solved in at most \(x\) times of the computational time that the fastest algorithm spends for each particular instance. For example, point \((1.0, 0.5)\) indicates that this algorithm is the fastest on 50% of the instances; and point \((2.5, 0.75)\) indicates that this algorithm does not take more than 2.5 times of the time that the fastest algorithm takes for 75% of the instances. According to Figure 1, the variant that prices out 2-cycle-free \(q\)-routes and separates rounded capacity inequalities gives the best overall performance on the test instances.

Tables 2 and 3 present detailed solution information for all six variants. Table 2 reports results on solving the root linear program (LP) of formulation (4), and Table 3 reports results on solving (4) to optimality. We use the following abbreviations in these tables:

- **EL/2C/QR**: each route \(r\) in (4) is an elementary routes/2-cycle-free \(q\)-routes/\(q\)-route.
- **-C**: rounded capacity inequalities are added
- **# sol**: percentage of instances in each family that the algorithm can solve within the time limit.
Figure 1: Comparison of six algorithms for solving the PRP instances.

- T: the average computational time for each family of instances. The time limit is used (3600 seconds for UK instances, and 7200 seconds for Solomon instances) in calculating the average when the time limit is hit.

- G: the average optimality gap for each family of instances. The upper bound used to calculate this optimality gap is the best upper bound obtained among all six variants within the time limit. The lower bound is given by the root LP in Table 2, and the best lower bound achieved within the time limit in Table 3 for each variant, respectively. When the root node is not even finished within the time limit, we use 100% in calculating the average.

Table 2 shows the tradeoff between computational time and relaxation strength for six different variants on solving the root LP of (4). We can see that options “QR” and “QR-C” give the best performance in terms of the percentage of root LPs solved within the time limit, the solution time, as well as the optimality gap. In addition, option “QR” takes less computational time than option “QR-C”, but yields a larger root gap. This effect is less significant for options “EL” and “EL-C”, and options “2C” and “2C-C”.

In terms of solving (4) to optimality, we can see from Table 3 that option “2C-C” gives the best performance in almost all test instances. Comparing Table 3 and Table 2, we see that although options “QR” and “QR-C” could finish solving more instances at root LPs, fewer instances are solved to optimality within the time limit than options “2C” and “2C-C”. This is somewhat expected since the relaxation bound given by options “QR” and “QR-C” are relatively weak, although these two options process each node faster.

Finally, we summarize the performance of our BCP algorithm using variant “2C-C” on all the
<table>
<thead>
<tr>
<th></th>
<th>EL</th>
<th>EL-C</th>
<th>2C</th>
<th>2C-C</th>
<th>QR</th>
<th>QR-C</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRPlib</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UK-A</td>
<td>45.0</td>
<td>50.0</td>
<td>70.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>T(s)/G(%)</td>
<td>2072/55.1</td>
<td>1957/50.0</td>
<td>1488/31.6</td>
<td>1571/30.4</td>
<td>67/10.9</td>
<td>496/5.0</td>
</tr>
<tr>
<td>UK-B</td>
<td>0.0</td>
<td>0.0</td>
<td>2.0</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
</tr>
<tr>
<td>T(s)/G(%)</td>
<td>568/10.4</td>
<td>590/10.3</td>
<td>2/0.6</td>
<td>4/0.4</td>
<td>2/1.3</td>
<td>4/0.7</td>
</tr>
<tr>
<td>UK-C</td>
<td>1.5</td>
<td>1.5</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>T(s)/G(%)</td>
<td>896/23.4</td>
<td>1027/23.3</td>
<td>73/2.2</td>
<td>142/1.5</td>
<td>8/5.0</td>
<td>22/2.8</td>
</tr>
<tr>
<td>C1</td>
<td>22.2</td>
<td>27.8</td>
<td>83.3</td>
<td>43.8</td>
<td>88.9</td>
<td>94.4</td>
</tr>
<tr>
<td>T(s)/G(%)</td>
<td>5759/78.5</td>
<td>5292/72.7</td>
<td>1416/18.7</td>
<td>1071/12.6</td>
<td>486/8.3</td>
<td>559/7.5</td>
</tr>
<tr>
<td>C2</td>
<td>6.2</td>
<td>6.2</td>
<td>43.8</td>
<td>43.8</td>
<td>56.2</td>
<td>56.2</td>
</tr>
<tr>
<td>T(s)/G(%)</td>
<td>6799/93.9</td>
<td>6785/93.9</td>
<td>4195/63.5</td>
<td>4327/63.5</td>
<td>3335/59.0</td>
<td>3353/58.8</td>
</tr>
<tr>
<td>R1</td>
<td>29.2</td>
<td>29.2</td>
<td>83.3</td>
<td>83.3</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>T(s)/G(%)</td>
<td>5147/70.9</td>
<td>5210/70.9</td>
<td>1542/18.8</td>
<td>1515/18.7</td>
<td>95/8.7</td>
<td>145/8.5</td>
</tr>
<tr>
<td>R2</td>
<td>18.2</td>
<td>18.2</td>
<td>31.8</td>
<td>31.8</td>
<td>31.8</td>
<td>31.8</td>
</tr>
<tr>
<td>T(s)/G(%)</td>
<td>6036/81.9</td>
<td>6032/81.9</td>
<td>5068/69.1</td>
<td>4437/60.9</td>
<td>5015/71.9</td>
<td>4691/67.4</td>
</tr>
<tr>
<td>RC1</td>
<td>62.5</td>
<td>68.8</td>
<td>93.8</td>
<td>93.8</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>T(s)/G(%)</td>
<td>2953/38.6</td>
<td>2708/32.7</td>
<td>557/10.5</td>
<td>615/10.5</td>
<td>12/4.8</td>
<td>20/4.6</td>
</tr>
<tr>
<td>RC2</td>
<td>50.0</td>
<td>56.2</td>
<td>62.5</td>
<td>62.5</td>
<td>62.5</td>
<td>62.5</td>
</tr>
<tr>
<td>T(s)/G(%)</td>
<td>3791/50.3</td>
<td>3465/44.1</td>
<td>2894/37.8</td>
<td>2781/37.8</td>
<td>2724/38.2</td>
<td>2774/38.1</td>
</tr>
</tbody>
</table>

Table 2: Detailed statistics for all six variants of the BCP algorithm on evaluating the root LP.

<table>
<thead>
<tr>
<th></th>
<th>EL</th>
<th>EL-C</th>
<th>2C</th>
<th>2C-C</th>
<th>QR</th>
<th>QR-C</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRPlib</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UK-A</td>
<td>40.0</td>
<td>50.0</td>
<td>35.0</td>
<td>52.5</td>
<td>7.5</td>
<td>67.5</td>
</tr>
<tr>
<td>T(s)/G(%)</td>
<td>2257/55.0</td>
<td>1965/50.0</td>
<td>2484/30.7</td>
<td>1858/47.5</td>
<td>3365/7.9</td>
<td>1402/15.8</td>
</tr>
<tr>
<td>UK-B</td>
<td>47.2</td>
<td>50.0</td>
<td>57.5</td>
<td>90.0</td>
<td>77.5</td>
<td>80.0</td>
</tr>
<tr>
<td>T(s)/G(%)</td>
<td>955/10.1</td>
<td>791/12.6</td>
<td>524/0.1</td>
<td>496/0.0</td>
<td>909/0.3</td>
<td>808/0.1</td>
</tr>
<tr>
<td>UK-C</td>
<td>55.0</td>
<td>57.5</td>
<td>57.5</td>
<td>65.0</td>
<td>35.0</td>
<td>42.5</td>
</tr>
<tr>
<td>T(s)/G(%)</td>
<td>1732/23.3</td>
<td>1612/27.6</td>
<td>1656/1.4</td>
<td>1300/6.1</td>
<td>2404/2.5</td>
<td>2078/4.1</td>
</tr>
<tr>
<td>C1</td>
<td>16.7</td>
<td>22.2</td>
<td>50.0</td>
<td>61.4</td>
<td>33.3</td>
<td>38.9</td>
</tr>
<tr>
<td>T(s)/G(%)</td>
<td>6281/78.1</td>
<td>5966/77.8</td>
<td>3674/17.5</td>
<td>2999/11.8</td>
<td>4845/7.1</td>
<td>4356/17.7</td>
</tr>
<tr>
<td>C2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>T(s)/G(%)</td>
<td>7200/93.9</td>
<td>7200/93.9</td>
<td>7200/68.1</td>
<td>7200/67.9</td>
<td>7200/83.9</td>
<td>7200/67.2</td>
</tr>
<tr>
<td>R1</td>
<td>29.2</td>
<td>29.2</td>
<td>62.5</td>
<td>62.5</td>
<td>62.5</td>
<td>62.5</td>
</tr>
<tr>
<td>T(s)/G(%)</td>
<td>5152/70.8</td>
<td>5217/70.8</td>
<td>3020/18.5</td>
<td>3096/18.5</td>
<td>294/8.0</td>
<td>2930/7.9</td>
</tr>
<tr>
<td>R2</td>
<td>9.1</td>
<td>9.1</td>
<td>4.5</td>
<td>4.5</td>
<td>4.5</td>
<td>4.5</td>
</tr>
<tr>
<td>T(s)/G(%)</td>
<td>6508/81.9</td>
<td>6580/86.4</td>
<td>6872/69.1</td>
<td>6570/65.5</td>
<td>6872/73.4</td>
<td>6872/74.1</td>
</tr>
<tr>
<td>RC1</td>
<td>43.8</td>
<td>37.5</td>
<td>50.0</td>
<td>50.0</td>
<td>43.8</td>
<td>50.0</td>
</tr>
<tr>
<td>T(s)/G(%)</td>
<td>4325/38.3</td>
<td>4544/38.6</td>
<td>3631/10.0</td>
<td>3616/10.0</td>
<td>4129/4.0</td>
<td>3617/4.0</td>
</tr>
<tr>
<td>RC2</td>
<td>37.5</td>
<td>41.2</td>
<td>43.8</td>
<td>25.0</td>
<td>18.8</td>
<td>18.8</td>
</tr>
<tr>
<td>T(s)/G(%)</td>
<td>4532/50.3</td>
<td>5179/50.2</td>
<td>4234/37.6</td>
<td>4117/37.6</td>
<td>5400/38.1</td>
<td>5850/50.2</td>
</tr>
</tbody>
</table>

Table 3: Detailed statistics for all six variants of the BCP algorithm on solving (4) to optimality.
PRP test instances in Table 4. We organize instances according to the number of customers for each instance family described in Section 6.1.2. For each instance family and size, Table 4 shows the number of instances that could be solved to optimality within the time limit, and the total number of instances is shown in the parenthesis. An entry with symbol ‘-' means that a particular instance size is not available for that instance family. We can see from Table 4 that, the BCP algorithm is capable of solving most instances to optimality within the time limit for instances with up to 20 customers, can solve 51.2% instances with 25 customers, and solves several instances with 50 customers. Note that the largest size of UK-A and UK-C instances ever solved to optimality before is 10, and the largest size of UK-B instances ever solved to optimality before is 20 (Fukasawa et al. 2015).

### 6.2 Comparison with the branch-and-cut method for the PRP in Fukasawa et al. (2015)

In the rest of this section, we use option “2C-C” as the default setting for our BCP algorithm, and compare its result with other solution approaches related to the PRP in the literature. The PRP can be formulated as a mixed-integer convex program, and solved in a branch-and-cut algorithm (Fukasawa et al. 2015). In this section, we compare our BCP algorithm with the algorithm proposed by Fukasawa et al. (2015) (labeled as BC) on UK instances with up to 20 customers. Note that only results on instances with 10 and 20 customers are reported in Fukasawa et al. (2015). Our algorithm can solve UK instances with up to 50 customers.

The first four columns of Table 5 show the results of both algorithms for instances with 10 customers. Both algorithms can solve these instances to optimality within the time limit. The column “Opt.” gives the optimal objective value for each instance. The next five columns show the performance of both algorithms on instances with 20 customers. Columns with label “Best” show the best solution found by the algorithm, and columns “GAP/time” give the total time (seconds) it takes to solve the instance to optimality, if the time limit is not hit, and give the ending optimality gap otherwise.

We can see from Table 5 that, the proposed BCP algorithm significantly outperforms the previous BC algorithm for all instances considered. The BCP algorithm is capable of solving 59 out of 60 instances within the time limit, and takes 6.48 seconds on average to solve these 59 instances. On the other hand, the BC algorithm can only solve 42 instances, takes 371.1 seconds on average to solve them, and the average optimality gap for the remaining 18 instances is as large as 17.6%. Note that there are four UK20-A instances and six UK20-C instances that the BC algorithm finds the optimal solution but could not prove its optimality.
Table 5: Comparison between the BCP algorithm and the BC algorithm in Fukasawa et al. (2015) on UK instances.

<table>
<thead>
<tr>
<th>Instance</th>
<th>BC Opt.</th>
<th>BC time</th>
<th>BCP time</th>
<th>Instance</th>
<th>Best</th>
<th>GAP/</th>
<th>Best</th>
<th>GAP/</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK10A-1</td>
<td>170.64</td>
<td>1354</td>
<td>0.4</td>
<td>UK20A-1</td>
<td>352.48</td>
<td>22.9%</td>
<td>351.82</td>
<td>13.9</td>
</tr>
<tr>
<td>UK10A-2</td>
<td>204.88</td>
<td>813</td>
<td>0.1</td>
<td>UK20A-2</td>
<td>365.77</td>
<td>20.7%</td>
<td>365.77</td>
<td>1.8</td>
</tr>
<tr>
<td>UK10A-3</td>
<td>202.56</td>
<td>1708</td>
<td>0.2</td>
<td>UK20A-3</td>
<td>230.49</td>
<td>23.6%</td>
<td>230.49</td>
<td>11.8</td>
</tr>
<tr>
<td>UK10A-4</td>
<td>189.88</td>
<td>844</td>
<td>0.2</td>
<td>UK20A-4</td>
<td>347.04</td>
<td>21.2%</td>
<td>347.04</td>
<td>43.5</td>
</tr>
<tr>
<td>UK10A-5</td>
<td>175.59</td>
<td>2649</td>
<td>0.2</td>
<td>UK20A-5</td>
<td>323.44</td>
<td>16.6</td>
<td>323.44</td>
<td>16.6</td>
</tr>
<tr>
<td>UK10A-6</td>
<td>214.48</td>
<td>1472</td>
<td>0.04</td>
<td>UK20A-6</td>
<td>364.23</td>
<td>18.1</td>
<td>364.23</td>
<td>18.1</td>
</tr>
<tr>
<td>UK10A-7</td>
<td>190.14</td>
<td>882</td>
<td>0.07</td>
<td>UK20A-7</td>
<td>253.61</td>
<td>0.00%</td>
<td>253.61</td>
<td>0.00%</td>
</tr>
<tr>
<td>UK10A-8</td>
<td>222.17</td>
<td>564</td>
<td>0.02</td>
<td>UK20A-8</td>
<td>301.51</td>
<td>12.7</td>
<td>301.51</td>
<td>12.7</td>
</tr>
<tr>
<td>UK10A-9</td>
<td>174.54</td>
<td>352</td>
<td>0.07</td>
<td>UK20A-9</td>
<td>362.56</td>
<td>19.5%</td>
<td>362.56</td>
<td>18.6</td>
</tr>
<tr>
<td>UK10A-10</td>
<td>189.82</td>
<td>211</td>
<td>0.1</td>
<td>UK20A-10</td>
<td>313.33</td>
<td>13.7</td>
<td>313.33</td>
<td>13.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Instance</th>
<th>BCP Opt.</th>
<th>BCP time</th>
<th>Instance</th>
<th>BCP Opt.</th>
<th>BCP time</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK10B-1</td>
<td>246.44</td>
<td>0.2</td>
<td>UK20B-1</td>
<td>469.35</td>
<td>0.2</td>
</tr>
<tr>
<td>UK10B-2</td>
<td>303.73</td>
<td>0.2</td>
<td>UK20B-2</td>
<td>477.05</td>
<td>0.2</td>
</tr>
<tr>
<td>UK10B-3</td>
<td>301.89</td>
<td>0.2</td>
<td>UK20B-3</td>
<td>454.46</td>
<td>0.2</td>
</tr>
<tr>
<td>UK10B-4</td>
<td>273.90</td>
<td>0.3</td>
<td>UK20B-4</td>
<td>523.59</td>
<td>0.3</td>
</tr>
<tr>
<td>UK10B-5</td>
<td>255.07</td>
<td>0.2</td>
<td>UK20B-5</td>
<td>447.33</td>
<td>0.2</td>
</tr>
<tr>
<td>UK10B-6</td>
<td>332.34</td>
<td>0.2</td>
<td>UK20B-6</td>
<td>511.78</td>
<td>0.2</td>
</tr>
<tr>
<td>UK10B-7</td>
<td>314.64</td>
<td>0.6</td>
<td>UK20B-7</td>
<td>379.02</td>
<td>0.6</td>
</tr>
<tr>
<td>UK10B-8</td>
<td>339.36</td>
<td>0.2</td>
<td>UK20B-8</td>
<td>451.31</td>
<td>0.2</td>
</tr>
<tr>
<td>UK10B-9</td>
<td>261.10</td>
<td>0.2</td>
<td>UK20B-9</td>
<td>548.68</td>
<td>0.1</td>
</tr>
<tr>
<td>UK10B-10</td>
<td>285.20</td>
<td>0.2</td>
<td>UK20B-10</td>
<td>410.32</td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Instance</th>
<th>BCP Opt.</th>
<th>BCP time</th>
<th>Instance</th>
<th>BCP Opt.</th>
<th>BCP time</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK10C-1</td>
<td>210.18</td>
<td>1.6</td>
<td>UK20C-1</td>
<td>432.82</td>
<td>6.8%</td>
</tr>
<tr>
<td>UK10C-2</td>
<td>271.95</td>
<td>72.1</td>
<td>UK20C-2</td>
<td>450.85</td>
<td>12.7%</td>
</tr>
<tr>
<td>UK10C-3</td>
<td>229.18</td>
<td>2.5</td>
<td>UK20C-3</td>
<td>287.04</td>
<td>9.9%</td>
</tr>
<tr>
<td>UK10C-4</td>
<td>230.52</td>
<td>5.2</td>
<td>UK20C-4</td>
<td>434.23</td>
<td>7.3%</td>
</tr>
<tr>
<td>UK10C-5</td>
<td>205.49</td>
<td>7.6</td>
<td>UK20C-5</td>
<td>382.46</td>
<td>10.9%</td>
</tr>
<tr>
<td>UK10C-6</td>
<td>255.82</td>
<td>17.5</td>
<td>UK20C-6</td>
<td>444.35</td>
<td>14.9%</td>
</tr>
<tr>
<td>UK10C-7</td>
<td>217.79</td>
<td>14.3</td>
<td>UK20C-7</td>
<td>321.67</td>
<td>17.3%</td>
</tr>
<tr>
<td>UK10C-8</td>
<td>251.29</td>
<td>4.7</td>
<td>UK20C-8</td>
<td>410.35</td>
<td>1224</td>
</tr>
<tr>
<td>UK10C-9</td>
<td>186.04</td>
<td>5.8</td>
<td>UK20C-9</td>
<td>421.39</td>
<td>3347</td>
</tr>
<tr>
<td>UK10C-10</td>
<td>211.02</td>
<td>3.8</td>
<td>UK20C-10</td>
<td>390.68</td>
<td>7.2%</td>
</tr>
</tbody>
</table>

Table 6 shows the computational performance of our BCP algorithm and the BP algorithm in Dabia et al. (2015) (labeled as BPF) on Solomon instances with up to 50 customers. We only show instances for which both algorithms can provide lower and upper bounds. Columns labeled “LP” show the root LP relaxation bound for each instance, and columns labeled “BUB” report the best upper bound found by the algorithm within the time limit. Columns labeled “GAP/time” report the final optimality gap if an instance is not solved to optimality within the time limit, and report the total computational time (seconds) otherwise.

There are in total 112 instances (56 instances with 25 nodes and 56 instances with 50 nodes) in the Solomon instance family. For the root LP relaxation, the BCP algorithm can solve 70 instances, while the BPF algorithm can solve 74 instances. Within the time limit, the BCP algorithm can solve 35 instances to optimality, while the BPF algorithm solves 56 instance to optimality among these instances. The BPF algorithm performs slightly better in terms of computational time, due to the important restriction that the vehicle speed is kept constant over any given route, which reduces the number of feasible speed vectors considerably.

On the other hand, the optimal objective values of the PRP variant in Dabia et al. (2015) are always higher than the corresponding optimal objective values of the PRP. Even when the BCP algorithm does not solve the instance to optimality, the objective of obtained best feasible solutions are almost always lower than the LP relaxation bound given by the BPF algorithm (except instances...
Table 6: Comparison between the BCP algorithm with the BP algorithm in Dabia et al. (2015) on Solomon instances.
r109-50, r201-50, and rc201-50). It implies that the cost reduction gained by allowing department time at the depot to vary cannot in general compensate the cost increased by restricting speed to be constant over the entire route. The BCP algorithm is able to explore a larger combination of feasible routes and speeds, likely yielding a better solution for the PRP than that of the PRP variant considered in Dabia et al. (2015).

7 Conclusion

In this paper we explore the effect of simultaneously optimizing routing and speed decisions on fuel savings and emission reduction. The fuel consumption is modeled as a strictly convex function of vehicle speed. We propose a novel set partitioning formulation for this problem by exploiting the structure of optimal speeds over a route. With this new formulation, we design an efficient BCP algorithm capable of solving instances of sizes significantly larger than what has been reported in the literature. We next outline some directions that deserve further explorations.

Heterogeneous vehicles and varying depart time at the depot. This paper focuses on homogeneous vehicles. It has been shown in Koç et al. (2014) that additional fuel savings can be achieved by using a mixed fleet. It is not difficult to extend our formulation and algorithm to the case with heterogeneous vehicles. We can simply expand each column in the formulation to a set of columns, one for each vehicle type, and solve the pricing problem multiple times with different parameters from each vehicle type. In addition, our model assumes each vehicle always departs from the depot in the beginning. Varying departure time at the depot has been considered in Dabia et al. (2015), Kramer et al. (2015), leading to a reduced overall operational and environmental costs. It seems a nontrivial task though to apply our algorithm to the JRSP with varying departure time, since the implementation of the dominance rules in Section 5 would require solving nonconvex optimization problems in higher dimensions.

Refined fuel consumption and emission models. Our fuel consumption model assumes the speed-dependent part $G(v)$ in (1) is the same across the network. In practice, some parameters in $G(v)$ are related to the environment (such as pavement conditions, temperature, and traffic) and can vary from one arc to another. Note that in these cases the speed optimization problem remains a convex optimization problem, but the crucial property in Proposition 1 stating that the optimal speeds into and out of a nonactive customer are the same no longer holds. Therefore, it is unclear how to extend our results to deal with these extensions.

Dynamic models. This paper considers a deterministic and static model where all the parameters are known at the time of planning. In a practical setting, many parameters—such as traffic—are only available in real time. How to learn these parameters on the fly and jointly optimize routing and speed decisions is a challenging and promising direction for future research.

Speed is an important factor affecting fuel consumption and emissions. The complexities and benefits of transportation practices with speed control are worth exploring for us to better understand this trade-off. We hope our work in this paper can serve as a stepping stone.

A Proof of Proposition 1

We prove the statement by contradiction. Suppose the speed optimization problem has an optimal speed vector with three consecutive vertices $j, i$, and $k$, such that the service starting time at customer $i$ is $t_i$ with $t_i \in (a_i, b_i)$, and $v_{ji} \neq v_{ik}$. Then we can create a feasible speed vector that differs from the optimal speed vector only by the speeds over arcs $(j, i)$ and $(i, k)$, with a strictly lower cost.
Let the departure time at vertex $j$ be $t_j$ and the arriving time at vertex $k$ be $t_k$. We first assume that $v_{ji} < v_{ik}$. According to (2), the cost over arcs $(j, i)$ and $(i, k)$ which depends on speeds is $d_{ji}G(v_{ji}) + d_{ik}G(v_{ik})$. We create two new speeds $v'_{ji} = v_{ji} + d_{ik}\delta$ and $v'_{ik} = v_{ik} - d_{ji}\delta$ with

$$0 < \delta < \min\{\frac{v_{ik} - v_{ji}}{d_{ik}}, \frac{v_{ik} - v_{ji}}{d_{ji}}, \frac{v_{ik}^2 - v_{ji}^2}{v_{ji}d_{ik} + v_{ik}d_{ji}}\}.$$ 

We will show that with these two new speeds, the vehicle is able to depart vertex $j$ at time $t_j$, serve customer $i$ within its time window, arrive at node $k$ at time $t_k$, and incur a lower fuel cost.

According to the choice of $\delta$ and Assumption 3, we have $v'_{ji}, v'_{ik} \in (v_{ji}, v_{ik})$. Since the function $G(v)$ is strictly convex in $v$, then

$$\frac{G(v'_{ji}) - G(v_{ji})}{d_{ij}\delta} = \frac{G(v_{ji} + d_{ik}\delta) - G(v_{ji})}{d_{ik}\delta} < \frac{G(v_{ik}) - G(v_{ik} - d_{ji}\delta)}{d_{ji}\delta} = \frac{G(v_{ik}) - G(v'_{ik})}{d_{ji}\delta}.$$ 

Thus $d_{ji}G(v'_{ji}) + d_{ik}G(v'_{ik}) < d_{ji}G(v_{ji}) + d_{ik}G(v_{ik})$, which implies the new speeds incur less fuel cost on the route. Now we show that with the new speeds, the vehicle is able to leave vertex $j$ at $t_j$ and arrive at vertex $k$ at time $t_k$. The original travel time over arcs $(j, i)$ and $(i, k)$ is

$$T = \frac{d_{ji}}{v_{ji}} + \frac{d_{ik}}{v_{ik}},$$

and the travel time with the new speeds is

$$T' = \frac{d_{ji}}{v_{ji} + d_{ik}\delta} + \frac{d_{ik}}{v_{ik} - d_{ji}\delta}.$$ 

Then

$$T' - T = \frac{d_{ji}}{v_{ji} + d_{ik}\delta} + \frac{d_{ik}}{v_{ik} - d_{ji}\delta} - \left(\frac{d_{ji}}{v_{ji}} + \frac{d_{ik}}{v_{ik}}\right) = d_{ji}d_{ik}\delta\frac{v_{ji}d_{ik} + v_{ik}d_{ji} - (v_{ik}^2 - v_{ji}^2)}{v_{ik}v_{ji}(v_{ik} - d_{ji}\delta)(v_{ji} + d_{ik}\delta)} < 0.$$ 

The last inequality follows from the choice of $\delta$. Since the total travel time between vertex $j$ and vertex $k$ decreases with the new speeds, with sufficiently small $\delta$ the vehicle is able to satisfy the time window constraints for all customers while incurring a lower cost. The case with $v_{ji} > v_{ik}$ can be proven in a similar way.

**B Proof of Proposition 3**

Suppose there exists $L^1$ satisfying conditions 2–4. Let $(P, v_P) \in E(L^2)$. For any $v_2 \in S^2$ such that $(\tilde{v}^2, v_2, v_P)$ is a feasible speed vector over $P^2 \oplus P$, conditions 1, 2, 3, and (4-a) ensure that there exists $v_1 \in S^1$ such that $(\tilde{v}^1, v_1, v_P)$ is a feasible speed vector over $P^1 \oplus P$. Therefore, $(P, v_P) \in E(L^1)$.

We are going to show that for any triple $(P^2 \oplus P, P^2, s^2)$ induced by $v_2 \in S^2$ that leads to a feasible speed vector $(\tilde{v}^2, v_2, v_P)$ on $P^2 \oplus P$, there exists a triple $(P^1 \oplus P, P^1, s^3)$ with a smaller reduced cost. It is sufficient to consider the optimal speed vector $v^*$ over $P^2 \oplus P$ that is consistent with label $L^2$, i.e., the components of $v^*$ from the depot to vertex $i_1$ are $\tilde{v}^2$, and customers after $i_1$ and before $i_2$ (including $i_2$) are all seamless. Let $v^*_{\tilde{v}^2}$ be the component of $v^*$ corresponding to
the speed over the arc entering vertex $i^2$. Then the vehicle travels along $P^2 \oplus P$ in the following manner: start from the depot to vertex $i_{w2}$ with speed $\bar{v}^2$, leave from vertex $i_{w2}$ to vertex $i^2$ with speed $v^*_2$, and leave from vertex $i^2$ back to the depot with speeds $v^*_P$, the components of $v^*$ that correspond to $P$. Assume that the vehicle returns to the depot at time $t_{P^2 \oplus P}$, and the total speed-dependent and load-dependent costs on $P$ is $F^*$. Note that the cumulative load $q(P)$ on $P$ also incurs an additional load-dependent cost $\bar{D}^2_{0,i^2}q(P)$ when carried through $P^2$. Thus the total cost of route $P^2 \oplus P$ with speed vector $v^*$ is

$$C^2(v_2) + \bar{D}^2_{0,i^2}q(P) + F^* + pt_{P^2 \oplus P}.$$ 

Now we show how the route $P^1 \oplus P$ admits a triple $(P^1 \oplus P, P^3, s^3)$ with a better reduced cost than $(P^2 \oplus P, P^2, s^2)$ corresponding to the speed vector $v^*$. Since $v_2 \in S^2$, by (4-a) there exists $v_1 \in S^1$ such that $T^1(v_1) \leq T^2(v_2)$. Therefore, it is feasible to travel along $P^1 \oplus P$ in the following manner: travel from the depot to vertex $i_{w1}$ along $P^1$ with speed vector $\bar{v}^1$, travel from vertex $i_{w1}$ to vertex $i^1$ with speed $v_1$, finish serving vertex $i^1$ at $T^1(v_1)$, leave vertex $i^1$ at time $T^2(v_2)$, travel along $P$ using speeds given by the corresponding components in $v^*$, and return to the depot at time $t_{P^2 \oplus P}$. Then the total cost of traveling along route $P^1 \oplus P$ in the above manner is

$$C^1(v_1) + \bar{D}^1_{0,i^1}q(P) + F^* + pt_{P^2 \oplus P}. \tag{14}$$

Figure 2 illustrates the comparison between $P^1 \oplus P$ and $P^2 \oplus P$.

Figure 2: An illustration of comparison between $P^1 \oplus P$ and $P^2 \oplus P$.

Note that traveling along $P^1 \oplus P$ in the above manner does not necessarily respect the set of active customers in $L^1$. Nonetheless it is a feasible way to travel along $P^1 \oplus P$. Then there must exists a triple $(P^1 \oplus P, P^3, s^3)$ such that $c_{P^1 \oplus P, P^3, s^3}$ is at most the total cost in (14). Thus

$$c_{P^1 \oplus P, P^3, s^3} - c_{P^2 \oplus P, P^2, s^2} \leq [C^1(v_1) + \bar{D}^1_{0,i^1}q(P) + F^* + pt_{P^2 \oplus P} - (\mu(P^1) + \mu(P) + \nu)]$$

$$- [C^2(v_2) + \bar{D}^2_{0,i^2}q(P) + F^* + pt_{P^2 \oplus P} - (\mu(P^2) + \mu(P) + \nu)]$$

$$= [C^1(v_1) + \bar{D}^1_{0,i^1}q(P) - \mu(P^1)] - [C^2(v_2) + \bar{D}^2_{0,i^2}q(P) - \mu(P^2)] < 0.$$ 

The last inequality follows from conditions 3 and (4-b).
C Proof of Lemma 1

Since $H(v) = F_1^{\text{load}} + F_1^{\text{speed}} + D_1 G(\beta(v)) - F_2^{\text{load}} - F_2^{\text{speed}} - D_2 G(v)$, its derivative

$$H'(v) = D_1 G'(\beta(v)) \frac{D_1 D_2}{(D_2/v + \delta)^2 v^2} - D_2 G'(v)$$

$$= D_2 G'(\beta(v)) \left( \frac{D_1}{D_2/v + \delta} \right)^2 \frac{1}{v^2} - D_2 G'(v)$$

$$= D_2 G'(\beta(v)) \frac{(\beta(v))^2}{v^2} - D_2 G'(v).$$

Thus $H'(v) = 0$ if and only if $G'(\beta(v)) (\beta(v))^2 - G'(v)v^2 = 0$. The function $G'(v)v^2$ is strictly increasing in $v \in [l, u]$, since $G'(v)$ is non-negative and strictly increasing in $[l, u]$ due to Assumption 2 and $v^2$ is non-negative and strictly increasing in $[0, \infty)$. When $v \in [l, u]$, $G'(\beta(v)) (\beta(v))^2 - G'(v)v^2 = 0$ if and only if $\beta(v) = v$.

References


