Article publicat / *Published paper*:

Codina, E., Rosell, F. A heuristic method for a congested capacitated transit assignment model with strategies. "Transportation research. Part B: methodological", Desembre 2017, vol. 106, p. 293-320. Doi: 10.1016/j.trb.2017.07.008

A HEURISTIC METHOD FOR A CONGESTED CAPACITATED TRANSIT ASSIGNMENT MODEL WITH STRATEGIES. [1].

by

Esteve Codina, Francisca Rosell

Statistics and Operations Research Department
Universitat Politècnica de Catalunya
Campus Nord, Building C5, Office 216
C/ Jordi Girona, 1-3, Barcelona 08034, Spain

(Author's accepted manuscript)

Abstract

This paper addresses the problem of solving the congested transit assignment problem with strict capacities. The model under consideration is the extension made by Cominetti and Correa (2001) [Common-lines and passenger assignment in congested transit networks. Transportation Sci. 35(3):250-267], for which the only solution method capable of resolving large transit networks is the one proposed by Cepeda et al. (2006) [A frequency-based assignment model for congested transit networks with strict capacity constraints: characterization and computation of equilibria. Transportation Res. 40B(6):437-459]. This transit assignment model was recently formulated by the authors as both a variational inequality problem and a fixed point inclusion problem. As a consequence of these results, this paper proposes an algorithm for solving the congested transit assignment problem with strict line capacities. The proposed method consists of using an MSA-based heuristic for finding a solution for the fixed point inclusion formulation. Additionally, it offers the advantage of always obtaining capacity-feasible flows with equal computational performance in cases of moderate congestion and with greater computational performance in cases of highly congested networks. A set of computational tests on realistic small- and large-scale transit networks under various congestion levels are reported, and the characteristics of the proposed method are analyzed.

Keywords: Congested Transit Assignment; Variational Inequalities; Strategy-Based Transit Equilibrium; Heuristic Methods.

1 Introduction

Passenger transit assignment is a fundamental tool in evaluating and planning public transportation systems. Although the primary aim of such tools is to ascertain the passengers' route choice on transit networks, there is an increasing need to take into account the effects of congestion on the performance of the systems, specifically concerning the limitations on the maximum flows that the line segments may carry (Freijinger and Florian (2013)). Thus, this paper focuses on solving congested frequency-based transit assignment network models under the assumption that passengers make their decisions according to the concept of strategy from a static perspective. Initially, Chriqui and Robillard (1975) introduced the concept known as "set of attractive lines" at passenger stops. The extension to the case of multi-destination networks was done by Spiess (1984) and Spiess and Florian (1989), who introduced the notion of strategy as a choice of an attractive set of lines at each stop. The resulting model sought to minimize the expected value of the total travel time. Nguyen and Pallottino (1988) expressed the general concept of hyperpath for a transit network in terms of an acyclic directed graph that leads to a destination; they formulated a transit assignment model as a variational inequality model; and they proposed a solution algorithm based on an iterative MSA scheme. These initial models assumed that boarding may be possible when the first vehicle arrives at the stop; but they did so without taking into account limitations on the maximum number of passengers that can be transported, which consequently failed to consider that no passenger queueing may occur at stops. Gendreau (1984), Bouzayene-Ayari

[1]Research supported under Spanish Research Projects TRA2014-52530-C3-3-P

et al. (1995) initially studied the bus stop model. These bus stop models were integrated into transit assignment models by Bouzayene-Ayari et al. (1995) and Wu and Florian (1993). Bouzayene-Ayari et al. (2001) presented bus stop models in which waiting times at stops follow an additive law of the lines' effective frequencies. This model was analyzed under too restrictive assumptions which did not allow including delay formulas from queueing theory (bounded gradients for the delay formulas were required). Also, De Cea and Fernndez used BPR-like functions to formulate a congested transit assignment model for taking into account congestion effects at transfer nodes, but without imposing strict capacity limits; thus, their proposed model may present solutions that go beyond the stated capacity limits of transit lines.

The works by Karauchi et al. (2003) present a different approach for considering congestion, since they include in their formulation passengers' risk aversion to failing at first boarding. These ideas are extended later in the quasy-dynamic transit assignment model of Schmöcker et al. (2008).

The inclusion of strict capacity bounds has also been considered in the traffic assignment models. The classical fixed-demand diagonal traffic assignment model is also formulated using increasing volume-delay functions but their solution flows may override physical bounds (i.e., capacities). This may be the case when using volume-delay functions directly from calibration. The use of volume-delay functions reproducing the effects of waiting times from queueing theory, which increase to infinity at queue's capacity, has had a series of contributions, starting with the one of Daganzo (1977a,b) to more recent ones as in Nie et al. (2004). Basically all these contributions have relied on the well studied properties of the barrier and penalty methods as well as aumented lagrangian methods for mathematical programming problems. For the case of asymmetric problems (i.e., models based on variational inequalities), the inclusion of explicit capacity bounds is studied extensively in Larsson and Patriksson (1994), although they don't provide an algorithm for solving that model.

Nowadays, congestion effects and capacity limitations are taken into account virtually by any transit assignment model and not only by static deterministic frequency-based models but also by stochastic models such as that of Szeto et al. (2016) as well as by static schedule-based models and dynamic models, doing so by following either a macroscopic flow approach or a simulation approach (see, for instance, Verbas et al. (2016) and Cats et al. (2016), amongst others). Thus, Nguyen et al. (2001) developed a schedule-based model which takes into account the departure time and route choice jointly. The available capacity of the service is taken into account in order to enforce the FIFO rule of boarding passengers. In Hamdouch and Lawphongpanich (2008), the concept of strategies is included in a capacitated time table based model formulated as a variational inequality model; and, in Hamdouch et al. (2011), the problem of seat availability is considered in addition to the constraints imposed by line capacities. In Trozzi et al. (2013), a frequency-based capacitated dynamic model is developed by considering strategies and queueing at stops which observe the FIFO discipline.

Returning to static, deterministic frequency-based models that follow strategies, Cominetti and Correa (2001) analyzed the inclusion of congestion aspects and formulated the model conceptually as a fixed point problem. It must be noticed that, although the GSM model by Bouzayene-Ayari (2001) was initially developed under restricting conditions that did not allow to include queueing model results, these difficulties were overcome in the paper by Cominetti and Correa (2001) (see appendix A in that work). Cepeda et al. (2006) further studied the model initially set by Cominetti and Correa (2001) and approached it as a non-differentiable optimization problem that they solved heuristically by minimizing a gap function through an adaptation of the method of successive averages. Their model reflects the effects of congestion basically through the concept of effective frequencies experienced by passengers. The effective frequency of a line at a stop is a concept closely related to queueing, and it takes into account that a fraction of passengers may not board the first transportation unit that arrives; so, theoretically they must vanish when the stop is about to collapse. This model will be referred to in this paper as the C3F model. A remarkable feature of the C3F model is that, being based on the general stop model by Bouzayene-Ayari, allows to consider very complex models of passenger waiting times at stops which may originate from different headway distributions and can be used to incorporate queueing models for modelling passenger waiting times, which is specially useful in congested situations.

Babazadeh and Aashtiani (2005) proved that strategy-based transit assignment in congested networks is equivalent to a nonlinear complementarity problem formulated in the space of path flows. They also proposed a solution algorithm. Some other models for tackling the problem of congested transit networks can also be found in Castro and Leal (2003).

In Codina (2013), it is shown that the C3F model can also be reformulated as both a variational inequality problem and a fixed point inclusion. In that work, conditions for the existence of solutions are

established when effective frequency functions either do not vanish at capacity or when they do vanish at capacity (i.e., passenger waiting times are finite or infinite at capacity, respectively). In this last case, the effective frequency functions implicitly define capacities which may be considered explicitly formulated as upper bounds. This is equivalent to saying that delay formulas from queueing theory may be integrated in order to model the waiting queues of passengers. It must be noted that the C3F model is the only static, frequency-based, strategy-based transit assignment model with that feature.

This paper is a consequence of and is substantiated by the results in Codina (2013). It introduces a method for solving the C3F model in the special case when effective frequency functions become zero at capacity (infinite passenger waiting times) and also when they implicitly define sharp line capacities. Although a method was proposed in Cepeda et al (2006) to solve the problem, the empirical evidence in this paper shows that it may fail to find feasible solutions in cases of high congestion. Furthermore, based on the reformulation as a fixed-point inclusion in Codina (2013), several alternatives for solving it are discussed and presented, with the most efficient one being described in detail. The proposed method consists of finding a solution for the fixed point inclusion formulation by using an MSA-based heuristic. At each iteration, an uncongested capacitated strategy-based transit assignment problem is solved by means of the classical dual cutting-plane or outer-linearization method, which in turn uses the efficient hyper-path calculating algorithm of Spiess (1984). A remarkable feature of the proposed method is that, even in cases of very high congestion, its solutions are guaranteed to strictly verify capacity limits at all links in the network where these constraints may apply, provided that the network model verifies the assumptions concerning the existence of solutions for the variational inequality formulation, as stated in Codina (2013). The main advantage of the method is that it allows solving the problem with equivalent performance in cases of moderate congestion and with higher performance in cases of very high congestion, even if very high accuracy is required for the solution. This aspect is especially relevant not only for the direct application of the congested transit assignment model in planning studies, but also when transit assignment methods are used as an active part of frequency setting methodologies (e.g., in Noriega and Florian (2003)). This is particularly true for estimating origin-destination trip matrices in public transportation networks (e.g., Freijinger and Florian (2013)) or in methods for multimodal assignment. In these applications, the robustness of the assignment process may be very relevant, because large amounts of the passenger demand generated by the methodology's iterations may eventually need to be assigned in the transit network. However, in the end, the congestion level of the final solution may not be so severe. Since in this type of applications the computational performance of the assignment algorithms are very relevant, the new methods of Liu et al. (2009) for calculating step lengths in fixed-point iterations (recently extended by Di Gangi et al. (2015)) have been included in the algorithms. These methods are more efficient than the classic method of successive averages (MSA), widely used in transportation models, and also than the Ishikawa method (1974).

Summary of the main contributions.

- As a result of reformulating the C3F model into a fixed-point inclusion problem in Codina (2013), this paper presents the application of an MSA or Mann-type algorithm (Mann (1953)) to this model and provides further discussion on other related variants.
- The presented algorithm guarantees that solutions will always be feasible, specifically in contrast with the only method known up until the present (the method of Cepeda et al. (2006)), which may completely fail to find feasible solutions, precisely under high congestion conditions in which preserving feasibility would be more valuable.
- It is also shown that the degree of violation of capacity constraints is not reduced to a small percentage but can reach considerable values.
- In addition, the presented method proves to behave better and is computationally as efficient as the one in Cepeda et al. (2006), even in situations of very high congestion.

The outline of the paper is as follows. Section 2 provides a summarized description of the C3F model in terms of its formulation as a variational inequality problem. It also presents a summarized explanation of the results in Codina (2013), which led to the fixed point inclusion problem upon which the algorithm of this paper relies. Section 3 describes the algorithm and Section 4 describes an effective way to solve the capacitated uncongested transit assignment problems by using the well-known dual cutting-plane algorithm. In addition, these two sections contain a discussion on other solution methods that have been tried in solving these problems. Section 5 gives an interpretation of the capacitated problem solved in Section 4, and Section 6 describes how to evaluate the strategies used in the solution of the capacitated

congested transit assignment problem. Next, Section 7 shows the initial application of the algorithm. Then, for illustrative purposes, we use it first on small test networks in Subsection 7.1 before moving on to a medium-sized network and two realistic large-sized transit networks in Subsection 7.2. In those subsections, the performance of the algorithm and characteristics of the solutions are examined and compared to Cepeda et al. (2006). Finally, Section 8 discusses the conclusions and possible lines of future research. Appendices A and B contain a brief index of the notations used in the paper and the strategies used by the solution of a small example.

2 Description and summary of the model's formulation

2.1 Notation and network model

The transit network is represented by means of a directed graph $G = (N, A)$, N being the set of nodes and A the set of links. Generic nodes in N will be represented by the symbols i, j and a link will usually be represented by the index a as a shorthand for an ordered pair of indexes (i, j), both in N. The number of passenger trips from i to d will be denoted by g_i^d. The set of active origin-destination (O-D) pairs $\omega = (i, d)$ on the network will be denoted by W, i.e., $W \triangleq \{\ (i,d) \in N \times N \mid g_i^d > 0\ \}$. The set of destinations in the network will be denoted by D, i.e., $D \triangleq \{\ d \in N \mid \exists (i,d) \in W\ \}$ and the set of origin nodes for a fixed destination $d \in D$ will be denoted by $O(d)$, i.e., $O(d) \triangleq \{\ i \in N \mid (i,d) \in W\ \}$.

The set of nodes in the network excluding destination $d \in D$ will be denoted by N_d, i.e., $N_d = N \setminus \{d\}$. For a node $i \in N$, the set of emerging links will be denoted by $E(i)$, and the set of incoming links by $I(i)$. Figure 1 shows schematically the configuration of the network $G = (N, A)$. In this representation, transit stops are associated with a node. Some of the their outgoing links will be boarding links to a transit line. The set of these boarding links will be denoted by A_a. Some incoming links will be alighting links from a transit line (making up the set A_y). Also, flows on a line not alighting at the stop will be associated with links in A_x (dwell links). Each transit line with vehicles halting at the stop will have a single boarding link a from the stop, a single alighting link denoted by $y(a)$ and a single dwell link $x(a)$ to the stop. Non-boarding or non-alighting links at transit stops will model connections to other transportation modes, such as pedestrian routes or connections to other transit stops.

The following notation will be used:

- v_a^d will denote the flow at link $a \in A$ with destination $d \in D$ and $\mathbf{v}_i^d = (..., v_a^d, ...;\ a \in E(i)) \in \mathbb{R}_+^{|E(i)|}$, $i \in N$, $d \in D$ is the vector of flows with destination d at emerging links of node i.
- $v_i^d = \sum_{a \in E(i)} v_a^d$ is the total inflow through node $i \in N$ with destination $d \in D$.
- From previous scalar flows, the following flow vectors (expressed in bolds) can be defined: the per-destination flow vector $\mathbf{v}^d = (..., \mathbf{v}_i^d, ...;\ i \in N) \in \mathbb{R}_+^{|A|}$, $d \in D$ and the vector of link flows $\mathbf{v} = (..., \mathbf{v}^d, ...;\ d \in D) \in \mathbb{R}_+^{|A|\,|D|}$. From this, the vector of total flows on links $v = \sum_{d \in D} \mathbf{v}^d \in \mathbb{R}_+^{|A|}$ with components $v_a = \sum_{d \in D} v_a^d,\ a \in A$, with vector v of total link flows being the only one expressed without bold. However, if considered convenient, instead of v, the function $\mathrm{w}(\mathbf{v}) = \sum_{d \in D} \mathbf{v}^d$ for returning total link flows will be used.

The per-destination flow vector $\mathbf{v}^d$ will satisfy the flow balance constraints at nodes defining the following set V^d of feasible flows for each destination $d \in D$:

$$\mathrm{V}^d \triangleq \left\{ \mathbf{v}^d \in \mathbb{R}_+^{|A|} \;\middle|\; \sum_{a \in E(i)} v_a^d - \sum_{a \in I(i)} v_a^d = g_i^d,\ \ i \in N \right\},\ d \in D \tag{1}$$

Then, the feasible set $\mathcal{V}$ of total link flows v can be defined as:

$$\mathcal{V} \triangleq \left\{ v \in \mathbb{R}_+^{|A|} \;\middle|\; v_a = \sum_{d \in D} v_a^d \le c_a,\ a \in A_e\ \ \mathbf{v}^d \in \mathrm{V}^d \right\} \tag{2}$$

and the feasibility set for the congested transit equilibrium problem can be formulated as:

$$\mathrm{V} \triangleq \bigotimes_{d \in D} \mathrm{V}^d \tag{3}$$

Boarding links a from stop i are associated with a transit line and a non-negative effective frequency function $f_a(\cdot) : \mathcal{V} \to \mathbb{R}$ for that line (referred to as efff in the following), which will depend on the total flows of some neighbor links and on the total flow of link a. Effective frequencies are the inverse of the expected waiting times at the stop until boarding a vehicle on the line. Mean waiting time functions for a boarding (referred to as mwtf in the following) will be denoted by $\sigma_a(\cdot) = 1/f_a(\cdot)$. Travel times on links will be modeled by general functions $t_a(v)$, $a \in A$. The properties required by these functions are wide enough so as to deal with practical situations and are stated in assumptions 2.2, 2.3 and 2.4, in next Subsection 2.2.

Also, in the way that the C3F model has been formulated in Codina (2013) and in this paper, mean waiting times and travel times are weighted both with a factor of 1, whereas it is known that passengers perceive waiting times 2 to 3 times greater than travel times. This factor of perception, say κ, between travel times and mean waiting times could be incorporated more explicitely without any difficulty into the model by considering mwtf's of the form $\kappa\sigma_a(\cdot)$ instead of simply $\sigma_a(\cdot)$. We have prefered to use $\sigma_a(\cdot)$ for simplicity and wihtout any loss of generality.

The set of stops $\hat{N} \subseteq N$ can also be viewed as nodes for which emerging links exist with a bounded from above (finite) efff on $\mathcal{V}$.

$$\hat{N} \triangleq \{\, i \in N \mid \exists a \in E(i),\ f_a(\cdot) < +\infty \,\} \tag{4}$$

The sets $\hat{N}_d = \hat{N} \setminus \{d\}$, $d \in D$ and $\hat{A} = \{\, a \in A \mid \exists i \in \hat{N},\ a \in \hat{E}(i) \,\}$ will also be used. For nodes $i \in N$, the subset of emerging links with finite effective frequency will be denoted by $\hat{E}(i)$:

$$\hat{E}(i) \triangleq \begin{cases} \{\, a \in E(i) \mid f_a(\cdot) < +\infty \,\}, & i \in \hat{N} \\ \emptyset, & i \in N \setminus \hat{N} \end{cases} \tag{5}$$

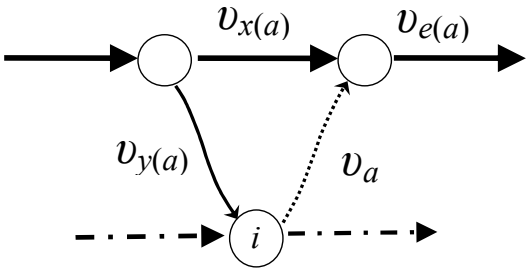


Figure 1: Links with total flows intervening in the efff $f_a(\cdot)$ for boarding link a.

Line segments as well as pedestrian, transfer and non-transit facilities will be represented by links $a \in A$ with either constant or flow-dependent travel time functions $t_a(\cdot)$ and infinite frequencies $f_a = +\infty$. The same will apply for links $a \in I(i)$ and $i \in \hat{N}$, which represent alighting at stops.

The following notation will be used for formulating a V.I. If $C \subseteq \mathbb{R}^n$ is a convex set, then for a function $\Phi : \mathbb{R}^n \to \mathbb{R}^n$, a V.I. will be stated in the Hartmann-Stampacchia form, i.e., find $x \in C$ so that $\Phi(x)^\top (x' - x) \geq 0,\ \forall x' \in C$. This VI will also be referred to for brevity as $\mathsf{VI}(\Phi, C)$. In expressions describing optimization problems immediately after a constraint, dual variables or multipliers for that constraint may appear after a bar "|" when considered convenient. Thus, $v \leq wr,\ | \theta$ indicates that θ is the dual variable for constraint $v \leq wr$. For the convex polyhedron V in (3), its set of vertices is denoted by $\mathsf{vertex}(\mathrm{V})$. Also, the convex hull of all vertices in the set V will be denoted by $\hat{\mathrm{V}} = \mathsf{conv}(\mathsf{vertex}(\mathrm{V}))$.

2.2 A brief excerpt from the formulation of the model as a variational inequality problem

Generally, functions that model effective frequencies $f_a(\cdot)$ at stops depend on total flows on links close to the boarding link a. When the C3F model must reflect the strict limitations of vehicle capacities, it seems logical that effective frequencies depend on total flows of links a, $e(a)$, $x(a)$ and $y(a)$, which are associated with boarding link $a \in A$, as shown in Figure 1 above. Passenger flows on these links should be non-negative and verify that:

$$v_a + v_{x(a)} = v_{e(a)} \tag{6}$$
$$v_{x(a)} \leq c_{e(a)} \tag{7}$$
$$v_{e(a)} \leq c_{e(a)} \tag{8}$$

where $c_{e(a)}$ is the total capacity associated with maximum passenger flows at in-vehicle link $e(a)$, usually calculated as the product of the vehicle's passenger capacity times the number of services. The term *capacity-feasible* will refer to a vector $v = \sum_{d \in D} \mathbf{v}^d$ of total flows on links originated by some per-destination link flow vectors $\mathbf{v}^d \in \mathrm{V}^d$, $d \in D$ verifying constraints (7) and (8).

In general, the ratio ρ_a at boarding link $a \in \hat{E}(i)$, $i \in \hat{N}$, can be interpreted as the loading or traffic factor of passenger queues that are boarding at stop node i. Using flows on links a, $e(a)$, $x(a)$ it will be expressed as

$$\rho_a(v) = \frac{v_a}{c_{e(a)} - v_{x(a)}} \tag{9}$$

For feasible link flows, $0 \leq \rho_a \leq 1$. If passenger flows verify constraint (7) but do not verify constraint (8), then $\rho_a > 1$. In this case, units arrive at the stop, still with capacity *-on average-* but not enough to allocate the amount of boardings, which overflows that capacity. If constraint (7) is violated (and consequently constraint (8) is violated too), then $\rho_a < 0$. In this case, units arrive *-on average-* to stops that already exceeded capacity (i.e., they are "more than full"). According to this, it seems reasonable to consider functional forms for effective frequencies of the type:

$$f_a(v) = \begin{cases} \phi_a(\rho_a(v)) & \text{if } \rho_a \geq 0 \\ \epsilon & \text{if } \rho_a < 0 \end{cases} \tag{10}$$

where ϵ is some pre-specified small number and $\phi_a(\cdot)$ is a non-negative decreasing function such that $\phi \geq \epsilon$ and $\phi(1) = \epsilon$. For example, Cepeda *et al.* (2006) consider in their numerical tests efff, for which $\phi_a(x) = \max\{\epsilon, \kappa(1 - x^\beta)\}$.

Thus, when capacity-feasible flows v are close to capacities, or equivalently when the slack variables in constraints (7) and (8) are small, effective frequency functions such as those defined in (10) act as barriers that prevent capacities from being exceeded by the solution of the C3F model. Based on this, the solutions provided by the algorithm in Cepeda et al. (2006) are expected to be capacity-feasible, despite there being no explicit mechanism preventing the violation of capacities. In this paper, it is shown that the algorithm in Cepeda et al. (2006), actually complies with these expectations, but only up to either a given degree marked by the level of congestion in the network or at some points in it. When the demand level is high, the solutions of Cepeda et al.'s algorithm may systematically violate to a great extent the capacity constraints at one or more bottlenecks. This is not the case with the algorithm presented in this paper, as it guarantees that solutions will always be capacity-feasible no matter the level of demand applied to the network. In addition, it will also verify equilibrium conditions.

Because constraints (8) are verified at any time by feasible flows, only explicit bounds on total flows on line segments will be considered in the formulation, i.e., $v_{e(a)} \leq c_{e(a)}$. Since constraints (7) are clearly redundant when (6) and (8) are verified, if A_a is the subset of boarding links, we will consider the polyhedron X

$$X \triangleq \{\, \mathbf{v} \in \mathbb{R}^{|D||A|} \mid v_{e(a)} \leq c_{e(a)},\ a \in A_a \,\} \tag{11}$$

and its relative interior will be defined by $\mathsf{rint} X = \{\, \mathbf{v} \in \mathbb{R}^{|D||A|} \mid v_{e(a)} < c_{e(a)},\ a \in A_a \,\}$.

Also, for a formal presentation of the model, let S_i^d be the simplex sets $S_i^d = \left\{ \alpha \in \mathbb{R}_+^{|\hat{E}(i)|} \mid \sum_{a \in \hat{E}(i)} \alpha_a = 1 \right\}$ associated with a stop node $i \in \hat{N}_d$ and destination $d \in D$ and let $S = \bigotimes_{d \in D} \bigotimes_{i \in \hat{N}_d} S_i^d$.

In Codina (2013), the C3F model without explicit capacity constraints is formulated as a variational inequality problem defined on $\mathrm{V} \times S$. As a consequence, the model can be reformulated as a fixed-point inclusion problem from which the MSA-based algorithm by Cepeda et al. (2006) can be derived directly. Also in Codina (2013), an extension of the model for the case of explicit capacity constraints is analyzed and -under some assumptions indicated below for completeness- it is proved that their solutions always remain below capacity. This model can also be formulated as a variational inequality, but defined on $(\mathrm{V} \cap X) \times S$, i.e., by explicitly taking into account the capacity constraints. As in the previous case without explicit capacity constraints, an MSA-based algorithm can be derived directly by considering the reformulation of this variational inequality as a fixed-point inclusion problem. A brief presentation of the theoretical framework for the model with explicit constraints follows:

The C3F model with explicit capacity constraints can be defined by means of a variational inequality $\mathit{VI}(F, (\mathrm{V} \cap X) \times S)$, where the functional $F(\cdot,\cdot) : (\mathrm{V} \cap X) \times S \to \mathbb{R}^m$ is defined as:

$$F(\mathbf{v},\zeta) = \begin{pmatrix} -\frac{F_v(\mathbf{v},\zeta)}{F_\zeta(\mathbf{v})}- \end{pmatrix} = \begin{pmatrix} t_a(v) + \zeta_a^d \sigma_a(v) & ;\ a \in \hat{E}(i),\ i \in \hat{N}_d,\ d \in D \\ t_a(v) & ;\ a \in E(i) \setminus \hat{E}(i),\ i \in N_d,\ d \in D \\ \hline -\sigma_a(v)\, v_a^d & ;\ a \in \hat{E}(i),\ i \in \hat{N}_d,\ d \in D \end{pmatrix} \tag{12}$$

where $t_a(\cdot)$ and $\sigma_a(\cdot)$, as defined in subsection (2.1), are functions for the mean travel time and the mean waiting time for a boarding link a at a stop. The dimension m of the variational inequality is given by $m = |D| \cdot |A| + \sum_{d \in D} \sum_{i \in \hat{N}_d} |\hat{E}(i)|$. This variational inequality problem will be referred to as the C3F VI problem in the following. In its Hartmann-Stampacchia form, it will be stated as:

$$\begin{gathered} \text{Find } (\mathbf{v}^*, \zeta^*) \in (\mathrm{V} \cap X) \times S \text{ such that:} \\ F_v(\mathbf{v}^*, \zeta^*)^\top (\mathbf{v} - \mathbf{v}^*) + F_\zeta(\mathbf{v}^*)^\top (\zeta - \zeta^*) \geq 0,\ \ \forall (\mathbf{v}, \zeta) \in (\mathrm{V} \cap X) \times S \end{gathered} \tag{13}$$

At this point it should be taken into account that the various sets of feasible flows defined previously in subsection 2.1 are polyhedra that not necessarily bounded. That depends mainly on the configuration of the expanded transit network. Polyhedron V in (3) is not bounded if there are cycles in the graph. Moreover, polyhedrons $\mathcal{V}$ (total link flows) and $\mathrm{V} \cap X$ may not be bounded, although they include the constraints $v_a \leq c_a,\ a \in A_e$, thus leading to a perception of them as bounded at first glance. For instance, although any line segment in the expanded transit network may have the corresponding capacity constraint, there are also other types of links; amongst them are those used to model connections or pedesterian movements, and it may be well the case that loops exist in the expanded transit network due to these links.

In Codina (2013) it is shown that the following assumptions 2.1 to 2.4 guarantee the existence of solutions for the C3F VI problem in (13).

Assumption 2.1 *We assume that* $\hat{\mathrm{V}} \cap \mathsf{rint}\, X \neq \emptyset$. *In other words, that acyclic, per-destination, capacity-feasible flows exist, such that the corresponding total link flows are below capacity at in-vehicle links* $a \in A_e$.

Assumption 2.2 *Functions* $t_a(\cdot),\ a \in A$ *and* $\mathsf{mwtf}\ \sigma_a(\cdot),\ a \in \hat{E}(i),\ i \in \hat{N}$ *are continuous and positive on* $\mathrm{V} \cap \mathsf{rint}\, X$.

Assumption 2.3 *For any* $\omega \in W$, *there exists at least a path* γ_ω *on* $G = (N, A)$ *so that any link* $a \in \gamma_\omega$ *verifies that travel time functions* $t_a(\cdot),\ a \in \gamma_\omega$, *and* mwtf*'s* $\sigma_a(\cdot),\ a \in \hat{A} \cap \gamma_\omega$ *are continuous and finite on* $\hat{\mathrm{V}} \cap X$.

In other words, this assumption requires that for each O-D pair there exist a path with the role of "blowoff valve", on which travel time functions and mwtf's at boarding links (if the path contains any) remain finite on the bounded polyhedron $\hat{\mathrm{V}} \cap X$ of acyclic, capacity-feasible flows. These paths will be referred to in this paper as the "escape paths".

Let us consider the following two subsets of links in the expanded network associated with the vector of per-destination flows $\mathbf{v}$:

$$A_\sigma^\infty(\mathbf{v}) \triangleq \{\ a \in \hat{A} \mid \sigma_a(v) = +\infty\ \} \tag{14}$$

$$A_t^\infty(\mathbf{v}) \triangleq \{\ a \in A \mid t_a(v) = +\infty\ \} \tag{15}$$

Assumption 2.4 *For any* $\mathbf{v} \in \{\mathbf{v}' \in \mathrm{V} \cap X \,|\, \exists a \in A_e,\ v'_a = c_a\}$, $A_t^{\infty}(\mathbf{v}) \neq \emptyset$ *and/or* $A_\sigma^{\infty}(\mathbf{v}) \neq \emptyset$.

In other words, if for a capacity-feasible per-destination vector of flows, its corresponding vector of total flows v is at capacity at some in-vehicle link $a \in A_e$, then at least a travel time function or a mwtf σ_a evaluated at $\mathbf{v}$ becomes ∞. The following theorem explains how the solutions of the model verify the capacity constraints and accordingly which of the previous assumptions 2.1 to 2.4 are verified.

Theorem 2.5 *(Codina (2013)) Under conditions stated in assumption (2.1), then*

1. *Suppose that travel time functions* $t_a(\cdot)$, $a \in A$, *and* efff *'s* $\sigma_a(\cdot)$, $a \in \hat{A}$ *are positive and continuous on* $\mathrm{V} \cap X$. *Then, the C3F VI problem in (13) has at least a solution* $\mathbf{v}^*$ *and any of them is such that* $\mathbf{v}^* \in \hat{\mathrm{V}} \cap X$.

2. *If assumptions* 2.2, 2.3 *and* 2.4 *also hold, then the C3F VI problem in (13) has at least one solution, and any of its solutions* $\mathbf{v}^*$ *is a) an equilibrium flow for the C3F model, is such that b)* $\mathbf{v}^* \in \hat{\mathrm{V}} \cap \mathsf{rint}\, X$.

In both cases origin-destination travel times are finite independently of the demands g_ω, $\omega \in W$.

The practical implications of the previous theorem 2.5 are discussed at the beginning of Section 3, next. The reader must notice the parallelism between the properties stated in previous theorem 2.5 for the C3F model under assumptions 2.1 to 2.4 and the barrier methods used for solving optimization problems, and particularly, diagonal traffic assignment problems, in Daganzo (1977a,b) and Nie et al. (2004).

As stated in Codina (2013), the proper gap function for the C3F VI problem in (13) must be $\hat{G}_{\texttt{C3F}} : \hat{\mathrm{V}} \cap X \to \mathbb{R}$, defined as:

$$\hat{G}_{\texttt{C3F}}(\mathbf{v}) \triangleq G^{(0}_{\texttt{C3F}}(\mathbf{v}) - \hat{G}^{(1}_{\texttt{C3F}}(v) \tag{16}$$

where $G^{(0}_{\texttt{C3F}}(\mathbf{v}) = \varphi(\mathbf{v}, v)$, $\quad \hat{G}^{(1}_{\texttt{C3F}}(v) = \mathsf{Min}_{\ \mathbf{u} \in \mathrm{V} \cap X}\ \varphi(\mathbf{u}, v)$ and function $\varphi(\cdot,\cdot)$ is defined as:

$$\varphi(\mathbf{u}, v) \triangleq \sum_{d \in D} \left[\sum_{a \in A} t_a(v) u_a^d + \sum_{i \in \hat{N}_d} \mathsf{Max}_{a \in \hat{E}(i)} \left\{ \frac{u_a^d}{f_a(v)} \right\} \right] \tag{17}$$

The following corollary 2.6 of the previous theorem 2.5 is also a direct consequence of the previous results and it is the cornerstone of the algorithm presented in this paper.

Corollary 2.6 (*Codina (2013). Characterization of C3F equilibrium as a fixed point inclusion*). *Under conditions of theorem* 2.5, $\mathbf{v}^*$ *is a solution of the C3F VI problem in (13) iff the following fixed point inclusion is verified:*

$$\mathbf{v}^* \in \mathsf{Sol\ Min}_{\ \mathbf{u} \in \mathrm{V} \cap X}\, \varphi(\mathbf{u}, \mathrm{w}(\mathbf{v}^*)) \tag{18}$$

where $\mathrm{w}(\mathbf{v}^*) = \sum_{d \in D} \mathbf{v}^{d*}$.

The problem defining the fixed point inclusion in (18) is the following linear problem $\texttt{[CapPL]}(1/\sigma(u), t(u))$. This problem can be viewed as a capacitated uncongested strategy-based transit assignment model. This problem $\texttt{[CapPL]}(r, t)$ is defined simply as:

$$\texttt{[CapPL]}(r,t) \quad \begin{array}{ll} \mathsf{Min}_{\ \mathrm{v},\, w} & \displaystyle\sum_{d \in D} \sum_{a \in A} t_a v_a^d + \sum_{d \in D} \sum_{i \in \hat{N}} w_i^d \\ s.t: & v_a^d \leq r_a w_i^d,\ a \in \hat{E}(i),\ i \in \hat{N},\ d \in D \\ & \mathrm{v} \in \mathrm{V} \cap X \end{array} \tag{19}$$

3 Solving the capacitated C3F problem

The previous corollary 2.6 states that solving the strict capacitated transit assignment problem is equivalent to finding a solution of the fixed point inclusion (18). The authors have considered several alternatives for solving it. On the one hand, these types of problems are widely solved with Krasnoselskii-Mann iterative schemes (Krasnoselskii (1954), Mann (1953)). An enhancement was proposed by Ishikawa (Ishikawa (1974)) that achieved a better rate of convergence and it consisted of a double step which required two evaluations of the point-to-set mapping at each iteration. These types of iterative schemes were initially developed by Blum (1954) and Robins and Monro (1951) and are also used for fixed-points of single-valued maps and for solving stochastic equations. In transportation modelling, they are widely known under the name Methods of Successive Averages (MSA). These iterative methods can be summarized as:

$$\mathbf{v}^{(k+1} = \mathbf{v}^{(k} + \alpha_k \left\{ T\left(\mathbf{v}^{(k} + \beta_k(\ T(\mathbf{v}^{(k}) - \mathbf{v}^{(k}\ )\right) - \mathbf{v}^{(k}\right\} \tag{20}$$

where the point-to-set map T in our case would be defined by the solutions of the parametrized linear program [CapPL]. The convergence of these iterative methods to a fixed-point has been proved under different hypotheses for the sequences α_k, β_k and some non-expansiveness properties of the map T (see, for instance Dunn (1978) and Maiti and Ghosh (1989)). For the case $\beta = 0$ the Krasnoselskii-Mann (or MSA) iterative scheme is obtained. In general, the Ishikawa iterative scheme has a better convergence rate than Mann's method in terms of the number of iterations. However, Ishikawa's method requires two evaluations of the point-to-set map per iteration, whereas Mann's method requires only one. In our problem, evaluating the point-to-set map is equivalent to solving the linear program [CapPL] in (19), which is computationally expensive. After some tests it was shown that Ishikawa's method did not present any significant improvement in terms of CPU time. On the other hand, another possibility that appeared as a natural option for obtaining the step length α_k would be to previously perform a line search on some merit function. Thus, we also implemented a variant of the iterative algorithm (20) (with $\beta_k = 0$) that included a line search to determine a value for α_k. Previously this method showed good results for the problem of minimizing the Auslander gap function that arises in an asymmetric traffic assignment problem (see Codina et al. (2015)). However, when applied to the $\hat{G}_{\texttt{C3F}}$ gap function it performed too many function evaluations resulting in an overall decrease of the algorithm performance.

After several tests, the most efective method for the problem under consideration has been the MSA method, which is described in detail below. Corollary 2.6 also states that the non-negative gap function $\hat{G}_{\texttt{C3F}}(\mathbf{v})$, which is shown in (16), vanishes at these solutions. Thus, a measure for monitoring the proximity of a point $\mathbf{v}$ to a solution of the problem will be the following relative gap:

$$\hat{g}(\mathbf{v}) = \hat{G}_{\texttt{C3F}}(\mathbf{v}) / \hat{G}^{(1}_{\texttt{C3F}}(v) \tag{21}$$

C3F type Algorithm for solving the strict capacitated transit assignment problem

Init : - For an initial value $f^{(0} = \phi_0$ of frequencies and $t^{(0}$of link travel times
solve [CapPL]$(f^{(0}, t^{(0})$ (Volumes $\mathbf{v}^{(0}$ are obtained on network links;)
- Calculate $\hat{g}(\mathbf{v}^{(0}) = \hat{G}_{\texttt{C3F}}(\mathbf{v}^{(0}) / \hat{G}^{(1}_{\texttt{C3F}}(v^{(0}); \ \ k = 0$

While : $(\ \hat{g}(\mathbf{v}^{(k}) \geq \epsilon\ )$ **do** :

1- Calculate new frequencies $f_a^{(k} = \phi_a(v^{(k}),\ a \in \hat{E}(b),\ b \in \hat{N}_G$;
and link travel times $t_a^{(k} = t_a(v^{(k})$

2- Solve [CapPL]$(f^{(k}, t^{(k}) \ \rightarrow$ (volumes $\hat{\mathbf{v}}^{(k}$ are obtained)

3- Update: $\mathbf{v}^{(k+1} = (1 - \alpha_k)\mathbf{v}^{(k} + \alpha_k \hat{\mathbf{v}}^{(k}$; (MSA step)

4- $k \leftarrow k + 1$; Evaluate $\hat{g}(\mathbf{v}^{(k}) = \hat{G}_{\texttt{C3F}}(\mathbf{v}^{(k}) / \hat{G}^{(1}_{\texttt{C3F}}(v^{(k})$

EndWhile

The rate of convergence of the Mann-type algorithm was determined in Dunn (1978) to be $\| \mathbf{v}^{(k} - \mathbf{v}^* \|_2 = O(k^{-1/2})$ if the step lengths follow the classical MSA $\alpha_k = 1/k$. The step length parameter at $k-$iteration, α_k, can be chosen in a variety of other ways. After some experimentation (not reported in Section 7), two of the possibilities introduced in Liu et al. (2009) were considered: a) the method of successive

weighted averages (MSWA) and b) the self-regulated MSA. The method of successive weighted averages determines at each iteration k the step length as:

$$\alpha_k = \frac{k^\nu}{\sum_{i=1}^{k} i^\nu} \tag{22}$$

with ν being a fixed integer. In taking the self-regulated MSA method of Liu et al. (2009) and adapting it to the congested transit assignment problem the step length at iteration k is as follows:

$$\begin{array}{ll} \textsf{If } \|\hat{\mathbf{v}}^{(k} - \mathbf{v}^{(k}\|_2 \geq \|\hat{\mathbf{v}}^{(k-1} - \mathbf{v}^{(k-1}\|_2 & \textsf{then } \beta_k = \beta_k + \Gamma; \\ & \textsf{else } \beta_k = \beta_k + \gamma; \\ \textsf{EndIf} & \\ \alpha_k = 1/\beta_k & \end{array}$$

where, as stated in Liu et al. (2009), $1.5 \leq \Gamma \leq 2$ and $0.01 \leq \gamma \leq 0.5$. As will be shown in Section 8, the self-regulated method performs better in the initial iterations but both methods have a similar performance if the accuracy requirements are high. Also, the rate of convergence of the Mann-type iteration with the step length (22) is illustrated empirically in Section 8.

It must be pointed out that in this case, the hyper-path calculating algorithm in Spiess and Florian (1989) cannot be used for solving the more complex [CapPL](r,t) due to capacity constraints. It will be shown through the computational tests in Section 7, that this does not imply any additional computational burden in solving the capacitated transit assignment problem. Notice that any vector flow generated by the algorithm verifies the capacity constraints, $v_a^{(k} \leq c_a,\ a \in A_e$.

In the end the previous algorithm can be viewed simply as a modification of the algorithm in Cepeda et al. (2006) for the congested transit assignment problem without explicit capacity constraints, where -at step 2 within the while loop- the uncapacitated problem [PL](r,t) is solved:

$$\texttt{[PL]}(r,t) \quad \begin{array}{ll} \textsf{Min}_{\ \mathbf{v},\, w} & \displaystyle\sum_{d\in D}\sum_{a\in A} t_a v_a^d + \sum_{d\in D}\sum_{i\in \hat{N}} w_i^d \\ s.t: & v_a^d \leq r_a w_i^d,\ a \in \hat{E}(i),\ i \in \hat{N},\ d \in D \\ & \mathbf{v} \in \mathrm{V} \end{array} \tag{23}$$

Because of this, in the following we will refer to the capacitated C3F algorithm and to the uncapacitated C3F algorithm, depending on whether the problem solved at step 2 is either problem [CapPL] or problem [PL] respectively.

Because the algorithm presented in this paper solves problem [CapPL] instead of problem [PL] any of its solutions is capacity-feasible. On the other hand, the method of Cepeda et al. (2006) may present serious difficulties in finding capacity-feasible solutions for large networks when the level of congestion is high, as shown in the computational experiments of Section 7. At initial iterations, some boarding links still have a frequency not far from the nominal (or uncongested) and then the subproblem provides extreme points in the space of flows of the uncapacitated polyhedron, which violate capacities. The higher the demand, the higher the violation of capacities, because the magnitude of these flows originated in the subproblems of the starting iterations. As the iterations progress, these extreme points lose weight in the resulting flows, but it may require of iterations to eliminate the effect of these initial extreme points. If the congestion level (or the demand level) is not too high, then these extreme points may dissappear from the flows more or less soon; but if the demand level is high it may be virtually impossible that they dissappear.

For the uncapacitated C3F algorithm, the following gap function can be defined:

$$G_{\texttt{C3F}}(\mathbf{v}) = G^{(0}_{\texttt{C3F}}(\mathbf{v}) - G^{(1}_{\texttt{C3F}}(v) \tag{24}$$

where $G^{(0}_{\texttt{C3F}}(\mathbf{v}) \equiv \hat{G}^{(0}_{\texttt{C3F}}(\mathbf{v})$ and $G^{(1}_{\texttt{C3F}}(v) = \textsf{Min}_{\ \mathbf{u}\in \mathrm{V}}\ \varphi(\mathbf{u}, v)$ and thus, analogously to (21), the corresponding relative gap function was considered:

$$g(\mathbf{v}) = G_{\texttt{C3F}}(\mathbf{v}) / G^{(1}_{\texttt{C3F}}(v) \tag{25}$$

For solving problems where the efff at saturated boarding links a vanishes when $\rho_a(v) \approx 1$, one approach can be to modify the efff as given in (10). However, we must consider what is shown in the

computational tests in Section 7, specially regarding very congested transit networks or simply networks with a subset of stops where the loading factors are close to 1: if the linear problem in step 2 is solved in its uncapacitated version, it may happen that no iterate $\mathbf{v}^{(k}$ can be found to verify capacity constraints (8) at these congested stops; thus, loading factors may take values much greater than 1 or even be negative. As a result, when solving congested transit assignment problems with strict capacities at transit lines, capacity constraints should be explicitly taken into account.

4 Solving [CapPL] with dual Lagrangian methods

One reason we have considered several candidates for solving the linear problem [CapPL] is that the efficiency in solving it has a direct impact on solving the main problem. In addition, we took into account the fact that, for the uncongested uncapacitated strategy-based model [PL] (in equation (22)), the algorithm by Spiess (1984) has an outstanding computational performance. Moreover, relying on the Spiess algorithm permits straightforward identification of strategies that make up the solution flows in the main problem. Thus, because a Lagrangian relaxation of the capacity constraints, allows problem [CapPL] can be solved by solving a sequence of uncapacitated linear problems [PL], Lagrangian methods seemed to be natural candidates. For maximizing the resulting Lagrangian problem we have tried the classical subgradient algorithm, the cutting-plane method and the bundle method of Lemaréchal (1974). Although our expectations were that the bundle method of Lemaréchal (1974) could more efficiently resolve the maximization of the dual Lagrangian function, the cutting-plane method yielded the best performance.

This section details how the linear problem [CapPL] in (19) is solved with the dual cutting-plane algorithm (see, for instance, Bazaraa (2006), Chapter 6) at each iteration of the capacitated C3F algorithm in Section 3.

This algorithm is applied to a Lagrangian relaxation of constraints $v_a \leq c_a, a \in A_e$, in order to maximize the corresponding dual Lagrangian function defined below in (26). Assume that, at the k-th iteration of the previous capacitated C3F algorithm, the vector of link volumes is $v^{(k}$. Then, the link travel times and frequencies that apply to problem [CapPL] must be $t_a(v^{(k})$, $a \in A$ and $f_a(v^{(k})$, $a \in A_e$. Now let $t_a^{(k} = t_a(v^{(k})$ and $f_a^{(k} = f_a(v^{(k})$. Thus, the corresponding dual Lagrangian function $\omega^{(k}(\cdot)$ that is defined on the vector of non-negative multipliers μ_a, $a \in A_e$ is associated with capacity constraints and it must be:

$$\begin{aligned} \omega^{(k}(\mu) = \mathsf{Min}_{\,\mathbf{v},w} \quad & \sum_{a\in A} t_a^{(k} v_a + \sum_{d\in D}\sum_{i\in N_d} w_i^d - \sum_{a\in A_e} \mu_a(c_a - v_a) \\ & v_a^d \leq f_a^{(k} w_i^d, \quad a \in \hat{E}(i), \quad i \in \hat{N},\ d \in D \\ & \mathbf{v} \in \mathrm{V} \end{aligned} \tag{26}$$

It should be noted that linear problem (26) which defines the dual Lagrangian function $\omega^{(k}(\cdot)$, consists of a fixed frequency transit assignment problem with non-negative costs $t_a(v^{(k}) + \mu_a$, $a \in A_e$. This can be solved efficiently by using the Spiess algorithm in Spiess (1984).

The dual cutting-plane algorithm must be started at an initial feasible point of the dualized and non-dualized constraints, i.e., a capacity-feasible flow vector in our case. When problem [CapPL] needs to be solved for the first time, and if the transit network model verifies assumption 2.3, a capacity-feasible starting point may be easily found by assigning O-D flows on the paths where travel time functions and efff's are continuous and remain finite. This can be easily accomplished if the transit network model contains a subset of pedestrian links connecting O-D pairs. Because the vector of per-destination flows $\mathbf{v}^{(k}$ is capacity-feasible in subsequent iterations of the capacitated C3F algorithm, it is easy to use it as a feasible point in which the next problem [CapPL], for which the total waiting time at stops $w_i^{d,(k+1}$ can always be properly obtained as: $w_i^{d,(k+1} = \mathsf{max}_{a\in\hat{E}(i)}\{\, v_a^{d,(k}/f_a(v^{(k})\,\},\ i \in \hat{N},\ d \in D$.

In the description of the dual cutting-plane algorithm for the $k-th$ iteration of the capacitated C3F algorithm, superscript k will be omitted for clarity from the variables v, w and multipliers μ. Thus, instead of $v^{k,\ell}, w^{d(k,\ell}$..., we will simply use $v^{\ell}, w^{d(\ell}$...,. Let also $\tilde{t_a} = t_a^{(k}$.

0. Let $v_a^{(0} \leq c_a$, $a \in A_e$, $w_i^{d(0}$, $d \in D$ be an initial feasible solution. Set ℓ=0. STOP=.false.

1. While $\ell \leq N_{max}$ & not STOP

(a) Master Problem:

$$z^{(\ell} = \mathsf{Max}_{z,\,\mu} \quad z$$
$$s.t.\ z \leq \sum_{a\in A} \tilde{t_a} v_a^{(j} + \sum_{d\in D}\sum_{i\in N_d} w_i^{d(j} - \sum_{a\in A_e} \mu_a (c_a - v_a^{(j)} \ \mid \ \lambda_j^{(\ell},\ \ j = 0,\ldots,\ell \qquad (27)$$
$$\mu_a \geq 0,\ \ a \in A_e$$

where $v_a^{(j}, w_i^{d(j}, j = 1, 2, \ldots, \ell$ are the solutions obtained in previous subproblems, and for $j = 0$, $v_a^{(0}$ is the initial capacity-feasible vector of total flows.

(b) Subproblem: Using multipliers $\mu_a^{(\ell}, a \in A_e$, which have been obtained after solving the previous master problem, the value of $\omega^{(k}(\mu^{(\ell})$ is calculated by solving linear problem (26). Its solution, $\mathbf{v}^{(\ell+1}, w_i^{d(\ell+1}$, $i \in \hat{N}_d$, $d \in D$, is a new vector of flows and the total waiting times at stops, which will define a new constraint for the next master problem at iteration $\ell + 1$.

(c) Let STOP $= \left(\dfrac{z^{(\ell} - \omega^{(k}(\mu^{(\ell})}{\omega^{(k}(\mu^{(\ell})} \ \leq \epsilon \right)$

After solving the last master problem at iteration $L \leq N_{max}$, an approximated primal solution to problem [CapPL]$(f^{(k}, t^{(k})$ can be obtained directly by using the Lagrange multipliers $\lambda^{(L}$ which are found in the last master problem solved (see, for instance Bazaraa (2006), Theorem 6.5.2, Chapter 6)

$$\hat{\mathbf{v}} = \sum_{j=0}^{L-1} \lambda_j^{(L} \mathbf{v}^{(j}. \qquad (28)$$

Notice that, because of Lagrange multipliers $\lambda_j^{(L}$ are associated to the inequality constraints in (27), then $\lambda_j^{(L} \geq 0$ and $\sum_{j=0}^{L-1} \lambda_j^{(L} = 1$. Because of that, they will be referred to as baricentric coordinates of the solution $\hat{\mathbf{v}}$.

We have considered other candidates for solving the linear problem [CapPL] because the efficiency on solving it has a direct impact in solving the main problem. Also, we took into account the fact that, for the uncongested uncapacitated strategy-based model [PL] (in equation (23)), the algorithm by Spiess (1984) has an outstanding computational performance. Moreover, relying on the Spiess algorithm permits straightforward identification of strategies that make up the solution flows in the main problem, as discussed next in Section 6. Thus, because a Lagrangian relaxation of the capacity constraints allows problem [CapPL] to be solved by solving a sequence of uncapacitated linear problems [PL], Lagrangian methods appeared as natural candidates. In addition to the cutting-plane method for solving the Lagrangian problem of maximizing the function $w^{(k}(\mu)$, we have also tried the subgradient approach and the bundle method of Lemaréchal (1974) combined with the proximity parameter updating method by Kiwiel (1990). Although our expectations were that the bundle method in Lemaréchal (1974) could more efficiently resolve the maximization of the dual Lagrangian function, the cutting-plane method yielded the best performance. For brevity, and also because they would be more appropriate for a journal dedicated to computational algorithms, we have not reported these tests in the paper.

5 An interpretation of problem [CapPL]

As stated earlier, the algorithm presented in this paper for solving the capacitated strategy-based transit assignment model allows for the inclusion of congestion effects by using queuing models for passenger delays at stops. Problem [CapPL] in particular can be interpreted in this framework if it is assumed that delays at stops are modeled by a bulk service queue. It is well known that if d_0 is the headway?s average value and C_τ is its coefficient of deviation, then the mean waiting time of a single arriving passenger is given by $d_0(1 + C_\tau^2)/2$. Then, if the average waiting time for a given non-null flow of arriving passengers is d, the ratio d/d_0 can be considered as a normalized waiting time. Figure 2 shows the normalized waiting time d/d_0 of clients in a bulk service queue $M/D^{[c]}/1$ as a function of its traffic factor ρ. Arrival of passengers to the stop are Poissonian and are picked up instantaneously by a server (transport unit) with maximum capacity for C passengers. Arrivals of these units are completely regular, with constant inter-arrival times of 5 minutes and a random number c of available seats (capacity of service), $0 \leq c \leq C$.

Figure 2 corresponds to values of $C = 200$ pax. and headway=5 min. Curves 1, 2 and 3 correspond to c, distributed accordingly to a truncated normal, $0 \leq c \leq 200$, with mean $\mu = 50$ and deviations $\sigma_i = 3, 6, 10$, for $i = 1, 2, 3$. Curve 4 corresponds to the case of c distributed uniformly in $[0, 200]$. Delay d would correspond to function $1/\phi(\rho)$, where $\phi(\rho)$ appears in formula (10). It can be observed that the effects of congestion arise only when traffic factors are very close to 1. Clearly, problem [CapPL] reflects an extreme situation in which units arrive at the stop with fixed capacity, and in this case congestion effects appear suddenly for a loading factor of ρ=1, for which the effective frequency becomes null

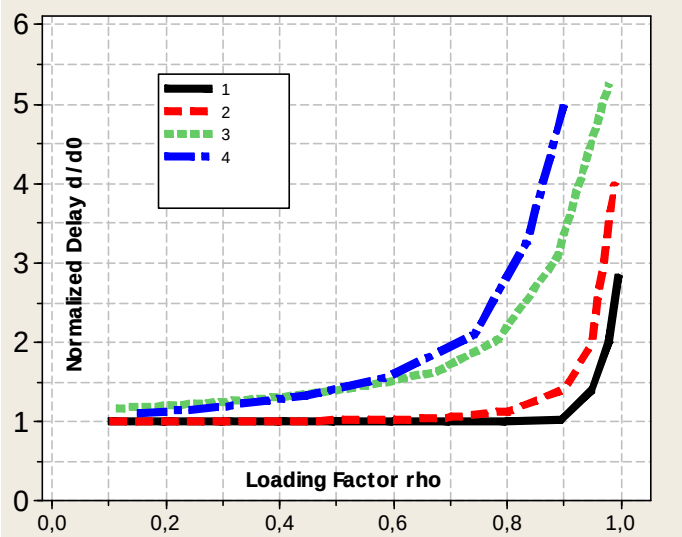


Figure 2: Delay d/d_0 versus loading factor ρ in a $M/D^{[c]}/1$ queue. Curves 1, 2 and 3 correspond to the server's capacity c distributed accordingly to a truncated normal, $0 \leq c \leq 200$, with mean $\mu = 50$ and deviations $\sigma_i = 3, 6, 10$, for $i = 1, 2, 3$. Curve 4 corresponds to the case of c distributed uniformly in $[0, 200]$.

6 The strategies used by the solution

A common requirement in planning studies is to perform path-analysis of the solutions that are generated by transit assignment models. In the case of strategy-based models, this requires using the hyper-paths on which the solution relies. An advantage of using the Lagrangian decomposition for solving problem [CapPL] is that it permits using the hyper-path-based algorithm of Spiess (1984) to calculate its solutions. The previous equation (28) expresses the solution v of problem [CapPL] in terms of the flows $\mathbf{v}^{(j}$, each of them corresponding to a set of strategies for the entire transit network.

This implies that an implementation of the transit assignment algorithm in Section 3, with the added capability of supporting path analysis, should keep track of the sequence of the various per-destination flows $\mathbf{v}^{(j}$ generated during the execution of the algorithm in Section 3, as well as their resulting weights. These weights should be obtained through weights $\lambda_j^{(L}$ after the maximization of the Lagrangian dual function (26) and the step lengths α_k used in step 3 of the algorithm in Section 3.

Any per-destination flow vector $\mathbf{v}_h^d$ generated by the hyperpath-generating linear problem [PL] in (23), can be expressed as:

$$\mathbf{v}_h^d = \sum_{i \in N_d} g_i^d \mathbf{v}_{i,h}^d \tag{29}$$

where $\mathbf{v}_{i,h}^d$ are flow vectors containing the fractions that are used in the links for the demand (i,d) corresponding to hypergraph h. However, as effective frequencies depend on volumes, then the fractions in vectors $\mathbf{v}_h^d$, although corresponding to common hypergraphs, are iteration-dependent.

Let now focus on a per-destination flow vector $\mathbf{v}^{d,(k}$ generated by the algorithm at iteration k. Let $\mathcal{H}_{i,d}^{(k}$ the set of hypergraphs generated by the algorithm solving one or more times problem (26) at each of the k master problems solved up to iteration k.

To see how the hyperpaths of a given solution of algorithm in Section 3 can be computed, assume, without loss of generality, that for a network consisting of a single O-D pair with demand g, the algorithm is started with a solution $\mathbf{v}^{(0} \equiv g\mathbf{v}_h^{(0}$ corresponding to some given hypergraph h. After the first iteration, with a resulting step length of α_1, the corresponding master problem is solved using the dual cutting plane method of Section 4, taking two iterations and yielding $\mathbf{v}_h^{(1}$ and $\mathbf{v}_{h'}^{(1}$ as solutions of the corresponding problems (26) solved at each iteration, with baricentric coordinates λ_1 and λ_1' (i.e., the solution $\hat{\mathbf{v}}^{(1}$ of problem [CapPL] at iteration 1 is then $\hat{\mathbf{v}}^{(1} = g \cdot (\lambda_1 \mathbf{v}_h^{(1} + \lambda_1' \mathbf{v}_{h'}^{(1})$). Next, in a second iteration, assume that the corresponding solutions yielded by problem (26) as part of the master problem of that iteration are: $\mathbf{v}_h^{(2}$, $\mathbf{v}_{h'}^{(2}$, $\mathbf{v}_{h''}^{(2}$ and $\mathbf{v}_{h'''}^{(2}$, with baricentric coordinates λ_2, λ_2', λ_2'' and λ_2''' and $\hat{\mathbf{v}}^{(2} = g \cdot (\lambda_2 \mathbf{v}_h^{(2} + \lambda_2' \mathbf{v}_{h'}^{(2} + \lambda_2'' \mathbf{v}_{h''}^{(2} + \lambda_2''' \mathbf{v}_{h'''}^{(2})$. Then, after the first iteration the solution provided by the algorithm in Section 3 will be expressed as:

$$\mathbf{v}^1 = g \cdot (\mathbf{v}_h^{(0} + \alpha_1(\lambda_1 \mathbf{v}_h^{(1} + \lambda_1' \mathbf{v}_{h'}^{(1} - \mathbf{v}_h^{(0})) = g \cdot ((1-\alpha_1)\lambda_0 \mathbf{v}_h^{(0} + \lambda_1 \mathbf{v}_h^{(1}) + g \cdot \lambda_1' \mathbf{v}_{h'}^{(1} \tag{30}$$

where $\lambda_0 \triangleq 1$. Next,

$$\begin{aligned} \mathbf{v}^2 = \mathbf{v}^1 + \alpha_2(\hat{\mathbf{v}}^2 - \mathbf{v}^1) = ... = \quad & g \cdot ((1-\alpha_1)(1-\alpha_2)\lambda_0 \mathbf{v}_h^{(0} + (1-\alpha_2)\lambda_1 \mathbf{v}_h^{(1} + \alpha_2 \lambda_2 \mathbf{v}_h^{(2}) + \\ & + g \cdot ((1-\alpha_2)\lambda_1' \mathbf{v}_{h'}^{(1} + \alpha_2 \lambda_2' \mathbf{v}_{h'}^{(2}) + \\ & + g \cdot \alpha_2 \lambda_2'' \mathbf{v}_{h''}^{(2} + \\ & + g \cdot \alpha_2 \lambda_2''' \mathbf{v}_{h'''}^{(2} \end{aligned} \tag{31}$$

and in general, for a network with several O-D pairs:

$$\mathbf{v}^k = \sum_{(i,d) \in W} g_i^d \sum_{h \in \mathcal{H}_{i,d}^{(k}} \left[\sum_{\ell=\kappa(h)}^{k} \lambda_\ell^h \left(\prod_{\nu=\ell+1}^{k} (1-\alpha_\nu) \right) \alpha_\ell^{\max\{0,\ell-k+1\}} \mathbf{v}_{i,h}^{d,(\ell} \right] \tag{32}$$

where $\kappa(h)$ is the iteration number in which hypergraph h appears by first time and it is assumed that $\prod_m^n = 1$ if $m > n$. Also, it is assumed that $|\mathcal{H}_{i,d}^{(0}| = 1$, $(i,d) \in W$ and that $\lambda_0^h \triangleq 1$ if $\kappa(h) = 0$. Notice that the term enclosed in braquets in equation (32) is associated to a given hypergraph h for an O-D pair corresponding to a strategy. Let $\tilde{\mathbf{v}}_{i,h}^d$ be the vector corresponding to the braquet in (32) :

$$\tilde{\mathbf{v}}_{i,h}^d \triangleq \sum_{\ell=\kappa(h)}^{k} \lambda_\ell^h \left(\prod_{\nu=\ell+1}^{k} (1-\alpha_\nu) \right) \alpha_\ell^{\max\{0,\ell-k+1\}} \mathbf{v}_{i,h}^{d,(\ell} \tag{33}$$

Then, let $\gamma_{i,h}^d$ and $\mathbf{s}_{i,h}^d$ be defined respectively as:

$$\gamma_{i,h}^d \triangleq \sum_{a \in E(i)} (\tilde{\mathbf{v}}_{i,h}^d)_a \tag{34}$$

$$\mathbf{s}_{i,h}^d \triangleq (\gamma_{i,h}^d)^{-1} \tilde{\mathbf{v}}_{i,h}^d \tag{35}$$

It must be noticed that, since $\sum_{a \in E(i)} \mathbf{v}_{i,h}^{d,(\ell} > 0$, then $\gamma_{i,h}^d > 0$ and this permits that relationship (32) can be rewritten as:

$$\mathbf{v}^k = \sum_{(i,d) \in W} g_i^d \sum_{h \in \mathcal{H}_{i,d}^{(k}} \beta_{i,h}^d \mathbf{s}_{i,h}^d \tag{36}$$

And this implies that $\gamma_{i,h}^{d}$ is the fraction of users for O-D pair $(i,d) \in W$, which choose the strategy marked by the hypergraph $h \in \mathcal{H}_{i,d}^{(k}$ and the effective frequencies at the congestion level corresponding to the vector of link volumes v^k.

Another aspect related to the hyperpaths used by congested strategy-based transit assignment models is that of the number of hyperpaths which make up a solution when the congestion level increases. It is known (see Cominetti and Correa (2001)) that the flows per O-D pair of these models may be composed by several hyperpaths, whereas for the uncongested case only a single hyperpath exists in the solution. As an example, Appendix B illustrates the calculation of strategies used in a small problem based on the classical transit network used in Spiess and Florian (1989). From this small example it is clear that the solution flows are composed by several hyperpaths. Also, an example of the evolution of the number of hyperpaths for increasing congestion levels is shown at the end of Section 7 on one of the large test networks used in the computational results Section.

7 Computational results

This section presents the results of the computational tests that compared the performance and characteristics of the solutions obtained from using both the algorithm in Cepeda et al. (2006) and its capacitated version in Section 3. Two small example networks are initially solved in Subsection 7.1. In Subsection 7.2, a small-medium-sized and two large networks are examined. Initial feasible solutions are obtained in all the tests by assigning all the demand to the non-transit links, i.e., by imposing an almost null frequency on the boarding links.

7.1 Illustrative examples with small test networks

These examples were solved using the AMPL/CPLEX system in its student version and the classical MSA step $\alpha_k = 1/(k+1)$, which were applied to both the capacitated and uncapacitated C3F algorithms.

Example 1. The following simple example with a single O-D pair in Figure 3 shows two transit lines, each of them with a total capacity of $c = 9,600$ passengers during the whole period and a demand of passenger $\tau \cdot 2,011$ trips for O-D pair $1 \rightarrow 2$, τ being a factor that will be changed in different runs to see the effects of congestion. efff for boarding links (1,3) and (1,4) are of the type shown in (9), where $f_a(v) = max\{\epsilon, \phi_0(1-\rho_a^2(v))\}$ and ρ_a is as given in (10). Table 1 below shows the network parameters for this example.

	t_a	ϕ_0
$(1,3)$	0.5	0.2
$(1,4)$	0.5	0.2
$(1,2)$	45	–
$(3,2)$	35	–
$(1,2)$	35	–

Table 1: Link parameters for the test network in Example 1 shown in Figure 3

By setting $\tau = 5$ in Example 1, the evolution of the uncapacitated C3F algorithm is good in terms of the relative gap $g(\mathbf{v})$. After 4,000 iterations a value of `3.8868E-05` is obtained for $\hat{g}$. The situation changes when the demand is increased by $\tau = 20$ and $\tau = 50$, where high levels of saturation are reached. In both cases, the uncapacitated C3F algorithm converges to a capacity-feasible point, although the evolution of the gap is irregular at the beginning because the solution to problem `[PL]`$(f^{(k}, t^{(k})$ may be capacity infeasible at links (1,3) and (1,4); thus the initial iterates may also be capacity infeasible. In fact, this is what happens during the first 24 iterations in the case of $\tau = 50$, with all subsequent iterates being capacity-feasible. After 4,000 iterations the obtained relative gap $\hat{g}$ is `5.76516E-05`. If the capacitated C3F algorithm shown in Section 3 is used, this initial period of infeasibility is avoided and shows greater stability by using better values of the relative gap during the starting iterations (compare graphics in

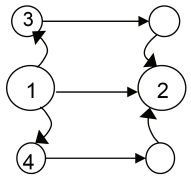


Figure 3: The test network in Example 1.

i	j	$t_{i,j}$
1	2	21.120
2	3	45.000
3	4	26.432
4	3	28.320
3	2	45.000
2	1	19.712
1	101	0.500
101	2	14.080
2	102	0.500
102	1	14.080
2	201	0.500
201	3	19.180
3	202	0.500
202	2	19.180

i	j	$t_{i,j}$
3	301	0.500
301	4	18.880
4	302	0.500
302	3	18.880
1	401	0.500
401	402	14.080
402	3	19.180
3	403	0.500
403	404	19.180
404	1	14.080
2	402	0.500
402	2	0.000
2	404	0.500
404	2	0.000

i	j	$t_{i,j}$
1	501	0.500
501	3	22.240
3	502	0.500
502	1	22.240
1	901	0.500
901	902	22.240
902	4	18.880
4	903	0.500
903	904	18.880
904	1	22.240
3	902	0.500
902	3	0.000
3	904	0.500
904	3	0.000

i	j	$t_{i,j}$
1	1001	0.500
1001	1002	14.080
1002	4	27.040
4	1003	0.500
1003	1004	27.040
1004	1	14.080
2	1002	0.500
1002	2	0.000
2	1004	0.500
1004	2	0.000
1	1101	0.500
1101	4	28.060
4	1102	0.500
1102	1	28.060

Table 2: Link travel times for test network in Example 2.

Figure 4, top and bottom). The run stops at iteration 2,167 with a relative gap $\hat{g}$=`8.98714E-09`, which is lower than the stopping tolerance of `1.0E-08`. At iteration 6, 13,.. the relative gap is $\hat{g} \approx$`4.66E-05`. Also, link (1,2) already has positive flow with little overcongestion; when $\tau = 50$, the network is absolutely saturated. When using the uncapacitated C3F algorithm without explicitly considering capacities, the behaviour is very sensitive to the total travel demand in the network. The capacitated C3F algorithm in Section 3 behaves better and converges notably faster.

Example 2. The transit network in Example 2 is made up of eight transit lines, and the expanded network is shown in Figure 5 with the x-type links suppressed for clarity. As in Example 1, efff for boarding links are also of the type $f_a(v) = max\{\epsilon, \phi_0(1-\rho_a^2(v))\}$. $\phi_0 = 0.2$, while capacity c at boarding links is also 9,600 passengers. Link travel times are shown in Table 2. Boarding links (i, j) are those whose i-node is either 1, 2, 3 or 4.

The various demand levels used for this test network were obtained by multiplying the O-D trip matrix shown in Table 3 by an augmenting factor τ. For this test network, the uncapacitated C3F algorithm converges for values of $\tau = 1.0, 1.2$ and 1.3. No points violating capacity constraints are generated for $\tau = 1.0 and 1.2$, although they do appear for $\tau = 1.3$. Notwithstanding this, the algorithm converges for $\tau = 1.3$. However this algorithm has unstable behaviour for $\tau = 1.6, 2.0$ and larger values, because capacity-infeasible iterates are generated and this causes the relative gap function to suffer sudden jumps. After 4,000 iterations, the minimum relative gap obtained at capacity-feasible points is $\hat{g} = 0.011$.

The capacitated C3F algorithm converges with $\tau = 1.6$ and reaches a relative gap of $\hat{g} =$ `2.43153E-04` after 50,000 iterations, doing so without such unstable behaviour. This run was made in order to obtain an accurate value of the solution flows. With 1,000 iterations the relative gap reaches $\hat{g} =$ `8.85104E-04`. The lowest relative gap of $\hat{g} =$ `8.36174E-05` is obtained at iteration 49,858. Figure 6 shows the evolution of the relative gap during the first 1,000 iterations. It must be remarked however that the algorithm

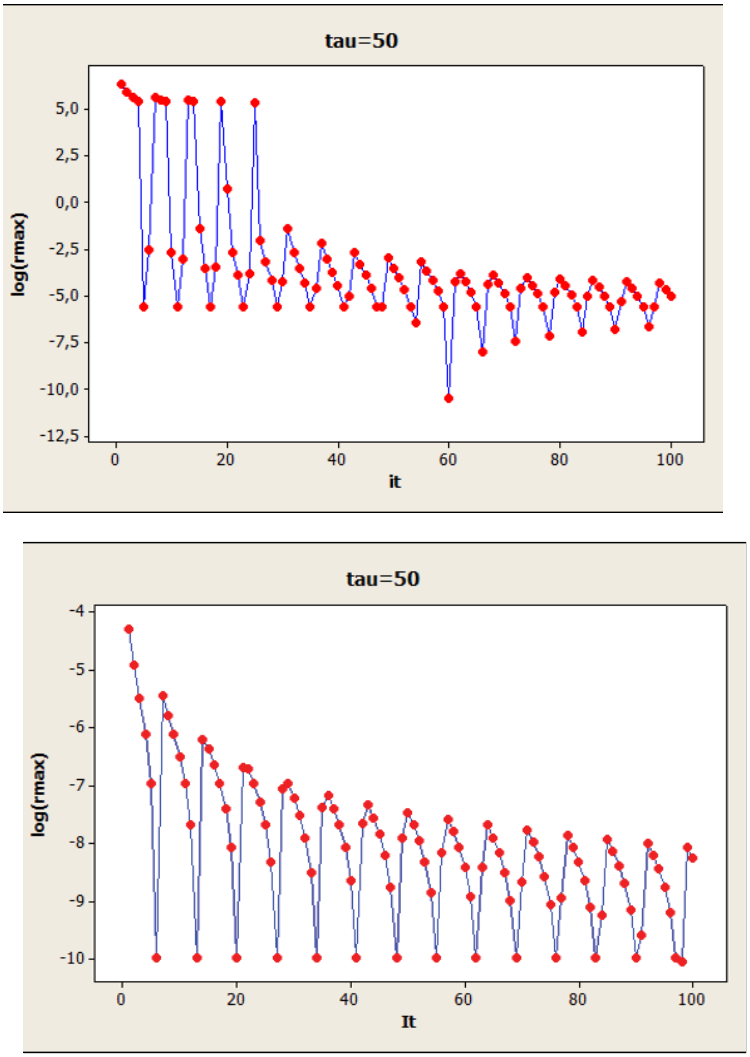


Figure 4: Evolution of the uncapacitated C3F algorithm (top) and the capacitated C3F algorithm(bottom) using the classical MSA step for the test network in Example 1. $\tau = 50$. (Natural logarithm of relative gap $\hat{g}$ on y-axis).

oscillates around the optimal value.

For $\tau = 2.0$, the capacitated C3F algorithm provides a relative gap of $\hat{g} =$ `8.304076E-04` at iteration 50,000, whereas the best relative gap reached is `7.68811E-04` at iteration 1,367 with similar oscillations as in the case of $\tau = 1.6$. At iteration 1,399 the uncapacitated C3F algorithm is able to reach a relative gap of $g =$ `1.952e-03` which seems to be the smallest one over 50,000 iterations.

The self-regulated MSA step in Liu *et al.* (2009) was also used in this example, with $\tau = 2$ for the capacitated and uncapacitated C3F algorithms. At iteration 50,000, a relative gap of $\hat{g}$ of `2.15858E-05` was obtained for the capacitated C3F algorithm with the lowest relative gap of `1.91788E-05` obtained at iteration 49,726. The run for the uncapacitated C3F algorithm also has an unstable behaviour as in the case of $\tau = 1.6$, which is depicted in Figure 6. The lowest relative gap of $g(\mathbf{v})$ obtained was `1.250E-03` at iteration 2,054. Figure 7 compares the evolution of the self-regulated MSA step and the classical MSA step for the capacitated C3F algorithm showing that the use of the self-regulated step can be advantageous for the congested transit assignment problem in the initial iterations. However when the number of iterations is very large both MSA methods show similar performance, while the level of accuracy in terms of the gap reduction is approximately the same.

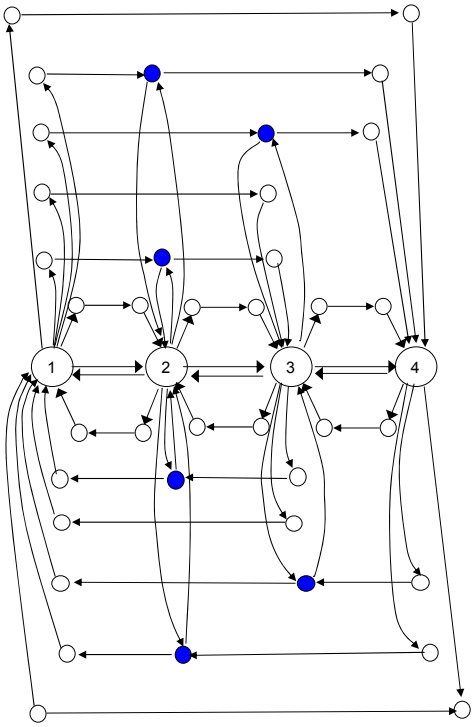


Figure 5: Test network in Example 2: expanded graph.

7.2 Examples with larger networks

In this subsection computational tests are performed on three networks, with two of them which are large dimensions. These correspond to realistic cases under different levels of congestion, such that we can illustrate the performance of the algorithm described in Section 3. The capacitated and uncapacitated C3F algorithms for solving the transit assignment model were implemented in C, using CPLEX V12.5. The tests were carried out on a R5500 workstation using an Intel(R) Xeon(R) CPU E5645 with 2.40 GHz and 48 Gb RAM. The three networks used for the tests are a) the public transit network for the city of Winnipeg provided by the EMME transportation planning system; b) a model of the bus network for the Spanish city of Zaragoza; and c) a model of an auxiliary bus system during disruptions on the Metro Line 1 in Barcelona, as detailed in Codina et al. (2013). The network model of Zaragoza can be considered poorly conditioned in the sense that it presents strong bottlenecks for very high demand, a situation which is very difficult to solve. The network models of Barcelona and Winnipeg present milder bottlenecks in cases of high demand.

These network models were adapted conveniently for the purposes of this paper, i.e., strict capacity limitations were added on line segments as well as effective frequency functions. The dimensions of the expanded networks are shown in Table 4 (number of nodes, links, number of O-D pairs, number of destinations in Dest. column, number of lines and number of stops). Columns **a**, **y**, **x**, **e** and **total** correspond to the number of the various link types, as shown in Figure 1. It must be noted that these models contain a subset of links for transfer movements and for movements that do not use the transit network. This subset of links comprises the so-called auxiliary transit network. The network models that are used in the tests consist of a rather accurate representation of the street map of the city. For the test networks considered here, the "escape paths" needed for assumption 2.3 can be adopted from this subnetwork. Thus, according to Table 4, Zaragoza's network model has 19,585 - 15,394 = 4,191 links of

	1	2	3	4	Total Or.	pax/min
1	0	2,011	22,097	368	24,476	135.98
2	170	0	3,066	230	3,466	19.25
3	4,386	150	0	170	4,706	26.14
3	2,504	150	2,438	0	5,092	28.28
Total Dest.	7,060	2,311	27,601	768	**37,740**	–
pax/min	39.22	12.40	153.34	4.26	–	–

Table 3: Basic O-D Trip table for a period of 180 minutes used in Example 2. The last row and column are, respectively, average arrival and departure rates of passengers at bus stops.

this type.

	Arcs						**O – D**			
Network	**a**	**y**	**x**	**e**	**Total**	**Nodes**	**Pairs**	**Dest.**	**Lines**	**Stops**
WINNIPEG	950	950	817	950	6,642	2,957	5,533	126	66	327
ZARAGOZA	3,864	3,864	3,802	3,864	19,585	9,966	4,156	129	31	1,176
BARCELONA	142	142	132	142	642	310	88	10	5	18

Table 4: Dimensions of the test networks.

The acronyms for the tests using the capacitated and uncapacitated C3F algorithms and the algorithm in Cepeda et al. (2006) are shown in Table 5 below. Three levels of congestion on each network were set to test the algorithms. For each test network a basic origin-destination trip matrix was used, and it was augmented by a multiplicative factor τ (shown in Table 6) in order to increase the trip demand. Other parameters also included in Table 6 are the exponent β for the effective frequency function (set with the same value for any boarding link), the capacity limitation on line segments and the tolerance to stop the cutting-plane algorithm in order to solve the capacitated linear problem in (19). In these tests the step length calculation was performed using the method of successive weighted averages given in (22). After some preliminary experimentation, the value of $\nu = 2$ used in the MSWA step length in (22) seemed to perform better and was adopted in all the tests.

Acronym	**Description**
LC-NLC	Low congestion using the uncapacitated C3F alg.
LC-WCL	Low congestion using the capacitated C3F alg.
HC-NLC	High congestion using the uncapacitated C3F alg.
HC-WCL	High congestion using the capacitated C3F alg.
VC-NLC	Very High congestion using the uncapacitated C3F alg.
VC-WCL	Very High congestion using the capacitated C3F alg.

Table 5: Acronyms according to the level of congestion and the algorithm used.

Each of the possible combinations (congestion level-capacity limitations) shown in Table 5 was tested on the three networks. The mean travel times per user in minutes appear in Table 7 for each test. These were calculated as the quotient between the total travel time (i.e., waiting + onboard + walking time = value of function $G^{(0}_{\texttt{C3F}}(\mathbf{v})$ in the best solution) and the total number of passenger trips:

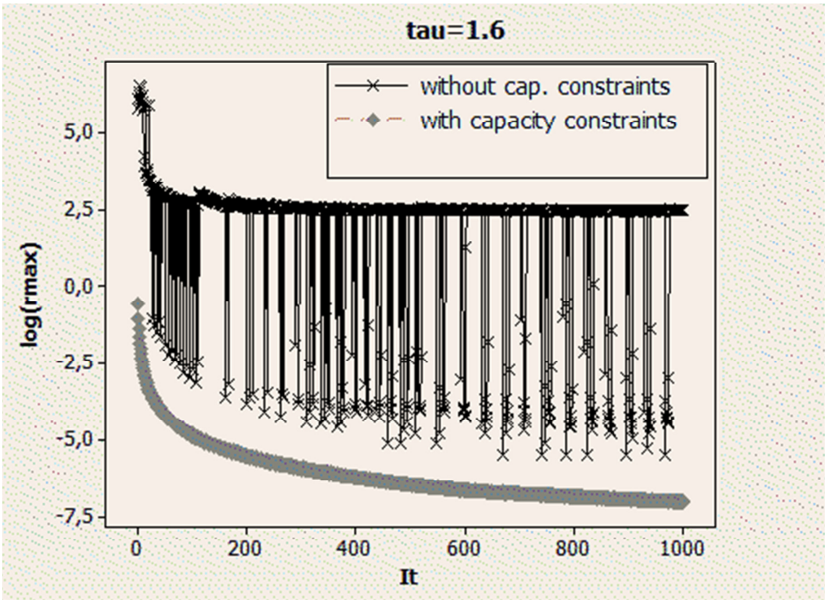


Figure 6: Example 2 for $\tau = 1.6$. Evolution of the relative gap $\hat{g}$ and g for capacitated and uncapacitated C3F algorithms (with and without explicit capacity constraints) using the classical MSA iteration. (Natural logarithm of relative gap on y-axis).

$$\frac{\sum_{a\in A} t_a v_a + \sum_{d\in D}\sum_{i\in N_d} w_i^d}{\sum_{d\in D}\sum_{i\in N} g_i^d} \tag{37}$$

The computational performance of both algorithms is summarized in Table 8. Tests for the Zaragoza and Winnipeg networks were limited to approximately 1,000 seconds of running time, whereas an approximate limit of 6 seconds was set for the Barcelona network. Column **# Total Iter** shows the total number of major iterations until the end of execution. Column **# Total Spiess** shows the total number of times that the Spiess algorithm (Spiess and Florian 1989) was executed throughout the test. Column **Average Spiess it.** shows the average number of times that the Spiess algorithm was solved per iteration. Column **Total CPU** shows the total CPU time in seconds during the process. The columns in the block **CPU per iteration average** show: mean CPU time used per iteration (**it.**), mean CPU time used for solving the Spiess algorithm (**Spiess**) and mean CPU time used for solving the master problem, (**m.prob.**). Finally, in the block of columns **best rel GAP**: column ($g(\mathbf{v})$ or $\hat{g}(\mathbf{v})$) shows the best relative gap value obtained during the execution ($g(\mathbf{v})$ or $\hat{g}(\mathbf{v})$)Fin. Notice that the best relative gap is not achieved in the final iterations but a little bit sooner. Finally column (**CPU**) shows the total CPU time used until achieving the best relative gap value.

Figure 8 illustrates the convergence to the solution for the capacitated C3F algorithm using the method of successive weighted averages in (22). It displays the evolution of the sequence of values indicating the distances to the solution, $\|\mathbf{v}^{(k} - \mathbf{v}^*\|_2$ (per-destination flow vectors) and $\|v^{(k} - v^*\|_2$ (total flows) versus the major iteration number (registered in column **# Total Iter** of Table 8), for the case of the Barcelona test network under the three levels of congestion. As the value for the solution to the problem (v^* or v^*), the flows providing the smaller relative gap have been used. These flows were obtained at iteration numbers 989 for the case of low congestion, 58 for the case of high congestion and 99 for the case of very high congestion. For the case of low level of congestion the errors $\|v^{(k} - v^*\|_2$ seem to follow a pattern given by $O(k^{-\xi})$, ($\xi \approx 1.2$), although in the case of high and very high congestion it seems to be a little

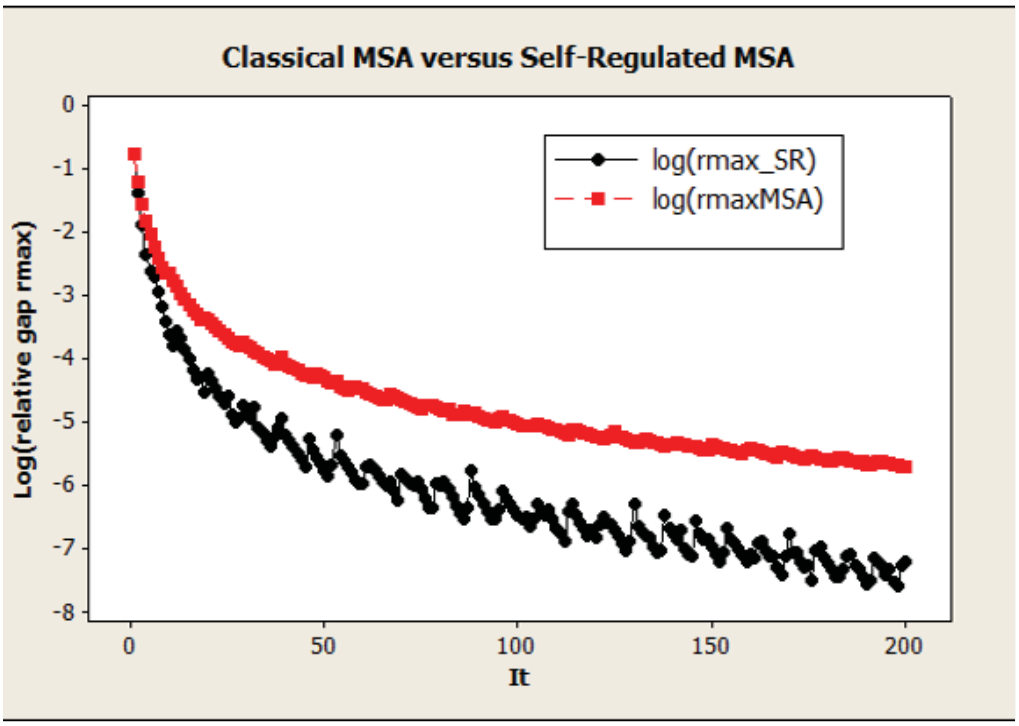


Figure 7: Evolution of relative gap $\hat{g}$ for the capacitated C3F algorithm using the classical MSA step length in comparison to the self-regulated MSA step length in Example 2. $\tau = 2.0$. (Natural logarithm of relative gap $\hat{g}$ on y-axis).

better.

Figures 9 and 12 complement the information given in Table 8 by describing the evolution of iterations and the character of the solutions obtained in each test. Figure 9 shows the evolution of the relative gap defined in (25) for the uncapacitated C3F algorithm and in (21) for the capacitated C3F algorithm versus the CPU time used. In cases of congestion and very high congestion, Figure 9 also reports whether a capacity-feasible solution is found by the uncapacitated C3F algorithm after an iteration, and it does it so by marking it with a circle. In cases of low congestion, the computational performance of both algorithms is comparable, with a slight advantage for the uncapacitated C3F algorithm, which is able to obtain feasible solutions to the problem after every iteration. It must be remarked that a smaller tolerance ϵ was set for the Zaragoza test network, as shown in Table 8. Furthermore, the maximum number of iterations N allowed in the cutting-plane algorithm is higher (40 iterations), whereas it is just $N = 20$ iterations for the Barcelona and Winnipeg test networks. In fact, the low congestion case for the Zaragoza network can also be solved satisfactorily with $\epsilon = 5 \cdot 10^{-4}$ and $N = 20$, by which both the capacitated and uncapacitated algorithms show equal performance. For high congestion, the capacitated C3F algorithm outperforms the uncapacitated one in the Zaragoza network, while feasible solutions can be found in the other two networks with a smaller gap using the capacitated C3F algorithm. However in the case of very high congestion (VC), the uncapacitated C3F algorithm is unable to reach a feasible solution at any iteration and is clearly outperformed by the capacitated C3F algorithm proposed in Section 3. The evolution of the number of infeasible arcs in the uncapacitated C3F algorithm is also shown in Figure 12 confirming these facts. Notice also that in Table 7 the uncapacitated C3F algorithm provides a slight overestimation of mean travel time, which occurs to a higher degree as long as the level of infeasibility increases.

Tables 9 and 10 and Figures 10 and 11 describe the solutions reported in Table 8 by using the

	M.Factor τ	γ Value	Cap. Lim.?	Max. Cap.	Mast.prob. Tolerance	Total Trips
Winnipeg						
LC-NLC	1.00	0.50	No	960.00	-	18,210.55
LC-WCL	1.00	0.50	Yes	960.00	5.00e-04	18,210.55
HC-NLC	3.00	0.50	No	960.00	-	54,631.66
HC-WCL	3.00	0.50	Yes	960.00	5.00e-04	54,631.66
VC-NLC	5.00	0.50	No	960.00	-	91,052.76
VC-WCL	5.00	0.50	Yes	960.00	5.00e-04	91,052.76
Zaragoza						
LC-NLC	0.20	0.50	No	1,500.00	-	74,715.70
LC-WCL	0.20	0.50	Yes	1,500.00	5.00e-05	74,715.70
HC-NLC	0.40	0.50	No	1,500.00	-	149,431.39
HC-WCL	0.40	0.50	Yes	1,500.00	5.00e-05	149,431.39
VC-NLC	0.60	0.50	No	1,500.00	-	224,147.09
VC-WCL	0.60	0.50	Yes	1,500.00	5.00e-05	224,147.09
Barcelona						
LC-NLC	1.00	0.50	No	960.00	-	37,992.00
LC-WCL	1.00	0.50	Yes	960.00	5.00e-04	37,992.00
HC-NLC	3.00	0.50	No	960.00	-	113,976.00
HC-WCL	3.00	0.50	Yes	960.00	5.00e-04	113,976.00
VC-NLC	5.00	0.50	No	960.00	-	189,960.00
VC-WCL	5.00	0.50	Yes	960.00	5.00e-04	189,960.00

Table 6: Main parameters used in the tests

distribution on the test networks of the loading factors of queues at the stops given by formula (9), as well as the associated waiting times. These tables and figures show that congestion is highly concentrated at a relatively small number of points in the networks. Every column in Table 9 shows the total number of boarding arcs comprising the loading factor value within the limits of the range indicated on the header, as well as its percentage of the total number of boarding arcs.

It must be remarked that Table 9 shows the results for the solution reported in Table 8, which shows when the solutions obtained by the uncapacitated C3F algorithm are capacity infeasible. In addition, the loading factor's minimum and maximum value for every test case are shown in Table 10.

It must be noted that negative loading factors appear in the Winnipeg network and especially in the Zaragoza network. This indicates that constraints (7) were violated, even in the instance of high congestion (HC). When the uncapacitated C3F algorithm was used, the number of boarding links with a loading factor greater than one does not seem to be very high, as shown in Table 9. Furthermore, this remains consistent in instances that are solved using the capacitated C3F algorithm, in which there is a reduced number of boarding links with loading factors close to those in Table 9. However, as shown in Table 10, the maximum values of these loading factors that are greater than 1 or negative can be extremely high in the solutions obtained by the uncapacitated C3F algorithm. In other words, the degree of violation of the capacity constraints is very high at the points where this occurs, and it is mainly due to this fact that the solutions provided by the uncapacitated C3F algorithm are invalid. This shows that despite the fact that both the Zaragoza and Winnipeg network models may behave well in low or moderate congestion, they may present strong bottlenecks with a high degree of violation of the constraints, which is what occurs in cases of high and very high congestion. The uncapacitated C3F algorithm comparatively behaves much better on the Barcelona network than on the other two networks.

In order to analyze the waiting time of passengers in queues compared to the uncongested waiting time at the stops $i \in \hat{N}$ of the transit network, the ratio r_i defined in (38) is used. It must be noticed that this ratio applies for all queues corresponding to the various strategies that may operate on the stop and thus, it may underestimate the increase in the waiting times due to congestion.

	Winnipeg	Zaragoza	Barcelona
LC-NCL	32.24	42.11	30.37
LC-WCL	32.24	42.11	30.36
HC-NCL	37.42	43.58	36.17
HC-WCL	37.36	43.35	36.14
VC-NCL	43.22	44.75	40.26
VC-WCL	42.95	44.55	39.51

Table 7: Mean travel times (minutes)

	# Total Iter	# Total Spiess	Average Spiess it.	Total CPU	CPU by iteration - average it.	Spiess	m. prob.	best rel. GAP $g(\mathbf{v})$, $\hat{g}(\mathbf{v})$	CPU(s)
Winnipeg									
LC-NLC	1,749	1,749	1.00	1,000.39	$5.7e-01$	$5.7e-01$	–	$4.12e-06$	902.64
LC-WCL	1,511	1,511	1.00	1,000.01	$6.6e-01$	$5.1e-01$	$3.2e-04$	$5.43e-06$	997.36
HC-NLC	1,823	1,823	1.00	1,000.07	$5.5e-01$	$5.5e-01$	–	$1.79e-04$	968.88
HC-WCL	154	1,924	12.49	1,002.43	$6.5e+00$	$4.9e-01$	$5.5e-03$	$1.08e-03$	838.02
VC-NLC	(∗) 1,869	1,869	1.00	1,000.30	$5.4e-01$	$5.4e-01$	–	$8.97e-03$	807.28
VC-WCL	100	2,000	20.00	1,007.99	$1.0e+01$	$4.6e-01$	$2.4e-02$	$2.10e-03$	947.79
Zaragoza									
LC-NLC	545	545	1.00	1,085.77	$1.8e+00$	$1.8e+00$	–	$4.87e-05$	964.18
LC-WCL	250	521	2.08	1,007.80	$4.0e+00$	$1.6e+00$	$1.4e-03$	$9.94e-05$	981.21
HC-NLC	(∗) 544	544	1.00	1,008.99	$1.8e+00$	$1.8e+00$	–	$1.76e-03$	888.59
HC-WCL	15	600	40.00	1,051.17	$7.0e+01$	$1.7e+00$	$1.7e-01$	$1.15e-02$	980.45
VC-NLC	(∗) 548	548	1.00	1,017.13	$1.8e+00$	$1.8e+00$	–	$1.35e-01$	743.63
VC-WCL	15	600	40.00	1,040.25	$6.9e+01$	$1.6e+00$	$3.3e-01$	$1.30e-02$	900.74
Barcelona									
LC-NLC	2,765	2,765	1.00	6.12	$2.2e-03$	$2.2e-03$	–	$1.76e-05$	5.88
LC-WCL	1,117	2,496	2.23	6.07	$5.4e-03$	$1.7e-03$	$6.0e-04$	$1.39e-05$	4.96
HC-NLC	2,590	2,590	1.00	6.05	$2.3e-03$	$2.3e-03$	–	$4.73e-04$	5.94
HC-WCL	118	2,360	20.00	6.04	$5.1e-02$	$1.6e-03$	$1.5e-02$	$1.87e-03$	5.93
VC-NLC	(∗) 2,574	2,574	1.00	6.07	$2.3e-03$	$2.3e-03$	–	$1.27e-03$	5.88
VC-WCL	112	2,240	20.00	6.01	$5.4e-02$	$1.5e-03$	$1.8e-02$	$2.42e-03$	5.48

Table 8: Summary of results; CPU - GAP. (*) = no capacity-feasible solution found.

$$r_i = \frac{\sum_{v_a^\star>0} f_a(0)}{\sum_{v_a^\star>0} f_a(v^\star)}\ ,\ \ (a = \text{ boarding links at a stop}),\ i \in \hat{N} \tag{38}$$

For cases with high and very high congestion, Figure 11 shows how the relative average waiting time r_i is distributed in the solution given by the capacitated C3F algorithm. In the tests of the uncapacitated C3F algorithm, there appeared stops with negative loading factors or loading factors greater than one; and it must be noted that when these bounds are enforced by applying the algorithm in Section 3, they result in very large values for r_i.

Finally, in order to illustrate the evolution of the number of hyperpaths for increasing congestion levels in a large test network, we have selected the Winnipeg network in its low-congestion level. The identification of hyperpaths has been done using four hashing functions, based on the structure and the labels of the hypergraphs. We have selected a given O-D pair $(i,d) \in W$ and for this O-D pair its demand g_i^d has been increased successively, multiplying it by increasing factors, to illustrate how the resulting set of hyperpaths are changing. All the remaining O-D pair demands have been set constant. Figure 13

	Loading factor distribution at boarding arcs							
	$< \mathbf{0}$	$= \mathbf{0}$	**(0,0.5]**	**(0.5,0.7]**	**(0.7,1]**	**(1.0,1.5]**	**(1.5,2]**	$> \mathbf{2}$
Winnipeg								
LC-NLC	0	433	517	0	0	0	0	0
	0.00%	45.58%	54.42%	0.00%	0.00%	0.00%	0.00%	0.00%
LC-WCL	0	433	517	0	0	0	0	0
	0.00%	45.58%	54.42%	0.00%	0.00%	0.00%	0.00%	0.00%
HC-NLC	0	392	521	32	5	0	0	0
	0.00%	41.26%	54.84%	3.37%	0.526%	0.00%	0.00%	0.00
HC-WCL	0	386	527	32	5	0	0	0
	0.00%	40.63%	55.47%	3.37%	0.526%	0.00%	0.00%	0.00%
VC-NLC	1	359	463	53	73	0	1	0
	0.105%	37.79%	48.74%	5.58%	7.68%	0.00%	0.105%	0.00%
VC-WCL	0	338	492	58	62	0	0	0
	0.00%	35.58%	51.79%	6.11%	6.53%	0.00%	0.00%	0.00%
Zaragoza								
LC-NLC	0	1,849	2,009	6	0	0	0	0
	0.00%	47.85%	51.99%	0.155%	0.00%	0.00%	0.00%	0.00
LC-WCL	0	1,847	2,013	4	0	0	0	0
	0.00%	47.80%	52.10%	0.104%	0.00%	0.00%	0.00%	0.00%
HC-NLC	1	1,543	2,211	78	29	1	0	1
	0.0259%	39.93%	57.22%	2.02%	0.751%	0.0259%	0.00%	0.0259%
HC-WCL	0	1,468	2,305	76	15	0	0	0
	0.00%	37.99%	59.65%	1.97%	0.388%	0.00%	0.00%	0.00%
VC-NLC	42	1,308	2,220	179	92	14	2	7
	1.09%	33.85%	57.45%	4.63%	2.38%	0.362%	0.0518%	0.181
VC-WCL	0	1,303	2,292	231	38	0	0	0
	0.00%	33.72%	59.32%	5.98%	0.983%	0.00%	0.00%	0.00%
Barcelona								
LC-NLC	0	43	72	19	8	0	0	0
	0.00%	30.28%	50.70%	13.38%	5.63%	0.00%	0.00%	0.00%
LC-WCL	0	43	72	19	8	0	0	0
	0.00%	30.28%	50.70%	13.38%	5.63%	0.00%	0.00%	0.00%
HC-NLC	0	33	21	22	66	0	0	0
	0.00%	23.24%	14.79%	15.49%	46.48%	0.00%	0.00%	0.00
HC-WCL	0	31	21	24	66	0	0	0
	0.00%	21.83%	14.79%	16.90%	46.48%	0.00%	0.00%	0.00%
HC-NLC	0	31	16	6	86	3	0	0
	0.00%	21.83%	11.27%	4.23%	60.56%	2.11%	0.00%	0.00
HC-WCL	0	31	20	4	87	0	0	0
	0.00%	21.83%	14.08%	2.82%	61.27%	0.00%	0.00%	0.00%

Table 9: Distribution of the loading factor ρ at solutions $\mathbf{v}^{(k}$ with best relative gap ($g(\mathbf{v}^{(k})$ and $\hat{g}(\mathbf{v}^{(k})$ for uncapacitated and capacitated C3F algorithms, respectively).

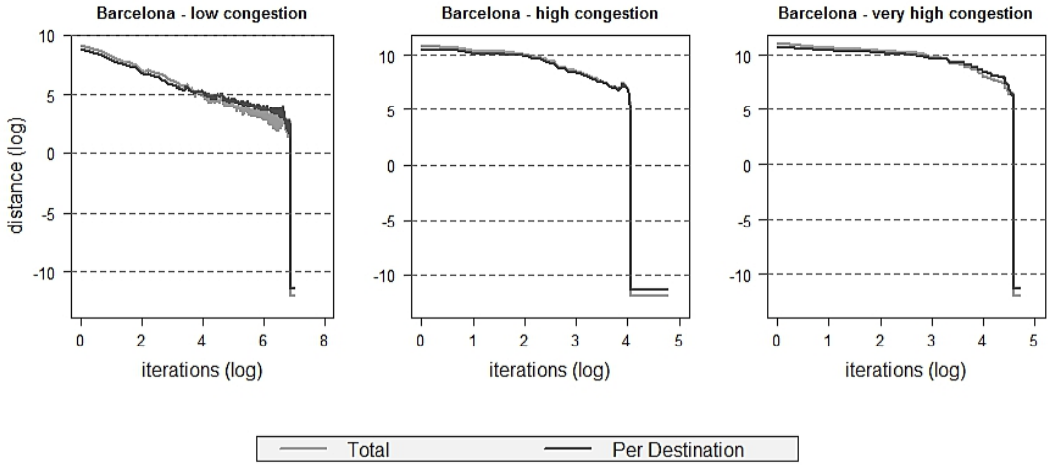


Figure 8: Evolution of errors $\|v^{(k} - v^*\|_2$ (Total) and $\|\mathbf{v}^{(k} - \mathbf{v}^*\|_2$ (Per Destination) versus major iteration number for the capacitated C3F algorithm using MSWA step length on the Barcelona test network in logarithmic scale (natural logarithm).

shows the composition in hyperpaths for flows of that O-D pair for factors of 1, 10, 11, 12, 13, 14, 15 and 20. Hyperpath h_1 was the only one that appeared for factors between 1 and ~ 13. Only when the level of demand is increased by a factor of ~ 13 a new hyperpath h_2 appears. Again, a third hyper path h_3 appears when the factor is 14. Finally, for a factor of 20, 8 hyperpaths appear in the solution: h_1, h_2, h_7, h_8, h_9, h_{10}, h_{11} and h_{12}. The fraction of use for each hyperpath is expressed by the length of the bar segments in the diagram of Figure 13.

8 Conclusions

The authors showed in a previous paper (Codina (2013)) that the congested strategy-based transit assignment problem formulated by Cominetti and Correa (2001) and Cepeda, Cominetti and Florian (2006) can be reformulated as a variational inequality problem and also as a fixed point inclusion problem. They also stated sufficient conditions for the existence of solutions to this problem, not only in cases where effective frequency functions do not define strict capacity limitations but also in cases where capacity bounds should be strictly observed. As a consequence of these results, this paper proposes an algorithm

	Load. Factor	
	Min.	Max.
Winnipeg		
LC-NLC	0.00	0.47
LC-WCL	0.00	0.47
HC-NLC	0.00	0.93
HC-WCL	0.00	0.93
VC-NLC	−0.91	1.79
VC-WCL	0.00	0.96

	Load. Factor	
	Min.	Max.
Zaragoza		
LC-NLC	0.00	0.61
LC-WCL	0.00	0.61
HC-NLC	−0.79	4.05
HC-WCL	0.00	0.87
VC-NLC	−148.10	25.89
VC-WCL	0.00	0.87

	Load. Factor	
	Min.	Max.
Barcelona		
LC-NLC	0.00	0.83
LC-WCL	0.00	0.83
HC-NLC	0.00	0.97
HC-WCL	0.00	0.96
VC-NLC	0.00	1.04
VC-WCL	0.00	0.97

Table 10: Minimum and maximum values for the loading factor ρ in solutions with best relative gap.

for solving the congested transit assignment problem with strict line capacities. The proposed method consists of finding a solution to the fixed point inclusion formulation by using an MSA-based heuristic. At each iteration an uncongested capacitated strategy-based transit assignment problem must be solved.

The authors have tested the use of Lagrangian duality based algorithms to solve the specified transit assignment problem. Amongst them, the most efficient one has been the cutting-plane method, which in turn uses the efficient hyper-path calculating algorithm of Spiess (1984). The proposed method presents the advantage of always guaranteeing capacity-feasible flows at each iteration. In case with low or moderate congestion the proposed method's computational performance is very similar to that in the method of Cepeda, et al. (2006). However, for very high levels of congestion the computational performance of the proposed method is higher. Additionally, the computational tests show that the Cepeda et al. method has serious difficulties in obtaining capacity-feasible solutions if the level of congestion is very high, whereas the feasibility of solutions in the proposed method is always ensured. These conclusions are substantiated by computational tests carried out on small and large realistic transit networks under various congestion levels. For cases with large transit networks, the authors choose the MSWA step length with slightly better performance than the classical MSA. Furthermore, the self regulated MSA step length was tested and it showed better performance if the problems are solved only up to moderately small relative gap values.

Finally, as further research, other options for solving problem [CapPL] could also be considered, such as for instance, interior point methods, (although computing the strategies involved would be then add a nontrivial problem to tackle). This would result in better computational performance of the method.

Acknowledgments. This work has been supported Research Grants TRA2014-52530-C3-3-P, TRA2016-76914-C3-1-P, of the Spanish Ministery of Education (MINECO).

Appendix A. Glossary of Notations

- $G = (N, A)$ - a general directed graph with N the set of nodes and A the set of links. In this paper it refers to a detailed transit network model in which possible passenger movements are schematically depicted in Figure 1. Nodes in N are labeled using i, j, d and magnitudes or elements labeled with i, j, d are associated to some node in N. Links in A are labeled using a.
- $E(i)$, $I(i)$ - the subsets of emerging and incoming links to node $i \in N$, respectively.
- W - the set of origin-destination pairs on the graph which capture passenger trips. Formally, origins and destinations are nodes in N.
- D - the set of destinations in the network; $D \subseteq N$.
- $\hat{N}$, $\hat{N}_d$ - $\hat{N}$ is the (sub)set of nodes which play the role of stops since, at least, one of its emerging links modelling boarding to a transit line has associated to it a finite effective frequency function. $\hat{N}$ is defined in (4). From it, $\hat{N}_d \triangleq \hat{N} \setminus \{d\}$ for a given destination $d \in D$.
- $\hat{E}(i)$ - links in $E(i)$ corresponding to boardings to a line with a finite effective frequency function. It is defined in (5).
- $\hat{A}$ the (sub)set of links made up by all those who are in $\hat{E}(i)$, for some $i \in \hat{N}$.

- A_a, A_x, A_e - subsets of links for boardings, dwellings and in-vehicle movements of passengers on the graph G; $A_a, A_x, A_e \subset A$.
- $a, e(a), x(a), y(a)$ - when referring to a boarding link $a \in A_a$, $e(a)$ is the associated in-vehicle link, $x(a)$ is the corresponding dwelling link and $y(a)$ the corresponding alighting link. Figure 1 depicts flows on links which are indexed using this notation.
- c_a - line capacity for $a \in A_e$ since in-vehicle links are associated necessarily to some transit line; $c_{e(a)}$ for some $a \in A_a$ can be also used (see, for instance equations (7) and (8)).
- $\mathbf{v}, \mathbf{v}^d, \mathbf{v}_i^d$; $\mathbf{u}, \mathbf{u}^d, \mathbf{u}_i^d$ - flow vectors. Network flows for passenger movements follow in the paper a multicommodity formulation, being each destination $d \in D$ one such commodity. In general, flows variables are designated using letter v. If it is required, as in for instance in equations (17) and (18), letters u and v are used for two different flow variables. See Subsection 2.1, after the second paragraph, for a concrete meanings of $\mathbf{v}, \mathbf{v}^d, \mathbf{v}_i^d$.
- v_a^d, v_a, v_i^d; u_a^d, u_a, u_i^d - scalar flows associated to a link $a \in A$ or a given node $i \in N$. Same observations as for flow vectors. See Subsection 2.1, after the second paragraph, for a concrete meanings of v_a, v_a^d, v_i^d.
- v, u - vectors of total link flows. Notice that they are not written in bold. $v = (..., v_a, ...; a \in A)$. Same observations as for flow vectors. See Subsection 2.1, after the second paragraph, for a concrete meanings of v_a, v_a^d, v_i^d.
- w_i^d - variable for total waited time of passengers with destination d at stop $i \in \hat{N}$. It appears in the formulation of the strategy-based uncongested transit assignment problem as a linear program, (23), and also in the formulation of problem [CapPL] in (19).
- $\mathrm{V}^d, \mathrm{V}, \mathcal{V}, \hat{\mathrm{V}}$ - for a given destination $d \in D$, V^d is the polyhedron of per-destination vector flows $\mathbf{v}^d$, as defined in (1). V is the polyhedron of vector flows $\mathbf{v}$, as defined in (3); $\mathcal{V}$ is the polyhedron of total capacity-feasible link flows v, as defined in (2) and $\hat{\mathrm{V}}$ is the polyhedron made up by the convex hull of the set of vertexes in V.
- $f_a(\cdot)$, $\sigma_a(\cdot)$ - the effective frequency function and its reciprocal, the mean waiting time function, $f_a = 1/\sigma_a$. Both apply only for boarding links, $a \in A_a$.
- $\rho_a(\cdot)$ - the loading factor of boardings to a line (the one associated with link $a \in A_a$). It is defined in (9) as the ratio between mean total boarding flow v_a at $a \in A_a$ and mean capacity available $c_{e(a)} - v_{x(a)}$ at the dwelling link $x(a)$.
- mwtf, efff - acronyms for $\sigma_a(\cdot)$ and $f_a(\cdot)$ respectively.
- $t_a(\cdot)$ - mean travel time function for a link $a \in A$.
- μ_a - Lagrange multiplier for constraints $v_a \leq c_a,\ a \in A_e$.
- $A_\sigma^\infty(\mathbf{v})$, $A_t^\infty(\mathbf{v})$ - for a given flow vector $\mathbf{v}$, $A_\sigma^\infty(\mathbf{v}) \subseteq A_a$ is the (sub)set of boarding links that present infinite mean waiting times (i.e., saturated queue) when the corresponding total flow vector v has components such that $v_a = c_a$, $a \in A_e$. Also, for a given flow vector $\mathbf{v}$, $A_t^\infty(\mathbf{v}) \subseteq A$ is the (sub)set of links that present infinite travel times when the corresponding total flow vector v has components such that $v_a = c_a$, $a \in A_e$. Both sets are defined in (14) and (15).

Appendix B. Calculation of strategies in a small example

In this appendix the calculation of the strategies used in a solution given by the algorithm in Section 3 is illustrated using an adaptation of the example network in Spiess and Florian (1989). The test network is shown in Figure 14. Notice that three walking links have been added: links $(1,2)$, $(2,3)$ and $(3,4)$ with costs 175, 135 and 150 respectively. O-D pairs are $(1,4)$ and $(2,4)$ with demands 1000 and 500 respectively. Effective frequency functions applying to the four lines L1 to L4 are of the type $f_a(v) = \kappa_a(1 - \rho_a(v)^\beta)$, where $\rho_a(v)$ is as given in (9) and κ_a is an uncongested frequency value set equally for all the boarding links a in a line. For boarding links in line L1, L2, L3 and L4 κ_a has been set to 1/6, 1/6, 1/15 and 1/3 respectively. Finally, the exponent β has been set common to all lines and equal to 0.5.

Assume that two iterations of the algorithm are performed and that we need to know the hyperpaths that make up the solution obtained after these two iterations. In the first iteration, the MSWA step length is $\alpha_1 = 0,8$, whereas in the second iteration is $\alpha_2 = 0.643$. For O-D pair (1,4), a total of 24 hypergraphs

can be found and for O-D pair (2,4) the total number of hyperpaths that can be found is 18. From them, only those hypergraphs depicted in Figure 15 appear in the two iterations for O-D (1,4), whereas those for O-D (2,4) appear in Figure 16. Tables 11 and 12 show the vector flows $\mathbf{v}_{i,h}^{d,(\ell}$ for O-D pair (1,4) and (2,4) respectively which are obtained at each of the subproblems in the cutting plane algorithm of Section 4 for solving problem [CapPL]. The row marked with λ shows the baricentric coordinates that correspond to these vectors as components of the solution for problem [CapPL]. Thus, after iteration 1, for O-D pair (1,4), the solution of problem [CapPL] is made up by vector flows $\mathbf{v}_{i,h}^{d,(\ell}$ on hypergraphs h_{13}, h_{15} and h_{16} (hypergraphs h_{12}, h_{14} and h_{15} have null baricentric coordinate). The row marked by **it-subproblem** specifies the iteration number of the main algorithm and the iterations of cutting plane algorithm. Thus, at iteration 1 of the main algorithm, problem [CapPL] has been solved approximately with 6 iterations. In the second iteration of the main algorithm, the cutting plane has solved the corresponding problem [CapPL] using 7 iterations. Finally, the elements intervening in equation (36) appear in Tables 13 and 14, i.e., the seven vectors $\mathbf{s}_{i,h}^{d}$ (the hyperpaths) that make up the solution found after two iterations and the fractions $\gamma_{i,h}^{d}$ in which they are used (row marked with γ).

Notice that for O-D pair $(1, 4)$, the solution hyperpath for this test network in the case of no congestion (constant effective frequencies) corresponds to hypergraph h_{17} in Figure 15 and for O-D pair $(2, 4)$ corresponds to the hyperpath associated to hypergraph h_{22}. The network is very congested (notice that, for instance, link (221,23) carries a total flow of 510.715 and that total boardings in link (2,223) are 443.23. Thus, at stop 2, the queue for boarding to line L2 has a loading factor $\rho = 0.9185$) and this makes that a multiplicity of hyperpaths for each O-D pair are composing the solution.

MSA: $\alpha \backslash 1-\alpha$	–	$0.8 \backslash 0.2$						$0.643 \backslash 0.357$						
λ	1	–	0.1	–	0.13	0.77	–	0.16	–	–	0.30	0.23	0.30	–
it-subproblem	**0-1**	**1-1**	**1-2**	**1-3**	**1-4**	**1-5**	**1-6**	**2-1**	**2-2**	**2-3**	**2-4**	**2-5**	**2-6**	**2-7**
arc \ hypergraph	$\mathbf{h_{11}}$	$\mathbf{h_{12}}$	$\mathbf{h_{13}}$	$\mathbf{h_{14}}$	$\mathbf{h_{15}}$	$\mathbf{h_{16}}$	$\mathbf{h_{15}}$	$\mathbf{h_{12}}$	$\mathbf{h_{15}}$	$\mathbf{h_{13}}$	$\mathbf{h_{13}}$	$\mathbf{h_{16}}$	$\mathbf{h_{17}}$	$\mathbf{h_{16}}$
$1-11$	–	0.5	1	–	0.5	0.5	0.5	0.356	0.356	1	1	0.356	0.356	0.356
$11-14$	–	0.5	1	–	0.5	0.5	0.5	0.356	0.356	1	1	0.356	0.356	0.356
$14-4$	–	0.5	1	–	0.5	0.5	0.5	0.356	0.356	1	1	0.356	0.356	0.356
$1-21$	–	0.5	–	1	0.5	0.5	0.5	0.644	0.644	–	–	0.644	0.644	0.644
$21-220$	–	0.5	–	1	0.5	0.5	0.5	0.644	0.644	–	–	0.644	0.644	0.644
$220-221$	–	0.5	–	–	–	–	–	0.644	–	–	–	–	–	–
$221-23$	–	0.5	–	–	–	0.357	–	0.644	–	–	–	0.339	0.339	0.339
$23-3$	–	0.5	–	–	–	0.357	–	0.644	–	–	–	0.339	0.339	0.339
$220-2$	–	–	–	1	0.5	0.5	0.5	–	0.644	–	–	0.644	0.644	0.644
$2-221$	–	–	–	–	–	0.357	–	–	–	–	–	0.339	0.339	0.339
$2-32$	–	–	–	1	0.5	0.143	0.5	–	0.644	–	–	0.305	0.305	0.305
$32-330$	–	–	–	1	0.5	0.143	0.5	–	0.644	–	–	0.305	0.305	0.305
$330-331$	–	–	–	1	0.5	0.143	0.5	–	0.644	–	–	–	0.305	–
$330-3$	–	–	–	–	–	–	–	–	–	–	–	0.305	–	0.305
$3-331$	–	0.083	–	–	–	0.060	–	0.263	–	–	–	0.263	0.139	0.263
$331-34$	–	0.083	–	1	0.5	0.202	0.5	0.263	0.644	–	–	0.263	0.444	0.263
$34-4$	–	0.083	–	1	0.5	0.298	0.5	0.263	0.644	–	–	0.263	0.444	0.263
$3-43$	–	0.417	–	–	–	0.298	–	0.381	–	–	–	0.381	0.200	0.381
$43-44$	–	0.417	–	–	–	0.298	–	0.381	–	–	–	0.381	0.200	0.381
$44-4$	–	0.417	–	–	–	0.298	–	0.381	–	–	–	0.381	0.200	0.381
$1-2$	1	–	–	–	–	–	–	–	–	–	–	–	–	–
$2-3$	1	–	–	–	–	–	–	–	–	–	–	–	–	–
$3-4$	1	–	–	–	–	–	–	–	–	–	–	–	–	–

Table 11: Column vectors $\mathbf{v}_{i,h}^{d,(\ell}$ generated by the capacitated C3F algorithm in Section 3 at iterations $\ell = 0, 1, 2$ for O-D pair $(i, d) = (1, 4)$ and for the hypergraphs h generated during each problem [CapPL] solved.

MSA: $\alpha \setminus 1-\alpha$	–	$0.8 \setminus 0.2$						$0.643 \setminus 0.357$						
λ	1	–	0.1	–	0.13	0.77	–	0.16	–	–	0.30	0.23	0.30	–
it-subproblem	**0-1**	**1-1**	**1-2**	**1-3**	**1-4**	**1-5**	**1-6**	**2-1**	**2-2**	**2-3**	**2-4**	**2-5**	**2-6**	**2-7**
arc \ hypergraph	$\mathbf{h_{21}}$	$\mathbf{h_{22}}$	$\mathbf{h_{23}}$	$\mathbf{h_{23}}$	$\mathbf{h_{23}}$	$\mathbf{h_{22}}$	$\mathbf{h_{23}}$	$\mathbf{h_{22}}$	$\mathbf{h_{23}}$	$\mathbf{h_{24}}$	$\mathbf{h_{22}}$	$\mathbf{h_{25}}$	$\mathbf{h_{22}}$	$\mathbf{h_{25}}$
$2-221$	–	0.714	–	–	–	0.714	–	0.527	–	–	0.527	0.527	0.527	0.527
$221-23$	–	0.714	–	–	–	0.714	–	0.527	–	–	0.527	0.527	0.527	0.527
$23-3$	–	0.714	–	–	–	0.714	–	0.527	–	–	0.527	0.527	0.527	0.527
$2-32$	–	0.286	1	1	1	0.286	1	0.473	1	1	0.473	0.473	0.473	0.473
$32-330$	–	0.286	1	1	1	0.286	1	0.473	1	1	0.473	0.473	0.473	0.473
$330-331$	–	0.286	1	1	1	0.286	1	0.473	1	–	0.473	–	0.473	–
$330-3$	–	–	–	–	–	–	–	–	–	1	–	0.473	–	0.473
$3-331$	–	0.119	–	–	–	0.119	–	0.215	–	–	0.215	0.409	0.215	0.409
$331-34$	–	0.405	1	1	1	0.405	1	0.689	1	–	0.689	0.409	0.689	0.409
$34-4$	–	0.405	1	1	1	0.405	1	0.689	1	–	0.689	0.409	0.689	0.409
$3-43$	–	0.595	–	–	–	0.595	–	0.311	–	1	0.311	0.591	0.311	0.591
$43-44$	–	0.595	–	–	–	0.595	–	0.311	–	1	0.311	0.591	0.311	0.591
$44-4$	–	0.595	–	–	–	0.595	–	0.311	–	1	0.311	0.591	0.311	0.591
$2-3$	1	–	–	–	–	–	–	–	–	–	–	–	–	–
$3-4$	1	–	–	–	–	–	–	–	–	–	–	–	–	–

Table 12: Column vectors $\mathbf{v}_{i,h}^{d,(\ell}$ generated by the capacitated C3F algorithm in Section 3 at iterations $\ell = 0, 1, 2$ for O-D pair $(i, d) = (2, 4)$ and for the hypergraphs h generated during each problem [CapPL] solved.

γ	0.071	0.105	0.222	–	0.037	0.370	0.194		$g_1^4 = 1000$
	hyperpath								
arc	$\mathbf{h_{11}}$	$\mathbf{h_{12}}$	$\mathbf{h_{13}}$	$\mathbf{h_{14}}$	$\mathbf{h_{15}}$	$\mathbf{h_{16}}$	$\mathbf{h_{17}}$	$\sum \gamma_{i,h}^d \mathbf{s_{i,d}^d}$	$g_i^d \sum \gamma_{i,h}^d \mathbf{s_{i,d}^d}$
$1-11$	–	0.356	1	–	0.500	0.442	0.356	0.511	510.714
$11-14$	–	0.356	1	–	0.500	0.442	0.356	0.511	510.714
$14-4$	–	0.356	1	–	0.500	0.442	0.356	0.511	510.714
$1-21$	–	0.644	–	–	0.500	0.558	0.644	0.418	417.857
$21-220$	–	0.644	–	–	0.500	0.558	0.644	0.418	417.857
$220-221$	–	0.644	–	–	–	–	–	0.067	67.485
$221-23$	–	0.644	–	–	–	0.350	0.339	0.263	262.866
$23-3$	–	0.644	–	–	–	0.350	0.339	0.263	262.866
$220-2$	–	–	–	–	0.500	0.558	0.644	0.350	350.372
$2-221$	–	–	–	–	–	0.350	0.339	0.195	195.381
$2-32$	–	–	–	–	0.500	0.209	0.305	0.155	154.991
$32-330$	–	–	–	–	0.500	0.209	0.305	0.155	154.991
$330-331$	–	–	–	–	0.500	0.085	0.305	0.109	109.116
$330-3$	–	–	–	–	–	0.124	–	0.046	45.876
$3-331$	–	0.263	–	–	–	0.142	0.139	0.107	107.225
$331-34$	–	0.263	–	–	0.500	0.227	0.444	0.216	216.341
$34-4$	–	0.263	–	–	0.500	0.284	0.444	0.237	237.293
$3-43$	–	0.381	–	–	–	0.331	0.200	0.202	201.517
$43-44$	–	0.381	–	–	–	0.331	0.200	0.202	201.517
$44-4$	–	0.381	–	–	–	0.331	0.200	0.202	201.517
$1-2$	1	–	–	–	–	–	–	0.071	71.429
$2-3$	1	–	–	–	–	–	–	0.071	71.429
$3-4$	1	–	–	–	–	–	–	0.071	71.429

Table 13: Vectors $\mathbf{s_{1,h}^4}$ for the generated hypergraphs; O-D pair (1,4)

γ	0.071	0.712	0.066	–	0.150		$g_2^4 = 500$
	hyperpath						
arc	$\mathbf{h_{21}}$	$\mathbf{h_{22}}$	$\mathbf{h_{23}}$	$\mathbf{h_{24}}$	$\mathbf{h_{25}}$	$\sum \gamma_{i,h}^d \mathbf{s_{i,d}^d}$	$g_i^d \sum \gamma_{i,h}^d \mathbf{s_{i,d}^d}$
$2-221$	–	0.585	–	–	0.527	0.496	247.849
$221-23$	–	0.585	–	–	0.527	0.496	247.849
$23-3$	–	0.585	–	–	0.527	0.496	247.849
$2-32$	–	0.415	1.000	–	0.473	0.433	216.437
$32-330$	–	0.415	1.000	–	0.473	0.433	216.437
$330-331$	–	0.415	1.000	–	–	0.362	180.818
$330-3$	–	–	–	–	0.473	0.071	35.619
$3-331$	–	0.186	–	–	0.409	0.194	96.889
$331-34$	–	0.601	1.000	–	0.409	0.555	277.707
$34-4$	–	0.601	1.000	–	0.409	0.555	277.707
$3-43$	–	0.399	–	–	0.591	0.373	186.579
$43-44$	–	0.399	–	–	0.591	0.373	186.579
$44-4$	–	0.399	–	–	0.591	0.373	186.579
$2-3$	1.000	–	–	–	–	0.071	35.714
$3-4$	1.000	–	–	–	–	0.071	35.714

Table 14: Vectors $\mathbf{s_{2,h}^4}$ for the generated hypergraphs; O-D pair (2,4)

References

Babazadeh, A. and Aashtiani, H. Z., 2005. Algorithm for equilibrium transit assignment problem.Transportation Research Record, 1923, pp. 227-235.

Bazaraa, M.S., Sherali, H., Shetty, C.M. Nonlinear programming. Theory and Algorithms. 3rd Edition (2006). John Wiley

Blum, J.R. (1954) Multidimensional Stochastic Approximations Methods. The Annals of Mathematical Statistics, Volume 25, Number 4 (1954), 737-744.

Bouzayene-Ayari, B. Gendreau, M. and Nguyen, S. (2001) Modelling bus stops in transit networks: a survey and new formulations. Transportation Science, 35, pp. 304-321.

Bouzayene-Ayari, B., Gendreau, M., and Nguyen, S. (1995). An Equilibrium-Fixed Point Model for Passenger Assignment in Congested Transit Networks (No. Publication CRT-95-97), Montréal: Université de Montréal.

Castro, F.R., Leal, J.E., 2003. Alocação de fluxos de passageiros em uma rede de transporte público de grande porte formulado como um problema de inequações variacionais. Pesquisa Operacional, v.23,n.2, p.235-264, 2003. (In portuguese)

Cats, O., Jens West, J. and Eliasson, J. (2016) A dynamic stochastic model for evaluating congestion and crowding effects in transit systems. Transportation Research Part B 89 (2016) 43-57.

Chriqui, C. and Robillard, P., 1975. Common bus lines, Transportation Science 9, 115-121.

Cepeda, M., Cominetti, R. and M. Florian, M., 2006. A frequency-based assignment model for congested transit networks with strict capacity constraints: characterization and computation of equilibria, Transportation Research B 40 437-459 (2006).

Codina, E., 2013. A Variational Inequality Reformulation of a Congested Transit Assignment Model by Cominetti, Correa, Cepeda, and Florian. Transportation Science 47 (2), 231-246.

Codina, E, Marín, A. and López, F, 2013. A model for setting services on auxiliary bus lines under congestion Top 21 (1), 48-83

Codina, E., Ibáñez, G. and Barceló, J. (2015) Applying projection-based methods to the asymmetric traffic assignment problem. Computer-Aided Civil and Infrastructure Engineering 30, 2, 103-119.

Cominetti, R. and Correa, J., 2001. Common-lines and passenger assignment in congested transit networks, Transportation Science 35 (3) 250-267.

Daganzo, C.F., 1977a. On the traffic assignment problem with flow dependent costs-I. Transportation Research 11, 433-437.

Daganzo, C.F., 1977b. On the traffic assignment problem with flow dependent costs-II. Transportation Research 11, 439-441.

De Cea, J. and Fernández, E., 1993. Transit assignment for congested public transport systems: an equilibrium model, Transportation Science, 27 (2) 133-147. (1993). 17

Dunn, J.C. (1978) Iterative construction of fixed points for multivalued operators of the monotone type. Journal of Functional Analysis, 27, pp. 38-50

Freijinger, E. and Florian, M., 2013. Large scale congested transit assignment: achievements and challenges. AITPM National Traffic and Transport Conference, Perth., Australia.

Di Gangi, M., Cantarella, G.E., Vitetta, A., 2015. MSA Algorithms for solving SUE in Urban Transit Networks. Proceedings of the 4th hEART Conference. Lyngby, Denmark.

Gendreau, M. 1984. Une étude approfondie d'un modèle d'équilibre pour l'affectation des passagers dans les réseaux de transport en commun, Ph.D. thesis, Département d'informatique et de recherche opérationnelle, Université de Montréal, Montréal, Canada

Goh, C.J. and Yang, X.Q., 2002. Duality in optimization and variational inequalities. Optimization Theory and Applications Series, Taylor & Francis.

Hamdouch, Y., and Lawphongpanich, S. (2008). Schedule-based transit assignment model with travel strategies and capacity constraints. Transportation Research Part B: Methodological, 42, 663-84.

Hamdouch, Y., Ho, H. W., Sumalee, A., and Wang, G. (2011). Schedule-based transit assignment model with vehicle capacity and seat availability. Transportation Research Part B, 45, 1805-30.

Ishikawa, S. Fixed points by a new iteration method. Proc. Amer. Math. Soc. 44 (1974), 147-150.

Jiang, Y. and Szeto, W.Y. (2016) Reliability-based stochastic transit assignment: Formulations and capacity paradox. Transportation Research Part B 93 (2016) 181-206

Karauchi, F., Bell, M.G.H. and Schmeker, J.D., 2003. Capacity constrained transit assignment with common lines. Journal of Mathematical Modelling and Algorithms, 2 pp. 309-327.

Kiwiel C.K. 1990. Proximity control in bundle methods for convex non-differentiable minimization. Mathematical Programming, 46 pp. 105-122.

Krasnoselskii, M.A. Two remarks on the method of successive approximations, Uspekhi Mat. Nauk 10 (1955), 123-127.

Larson, T., Patriksson, M. 1994. Equilibrium characterizations of solutions to side constrained asymmetric traffic assignment models. Le Mathematiche Vol. XLIX (1994) - Fasc. II, pp. 249-280

Lemaréchal, C. 1974. An algorithm for minimizing convex functions. In J.L. Rosenberg, Editor, Information Processing'74, pp.552-556, North Holland.

Liu, H.X., He, X. and He, B., (2009). Method of successive weighted averages and self-regulated averaging schemes for solving stochastic user equilibrium problems. Networks and Spatial Economics 9 485-503.

Mann, W. R., Mean value methods in iteration, Proc. Amer. Math. Soc. 4 (1953), 506-510.

Maiti, M. and Ghosh, M.K. (1989) Approximating fixed points by Ishikawa Iterates. Bull. Austral. Math. Soc., 40 (1989) pp. 113-117

Nie, Y., Zhang, H.M. and Lee, D-H. 2004. Models and algorithms for the traffic assignment problem with link capacity constraints. Transportation Research Part B 38 285-312 (2004)

Noriega Y. and Florian, M., 2003. Loptimisation des fréquences d'un réseau de transport en commun moyennement congestionné. INFOR vol 41, no 2. 129-153.

Robins H., Monro, S. (1951) A Stochastic Approximation Method. The Annals of Mathematical Statistics, Vol. 22, No. 3 (Sep., 1951), pp. 400-407

Schmöcker, J.-D., Bell, M. G. H., and Kurauchi, F. (2008). A quasi-dynamic capacity constrained frequency-based transit assignment model. Transportation Research Part B, 42, 925-45.

Spiess, H. Contribution à la théorie et aux outils de planification des réseaux de transport urbains. Ph. D. Thesis, Département d'Informatique et Récherche Opérationnelle, Publication 382, CRT, U. de Montréal (1984)

Spiess, H. and Florian, M., 1989. Optimal strategies: a new assignment model for transit networks, Transportation Research Part B 23 (2) 83-102.

Trozzi, V., Gentile,G., Bell, M.G.H. and Kaparias, I. (2013) Dynamic user equilibrium in public transport networks with passenger congestion and hyperpaths. Transportation Research Part B 57 (2013) 266-285.

Verbas, Ö., Mahmassani, H.S. and Hyland, M.F. (2016) Gap-based transit assignment algorithm with vehicle capacity constraints: Simulation-based implementation and large-scale application. Transportation Research Part B 93 (2016) 1-16.

Wu, J.H. and Florian (1993) M. A simplicial decomposition method for the transit equilibrium assignment problem. Annals of Operatins Research 44, 1993 pp. 245-260.

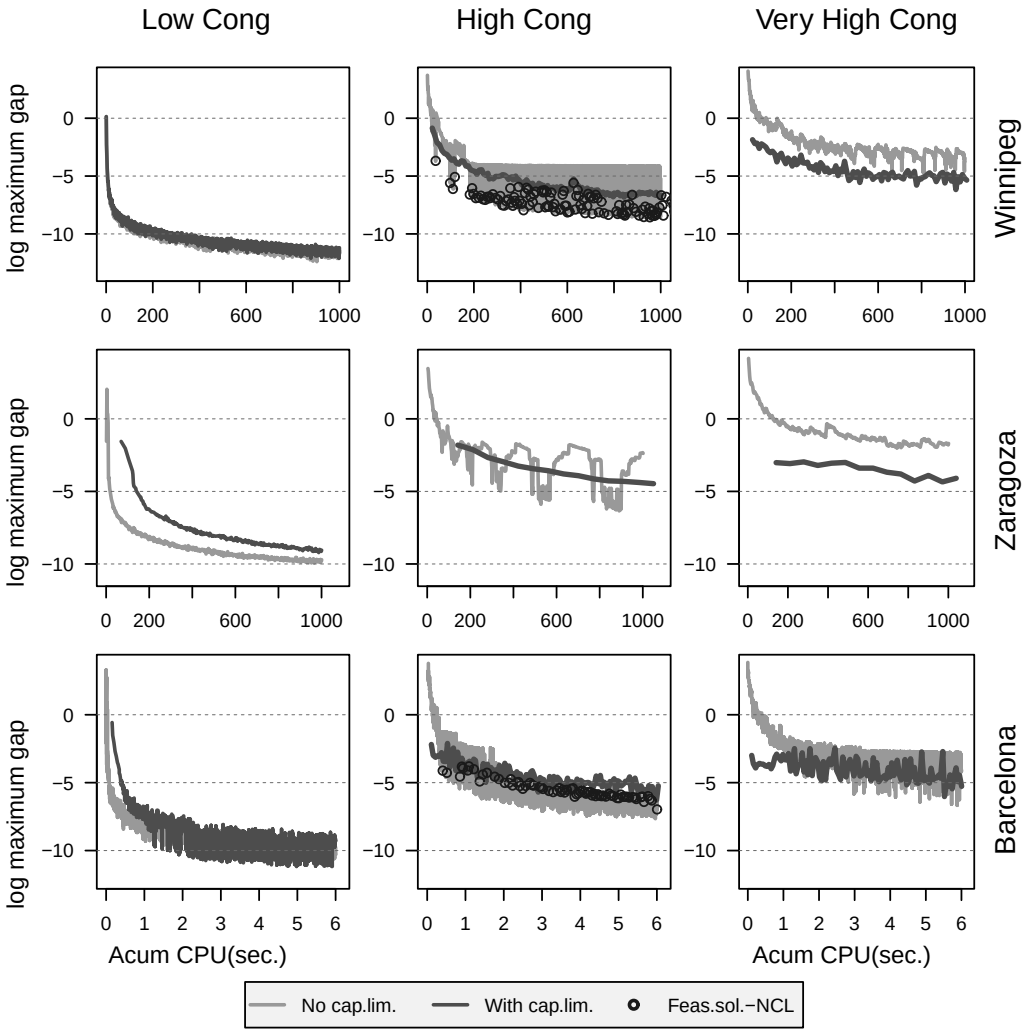


Figure 9: Comparative evolution of the relative gaps $g(\mathbf{v})$ and $\hat{g}(\mathbf{v})$ vs CPU time in the tests with the uncapacitated and capacitated C3F algorithms (No cap. lim. and With cap. lim. respectively). For Low Cong tests, the uncapacitated C3F algorithm finds capacity-feasible solutions at almost all iterations. For the High Cong. tests, the uncapacitated C3F algorithm finds capacity-feasible solutions only at iterations marked with a circle. For the Very High Cong. tests, the uncapacitated algorithm is unable to find capacity-feasible solutions.

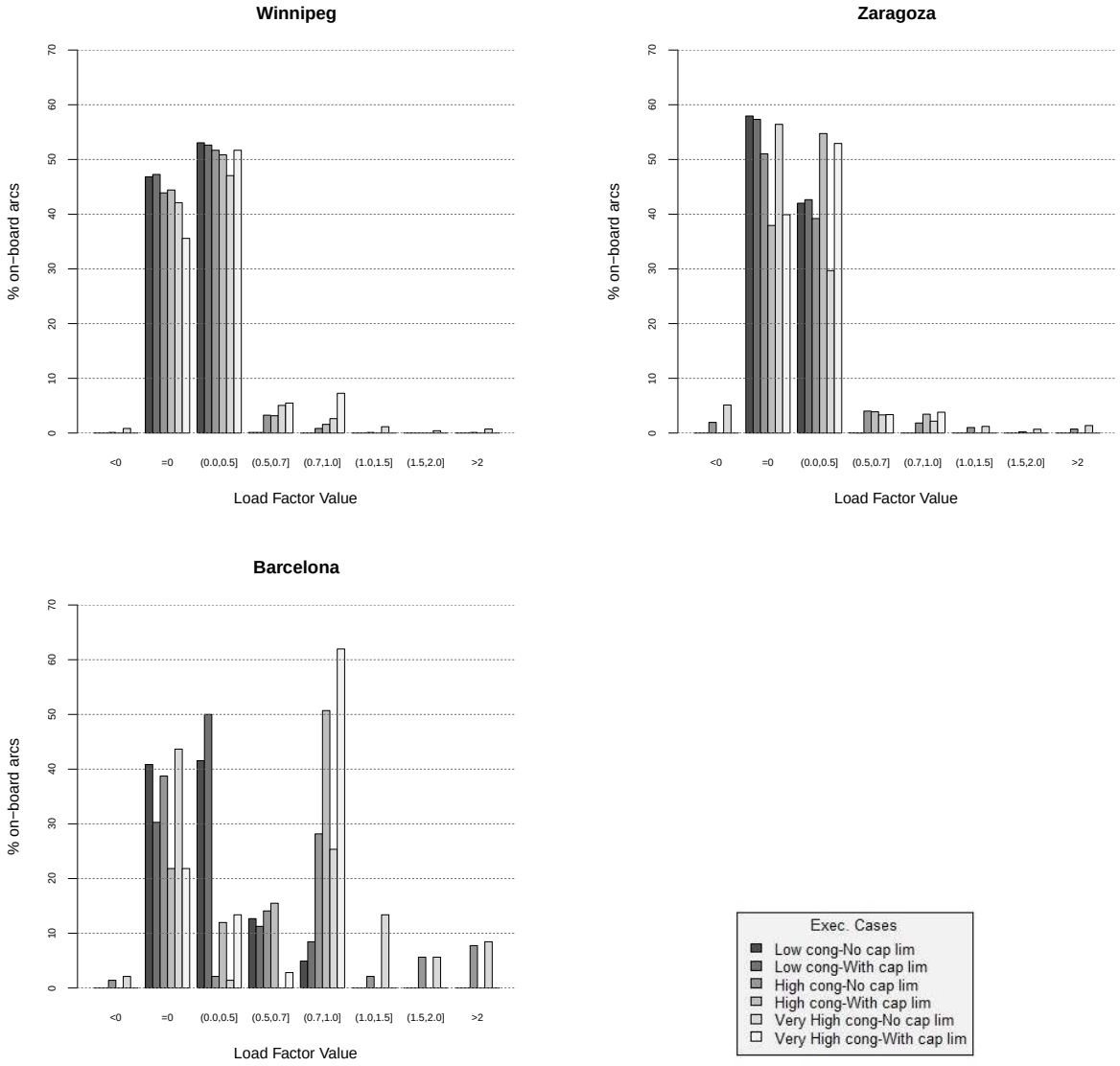


Figure 10: Distribution of the loading factor ρ at solutions found by the uncapacitated and capacitated C3F algorithms.

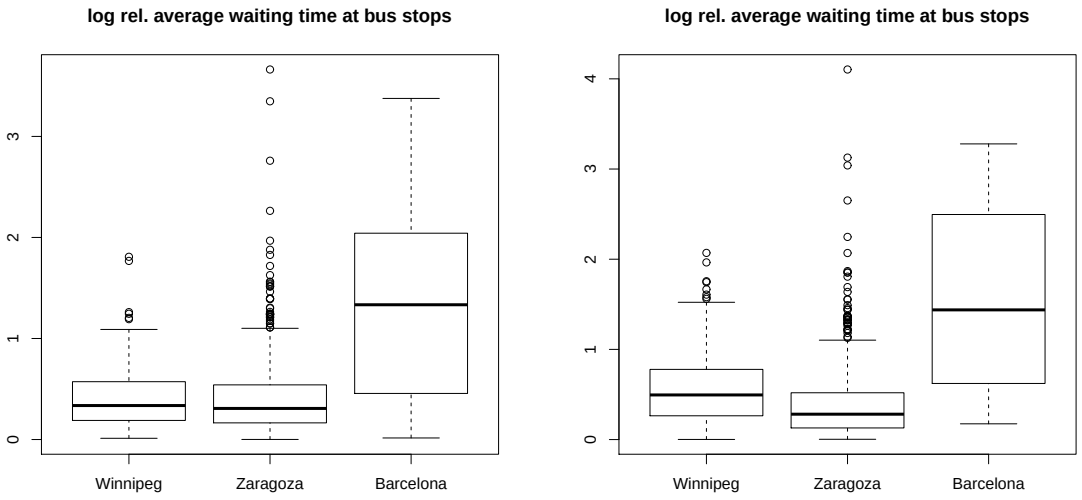


Figure 11: Boxplots for the relative average waiting time at bus stops as given by (38); High Congestion (left) and Very High Congestion (right) obtained by the capacitated C3F algorithm.

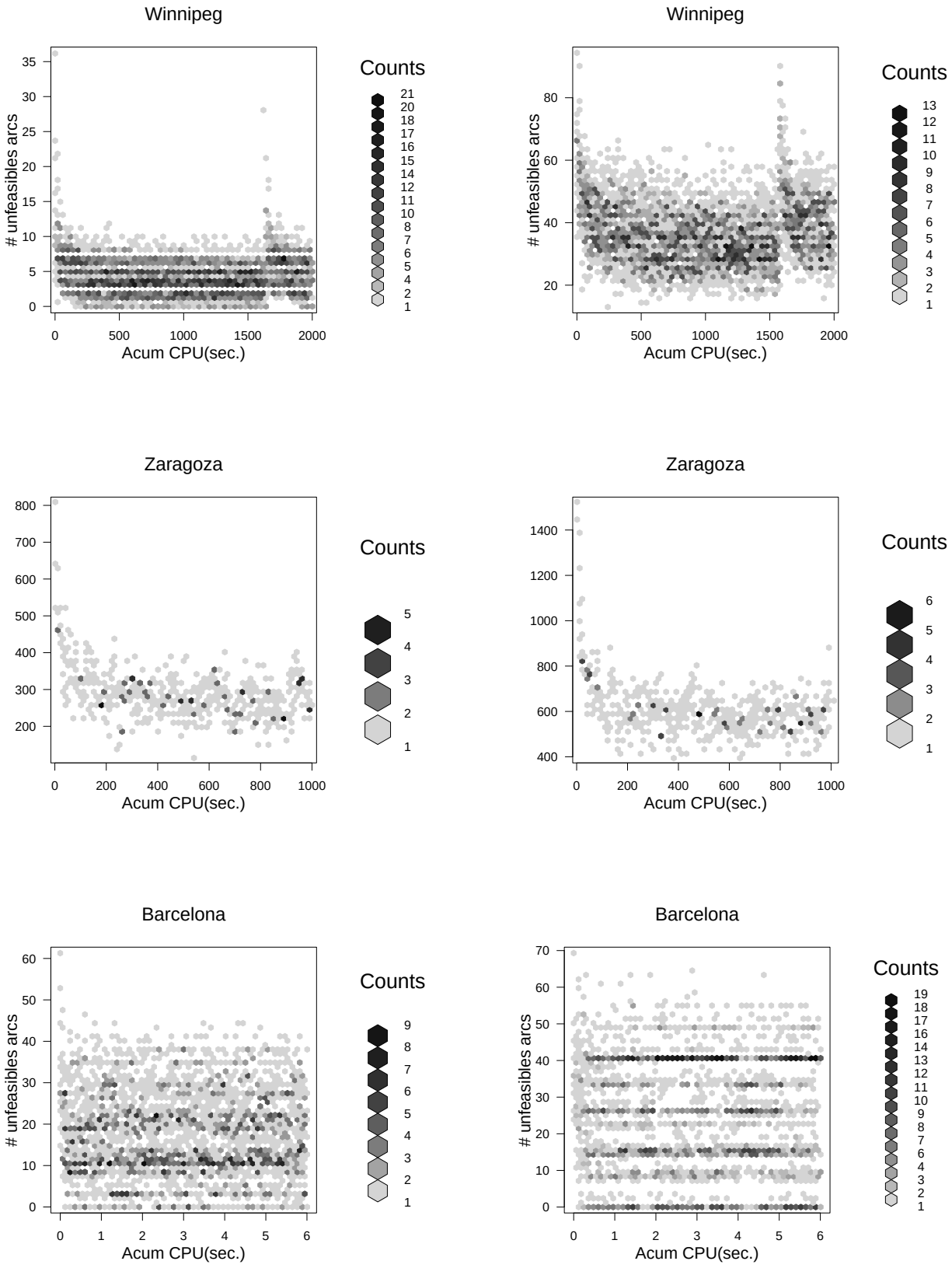
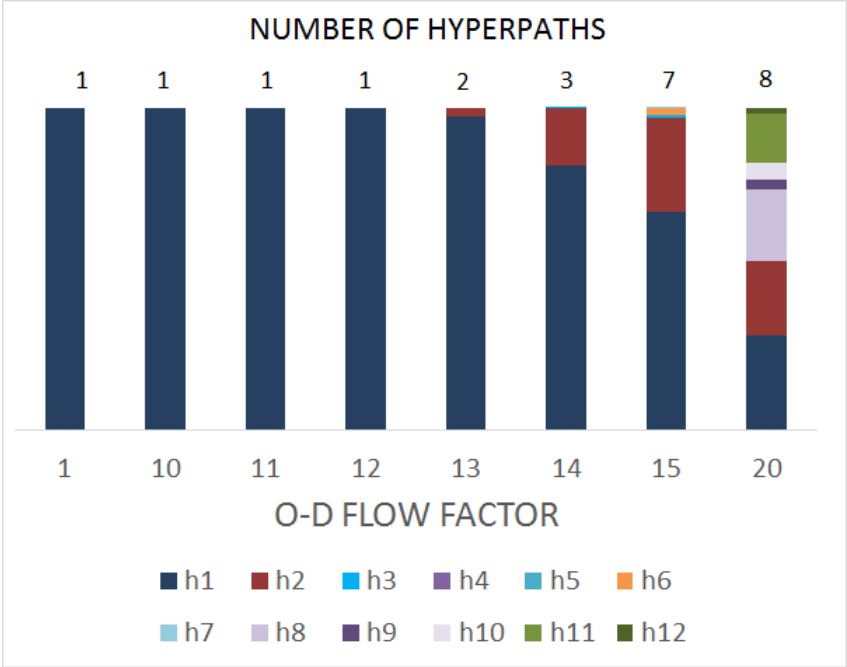


Figure 12: Evolution of the number of link flows beyond capacity limits in the uncapacitated C3F algorithm - cases of High Congestion case (left) and Very High congestion (right).



S

Figure 13: Winnipeg network. Hyperpath composition of flows for a given O-D pair (i, d) when its demand g_i^d is increased by different factors. All other O-D pair flows are set constant.

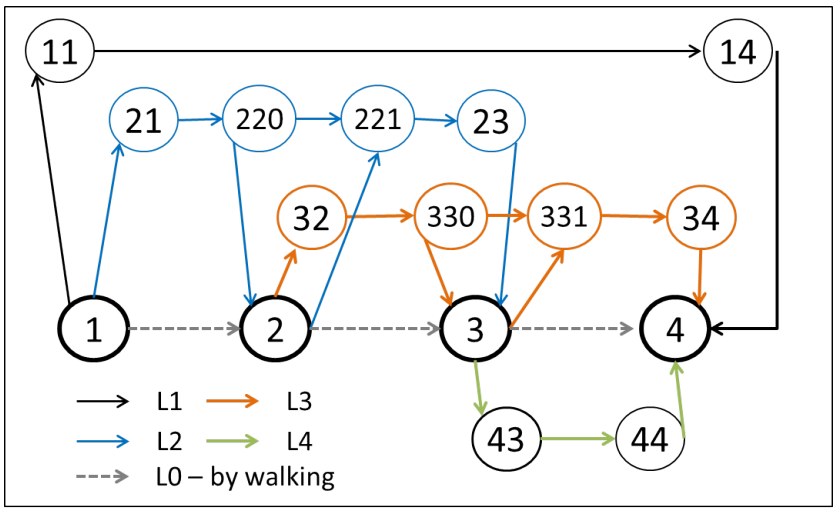


Figure 14: A slightly modified version of the Spiess and Florian test network.

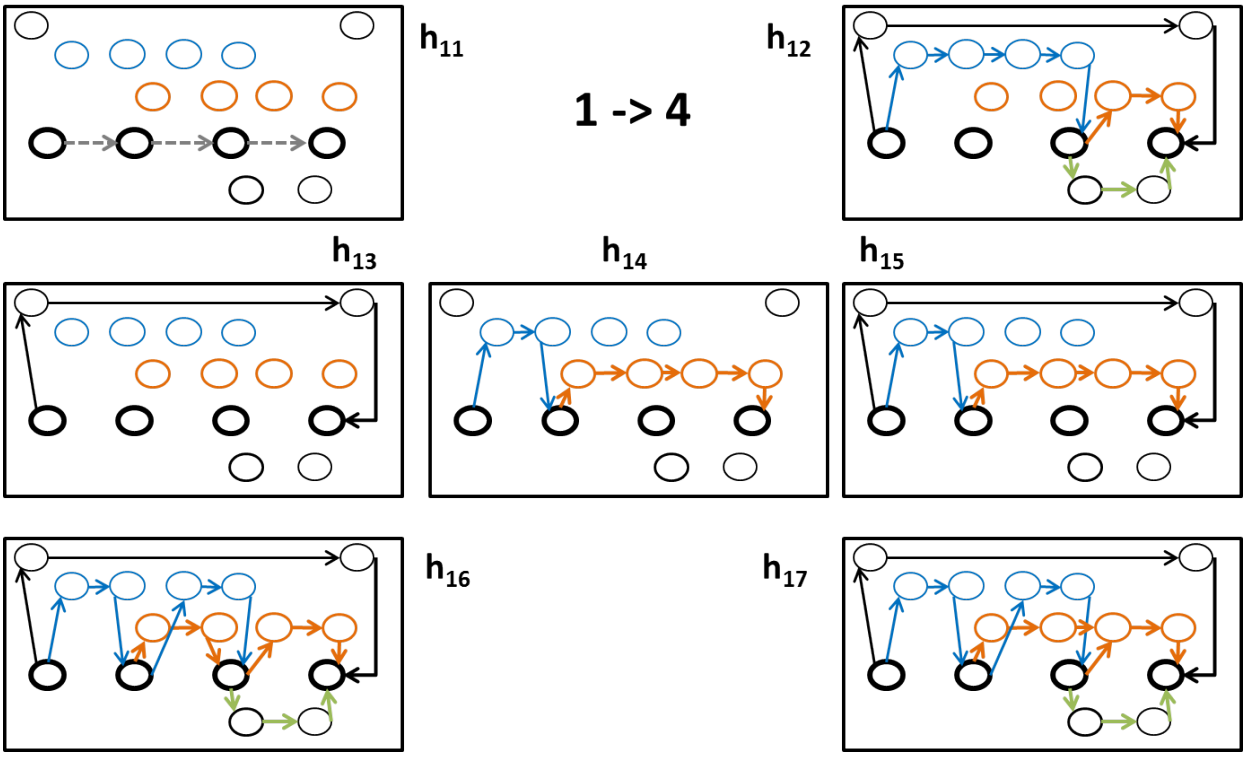


Figure 15: Hypergraphs h11 to h17 for the first O-D pair (1,4)

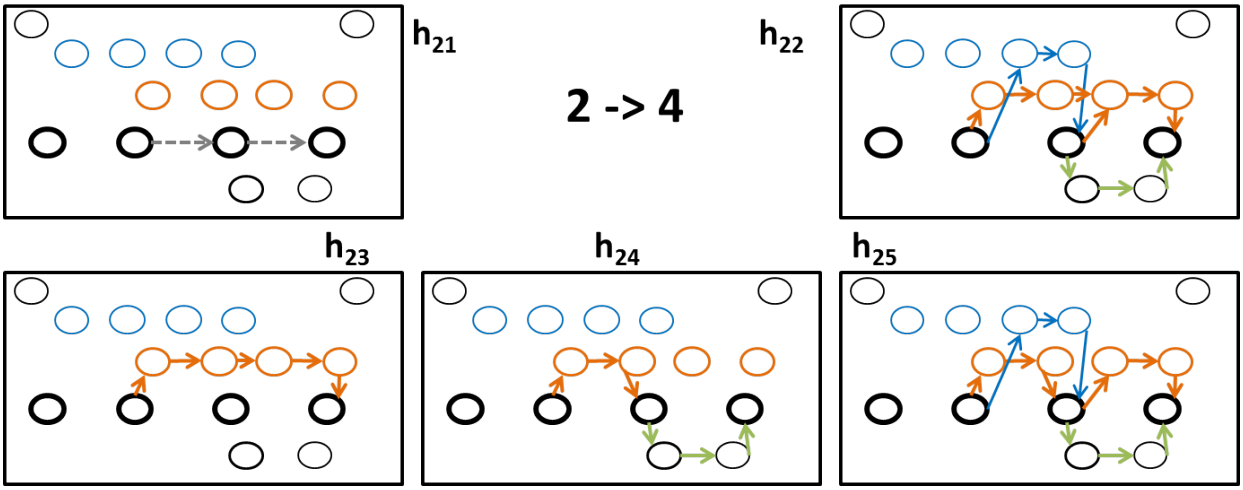


Figure 16: Hypergraphs h21 to h25 for the second O-D pair (2,4)