Transportation System Management under Congestion: A Bilevel Continuum Model

Yikang Hua

Department of Industrial and Systems Engineering, University of Wisconsin-Madison, yhua24@wisc.edu

Xin Wang

Department of Industrial and Systems Engineering, University of Wisconsin-Madison, xin.wang@wisc.edu

Abstract

Under the paradigm shift of smart city and autonomous vehicle technologies, modern transportation systems face new chances in traffic management. Given large-scale travel demand information, it empowers government agencies to make accurate and adaptive capacity planning decisions and traffic management operations. In light of this, we present an innovative mathematical framework for traffic management, where a bilevel continuum method is developed rather than using traditional discrete networks. In particular, the transportation system is modeled in a macroscopic view, where the road segment is infinitesimal and the travel time is evaluated on unit zones, or continuum, to characterize traffic congestion. The traffic management decisions are made in the upper level, which is optimized considering the lower level travelers' routing decisions, formulated as a Wardrop equilibrium. This helps agencies make fast and complex decisions over a large-scale transportation network to efficiently reduce congestion. With the help of our continuum model, analytical decisions are obtained for some simple symmetric transportation systems. A systematic solution algorithm for a general transportation system is also developed based on an iterative trust-region method. Through a series of numerical studies, the effectiveness and efficiency of our model are shown by comparing it with a traditional link-based network design model. Many interesting managerial insights are presented as well.

Key word: Continuum approximation; Traffic management; Wardrop equilibrium; Bilevel optimization

1 Introduction

The accelerating progress of urban expansion and the increasing mobility demand has resulted in a significant growth in the number of cars. The immediate side effect is the exhausting traffic congestion, especially for those in metropolitan areas. According to the recent research from a transport data company, INRIX (INRIX 2018), congestion is still a serious drain on the economy and caused nearly $87 billion of productivity loss in

2018 across the US, where the three worst cities are Boston, Chicago and Washington DC. As a result, many transportation systems in congested cities face the challenge that how to achieve efficient traffic management under rapidly changing mobility demand, considering the safety of travelers, the quality of life, and the environment benefit.

To reduce the traffic congestion in the new era, various advanced congestion mitigation mechanisms and innovative smart-city projects have been developed. New technologies are expected to change the game through a combination of advanced wireless communication, Internet of Things (IoT), big data, and autonomous vehicles. In an intelligent transportation system, traffic flow can be constantly monitored and managed by various instruments, such as smart traffic signals, viable and managed lanes, smart ramp metering, etc. Different from their traditional counterparts, these new smart instruments can operate in response to real-time traffic conditions or even coordinate over the whole transportation network. One typical example is the Variable Speed Limit (VSL) control, which can relieve the congestion at highway bottlenecks by performing different speed limits under different situations.

Although the technology, to some extent, has laid the platform required for network coordination, it still lacks of effective coordination strategies over a whole transportation system due to its large-scale nature. The traditional approach is to identify individual models independently for different management instruments (for example, reversing the direction of viable lanes, controlling the inflow of ramps, and alteration of road segments). Even if some demand-based transportation planning models are designed in a flexible way (Patriksson 2015), which allow their applications to a wide variety of planning problems in theory, it is impossible for scaling them up to coordinate multiple instruments over a large transportation network. This is mainly due to the curse of dimensionality in a discrete network setting, where instances of thousands of roads are easily seen in the analysis of city-scale problems. Detailed modeling for each road segment, e.g., as a link in a graph, normally makes the problem structure NP-hard (Fotakis et al. 2002) and leads to the intractability issue with reasonable computational resources. To make full use of the large-scale travel demand information and the flexibility of management brought by the intelligent transportation system, the government agencies are in urgent need of strategic and macroscopic tools that can guide their decision processes.

To fill this gap, we propose a continuum bilevel planning model to help government agencies make synthetic traffic management decisions with a macroscopic view. In our model, the government agency plays the role of the leader by controlling the road conditions of a traffic network (e.g., road capacity), and travelers make their corresponding routing decisions to minimize their individual travel costs (e.g., travel time). The travelers' joint routing behaviors are captured as a so-called Wardrop equilibrium (Wardrop 1952) to provide evaluation criteria of the traffic congestion. Compared with traditional traffic equilibrium models, where the congestion effect is captured by the increasing travel time on each road segment, our model features a continuous setting where the road conditions and travel time are both evaluated on unit zones, or "continuum", i.e., each road segment is infinitesimal. In our continuous setting, a large metropolitan

area is approximated as a bounded continuous plane where traffics move freely, and the local travel time is dependent on the local flow intensity and road configuration, but not upon direction (Yang et al. 1994; Yang and Wong 2000; Carlier et al. 2008). Our model provides an approach to model a large-scale urban traffic network from a macroscopic point of view when deciders only care about the average congestion level in different zones over a large area.

The bilevel continuum model that we propose can be applied to many traffic management and planning problems. First, our model can help design the road network for a new city given a predicted transport plan. Second, our mathematical framework considers various traffic conditions (e.g., road capacity, free-flow travel time) simultaneously, and thus is able to integrate different instruments and provides synthetic daily operational decisions (e.g., how to set variable speed limit, how to regulate viable lanes) for the modern transportation control system. Moreover, our model can serve as a building block, and help with more complicated problems that are congestion related. For example, as autonomous vehicles can reduce the headways and thus can increase the road capacity, our model can be integrated to help guide design and operations of autonomous vehicle managed lanes.

Our contributions are summarized as follows:

- We present an innovative bilevel integration model to solve large-scale traffic network design problems under the Wardrop equilibrium. The model builds upon the continuous congestion theory. This framework can serve as a building block for various complicated traffic management problems that are congestion related.
- We have obtained analytical solution for special cases (e.g., scenarios under symmetric city structures), and developed a fast numerical solution method to solve general large-scale instances. The bilevel structure of the problem is reduced to a convex optimization problem based on a local linear approximation of the conjugated dual lower level objective function.
- We conduct a series of analytical and numerical studies gradually to show the power of the model. We perform several symmetric cases in a circular city structure to show analytical patterns of the solution and their insights. Then, we compare the solution to its discrete counterpart for validation, and apply the algorithm to large-scale instances which are traditionally intractable. The results show the strong performance and effectiveness of our algorithm under tolerable accuracy sacrifice due to continuum approximation.

The remainder of this paper is organized as follows. §2 provides a literature review for some applications of the bilevel programming in the field of transportation and their solution algorithms, as well as the recent development of continuous congestion theory. In §3, we present a detailed introduction to our bilevel model under the continuous setting. Then in §4, we provide a trust-region iterative algorithm to solve the model we proposed. We provide the analytical solution for two representative cases and some managerial insights

are given in §5. The validation and efficiency of our model and solution method are supported by several numerical studies in §6. Finally, §7 concludes our work. The proofs of the lemmas and propositions in this paper are provided in the appendix.

2 Literature Review

Bilevel optimization problems are popular in many practical situations where the decision makers act according to a certain hierarchical order, and the upper level optimization problem can only be figured out by taking the lower level decisions into account. In the field of transportation, bilevel optimization models play an important role in solving strategic traffic management problems like congestion toll price setting (Bergendorff et al. 1997), transit line frequency (Gao et al. 2004) and road expansion (Gao et al. 2005). In most settings, the lower level model solves a traffic assignment problem where the travelers' path-choosing behavior is assumed to follow the Wardrop user-equilibrium principle. Soon after Wardrop formulated his principles, Beckmann et al. (1956) discovered that Wardrop equilibrium has a variational characterization, and thus can be converted to a convex optimization problem with the objective of optimizing a global criterion, taking into account the total congestion. De Palma et al. (1998) precisely addressed the relation between the concepts of equilibrium and optimum in static transportation networks, focusing in particular the theory of variational inequalities. Patriksson and Rockafellar (2002) formally formulated and analyzed the bilevel traffic management model as a mathematical program with Wardrop equilibrium conditions cast as a variational inequality problem, and provided a descent algorithm to solve it.

Although bilevel optimization problems with equilibrium constraints have many applications, they are hard to solve in general, because most of them are non-smooth, non-convex optimization problems. Lignola and Morgan (2001) proved the existence of solutions to the bilevel problems with equilibrium constraints under suitable assumptions, but raise the question of how to construct converging algorithms as an open problem. Maugeri and Raciti (2009) further studied the existence theorems based on monotonicity and give some necessary conditions for the solvability of the variational inequality problems. Although some progress has been made for solving mathematical programming with equilibrium constraints (Aussel and Lalitha 2018), in general all existing methods highly rely on fully convex setting with specific forms of objective functions, and until now it is still hard to obtain the exact solution for these problems. As an alternative, people have to compromise and take the second best, and specific heuristic methods with satisfactory performance have been developed for specific problems. Yang (1995) presented a sensitivity-analysis-based algorithm for a bilevel estimation model of origin-destination matrix in congested networks and showed the efficiency theoretically. A similar idea is presented in the work of Gao et al. (2004) in which they presented a bilevel transit network design model and solved it via a subgradient algorithm based on a linear approximation of the nonlinear and implicit relation between the changes of upper variables and the corresponding changes of lower variables. Gao et al. (2005) developed the bilevel network design problem

that deals with the selection of links to an existing road network, and solved the model by taking advantage of a mathematical relationship between the lower and upper level objective functions and combining it with the concept of the support functions in generalized Benders decomposition. Sun et al. (2008) presented a mixed-integer bilevel programming model to help seek the optimal location for logistics distribution centers, where the lower gives an equilibrium demand distribution, and proposed a simple heuristic method based on a linear approximation of the relationship between upper and lower variables. Chiou (2009) proposed a bilevel logistics network design problem and theoretically developed a subgradient method where the directional derivatives are obtained via the first-order sensitivity analysis of equilibrium constraints.

However, these conventional models feature integer variables to address discrete location and network related decisions, and thus can only be applied to moderate-scale instances. This motivates alternative approaches to formulate large-scale discrete problems with their asymptotically continuous counterpart. Yang et al. (1994); Yang (1996) first formulated a continuum equilibrium model to understand the traffic assignment problem under a continuous setting in which travel cost is a linear function of the traffic flow density. A dual-based solution method is developed and the model is solved by finite element method techniques. Wong and Yang (1999); Yang and Wong (2000) further extended this continuous user equilibrium by considering a more general setting, e.g., nonlinear cost functions, elastic customer demands and market attractiveness, and the customers' destination and path choices are captured by nonlinear partial differential equations. Maugeri (2001) studied the traffic equilibrium problem in the continuum case and expressed the equilibrium conditions by means of variational inequalities. Carlier et al. (2008) extended the classical Monge-Kantorovich problem by considering congestion and proved the existence and variational characterization of Wardrop equilibrium in a continuous space setting. In particular, they introduce the concept of flow intensity which can be viewed as the path-dependent analogue of the transport density in Monge's problem. Carlier and Santambrogio (2012) further provided a dual and tractable formulation for the short-term continuous static Wardrop user-equilibrium problem which can be solved over a discrete grid via a subgradient marching algorithm provided by Benmansour et al. (2010).

Besides the continuous congestion theory, a school of continuous approximation (CA) models (Newell 1971, 1973; Daganzo and Newell 1986; Ouyang and Daganzo 2006) has been developed as an alternative to provide near-optimal solutions to large-scale network design (e.g., facility location design) problems. These models are built upon the special case of an infinite homogeneous plane and have been applied to various location and routing problems. Li and Ouyang (2010) developed a CA model to analyze the facility location design problem where facilities are subject to risks of disruptions. Cui et al. (2010) validated the efficiency and accuracy of the CA model by comparing with the discrete model through a series of numerical experiments. Wang and Ouyang (2013) extended the CA approach to deal with facility location problems with spatial competition involved. Ouyang et al. (2015) further extended the topic by considering congestion effect in a continuous space, attributed with a static user equilibrium to characterize customers' route choice behavior. These CA models provide promising approaches to integrate both network design decisions and

traffic equilibrium while avoiding the excessive computational burden from the discrete models.

Our work bridges the gap, inherits and extends the single-level continuous congestion theory, and develops a CA-based bilevel optimization structure to conquer the challenges for intractable large-scale bilevel traffic management problems.

3 Model Formulation

To facilitate our modeling work, in Subsection 3.1, we first develop the concept of continuous congestion theory based on (Carlier et al. 2008) from a new angle. In particular, each road segment is treated as infinitesimal and congestion effect is evaluated on small continuous zones, or "continuum". We inherit and develop several fundamental lemmas, which are proved with the variational inequality tools. Then we establish our planning framework on these continuums in Subsection 3.2, which is a bilevel optimization problem. In particular, the government agencies take the lead to decide and control the road conditions, while travelers react to them in routing.

3.1 Continuous Traffic Equilibrium

Demand Pattern. The transportation system is assumed to be located in a bounded continuous region $\Omega \subset \mathbb{R}^2$ (e.g., a continuous region within a jurisdictional boundary of a city). Within the system, the demand of all origin-destination (OD) pairs is denoted by a function γ on $\Omega \times \Omega \mapsto \mathbb{R}_+ \cup \{0\}$, i.e., $\gamma(x, y)$ denotes the travel demand from point x to point y, with $x, y \in \Omega$. For modeling convenience, we assume that the region Ω is strongly connected so that there are always multiple routes between any OD pairs. We call such demand function γ as a *demand pattern* and consider it to be exogenous to our model. A demand pattern can usually be obtained by historical data or prediction models.

Route Pattern. Suppose any continuous curves within Ω are feasible paths, or rigorously, all paths in Ω are admissible. Let a continuous function $\sigma : [0, 1] \mapsto \Omega$ be a particular path, where $\sigma(0)$ and $\sigma(1)$ represent the corresponding origin and destination, respectively. Hence, we can define $S^{x,y} := \{\sigma | \sigma(0) = x, \sigma(1) = y\}$ to be the set of all paths connecting $x, y \in \Omega$, and $S := \{\sigma | \sigma(0) \in \Omega, \sigma(1) \in \Omega\}$ be the set of all paths. Further, we define the travel flow function $f : S \mapsto \mathbb{R}_+ \cup \{0\}$, i.e., the travel flow over each path $\sigma \in S$ is $f(\sigma)$. Then, the total travel flow with OD pair (x, y) is $\int_{S^{x,y}} f(\sigma)\, d\sigma$. For demonstration convenience, we call the travel flow across the entire S, denoted by $\boldsymbol{f} := \{f(\sigma)\}_{\sigma \in S}$, the *route pattern*, and may use $f(S^{x,y})$ instead of $\int_{S^{x,y}} f(\sigma)\, d\sigma$ in abbreviation without causing conflicts.

Flow Pattern. Similarly, we define the travel flow density on Ω as a function $\lambda : \Omega \mapsto \mathbb{R}_+ \cup \{0\}$, i.e., the flow density near point x is $\lambda(x)$. Hence, for an arbitrary sub-region $A \subseteq \Omega$, $\int_A \lambda(x)\, dx$ represents the total accumulated traffic flows in A. We define $\boldsymbol{\lambda} := \{\lambda(x)\}_{x \in \Omega}$ as a *flow pattern* on Ω, namely, the entire traffic flow profile over Ω.

Congestion & Infrastructure Pattern. To capture the congestion effect on a continuum, we let $g\left(c\left(x\right),\lambda\left(x\right)\right)$ be the travel time passing through a unit length near $x \in \Omega$, where $c(x)$ indicates the road conditions near x (e.g., the road capacity and/or the free-flow travel time). Note that the form of g is the same across Ω, while its variation over space is all captured by parameters $c\left(x\right)$. Here we ignore the direction of the traffic passing through x (isotopic case), i.e., all traffic with different directions share the same travel time near x. We call $\boldsymbol{\xi} := \left\{\xi(x)|\xi(x) = g\left(c\left(x\right),\lambda\left(x\right)\right)\right\}_{x\in\Omega}$ as a *congestion pattern*, and call $\boldsymbol{c} := \left\{c\left(x\right)\right\}_{x\in\Omega}$ as an *infrastructure pattern*. Under this setting, a traveler from x to y $(x, y \in \Omega)$ choosing a path $\sigma \in S^{x,y}$ spends time

$$T\left(\sigma\right) = \int_0^1 g\left(c\left(\sigma\left(t\right)\right),\lambda\left(\sigma\left(t\right)\right)\right)\left|\dot{\sigma}\left(t\right)\right| dt,$$

which is the line integral of $g\left(c\left(x\right),\lambda\left(x\right)\right)$ over the path σ.

With the above definition, given γ, we immediately have the flow conservation constraints for any $\boldsymbol{f}$,

$$f\left(S^{x,y}\right) = \gamma\left(x,y\right), \forall\left(x,y\right) \in \Omega\times\Omega, \tag{1}$$

with non-negativity,

$$f\left(\sigma\right) \geq 0, \forall\sigma \in S. \tag{2}$$

Hence, we denote $\mathcal{F}\left(\gamma\right)$ be the set of all feasible route patterns given γ,

$$\mathcal{F}\left(\gamma\right) := \left\{\boldsymbol{f}|\boldsymbol{f} \text{ satisfies Constraints (1) and (2)}\right\}.$$

As the flow pattern and the route pattern are just two different measures to represent the aggregated flow over the whole region, they are highly correlated. Specifically, Once a route pattern $\boldsymbol{f}$ is determined, the flow pattern $\boldsymbol{\lambda}$ should also be determined, while the same flow pattern can be resulted from multiple route patterns. To represent this, we call $\boldsymbol{f}$ and $\boldsymbol{\lambda}$ are compatible if

$$\int_\Omega \varphi\left(x\right)\lambda\left(x\right)dx = \int_S f\left(\sigma\right)\left(\int_0^1 \varphi\left(\sigma\left(t\right)\right)\left|\dot{\sigma}\left(t\right)\right| dt\right) d\sigma, \forall\varphi \in C_0\left(\Omega,\mathbb{R}_+\right), \tag{3}$$

where $\varphi\left(x\right)$ can be any arbitrary zero-order continuous functions on Ω. Especially, if φ represents the travel time near x, i.e., replacing φ by g, both the left hand side and the right hand side are the total travel time of all traffic across the region, which should be identical intuitively.

Similarly, we denote $\Lambda\left(\gamma\right)$ to be the set of all feasible flow patterns given γ:

$$\Lambda\left(\gamma\right) := \left\{\boldsymbol{\lambda}|\exists\boldsymbol{f} \in \mathcal{F}\left(\gamma\right) \text{ such that Constraints (3) are satisfied}\right\},$$

that is, $\boldsymbol{\lambda}$ is compatible to a certain feasible route pattern resulting from γ.

Lemma 1. *Given γ, the set $\Lambda\left(\gamma\right)$ is bounded, closed and convex.*

Wardrop Equilibrium. The aggregated behavior of travelers is captured following the principle of Wardrop equilibrium: first, all used paths that connect the same OD must provide the same travel time;

second, all the other possible but unused paths must provide a longer travel time. In other words, in a continuous setting, only the "geodesic paths", or the path with the shortest travel time, between any two points are used. A continuous version of Wardrop equilibrium conditions can be defined as follows: If $\boldsymbol{f}^* \in \mathcal{F}(\gamma)$ is a route pattern under Wardrop equilibrium, with its corresponding compatible flow pattern $\boldsymbol{\lambda}^* \in \Lambda(\gamma)$, then there exists $\{u^*(x,y)\}_{(x,y)\in\Omega\times\Omega}$ such that

$$\int_0^1 g\left(c\left(\sigma\left(t\right)\right), \lambda^*\left(\sigma\left(t\right)\right)\right)\left|\dot{\sigma}\left(t\right)\right| dt \geq u^*\left(x,y\right), \forall \sigma \in S^{x,y}, (x,y) \in \Omega \times \Omega, \tag{4}$$

$$f^*\left(\left\{\sigma \in S^{x,y} : \int_0^1 g\left(c\left(\sigma\left(t\right)\right), \lambda^*\left(\sigma\left(t\right)\right)\right)\left|\dot{\sigma}\left(t\right)\right| dt > u^*\left(x,y\right)\right\}\right) = 0, \forall\left(x,y\right) \in \Omega \times \Omega. \tag{5}$$

where $u^*(x,y)$ is the actual travel time between x and y in equilibrium. Constraints (4) indicate that the travel time of any paths connecting each OD pair (x,y) have to be longer or equal to their actual travel time. Constraints (5) (written with a similar abbreviation as Constraints (1)) guarantees that there is no positive flow on those paths with longer travel time. Similar to the traditional link-path model, the continuous Wardrop equilibrium conditions can be characterized by variational principles with some reasonable assumptions on $g\left(c\left(x\right), \lambda\left(x\right)\right)$.

Lemma 2. *Given $\boldsymbol{c} \in \mathcal{C}$, if the congestion function $g\left(c\left(x\right), \lambda\left(x\right)\right)$ is convex, continuous and strictly increasing with respect to $\lambda(x)$ for $\forall x \in \Omega$, then the Wardrop equilibrium flow pattern $\boldsymbol{\lambda}^*$ coincides with the set of solutions to the following optimization problem:*

$$\min_{\boldsymbol{\lambda}\in\Lambda(\gamma)} \int_\Omega h\left(c\left(x\right), \lambda\left(x\right)\right) dx, \tag{6}$$

where $h\left(c\left(x\right), \lambda\left(x\right)\right)$ satisfies the relationship that $\frac{\partial h(c(x),\lambda(x))}{\partial \lambda(x)} = g\left(c\left(x\right), \lambda\left(x\right)\right)$.

Remark. The typical link performance function is convex, continuous and strictly increasing in terms of link flow (Sheffi 1985), so it is reasonable to make the assumption that the function $g\left(c\left(x\right), \lambda\left(x\right)\right)$ is convex, continuous and strictly increasing in terms of $\lambda(x)$ for $\forall x \in \Omega$ under the continuous setting. To avoid reiterated explanations, we will assume this point in the sequel without further mention it. Such continuous version of traffic assignment problem can be numerically solved (Carlier and Santambrogio 2012).

Lemma 3. *The Wardrop equilibrium point flow pattern $\boldsymbol{\lambda}^*$ exists and is unique.*

3.2 Continuous Transportation Planning Model

With the above building blocks, we now derive our transportation planning models. First, we consider that the government agency focuses on infrastructure planning decisions $\boldsymbol{c}$ to meet a desired goal. A common goal is to improve the overall system efficiency that can be reflected by the total travel time, i.e, $\int_S f(\sigma) T(\sigma) d\sigma$. By Constraints (3), the system travel time can be rewritten as $\int_\Omega g\left(c\left(x\right), \lambda\left(x\right)\right) \lambda(x) dx$. Moreover, a

modification of the infrastructure pattern normally incurs a certain cost, which should also be considered in the objective. Specifically, we use $L(\boldsymbol{c})$ to denote the planning cost for an infrastructure pattern $\boldsymbol{c}$, hence, the government agency should seek trade-off between investment costs and the system performance, which is described as follows,

$$\min_{\boldsymbol{c}\in\mathcal{C}} \int_{\Omega} g\left(c\left(x\right),\lambda\left(x\right)\right)\lambda\left(x\right)dx + L\left(\boldsymbol{c}\right), \tag{7}$$

where $\mathcal{C}$ denotes the set of all feasible infrastructure patterns. We should note that, the flow pattern $\boldsymbol{\lambda}$ here is the result of a certain route pattern $\boldsymbol{f}$, which captures travelers' routing decisions given $\boldsymbol{c}$ and γ. We would like to stress at this stage that, we shall be concerned only with the case of static equilibrium conditions and ignore possible time-dependent (dynamic) effects in this work. We assume that the aggregated behavior of travelers follow the principle of Wardrop equilibrium, and thus the lower level problem is a typical traffic assignment equilibrium problem, which can be described as an optimization problem by Lemma 2,

$$\min_{\boldsymbol{\lambda}\in\Lambda(\gamma)} \int_{\Omega} h\left(c\left(x\right),\lambda\left(x\right)\right)dx, \tag{8}$$

where $h\left(c\left(x\right),\lambda\left(x\right)\right)$ satisfies the relationship that $\frac{\partial h(c(x),\lambda(x))}{\partial\lambda(x)} = g\left(c\left(x\right),\lambda\left(x\right)\right)$. The existence and uniqueness of its solution $\boldsymbol{\lambda}^*$ follows Lemma 3.

With previous discussion, we are now able to formulate our continuum bilevel transportation planning model.

Proposition 1. *The continuum transportation planning model (CTPM) can be formulated as the following*

$$\text{(CTPM)} \qquad \begin{aligned} &\min_{\boldsymbol{c}\in\mathcal{C}} \int_{\Omega} g\left(c\left(x\right),\lambda^*\left(x\right)\right)\lambda^*\left(x\right)dx + L\left(\boldsymbol{c}\right), \\ &s.t. \quad \boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}\in\Lambda(\gamma)} \int_{\Omega} h\left(c\left(x\right),\lambda\left(x\right)\right)dx. \end{aligned} \tag{9}$$

4 Solution Algorithm

The general bilevel model (9) formulated in Section 3 is difficult to solve due to two reasons. First, the feasible region $\Lambda\left(\gamma\right)$ in the lower level problem is highly implicit and lack of feasibility information. Second, it is difficult to solve a bilevel programming problem with optimization algorithms in general since it is NP-hard and normally non-convex. In this section, we first reformulate the model based on the duality of the lower level problem, which features a greatly-simplified feasible region, and then provide a trust-region-based method to solve it.

4.1 Lower Level Reformulation

We will consider a reformulation of the lower level problem of (9) from the decision of flow pattern $\boldsymbol{\lambda}$ into an equivalent form to the decision of congestion patterns. The conversion is a direct application of the conjugate

dual reformulation by Carlier and Santambrogio (2012). To achieve that, we need some further definitions.

1) Recall $\xi(x) = g(c(x), \lambda(x))$ represents the travel time under congestion for $\forall x \in \Omega$. Given a fixed infrastructure pattern $\boldsymbol{c}$, due to the monotonicity of g, there exists a bijection between the flow pattern $\boldsymbol{\lambda}$ and the congestion pattern $\boldsymbol{\xi}$. Hence we simply denote $\lambda(x) = g^{-1}(c(x), \xi(x))$ to represent this inverse relationship.

2) Given a congestion pattern $\boldsymbol{\xi}$, we define

$$u_{\boldsymbol{\xi}}(x, y) := \inf_{\sigma \in S^{x,y}} \int_0^1 \xi(\sigma(t)) |\dot{\sigma}(t)| \, dt$$

as the geodesic distance (travel time) between any two points $x, y \in \Omega$. Given fixed $x, y \in \Omega$, the mapping $\boldsymbol{\xi} \to u_{\boldsymbol{\xi}}(x, y)$ is concave as the minimum of linear functions of $\boldsymbol{\xi}$(Benmansour et al. 2010).

3) We define

$$\bar{h}(c(x), \xi(x)) := \sup_{\lambda(x) \geq 0} \{\xi(x) \lambda(x) - h(c(x), \lambda(x))\}$$

as the conjugate function of $h(c(x), \lambda(x))$ related to $\lambda(x)$.

4) We define a set

$$\Xi(\boldsymbol{c}) : \{\boldsymbol{\xi} : \xi(x) \geq g(c(x), 0) \text{ for } \forall x \in \Omega\}$$

which includes all feasible congestion patterns $\boldsymbol{\xi}$. Then the following proposition holds.

Proposition 2. *The CTPM (9) can be reformulated as its dual version,*

$$\text{(DCTPM)} \qquad \min_{\boldsymbol{c} \in \mathcal{C}} \int_\Omega g^{-1}(c(x), \xi^*(x)) \xi^*(x) \, dx + L(\boldsymbol{c}), \tag{10}$$

$$s.t. \quad \boldsymbol{\xi}^* = \arg\min_{\boldsymbol{\xi} \in \Xi(\boldsymbol{c})} \int_\Omega \bar{h}(c(x), \xi(x)) \, dx - \int_{\Omega \times \Omega} u_{\boldsymbol{\xi}}(x, y) \gamma(x, y) \, dx dy. \tag{11}$$

The lower level feasible region of reformulated model (10), i.e., $\Xi(\boldsymbol{c})$, is much simpler than that of (9), and this advantage provides the possibility for the optimal upper level decision $\boldsymbol{c}$ to be solved.

4.2 Linear Approximation Based Trust Region Method

The difficulty in solving the DCTPM lies in how to analytically evaluate the equilibrium congestion pattern $\boldsymbol{\xi}^*$ for a given upper level decision $\boldsymbol{c}$. Specifically, if we know a relatively explicit expression $\boldsymbol{\xi}^*(\boldsymbol{c})$ from solving the lower level problem, the DCTPM can be solved as a single level optimization. We notice that Benmansour et al. (2010) provides a subgradient marching algorithm to solve the lower level problem over a discrete grid, which calculates the subgradient of the concave mapping $\boldsymbol{\xi} \to u_{\boldsymbol{\xi}}(x, y)$. Therefore, we seek a linear approximation of $u_{\boldsymbol{\xi}}(x, y)$ based on this subgradient, which can greatly simplify the structure of the bilevel problem and further develop a trust-region numerical algorithm to solve the original DCTPM.

Flow Pattern Estimator. We denote $\boldsymbol{\delta}_{\boldsymbol{\xi}_0}^{x', y'}$ to be the subgradient at $\boldsymbol{\xi}_0$ of the mapping $\boldsymbol{\xi} \to u_{\boldsymbol{\xi}}(x', y')$ for a specific OD pair $x', y' \in \Omega$. For any point $x \in \Omega$, $\boldsymbol{\delta}_{\boldsymbol{\xi}_0}^{x', y'}(x)$ tells how much the geodesic distance between

x', y' is sensitive to a perturbation on $\xi_0(x)$. Then for all $\boldsymbol{\xi}$ that is close enough to $\boldsymbol{\xi}_0$, we can obtain a linear approximation of $u_{\boldsymbol{\xi}}(x', y')$ by performing its first order Taylor's expansion as follows,

$$u_{\boldsymbol{\xi}}(x', y') = u_{\boldsymbol{\xi}_0}(x', y') + \left\langle \boldsymbol{\delta}_{\boldsymbol{\xi}_0}^{x', y'}, \boldsymbol{\xi} - \boldsymbol{\xi}_0 \right\rangle, \tag{12}$$

where we define $\langle \cdot \rangle$ as a natural extension of inner product to the infinite-dimensional space, i.e., $\left\langle \boldsymbol{\delta}_{\boldsymbol{\xi}_0}^{x', y'}, \boldsymbol{\xi} - \boldsymbol{\xi}_0 \right\rangle = \int_\Omega \boldsymbol{\delta}_{\boldsymbol{\xi}_0}^{x', y'}(x) \cdot (\xi(x) - \xi_0(x))\, dx$. In this way, given a $\boldsymbol{\xi}_0$, for any $\boldsymbol{\xi}$, s.t. $\|\boldsymbol{\xi} - \boldsymbol{\xi}_0\| \leq \Delta$, we accept the following linear approximation for the term $\int_{\Omega\times\Omega} u_{\boldsymbol{\xi}}(x, y)\, \gamma(x, y)\, dxdy$ in the lower level problem (11)

$$\int_{\Omega\times\Omega} u_{\boldsymbol{\xi}}(x, y)\, \gamma(x, y)\, dxdy \approx \int_\Omega \alpha_\gamma(x)\, \xi(x)\, dx + \beta_\gamma, \tag{13}$$

where

$$\alpha_\gamma(x) = \int_{\Omega\times\Omega} \boldsymbol{\delta}_{\boldsymbol{\xi}_0}^{x', y'}(x)\, \gamma(x', y')\, dx'dy', \tag{14}$$

$$\beta_\gamma = \int_{\Omega\times\Omega} \left(u_{\boldsymbol{\xi}_0}(x', y') - \int_\Omega \boldsymbol{\delta}_{\boldsymbol{\xi}_0}^{x', y'}(x)\, \xi_0(x)\, dx \right) \gamma(x', y')\, dx'dy'.$$

The value of Δ should be interpreted as a "trust region" radius, literally meaning that the accuracy of this linear approximation can be trusted only for $\boldsymbol{\xi}$ within a distance of Δ near $\boldsymbol{\xi}_0$. Hence, fixing $\boldsymbol{\xi}_0$ and Δ, we have $\boldsymbol{\alpha}_\gamma := \{\alpha_\gamma(x)\}_{x\in\Omega}$ only depends on the transport plan γ, and is independent of the congestion pattern $\boldsymbol{\xi}$. By plugging equation (13) into (11), the lower level problem can be approximated locally near $\boldsymbol{\xi}_0$ as

$$\boldsymbol{\xi}^* = \arg\min_{\xi(x)\geq g(c(x),0)} \int_\Omega \left(\bar{h}(c(x), \xi(x)) - \alpha_\gamma(x)\, \xi(x) \right) dx, \forall x \in \Omega, \tag{15}$$

where we ignore the constant term β_γ since it will not affect the solution. In this way, the approximated lower level problem can be solved point-wisely for each $x \in \Omega$ and thus we can obtain an approximated evaluation function $\boldsymbol{\xi}^*(\boldsymbol{c})$ in a decomposable manner. The following proposition reveals the inner relationship between $\alpha_\gamma(x)$ and $\lambda(x)$ and directly give the solution to the lower level problem within the trust region.

Proposition 3. *If the convex function $g(c(x), \lambda(x))$ is super-linear with respect to $\lambda(x)$, we have $\xi^*(x) = g(c(x), \alpha_\gamma(x))$.*

Proposition 3 indicates that $\xi(x)$ achieves optimum in the lower level problem when its corresponding flow $\lambda(x)$ coincides with $\alpha_\gamma(x)$ for all $x \in \Omega$. In this sense we name $\boldsymbol{\alpha}_\gamma$ as the *flow pattern estimator.*

However, we should recall that the approximated bilevel model can only be trusted within a certain range, and if the congestion pattern $\boldsymbol{\xi}^*$ we obtained by solving (15) is not close enough to $\boldsymbol{\xi}_0$ based on which we build the linear approximation (13), its accuracy cannot be guaranteed. In light of this, we provide an algorithm to adjust $\boldsymbol{\xi}_0$ and Δ.

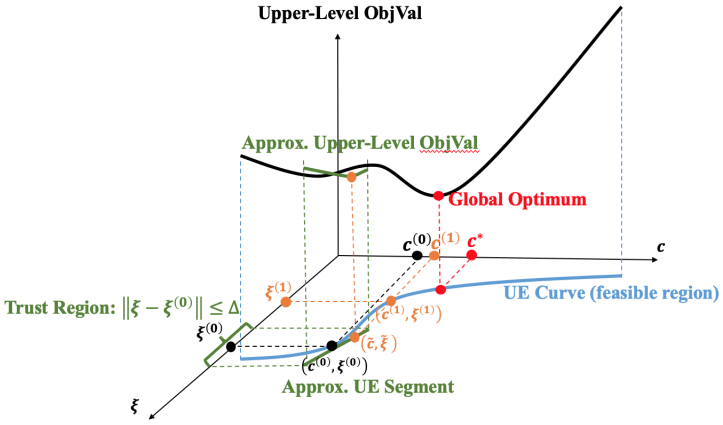


Figure 1: Illustration for the first iteration of the linear approximation based trust region method

We demonstrate the algorithm with the help of Figure 1. The UE curve illustrates the abstract mapping from a given infrastructure design $\boldsymbol{c}$ to its corresponding congestion pattern $\boldsymbol{\xi}$ under equilibrium, which indicates the lower-level optimization problem (with Lemma 2). The Upper-level ObjVal indicates the mapping from each feasible solution $(\boldsymbol{c}, \boldsymbol{\xi})$ determined by the lower-level UE curve, to its corresponding objective value of the upper-level optimization, expressed as the black curve. The global optimum point of the bi-level problem, and the corresponding optimal infrastructure design $\boldsymbol{c}^*$, is expressed in red in Figure 1. To approach the optimal design, we first construct an initial solution. Let $\boldsymbol{c}^{(0)}$ be the original infrastructure pattern (or any arbitrary pattern for a new city). Then we can obtain its corresponding equilibrium congestion pattern $\boldsymbol{\xi}^{(0)} := \boldsymbol{\xi}_{\boldsymbol{c}^{(0)}}$ which corresponds to a point on the UE curve. Then based on $\boldsymbol{\xi}^{(0)}$, we can calculate $\boldsymbol{\alpha}_\gamma$ by (14), which enables us to build a localized approximated bilevel program as follows,

$$\text{(Localized-DCTPM)} \qquad \begin{aligned} &\min_{\boldsymbol{c}\in\mathcal{C}} \int_\Omega g^{-1}\left(c\left(x\right), \xi^*\left(x\right)\right) \xi^*\left(x\right) dx + L\left(\boldsymbol{c}\right), \\ &\text{s.t.} \quad \boldsymbol{\xi}^* = \arg \min_{\boldsymbol{\xi}\in\Xi(\boldsymbol{c}), \left\|\boldsymbol{\xi}-\boldsymbol{\xi}^{(0)}\right\|\leq\Delta} \int_\Omega \left(\bar{h}\left(c\left(x\right), \xi\left(x\right)\right) - \alpha_\gamma\left(x\right)\xi\left(x\right)\right) dx. \end{aligned} \tag{16}$$

The above localized-DCTPM is a local approximation to the UE curve with estimated flow pattern $\boldsymbol{\alpha}_\gamma$, denoted by the approximated UE segment in Figure 12. This shifts the the original DCTPM to the localized-DCTPM. Notice that we add the trust-region constraints $\left\|\boldsymbol{\xi} - \boldsymbol{\xi}^{(0)}\right\| \leq \Delta$, which limits the deviation of the solution from the DCTPM to the localized-DCTPM with a tolerable accuracy. We would like to emphasize that the Localized-DCTPM should be much easier to solve than the original DCTPM, especially if the trust-region constraints $\left\|\boldsymbol{\xi} - \boldsymbol{\xi}^{(0)}\right\| \leq \Delta$ and the upper level constraints $\boldsymbol{c} \in \mathcal{C}$ are point-wisely independent. In §6.1 we will show the solution to the Localized-DCTPM under a little more specific setting, but at this stage

we simply denote its solution to be $\left(\tilde{\boldsymbol{c}}, \tilde{\boldsymbol{\xi}}\right)$, as shown in Figure 1.

Then, given the solution to the localized-DCTPM $\left(\tilde{\boldsymbol{c}}, \tilde{\boldsymbol{\xi}}\right)$, we hope to update the solution to the original DCTPM, $\left(\boldsymbol{c}^{(0)}, \boldsymbol{\xi}^{(0)}\right)$. Note that $\tilde{\boldsymbol{\xi}}$ is normally not feasible to DCTPM, since $\left(\tilde{\boldsymbol{c}}, \tilde{\boldsymbol{\xi}}\right)$ is on the approximated UE segment rather than the UE curve. To generate a feasible solution to DCTPM, we project the point $\left(\tilde{\boldsymbol{c}}, \tilde{\boldsymbol{\xi}}\right)$ to the UE curve with the same $\tilde{\boldsymbol{c}}$, by solving the traffic equilibrium given $\tilde{\boldsymbol{c}}$, denoted by $\left(\boldsymbol{c}^{(1)}, \boldsymbol{\xi}^{(1)}\right) := (\tilde{\boldsymbol{c}}, \boldsymbol{\xi}_{\tilde{\boldsymbol{c}}})$. With reasonable range of the trust region, we can expect $\left(\boldsymbol{c}^{(1)}, \boldsymbol{\xi}^{(1)}\right)$ to be better than $\left(\boldsymbol{c}^{(0)}, \boldsymbol{\xi}^{(0)}\right)$. Following this updating procedure, we can iteratively obtain a sequence of feasible solution to DCTPM, denoted by $\left\{\left(\boldsymbol{c}^{(k)}, \boldsymbol{\xi}^{(k)}\right)\right\}_{k=1,2,\ldots,n}$. If some stopping criterion is met (e.g., $\tilde{\boldsymbol{\xi}}$ and $\boldsymbol{\xi}_{\tilde{\boldsymbol{c}}}$ are sufficiently close), we stop and take the best solution among $\left(\boldsymbol{c}^{(k)}, \boldsymbol{\xi}^{(k)}\right)$ to be our final solution. We summarize the complete process with some stopping criteria in Algorithm 1.

Algorithm 1 Trust Region Algorithm

function TRA$(\epsilon_c, \epsilon_\xi, k_{\max})$:

Initialize: $I_{Stop} = 0, k = 0, \Delta^{(0)} = \Delta, 0 < \eta_1 \leq \eta_2 < 1, 0 < \zeta_1 < 1 < \zeta_2$

Set: $\boldsymbol{c}^{(k)} = \boldsymbol{c}^* = \boldsymbol{c}_0$,

Solve: $\boldsymbol{\xi}^{(k)} = \arg\min_{\boldsymbol{\xi}\in\Xi(\boldsymbol{c})} \int_\Omega \bar{h}\left(c^{(k)}(x), \xi(x)\right) dx - \int_{\Omega\times\Omega} c_{\boldsymbol{\xi}}(x,y)\,\gamma(x,y)\,dxdy.$

Set: $optVal = \int_\Omega g^{-1}\left(c^{(k)}(x), \xi^{(k)}(x)\right)\xi^{(k)}(x)\,dx + L\left(\boldsymbol{c}^{(k)}\right)$

While $(I_{Stop} == 0)$:

Calculate: $\boldsymbol{\alpha}_\gamma^{(k)} = \int_{\Omega\times\Omega} \boldsymbol{\delta}_{\boldsymbol{\xi}^{(k)}}^{x,y}\,\gamma(x,y)\,dxdy.$

Solve:

$$\left(\boldsymbol{c}^{(k+1)}, \tilde{\boldsymbol{\xi}}\right) = \arg\min_{\boldsymbol{c}\in\mathcal{C}} \int_\Omega g^{-1}\left(c(x), \xi^*(x)\right)\xi^*(x)\,dx + L(\boldsymbol{c}) \tag{17}$$
$$\text{s.t.} \quad \boldsymbol{\xi}^* = \arg\min_{\boldsymbol{\xi}\in\Xi(\boldsymbol{c}), \left\|\boldsymbol{\xi}-\boldsymbol{\xi}^{(k)}\right\| \leq \Delta^{(k)}} \int_\Omega \left(\bar{h}\left(c(x), \xi(x)\right) - \alpha_\gamma(x)\,\xi(x)\right) dx.$$

Solve: $\boldsymbol{\xi}^{(k+1)} = \arg\min_{\boldsymbol{\xi}\in\Xi(\boldsymbol{c})} \int_\Omega \bar{h}\left(c^{(k+1)}(x), \xi(x)\right) dx - \int_{\Omega\times\Omega} c_{\boldsymbol{\xi}}(x,y)\,\gamma(x,y)\,dxdy.$

Set: $curVal = \int_\Omega g^{-1}\left(c^{(k+1)}(x), \xi^{(k+1)}(x)\right)\xi^{(k+1)}(x)\,dx + L\left(\boldsymbol{c}^{(k+1)}\right)$

If $curVal > objVal$:

Set: $\boldsymbol{c}^* = \boldsymbol{c}^{(k+1)}$, $objVal = curVal$.

Calculate: $\rho^{(k+1)} = \frac{\left\|\xi^{(k+1)} - \tilde{\boldsymbol{\xi}}\right\|}{\left\|\xi^{(k+1)}\right\|}$.

If $\rho^{(k+1)} > \eta_2$:

Set: $\left(\boldsymbol{c}^{(k+1)}, \boldsymbol{\xi}^{(k+1)}\right) = \left(\boldsymbol{c}^{(k)}, \boldsymbol{\xi}^{(k)}\right), \Delta^{(k+1)} = \zeta_1\Delta^{(k)}$

If $\rho^{(k+1)} < \eta_1$:

Set: $\Delta^{(k+1)} = \zeta_2\Delta^{(k)}$

If one of the following criteria is met:

(1) $\left\|\boldsymbol{c}^{(k+1)} - \boldsymbol{c}^{(k)}\right\| < \epsilon_c$

(2) $\left\|\boldsymbol{\xi}^{(k+1)} - \tilde{\boldsymbol{\xi}}\right\| < \epsilon_\xi$

(3) $k > k_{\max}$

Set: $I_{Stop} = 1$

Set: $k = k + 1$.

Return: $\boldsymbol{c}^*$

5 Analysis of Representative Transportation System

In this section, we apply the proposed modeling framework in §3 to a series of simple symmetric cases where the geodesic routes can be easily analyzed.

We first consider a symmetric single-ring system. It can be regarded as an approximation to a circular city structure where residents live in several circular belts around the city center. Suppose a city area is

represented by $\Omega = \{(\rho, \theta) \in \mathbb{R}^+ \times [0, 2\pi) | \rho \leq R\}$ under the polar coordinate system, which is a circle centered at the origin in a 2D Euclidean plane. Here, all residents are assumed to be distributed on a ring with radius $r \leq R$, denoted by $X = \{(\rho, \theta) \in \mathbb{R}^+ \times [0, 2\pi) | \rho = r\}$, at a uniform density k. Hence, the total number of residents is $2\pi rk$. We consider a demand pattern where resident located at $x = (r, \theta) \in X$ travels to $y = (r_0, \theta) \in \Omega$ for any $\theta \in [0, 2\pi)$, where $r_0 < r$, as illustrated in Figure 2. The existing infrastructure pattern is assumed to be symmetric over the center, i.e., the infrastructure at position (ρ, θ) is independent of θ. In this way, residents at x will take the straight (i.e., shortest) path towards y under equilibrium due to the symmetry. Without loss of generality, we assume the congestion function follows the form of

$$g\left(c\left(\rho\right), \lambda\left(\rho\right)\right) = t_0\left(\rho\right)\left(1 + \kappa\left(\frac{\lambda\left(\rho\right)}{c\left(\rho\right)}\right)^4\right),$$

where $t_0\left(\rho\right)$ is an exogenous, continuous, and differentiable function, indicating the free-flow travel time, $\lambda\left(\rho\right)$ and $c\left(\rho\right)$ is the flow density and the capacity near the area with radial coordinate ρ, respectively. This function is inspired by the conventional BPR function, but we replace flow with density in the formulation which is believed to be more reasonable under the context of fundamental diagram. But we would like to mention that the specific formulation is not limited to the one we presented. The methodology framework applies as long as the congestion function satisfies the promise of being convex, continuous and strictly increasing with respect to $\lambda\left(x\right)$ for $\forall x \in \Omega$ in Lemma 2. Moreover, the formulation we adopted facilitates our discussion on the Sioux-Falls case and the Chicago case in §6.

By simple algebra we have $g^{-1}\left(c\left(\rho\right), \lambda\left(\rho\right)\right) = \left(\frac{\xi(\rho) - t_0(\rho)}{\kappa t_0(\rho)}\right)^{\frac{1}{4}} c\left(\rho\right)$ and $\bar{h}\left(c(\rho), \xi(\rho)\right) = \frac{4}{5} c\left(\rho\right)\left(\xi\left(\rho\right) - t_0\left(\rho\right)\right)\left(\frac{\xi(\rho) - t_0(\rho)}{\kappa t_0(\rho)}\right)^{\frac{1}{4}}$. The cost term is assumed to be linear with respect to the capacity improvement, i.e., $L\left(\boldsymbol{c}\right) = \int_0^R A\left(\rho\right)\left(c\left(\rho\right) - a\left(\rho\right)\right) \cdot 2\pi\rho d\rho$, where $A\left(\rho\right)$ and $a\left(\rho\right)$ are the unit improvement cost and original capacity, respectively. The feasible region $\mathcal{C}$ is defined as $\{\boldsymbol{c} | c\left(\rho\right) \geq a\left(\rho\right), 0 < \rho \leq R\}$. To determine the optimal capacity $c^*\left(\rho\right)$, we need to solve the following bilevel problem by Proposition 2,

$$\begin{aligned} \min_{c(\rho) \geq a(\rho)} & \int_0^R \left(\frac{\xi\left(\rho\right) - t_0\left(\rho\right)}{\kappa t_0\left(\rho\right)}\right)^{\frac{1}{4}} c\left(\rho\right) \xi\left(\rho\right) 2\pi\rho d\rho + \int_0^R A\left(\rho\right)\left(c\left(\rho\right) - a\left(\rho\right)\right) 2\pi\rho d\rho, \\ \text{s.t.} \boldsymbol{\xi} = \arg \min_{\xi(\rho) \geq t_0(\rho)} & \int_0^R \frac{4}{5} c\left(\rho\right)\left(\xi\left(\rho\right) - t_0\left(\rho\right)\right)\left(\frac{\xi\left(\rho\right) - t_0\left(\rho\right)}{\kappa t_0\left(\rho\right)}\right)^{\frac{1}{4}} 2\pi\rho d\rho - 2\pi rk \int_{r_0}^r \xi\left(\rho\right) d\rho, \end{aligned} \tag{18}$$

To ease the notations, we define $k_\rho := \frac{kr}{\rho}$ be the equivalent traveler density passing through the ring at radius ρ. Then such single-ring system can be solved by the following proposition.

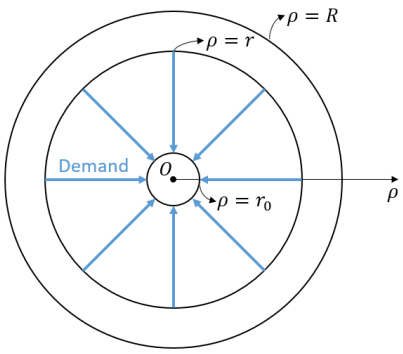


Figure 2: Illustration of the demand pattern in a symmetric single-ring system

Proposition 4. *The solution to (18) at nontrivial region, $\rho \in [r_0, r]$, is given by*

$$\xi^*(\rho) = t_0(\rho)\left(1 + \kappa\left(\frac{k_\rho}{c^*(\rho)}\right)^4\right),$$

$$c^*(\rho) = \max\left\{a(\rho), k_\rho\left(\frac{4\kappa t_0(\rho)}{A(\rho)}\right)^{\frac{1}{5}}\right\}.$$

Note that there is no positive flow outside the nontrivial region. The solution of $c^*(\rho)$ suggests that, if the original capacity a is not sufficient, the optimal capacity should linearly increase with respect to the demand density k, residential distance r, but linearly decrease with respect to the proximity to the center ρ. Meanwhile, less capacity is acceptable if the free-flow travel time decreases or the improvement cost becomes too high, in the order of 0.2. The impacts of key parameters $\{k, t_0, A, a\}$ are illustrated by perturbing one of them each time from a default setting as $R = r = 1.0, r_0 = 0.1, a(\rho) = 1.0, t_0(\rho) = 0.5, k = 1.0, A(\rho) = 5.0$. The plot of the optimal capacity under the default setting is shown in Figure 3 ($\kappa = 0.15$). The changes of $\xi^*(\rho)$ (indicated by y-axis) related to ρ (indicated by x-axis) are shown in Figure 4. We notice that the optimal capacity relieves congestion in such a way that the maximal congestion is related to the investment cost $A(\rho)$ and the original free-flow travel time $t_0(\rho)$, but is irrelevant to the demand density k and the original capacity $a(\rho)$.

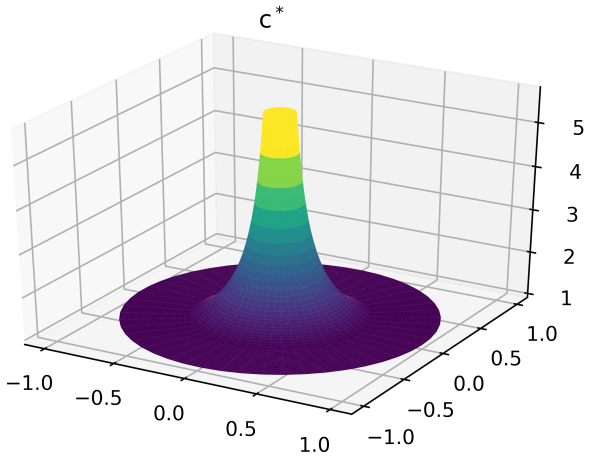


Figure 3: Plot of the optimal capacity under the default setting

Then, we can simply extend the single-ring system to multiple residential rings to capture the distribution of citizens in a more general setting. Consider a multi-ring system where residents are distributed over circles of radius $r_1, r_2, ..., r_N$ $(r_0 < r_1 < r_2 < ... < r_N < R)$ with uniform densities $k_1, ..., k_N$, respectively. Other settings remain the same. Then, the bilevel model to solve is

$$\min_{c(\rho)\geq a(\rho)} \int_0^R \left(\frac{\xi(\rho) - t_0(\rho)}{\kappa t_0(\rho)}\right)^{\frac{1}{4}} c(\rho)\,\xi(\rho)\,2\pi\rho d\rho + \int_0^R A(\rho)\,(c(\rho) - a(\rho))\,2\pi\rho d\rho, \tag{19}$$
$$\text{s.t.}\boldsymbol{\xi} = \arg\min_{\xi(\rho)\geq t_0(\rho)} \int_0^R \frac{4}{5} c(\rho)\,(\xi(\rho) - t_0(\rho)) \left(\frac{\xi(\rho) - t_0(\rho)}{\kappa t_0(\rho)}\right)^{\frac{1}{4}} 2\pi\rho d\rho - \sum_{i=1}^{N} 2\pi r_i k_i \int_{r_0}^{r_i} \xi(\rho)\, d\rho.$$

Similarly, for each $n \in [1, 2, ..., N]$, we define $k_{\rho,n} := \frac{k_n r_n}{\rho}$ be the equivalent traveler density from those demand at circles of radius r_n, passing through the ring at radius $\rho < r_n$.

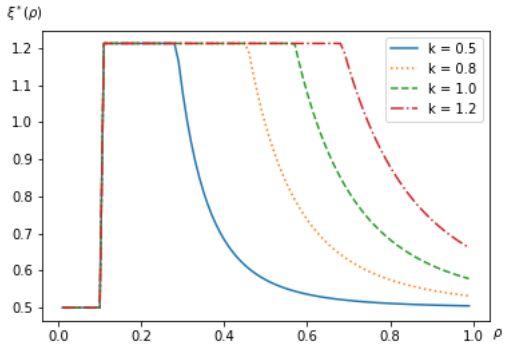
Corollary 1. *The solution to (19) at nontrivial region, $\rho \in [r_0, r_N)$, is given by*

$$\xi^*(\rho) = t_0(\rho)\left(1 + \kappa\left(\frac{\sum_{i=n}^{N} k_{\rho,i}}{c^*(\rho)}\right)^4\right), \rho \in [r_{n-1}, r_n), \forall n \in [1, 2, ..., N],$$

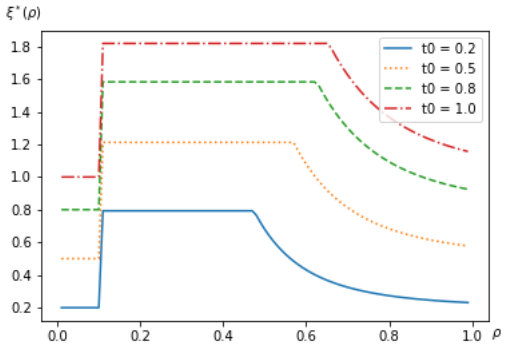
$$c^*(\rho) = \max\left\{a(\rho), \left(\sum_{i=n}^{N} k_{\rho,i}\right)\left(\frac{4\kappa t_0(\rho)}{A(\rho)}\right)^{\frac{1}{5}}\right\}, \rho \in [r_{n-1}, r_n), \forall n \in [1, 2, ..., N].$$

Note that $\sum_{i=n}^{N} k_{\rho,i}$ can be interpreted as the density due to total travelers passing through the ring at radius ρ from all demands outside, or "total passing demand". From Corollary 1, we can immediately tell the optimal capacity and travel time near circle n can be obtained by considering the total passing demand. Hence, with the same amount of residents, the more concentrated those residents distributed, the less capacity is needed.

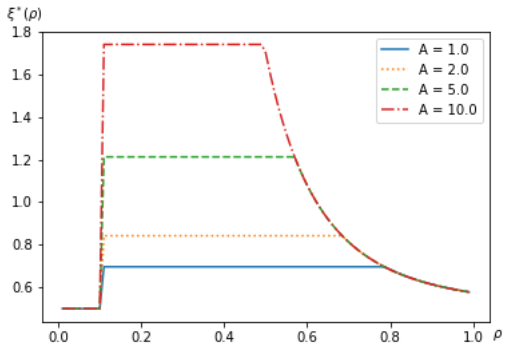
We further extend the multi-ring system by considering its continuous version, i.e., all residents are assumed to be continuously distributed between r_0 and R, denoted by $X' = \{(\rho, \theta) \in \mathbb{R}^+ \times [0, 2\pi) | r_0 \leq \rho \leq$



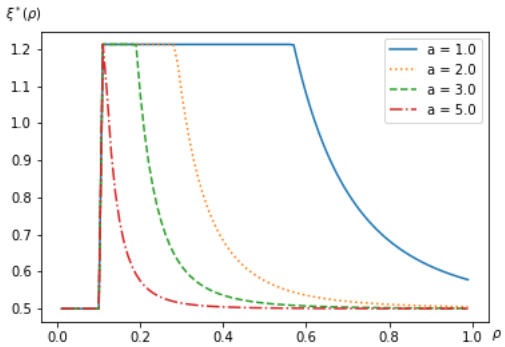
(a) Variance of k



(b) Variance of t_0



(c) Variance of A



(d) Variance of a

Figure 4: Plots of $\xi^*(\rho)$ in a ring system as each of the parameters $\{k, t_0, A, a\}$ perturbs from the default setting.

$R\}$, but with a homogeneous density near the same ρ, $k(\rho)$. Without loss of generality, we assume that $k(\rho)$ is an integrable function over $[r_0, R]$. Hence, the total number of residents is $\int_{r_0}^{R} k(\rho) 2\pi\rho d\rho$. Other settings remain the same. The bilevel model to solve is

$$\min_{c(\rho)\geq a(\rho)} \int_0^R \left(\frac{\xi(\rho)-t_0(\rho)}{\kappa t_0(\rho)}\right)^{\frac{1}{4}} c(\rho)\xi(\rho) 2\pi\rho d\rho + \int_0^R A(\rho)(c(\rho)-a(\rho)) 2\pi\rho d\rho, \tag{20}$$
$$\text{s.t.} \boldsymbol{\xi} = \arg\min_{\xi(\rho)\geq t_0(\rho)} \int_0^R \frac{4}{5} c(\rho)(\xi(\rho)-t_0(\rho)) \left(\frac{\xi(\rho)-t_0(\rho)}{\kappa t_0(\rho)}\right)^{\frac{1}{4}} 2\pi\rho d\rho - \int_{r_0}^R \left(2\pi\rho k(\rho) \int_{r_0}^{\rho} \xi(r) dr\right) d\rho$$

Corollary 2. *The solution to (20) at nontrivial region, $\rho \in [r_0, R]$, is given by*

$$\xi^*(\rho) = t_0(\rho)\left(1+\kappa\left(\frac{\int_\rho^R r k(r) dr}{\rho c^*(\rho)}\right)^4\right), \rho \in [r_0, R],$$

$$c^*(\rho) = \max\left\{ a(\rho), \left(\frac{\int_\rho^R rk(r)\,dr}{\rho}\right)\left(\frac{4\kappa t_0(\rho)}{A(\rho)}\right)^{\frac{1}{5}} \right\}, \rho \in [r_0, R].$$

Note that $\frac{\int_\rho^R rk(r)dr}{\rho}$ can be considered as the passing traveler density as well.

6 Numerical Study

In this section, we present several numerical examples to illustrate the model and its solution algorithm we have proposed in this paper. §6.1 describes the basic setting that we used for all cases in this section. The first hypothetical case described in §6.2 is used to illustrate the effectiveness of the proposed algorithm. The case is further studied in §6.3 to compare the performance under different resolutions. We then conduct a Sioux-Falls case study in §6.4 to further illustrate how to convert a discrete setting to a continuous setting so that it can fit into our continuum model, and how it performs as compared to the discrete model. Last, we apply our model to the large-scale network of Chicago in §6.5. We solve all models on a desktop computer with i5-6500 CPU @3.20GHz and 16.0GB RAM.

6.1 Problem Setting

Similar to the previous analytical case in §5, the congestion function $g(c(x), \lambda(x))$ is assumed to be

$$g(c(x), \lambda(x)) = t_0(x)\left(1 + \kappa \cdot \frac{\lambda^4(x)}{c_0^4(x)}\right), \tag{21}$$

where $c(x) = (t_0(x), c_0(x))$. In the formula (21), $t_0(x)$ and $c_0(x)$ respectively reflect the free-flow travel time and the desired capacity near x. By simple algebra, we have the expression of $g^{-1}(c(x), \xi(x)) = c_0(x)\left(\frac{\xi(x)-t_0(x)}{\kappa t_0(x)}\right)^{\frac{1}{4}}$ and $\bar{h}(c(x), \xi(x)) = \frac{4}{5}c_0(x)(\xi(x) - t_0(x))\left(\frac{\xi(x)-t_0(x)}{\kappa t_0(x)}\right)^{\frac{1}{4}}$.

The cost function $L(\boldsymbol{c})$ is assumed to have a quadratic relationship with the change of capacity c_0 and free-flow travel time t_0, i.e., it can be captured as

$$L(\boldsymbol{c}) = \int_\Omega l(c_0(x), t_0(x))\,dx, \tag{22}$$

such that

$$l(c_0(x), t_0(x)) = A(x)(c_0(x) - a(x))^2 + B(x)(t_0(x) - b(x))^2,$$

where $a(x)$, $b(x)$ are parameters that indicate the original setting of $c_0(x)$ and $t_0(x)$, and $A(x)$, $B(x)$ are parameters reflecting the unit cost to change $c_0(x)$ and $t_0(x)$ near x, respectively. We would like to mention that the form of $L(\boldsymbol{c})$ is not limited; it can be linear (as we have discussed in the analytical case), piecewise, polynomial or any other reasonable functions. In this section, we consider a quadratic case to reflect that a

cost is incurred when a modification (either increase or decrease of the capacity or the free-flow travel time) is conducted. The feasible region $\mathcal{C}$ is assumed to be $\mathcal{C} := \left\{\underline{c_0}(x) \le c_0(x) \le \bar{c_0}(x), t_0(x) \ge \underline{t_0}(x), \forall x \in \Omega\right\}$. We also assume the trust region constraints are in the form of $\left\|\boldsymbol{\xi} - \boldsymbol{\xi}^{(k)}\right\|_\infty \le \Delta^{(k)}$ where $\|\cdot\|_\infty$ represent the l_∞-norm function. Under this setting, the trust-region sub-problem (17) is fully decomposable as follows: given $\boldsymbol{\xi}^{(k)}$, for $\forall x \in \Omega$,

$$\left(c^{(k+1)}(x), \tilde{\xi}^{(k+1)}(x)\right) = \arg \min_{(c_0(x),t_0(x))\in\mathcal{C}} c_0(x)\left(\frac{\xi(x)-t_0(x)}{\kappa t_0(x)}\right)^{\frac{1}{4}}\xi(x) + l(c_0(x), t_0(x)), \tag{23}$$

$$\text{s.t.} \quad \xi(x) = \arg \min_{\xi(x)\ge t_0(x), \left\|\xi(x)-\xi^{(k)}(x)\right\| \le \Delta^{(k)}} \tag{24}$$

$$\frac{4}{5}c_0(x)(\xi(x)-t_0(x))\left(\frac{\xi(x)-t_0(x)}{\kappa t_0(x)}\right)^{\frac{1}{4}} - \alpha_\gamma^{(k)}(x)\,\xi(x).$$

Proposition 5. *The bilevel trust-region sub-problem (23) is equivalent to the following single-level nonlinear program (SNP),*

$$\begin{aligned}
\text{(SNP)} \qquad \min_{c_0(x),t_0(x),\xi(x),u_1,u_2,u_3} \; & c_0(x)\left(\frac{\xi(x)-t_0(x)}{\kappa t_0(x)}\right)^{\frac{1}{4}}\xi(x) + l(c_0(x), t_0(x)), \\
s.t. \; & (c_0(x), t_0(x)) \in \mathcal{C}, \\
& c_0(x)\left(\frac{\xi(x)-t_0(x)}{\kappa t_0(x)}\right)^{\frac{1}{4}} - \alpha_\gamma^{(k)}(x) = u_1 + u_2 - u_3, \\
& u_1(t_0(x) - \xi(x)) = 0, \\
& u_2\left(\xi^{(k)}(x) - \xi(x) - \Delta^{(k)}\right) = 0, \\
& u_3\left(\xi(x) - \xi^{(k)}(x) - \Delta^{(k)}\right) = 0, \\
& \xi(x) \ge t_0(x), \\
& \xi(x) \ge \xi^{(k)}(x) - \Delta^{(k)} \\
& -\xi(x) \ge -\xi^{(k)}(x) - \Delta^{(k)} \\
& u_1, u_2, u_3 \ge 0
\end{aligned}$$

With the bounded closed feasible region, the solution to the SNP exists as its objective value has a lower bound of 0. There are multiple ways to solve the SNP. We adopt the traditional method of projected subgradient descent method (Boyd et al. 2003) to obtain a near-optimal solution to the SNP.

6.2 Continuous Transportation Network Design Problem

We now show a hypothetical numerical case over a square region smashed into a uniform grid of size 20×20. We consider a river passing through the region with a bridge linking the two sides of the river, and there

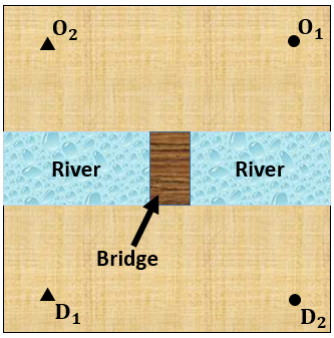
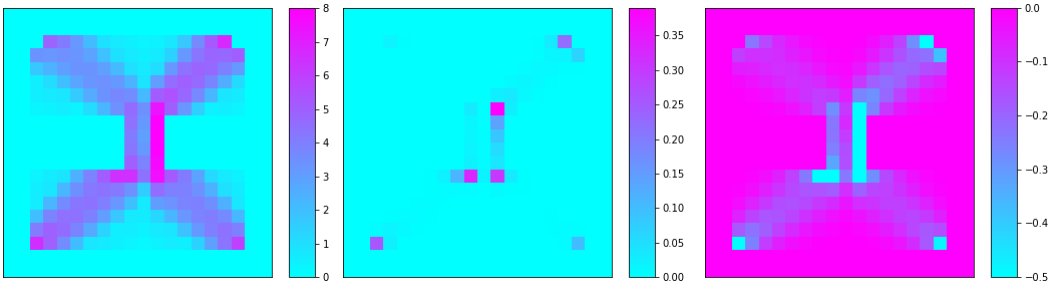


Figure 5: Geographic setting

Table 1: Parameter setting

Part	a	b	A	B
bridge	10	1	100	50
river	1	10	1000	100
road	10	1	2	20

are two origins (O_1 and O_2) and two destinations (D_1 and D_2) as shown in Figure 5. The parameter setting for each small zone over the region is shown in Table 1. Compared with other parts of the region, the river features a small passing capacity ($a = 1$) and a long free-flow travel time ($b = 10$), with a much larger investment cost ($A = 1000, B = 100$) to make a change. The bridge shares the same original traffic characteristics with normal road areas, but requires a relatively larger cost to change. We choose the traffic weights such that $\gamma(1,1) + \gamma(1,2) = 2(\gamma(2,1) + \gamma(2,2))$ and $\frac{\gamma(2,2)}{\gamma(2,1)} = \frac{\gamma(1,1)}{\gamma(1,2)} = 2$. Further, we assume that $\kappa = 0.15$. The result is shown in Figure 6.



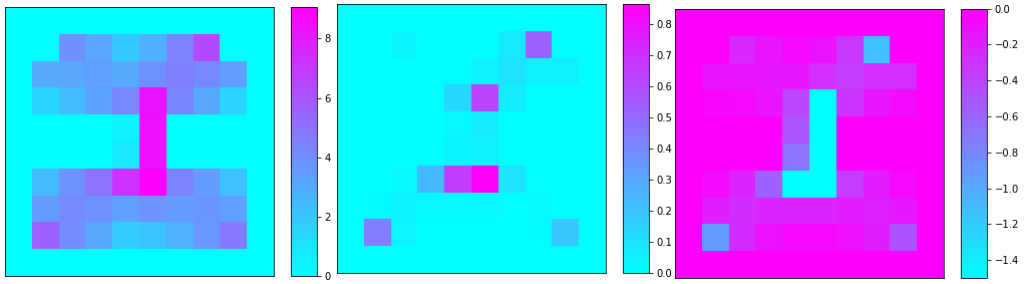
(a) $\boldsymbol{\lambda}^*$ (b) $c_0^* - a$ (c) $t_0^* - b$

Figure 6: Equilibrium flow and infrastructure changes

The optimal flow pattern $\boldsymbol{\lambda}^*$ is shown in Figure 6(a). The capacity should be increased and the free-flow time should be decreased along both sides of the bridge as shown in Figure 6(b) and (c). Moreover, From Figure 6(c), the free-flow travel time is suggested to be decreased over areas where flows exist. It can be seen that the bottleneck of the system which leads to most of the congestion is around the sides of the bridge, and thus improving the traffic infrastructure close to the bridge is the key to improving the performance of the whole system.

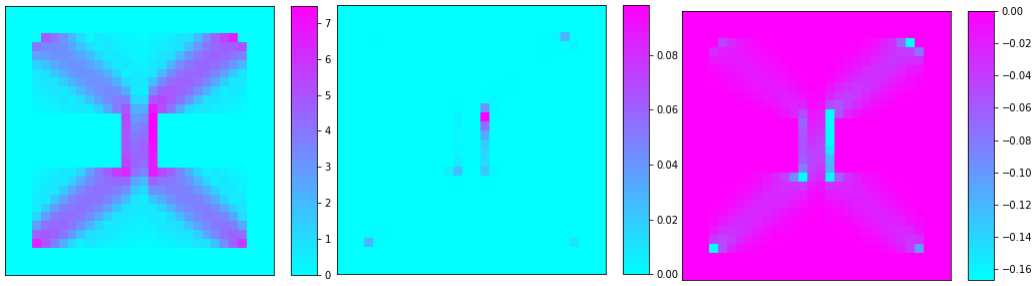
6.3 Performance Test under Multi-resolution

We now compare the performance of our model and algorithm under different resolutions. We take the same geographic setting as the previous river-bridge case, but instead of the 20×20 grid as we have tried in the previous section, we discretize the square region into uniform grids of sizes 10×10 and 30×30 with their solutions shown in Figure 7 and 8, respectively. The CPU running time under each case is compared in Table 2.



(a) $\boldsymbol{\lambda}^*$ (b) $c_0^* - a$ (c) $t_0^* - b$

Figure 7: Equilibrium flow and infrastructure changes over a 10×10 grid



(a) $\boldsymbol{\lambda}^*$ (b) $c_0^* - a$ (c) $t_0^* - b$

Figure 8: Equilibrium flow and infrastructure changes over a 30×30 grid

Resolution	Number of iteration	average CPU time (s)/iteration	Total CPU time (s)
10×10	6	0.75	4.5
20×20	4	2.95	11.8
30×30	6	7.08	42.5

Table 2: CPU time under different resolution

We can see that although the lower-resolution case is not as accurate as the higher-resolution case, it can give some insights from a macroscopic prospective. From the 10×10-resolution case, we can still identify

where traffic peaks appear and which zone is the bottleneck, although the solution is not as straightforward as the 20×20-resolution case. However, a higher resolution not necessarily indicates better accuracy. If we cut up the region too hard and break the integrity of road segments, we may have totally different strategies over a small zone, e.g., over the bridge, as shown in the 30×30-resolution case. Further, a higher resolution always accompanies with heavier computational burdens as shown in Table 2. In real applications, selecting a suitable resolution is important so that an accurate solution can be obtained within a reasonable time. We suggest that the width of each zone should be at least equal to the width of roads to achieve a consistency of strategy over each road segment, while it should not be too large so that different zones can still be well distinguished.

6.4 Sioux-Falls Network Design

In this section, we study the well-known Sioux-Falls network to show how to convert a discrete setting to a continuous setting so that our model can fit. Further, we show the effectiveness and efficiency of our bilevel continuum model by comparing its solution with that of a traditional discrete model.

The Sioux-Falls network includes 24 nodes and 76 links, where the links in the network are in pairs representing two-way directions. The original link performance function is in the same form of BPR function and the data on traffic characteristics, including capacity and free-flow travel time can be found on *Github* (The data is consistent with LeBlanc (1975), although in slightly different forms). To adapt the discrete parameters into our continuous setting, we smash the Sioux-Falls area into a 20×20 grid with 400 square zones. The capacity of each zone is taken as the average capacity of corresponding links passing through it, and is set to 1000 (a much smaller value) if there is no link passing through it. The free-flow travel time passing through each zone is taken as the average of the free-flow travel time over all segments of links inside it, and is set to 3 (a much larger value) if there is no link passing through it. By doing this, we ensure that the total free-flow travel time from an origin zone to a destination zone of any link is very close to the free-flow travel time of that link. In this way, we convert the discrete description of the network based on links into a continuous description of the whole area based on zones. We consider the same candidate projects as LeBlanc (1975) with the new capacity and free-flow time shown in Table 3. Different from the discrete version, we assume that partial improvement of each project is allowed, with a unit cost proportional to the square of changes, i.e., in the form of (22), and the values of A and B are set to make the total cost consistent with that of LeBlanc (1975), as shown in Table 3. Notice that we enforce a large unit cost for other zones so that they tend to be out of consideration for an improvement in the final solution. We obtain the solution of our continuum model following the proposed Algorithm 1, and compare it with the discrete solution obtained via an iterative algorithm provided by Gao et al. (2005).

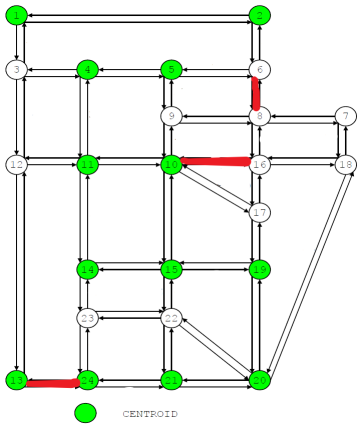
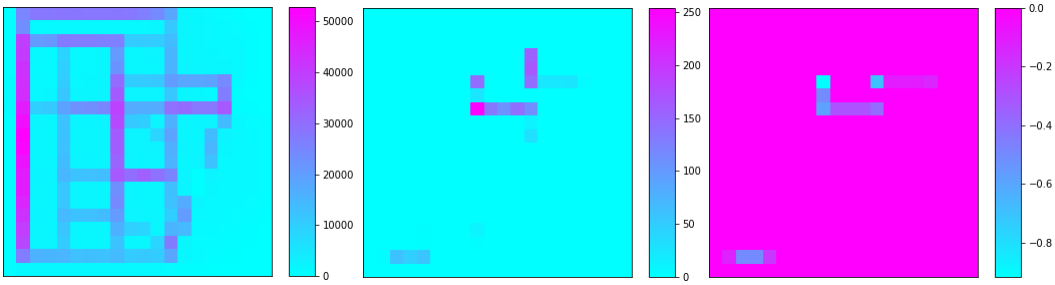


Figure 9: Solution of the traditional discrete model (Adopted projects marked in red)

Project No.	Arcs	A	B	New capacity	New free-flow time
1	$(6,8)$ and $(8,6)$	0.3	663,265	5,944.13	1.3
2	$(9,10)$ and $(10,9)$	0.075	159,439	15,958.88	1.6
3	$(13,24)$ and $(24,13)$	0.6	131,173	5,924.83	2.2
4	$(10,16)$ and $(16,10)$	0.5	355,030	5,946.04	2.7
5	$(7,8)$ and $(8,7)$	0.428	222,222	8,922.54	1.5
-	Others	5	5.0×10^7	-	-

Table 3: Proposed arc improvements and Costs



(a) Equilibrium flow (b) Change of capacity (c) Change of free-flow travel time

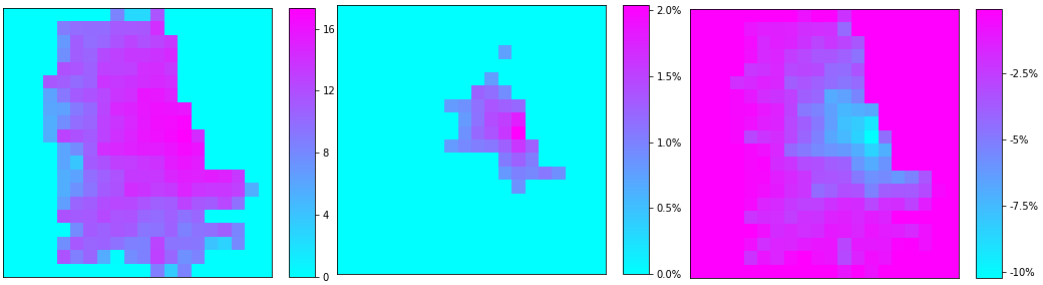
Figure 10: Solution of our continuum model

The solution based on the traditional link-based model is shown in Figure 9, i.e., project 1 (arc between node 6,8), project 3 (arc between node 13,24) and project 4 (arc between node 10,16) are selected. As a comparison, the solution of the continuum model is shown in Figure 10. The equilibrium flow over the network is shown in Figure 10(a), and the suggested modification of capacity and free-flow time is shown in

Figure 10(b) and (c) respectively. We can see that the solution of the continuum model presents a priority order for the five candidate projects that is consistent with the solution of the traditional model. Although the solution of the continuum model is not as straightforward as that of the traditional model, it contains more details and provides some insights that beyond the ability of the traditional model. For example, Figure 10(b) and (c) show that the capacity should be increased with free-flow travel time maintained for the arc between node 6 and 8, which indicates that the bottleneck over this area is its limited capacity, e.g., the width of roads. An interesting observation is that, it is strongly recommended from Figure 10(b) that the capacity of the arc between node 16 and 17 should be increased although it is punished with a very large investment cost. This shows one advantage of the continuum model that it can help identify the bottleneck of the system while the traditional discrete model cannot. We believe this particular case is enough to show the effectiveness and advantage of the continuum model, and it can be trusted to realistic cases with proper applications to local conditions.

6.5 Large-scale Network Design: Chicago Sketch Network

In this section, we apply our model to the Chicago sketch network as an illustration of the application to a large-scale case. The Chicago Sketch network has 933 nodes and 2950 links, and the traffic data can be found on Github. Following the same scheme as §6.4 we convert the discrete data into our continuous setting over a 20×20 grid. We assume that the investment cost over each zone is as follows: $A = 500, B = 10,000$ for the land area; $A = B = 100,000$ for the lake area. The solution is shown in Figure 11.



(a) Equilibrium flow under Log2 (b) Change of capacity (c) Change of free-flow travel time

Figure 11: Solution to the Chicago Sketch Network

7 Conclusion

The traditional traffic network design strategies by modeling upon links and path have been well studied, but it is impossible to scale them up as to coordinate multiple instruments over a large-scale transportation system due to extremely high computational complexity. This paper overcame such a challenge with a

continuous framework for designing large-scale transportation system. We have presented a continuum bilevel transportation network design model to help optimize the traffic characteristics of interest from a macroscopic view. The model is developed based on the continuous congestion theory, and it is the first attempt to combine the continuous theory with bilevel structure up to our knowledge.

In this paper, we formally built the continuum bilevel model by reorganizing and developing the concept of continuous congestion theory from a new angle. We inherit and develop several fundamental lemmas, which are derived based on our setting via the variational inequality tools. Then we presented an efficient approach to solve the model. The heart of our approach is a trust-region numerical scheme based on a partial linear approximation of the conjugated dual of the lower level objective function, which greatly simplifies the implicit relationship between the upper and lower level variables, and enables us to solve the complicated bilevel model in a decomposable manner. To show the effectiveness of the model, we conducted a series of symmetric analytical cases in which the equilibrium flow pattern can be deduced, and we found that the optimal road capacity for a circular city structure is linearly related to the passing travel demand under reasonable assumptions. Furthermore, we provided several numerical cases to show the effectiveness and efficiency of the proposed approach to more general models. By taking the well-known Sioux-Falls network as an example, we explained how to convert a discrete data set so that it can fit into our continuous setting, and compared our continuous solution with that of the traditional discrete model. We demonstrated that our continuum model can make synthetic strategy for multiple instruments and efficiently identify bottlenecks in a congested network, which is beyond the ability of traditional discrete models. Last, we studied the Chicago area as an illustration of how our model performs when being applied to a large-scale network.

Serving as a building block, this paper also motivates the research in many other directions. With proper adjustments, we believe that our model can help with other complicated problems related to congestion.

References

Aussel, D. and C. Lalitha (2018). *Generalized Nash Equilibrium Problems, Bilevel Programming and MPEC*. Springer.

Beckmann, M., C. B. McGuire, and C. B. Winsten (1956). Studies in the economics of transportation. Technical report.

Benmansour, F., G. Carlier, G. Peyré, and F. Santambrogio (2010). Derivatives with respect to metrics and applications: subgradient marching algorithm. *Numerische Mathematik 116*(3), 357–381.

Bergendorff, P., D. W. Hearn, and M. V. Ramana (1997). Congestion toll pricing of traffic networks. In *Network Optimization*, pp. 51–71. Springer.

Boyd, S., L. Xiao, and A. Mutapcic (2003). Subgradient methods. *lecture notes of EE392o, Stanford University, Autumn Quarter 2004*, 2004–2005.

Carlier, G., C. Jimenez, and F. Santambrogio (2008). Optimal transportation with traffic congestion and wardrop equilibria. *SIAM Journal on Control and Optimization 47*(3), 1330–1350.

Carlier, G. and F. Santambrogio (2012). A continuous theory of traffic congestion and wardrop equilibria. *Journal of Mathematical Sciences 181*(6), 792–804.

Chiou, S.-W. (2009). A bi-level programming for logistics network design with system-optimized flows. *Information Sciences 179*(14), 2434–2441.

Cui, T., Y. Ouyang, and Z.-J. M. Shen (2010). Reliable facility location design under the risk of disruptions. *Operations research 58*(4-part-1), 998–1011.

Daganzo, C. F. and G. F. Newell (1986). Configuration of physical distribution networks. *Networks 16*(2), 113–132.

De Palma, A., I. E. Nesterov, et al. (1998). Optimization formulations and static equilibrium in congested transportation networks.

Fotakis, D., S. Kontogiannis, E. Koutsoupias, M. Mavronicolas, and P. Spirakis (2002). The structure and complexity of nash equilibria for a selfish routing game. In *International Colloquium on Automata, Languages, and Programming*, pp. 123–134. Springer.

Gao, Z., H. Sun, and L. L. Shan (2004). A continuous equilibrium network design model and algorithm for transit systems. *Transportation Research Part B: Methodological 38*(3), 235–250.

Gao, Z., J. Wu, and H. Sun (2005). Solution algorithm for the bi-level discrete network design problem. *Transportation Research Part B: Methodological 39*(6), 479–495.

Idone, G. (2004). Variational inequalities and applications to a continuum model of transportation network with capacity constraints. *Journal of Global Optimization 28*(1), 45–53.

INRIX (2018). INRIX 2018 Global Traffic Scorecard. `http://inrix.com/scorecard/`. [Online; accessed 30-Oct-2019].

LeBlanc, L. J. (1975). An algorithm for the discrete network design problem. *Transportation Science 9*(3), 183–199.

Li, X. and Y. Ouyang (2010). A continuum approximation approach to reliable facility location design under correlated probabilistic disruptions. *Transportation research part B: methodological 44*(4), 535–548.

Lignola, M. B. and J. Morgan (2001). Existence of solutions to bilevel variational problems in banach spaces. In *Equilibrium Problems: Nonsmooth Optimization and Variational Inequality Models*, pp. 161–174. Springer.

Maugeri, A. (2001). Equilibrium problems and variational inequalities. In *Equilibrium Problems: Nonsmooth Optimization and Variational Inequality Models*, pp. 187–205. Springer.

Maugeri, A. and F. Raciti (2009). On existence theorems for monotone and nonmonotone variational inequalities. *J. Convex Anal 16*(3-4), 899–911.

Newell, G. F. (1971). Dispatching policies for a transportation route. *Transportation Science 5*(1), 91–105.

Newell, G. F. (1973). Scheduling, location, transportation, and continuum mechanics: some simple approximations to optimization problems. *SIAM Journal on Applied Mathematics 25*(3), 346–360.

Ouyang, Y. and C. F. Daganzo (2006). Discretization and validation of the continuum approximation scheme for terminal system design. *Transportation Science 40*(1), 89–98.

Ouyang, Y., Z. Wang, and H. Yang (2015). Facility location design under continuous traffic equilibrium. *Transportation Research Part B: Methodological 81*, 18–33.

Patriksson, M. (2015). *The traffic assignment problem: models and methods*. Courier Dover Publications.

Patriksson, M. and R. T. Rockafellar (2002). A mathematical model and descent algorithm for bilevel traffic management. *Transportation Science 36*(3), 271–291.

Sheffi, Y. (1985). Urban transportation networks.

Sun, H., Z. Gao, and J. Wu (2008). A bi-level programming model and solution algorithm for the location of logistics distribution centers. *Applied mathematical modelling 32*(4), 610–616.

Wang, X. and Y. Ouyang (2013). A continuum approximation approach to competitive facility location design under facility disruption risks. *Transportation Research Part B: Methodological 50*, 90–103.

Wardrop, J. G. (1952). Road paper. some theoretical aspects of road traffic research. *Proceedings of the institution of civil engineers 1*(3), 325–362.

Wong, S.-C. and H. Yang (1999). Determining market areas captured by competitive facilities: a continuous equilibrium modeling approach. *Journal of Regional Science 39*(1), 51–72.

Yang, H. (1995). Heuristic algorithms for the bilevel origin-destination matrix estimation problem. *Transportation Research Part B: Methodological 29*(4), 231–242.

Yang, H. (1996). A spatial price equilibrium model with congestion effects. *The Annals of Regional Science 30*(4), 359–371.

Yang, H. and S. C. Wong (2000). A continuous equilibrium model for estimating market areas of competitive facilities with elastic demand and market externality. *Transportation Science 34*(2), 216–227.

Yang, H., S. Yagar, and Y. Iida (1994). Traffic assignment in a congested discrete/continuous transportation system. *Transportation Research Part B: Methodological 28*(2), 161–174.

Appendix

Proof of Lemma 1

Proof. It is trivial that $\mathcal{F}(\gamma)$ is bounded and closed. Since the map from $\mathcal{F}(\gamma)$ to $\Lambda(\gamma)$ is surjective, $\Lambda(\gamma)$ is also bounded and closed. If $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \Lambda(\gamma)$ are different, then let $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \mathcal{F}(\gamma)$ such that Constraints (3) satisfies for $(\boldsymbol{\lambda}_1, \boldsymbol{f}_1)$ and $(\boldsymbol{\lambda}_2, \boldsymbol{f}_2)$ respectively. Then for $\forall \alpha \in [0,1]$, let $\boldsymbol{f}_\alpha = \alpha \boldsymbol{f}_2 + (1-\alpha)\boldsymbol{f}_1$ and $\boldsymbol{\lambda}_\alpha = \alpha \boldsymbol{\lambda}_2 + (1-\alpha)\boldsymbol{\lambda}_1$. It is easy to check that $\boldsymbol{f}_\alpha \in \mathcal{F}(\gamma)$ and $(\boldsymbol{\lambda}_\alpha, \boldsymbol{f}_\alpha)$ satisfies Constraints (3), and thus $\boldsymbol{\lambda}_\alpha \in \Lambda(\gamma)$, which indicates that $\Lambda(\gamma)$ is convex. □

Proof of Lemma 2

Proof. First we prove that Wardrop equilibrium flow pattern $\boldsymbol{\lambda}^*$ coincides with the solution to the following variational inequality (VI):

$$\int_\Omega g\left(c(x), \lambda^*(x)\right)\left(\lambda(x) - \lambda^*(x)\right) dx \geq 0, \forall \boldsymbol{\lambda} \in \Lambda(\gamma). \tag{25}$$

If $\boldsymbol{f}^* \in \mathcal{F}(\gamma)$ is a Wardrop equilibrium path flow with corresponding Wardrop equilibrium point flow $\boldsymbol{\lambda}^* \in \Lambda(\gamma)$, then for $\forall \boldsymbol{\lambda} \in \Lambda(\gamma)$, we have

$$\begin{aligned}
&\int_\Omega g\left(c(x), \lambda^*(x)\right)\lambda(x)\, dx \\
&= \int_S f(\sigma)\left(\int_0^1 g\left(c(\sigma(t)), \lambda^*(\sigma(t))\right)|\dot{\sigma}(t)|\, dt\right) d\sigma && \text{(By (3))} \\
&= \int_{\Omega\times\Omega}\left(\int_{S^{x,y}} f(\sigma)\left(\int_0^1 g\left(c(\sigma(t)), \lambda^*(\sigma(t))\right)|\dot{\sigma}(t)|\, dt\right) d\sigma\right) dxdy \\
&\geq \int_{\Omega\times\Omega}\left(\int_{S^{x,y}} f(\sigma)\, u^*(x,y)\, d\sigma\right) dxdy && \text{(By (4))} \\
&= \int_{\Omega\times\Omega}\left(\int_{S^{x,y}} f(\sigma)\, d\sigma\right) u^*(x,y)\, dxdy \\
&= \int_{\Omega\times\Omega}\left(\int_{S^{x,y}} f^*(\sigma)\, d\sigma\right) u^*(x,y)\, dxdy && \text{(By (1))} \\
&= \int_{\Omega\times\Omega}\left(\int_{S^{x,y}}\left(\int_0^1 g\left(c(\sigma(t)), \lambda^*(\sigma(t))\right)|\dot{\sigma}(t)|\, dt\right) f^*(\sigma)\, d\sigma\right) dxdy && \text{(By (5))} \\
&= \int_S f^*(\sigma)\left(\int_0^1 g\left(c(\sigma(t)), \lambda^*(\sigma(t))\right)|\dot{\sigma}(t)|\, dt\right) d\sigma \\
&= \int_\Omega g\left(c(x), \lambda^*(x)\right)\lambda^*(x)\, dx. && \text{(By (3))}
\end{aligned}$$

Thus, $\boldsymbol{\lambda}^*$ is a solution to the VI (25). Conversely, if $\boldsymbol{\lambda}^*$ is a solution to the VI (25), then we notice that $\boldsymbol{\lambda}^*$ should be a solution to the following optimization problem:

$$\begin{aligned}
&\min_{\boldsymbol{\lambda},\boldsymbol{f}} \int_{\Omega} g\left(c\left(x\right),\lambda^*\left(x\right)\right)\lambda\left(x\right)dx \\
&\text{s.t.} \int_{S^{x,y}} f\left(\sigma\right)d\sigma = \gamma\left(x,y\right), \forall\left(x,y\right)\in\Omega\times\Omega, \\
&\quad f\left(\sigma\right)\geq 0, \forall\sigma\in S, \\
&\quad \int_{\Omega}\varphi\left(x\right)\lambda\left(x\right)dx = \int_{S} f\left(\sigma\right)\left(\int_0^1 \varphi\left(\sigma\left(t\right)\right)\left|\dot{\sigma}\left(t\right)\right|dt\right)d\sigma, \forall\varphi\in C_0\left(\Omega,\mathbb{R}_+\right).
\end{aligned}$$

Consider the following Lagrangian:

$$\begin{aligned}
&\mathcal{L}\left(\boldsymbol{\lambda},\boldsymbol{f},\boldsymbol{\mu},\boldsymbol{\nu},\boldsymbol{\omega}\right) \\
=&\int_{\Omega} g\left(c\left(x\right),\lambda^*\left(x\right)\right)\lambda\left(x\right)dx \\
&+\int_{\Omega\times\Omega}\mu\left(x,y\right)\left(\int_{S^{x,y}} f\left(\sigma\right)d\sigma-\gamma\left(x,y\right)\right)dxdy \\
&-\int_{S}\omega\left(\sigma\right)f\left(\sigma\right)d\sigma \\
&+\int_{C_0(\Omega,\mathbb{R}_+)}\nu_{\varphi}\left(\int_{\Omega}\varphi\left(x\right)\lambda\left(x\right)dx-\int_{S} f\left(\sigma\right)\left(\int_0^1\varphi\left(\sigma\left(t\right)\right)\left|\dot{\sigma}\left(t\right)\right|dt\right)d\sigma\right)d\varphi,
\end{aligned}$$

where $\mu\left(x,y\right),\omega\left(\sigma\right)\left(\geq 0\right)$ and ν_{φ} are the Lagrangian multipliers associated with constraints (1), (2) and (3), respectively. From variational principal, let $\delta\boldsymbol{\lambda}$ be an arbitrary function on Ω and $\delta\boldsymbol{f}$ be an arbitrary function on S, we can easily show that

$$\begin{aligned}
\delta\mathcal{L} &= \mathcal{L}\left(\boldsymbol{\lambda}+\delta\boldsymbol{\lambda},\boldsymbol{f}+\delta\boldsymbol{f},\boldsymbol{\mu}+\delta\boldsymbol{\mu},\boldsymbol{\nu}+\delta\boldsymbol{\nu},\boldsymbol{\omega}+\delta\boldsymbol{\omega}\right)-\mathcal{L}\left(\boldsymbol{\lambda},\boldsymbol{f},\boldsymbol{\mu},\boldsymbol{\nu},\boldsymbol{\omega}\right) \\
&=\int_{\Omega}\left(g\left(c\left(x\right),\lambda^*\left(x\right)\right)+\int_{C_0(\Omega,\mathbb{R}_+)}\nu_{\varphi}\varphi\left(x\right)d\varphi\right)\delta\lambda\left(x\right)dx \\
&+\int_{S}\left(\mu\left(\sigma\left(0\right),\sigma\left(1\right)\right)-\omega\left(\sigma\right)-\int_{C_0(\Omega,\mathbb{R}_+)}\nu_{\varphi}\left(\int_0^1\varphi\left(\sigma\left(t\right)\right)\left|\dot{\sigma}\left(t\right)\right|dt\right)d\varphi\right)\delta f\left(\sigma\right)d\sigma \\
&+\int_{\Omega\times\Omega}\delta\mu\left(x,y\right)\left(\int_{S^{x,y}} f\left(\sigma\right)d\sigma-\gamma\left(x,y\right)\right)dxdy \\
&-\int_{S}\delta\omega\left(\sigma\right)f\left(\sigma\right)d\sigma \\
&+\int_{C_0(\Omega,\mathbb{R}_+)}\delta\nu_{\varphi}\left(\int_{\Omega}\varphi\left(x\right)\lambda\left(x\right)dx-\int_{S} f\left(\sigma\right)\left(\int_0^1\varphi\left(\sigma\left(t\right)\right)\left|\dot{\sigma}\left(t\right)\right|dt\right)d\sigma\right)d\varphi.
\end{aligned}$$

Since $\delta\boldsymbol{f}$ and $\delta\boldsymbol{\lambda}$ are arbitrary functions, the stationary point of the Lagrangian requires that

$$g\left(c\left(x\right),\lambda^{*}\left(x\right)\right)+\int_{C_{0}\left(\Omega,\mathbb{R}_{+}\right)}\nu_{\varphi}\varphi\left(x\right)d\varphi=0,\forall x\in\Omega, \tag{26}$$

$$\mu\left(\sigma\left(0\right),\sigma\left(1\right)\right)-\omega\left(\sigma\right)-\int_{C_{0}\left(\Omega,\mathbb{R}_{+}\right)}\nu_{\varphi}\left(\int_{0}^{1}\varphi\left(\sigma\left(t\right)\right)\left|\dot{\sigma}\left(t\right)\right|dt\right)d\varphi=0,\forall\sigma\in S. \tag{27}$$

We integrate both sides of (26) along an arbitrary path σ to get

$$\int_{0}^{1}g\left(c\left(\sigma\left(t\right)\right),\lambda^{*}\left(\sigma\left(t\right)\right)\right)\left|\dot{\sigma}\left(t\right)\right|dt+\int_{C_{0}\left(\Omega,\mathbb{R}_{+}\right)}\nu_{\varphi}\left(\int_{0}^{1}\varphi\left(\sigma\left(t\right)\right)\left|\dot{\sigma}\left(t\right)\right|dt\right)d\varphi=0,\forall\sigma\in S,$$

and then add with (27) to get

$$\mu\left(\sigma\left(0\right),\sigma\left(1\right)\right)+\int_{0}^{1}g\left(c\left(\sigma\left(t\right)\right),\lambda^{*}\left(\sigma\left(t\right)\right)\right)\left|\dot{\sigma}\left(t\right)\right|dt=\omega\left(\sigma\right),\forall\sigma\in S.$$

Since $\omega\left(\sigma\right)\geq0$, for every origin-destination pair $\left(x,y\right)$, we find a $u^{*}\left(x,y\right)=-\mu\left(x,y\right)$ such that condition (4) is satisfied. Further, by complementary conditions

$$\omega\left(\sigma\right)f\left(\sigma\right)=0,\forall\sigma\in S,$$

we have

$$\left(-u^{*}\left(\sigma\left(0\right),\sigma\left(1\right)\right)+\int_{0}^{1}g\left(c\left(\sigma\left(t\right)\right),\lambda^{*}\left(\sigma\left(t\right)\right)\right)\left|\dot{\sigma}\left(t\right)\right|dt\right)f\left(\sigma\right)=0,\forall\sigma,$$

and this indicates that

$$f^{*}\left(\left\{\sigma\in S:\int_{0}^{1}g\left(c\left(\sigma\left(t\right)\right),\lambda^{*}\left(\sigma\left(t\right)\right)\right)\left|\dot{\sigma}\left(t\right)\right|dt>u^{*}\left(\sigma\left(0\right),\sigma\left(1\right)\right)\right\}\right)=0,$$

which is exactly (5). This finishes the proof that $\boldsymbol{\lambda}^{*}$ is a Wardrop equilibrium point flow pattern.

Now we prove Lemma 2. If $\boldsymbol{\lambda}^{*}$is a solution to (25), since $g\left(c\left(x\right),\cdot\right)$ is strictly increasing, for $\forall\boldsymbol{\lambda}\in\Lambda\left(\gamma\right)$, $0\leq s\leq1$ we have

$$\left(g\left(c\left(x\right),\lambda^{*}\left(x\right)+s\left(\lambda\left(x\right)-\lambda^{*}\left(x\right)\right)\right)-g\left(x,c\left(x\right),\lambda^{*}\left(x\right)\right)\right)\left(\lambda\left(x\right)-\lambda^{*}\left(x\right)\right)\geq0,\forall x\in\Omega.$$

Integrate the above inequality over Ω and we get

$$\begin{aligned}&\int_{\Omega}g\left(c\left(x\right),\lambda^{*}\left(x\right)+s\left(\lambda\left(x\right)-\lambda^{*}\left(x\right)\right)\right)\left(\lambda\left(x\right)-\lambda^{*}\left(x\right)\right)dx\\ \geq&\int_{\Omega}g\left(c\left(x\right),\lambda^{*}\left(x\right)\right)\left(\lambda\left(x\right)-\lambda^{*}\left(x\right)\right)dx\\ \geq&0.\end{aligned}$$

Also notice that for $\forall x\in\Omega$, since $\frac{\partial h(c(x),\lambda(x))}{\partial\lambda(x)}=g\left(c\left(x\right),\lambda\left(x\right)\right)$, we have

$$\begin{aligned}&h\left(c\left(x\right),\lambda\left(x\right)\right)-h\left(c\left(x\right),\lambda^{*}\left(x\right)\right)\\ =&\int_{0}^{1}g\left(c\left(x\right),\lambda^{*}\left(x\right)+s\left(\lambda\left(x\right)-\lambda^{*}\left(x\right)\right)\right)\left(\lambda\left(x\right)-\lambda^{*}\left(x\right)\right)ds.\end{aligned}$$

Integrate the above equality over Ω and apply Fubini's theorem we get

$$\begin{aligned}
&\int_{\Omega} h\left(c\left(x\right),\lambda\left(x\right)\right)dx-\int_{\Omega} h\left(c\left(x\right),\lambda^{*}\left(x\right)\right)dx\\
=&\int_{\Omega}\left(\int_{0}^{1} g\left(c\left(x\right),\lambda^{*}\left(x\right)+s\left(\lambda\left(x\right)-\lambda^{*}\left(x\right)\right)\right)\left(\lambda\left(x\right)-\lambda^{*}\left(x\right)\right)ds\right)dx\\
=&\int_{0}^{1}\left(\int_{\Omega} g\left(c\left(x\right),\lambda^{*}\left(x\right)+s\left(\lambda\left(x\right)-\lambda^{*}\left(x\right)\right)\right)\left(\lambda\left(x\right)-\lambda^{*}\left(x\right)\right)dx\right)ds\\
\geq&0,
\end{aligned}$$

which indicates that the corresponding objective value of $\forall\boldsymbol{\lambda}\in\Lambda\left(\gamma\right)$ for the optimization problem (6) is equal to or larger than that of $\boldsymbol{\lambda}^{*}$ Thus, $\boldsymbol{\lambda}^{*}$ is a solution to (6). Conversely, if $\boldsymbol{\lambda}^{*}$ is a solution to (6), notice that the function $h\left(c\left(x\right),\cdot\right)$ is convex since $g\left(c\left(x\right),\cdot\right)$ is convex and strictly increasing, the first-order condition of convexity tells us that for $\forall\boldsymbol{\lambda}\in\Lambda\left(\gamma\right)$, the following inequality holds:

$$h\left(c\left(x\right),\lambda^{*}\left(x\right)\right)\geq h\left(c\left(x\right),\lambda\left(x\right)\right)+g\left(c\left(x\right),\lambda\left(x\right)\right)\left(\lambda^{*}\left(x\right)-\lambda\left(x\right)\right).$$

Integrate the above inequality over Ω and we get:

$$\begin{aligned}
&\int_{\Omega} g\left(c\left(x\right),\lambda\left(x\right)\right)\left(\lambda\left(x\right)-\lambda^{*}\left(x\right)\right)dx\\
\geq&\int_{\Omega} h\left(c\left(x\right),\lambda\left(x\right)\right)dx-\int_{\Omega} h\left(c\left(x\right),\lambda^{*}\left(x\right)\right)dx\\
\geq&0.
\end{aligned}$$

Notice that the function $g\left(c\left(x\right),\cdot\right)$ is continuous. For $\forall\alpha\in\left(0,1\right)$ let $\lambda_{\alpha}\left(x\right)=\lambda\left(x\right)+\alpha\left(\lambda^{*}\left(x\right)-\lambda\left(x\right)\right)$, by Proposition 1 we have $\boldsymbol{\lambda}_{\alpha}\in\Lambda_{\alpha}$ and thus

$$\begin{aligned}
0\leq&\int_{\Omega} g\left(c\left(x\right),\lambda_{\alpha}\left(x\right)\right)\left(\lambda_{\alpha}\left(x\right)-\lambda^{*}\left(x\right)\right)dx\\
=&\int_{\Omega} g\left(c\left(x\right),\lambda_{\alpha}\left(x\right)\right)\left(\lambda\left(x\right)+\alpha\left(\lambda^{*}\left(x\right)-\lambda\left(x\right)\right)-\lambda^{*}\left(x\right)\right)dx\\
=&\left(1-\alpha\right)\int_{\Omega} g\left(c\left(x\right),\lambda_{\alpha}\left(x\right)\right)\left(\lambda\left(x\right)-\lambda^{*}\left(x\right)\right)dx.
\end{aligned}$$

Thus, $\int_{\Omega} g\left(c\left(x\right),\lambda_{\alpha}\left(x\right)\right)\left(\lambda\left(x\right)-\lambda^{*}\left(x\right)\right)dx\geq 0$. Taking the limit as $\alpha\rightarrow 1$ we get $\int_{\Omega} g\left(c\left(x\right),\lambda^{*}\left(x\right)\right)\left(\lambda\left(x\right)-\lambda^{*}\left(x\right)\right)dx\geq 0$, i.e., $\boldsymbol{\lambda}^{*}$ is a solution to (25). □

Proof of Lemma 3

Proof. To show the existence of the Wardrop equilibrium solution $\boldsymbol{\lambda}^*$, it is sufficient to show that the condition (25) admits a solution. Inspired by the work of (Idone 2004), we will prove the existence result based on the following classical existence theorem:

"Let E be a real topological vector space and let $\mathbb{K} \subseteq E$ be convex, closed, bounded and nonempty. Let $C : \mathbb{K} \to E^*$ be given such that C is monotone and hemicontinuous along line segments. Then there exists $u \in \mathbb{K}$ such that $\langle C(u), v-u\rangle \geq 0, \forall v \in \mathbb{K}$"

To see a detailed proof for the existence theorem, refer to Corollary 3.7 in the work of (Maugeri and Raciti 2009). Set $\mathbb{K} := \Lambda$ and C be the congestion function $g(c(x), \lambda(x))$ and set $\langle C(u), v-u\rangle = \int_\Omega C(u(x))(u(x)-v(x))\,dx$, then it is trivial to check that all the assumptions of the theorem are satisfied, and the existence result naturally follows.

To show the uniqueness, assume that $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \Lambda(\gamma)$ are both Wardrop equilibrium solutions. Then for $\boldsymbol{\lambda}_\alpha = \alpha\boldsymbol{\lambda}_2 + (1-\alpha)\boldsymbol{\lambda}_1 \in \Lambda(\gamma)$ where $\alpha \in (0,1)$, we have

$$0 \leq \int_\Omega g(c(x), \lambda_1(x))(\lambda_\alpha(x) - \lambda_1(x))\,dx,$$

$$0 \leq \int_\Omega g(c(x), \lambda_2(x))(\lambda_\alpha(x) - \lambda_2(x))\,dx.$$

Since $g(c(x), \cdot)$ is monotone, we have

$$0 \leq \int_\Omega g(c(x), \lambda_\alpha(x))(\lambda_\alpha(x) - \lambda_1(x))\,dx = \alpha\int_\Omega g(c(x), \lambda_\alpha(x))(\lambda_2(x) - \lambda_1(x))\,dx,$$

$$0 \leq \int_\Omega g(c(x), \lambda_\alpha(x))(\lambda_\alpha(x) - \lambda_2(x))\,dx = (1-\alpha)\int_\Omega g(c(x), \lambda_\alpha(x))(\lambda_1(x) - \lambda_2(x))\,dx,$$

which implies that $\int_\Omega g(c(x), \lambda_\alpha(x))(\lambda_2(x) - \lambda_1(x))\,dx = 0$. Since $g(c(x), \lambda_\alpha(x)) > 0$, we have $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_2$. □

Proof of Proposition 2

Proof. All we need to show is: $\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}\in\Lambda(\gamma)} \int_\Omega h(c(x), \lambda(x))\,dx$ if and only if $\boldsymbol{\xi}^* = \arg\min_{\boldsymbol{\xi}\in\Xi(\boldsymbol{c})} \int_\Omega \bar{h}(c(x), \xi(x))\,dx - \int_{\Omega\times\Omega} u_\xi(x,y)\gamma(x,y)\,dxdy$. The following proof follows a similar argument of Carlier and Santambrogio (2012).

By the definition of $\bar{h}(c(x), \xi(x))$, for $\forall \boldsymbol{\xi} \in \Xi(\boldsymbol{c})$ we have

$$\bar{h}(c(x), \xi(x)) \geq \xi(x)\lambda(x) - h(c(x), \lambda(x)), \forall x \in \Omega.$$

Further, if the equality holds, we have

$$\xi(x)\lambda(x) - h(c(x), \lambda(x)) \geq \xi(x)\alpha(x) - h(c(x), \alpha(x)), \forall \alpha(x) \geq 0$$

$$\Longleftrightarrow h\left(c\left(x\right),\alpha\left(x\right)\right)\geq h\left(c\left(x\right),\lambda\left(x\right)\right)+\xi\left(x\right)\left(\alpha\left(x\right)-\lambda\left(x\right)\right),\forall\alpha\left(x\right)\geq 0$$

Since $h\left(c\left(x\right),\cdot\right)$ is convex, $\bar{h}\left(c\left(x\right),\xi\left(x\right)\right)=\xi\left(x\right)\lambda\left(x\right)-h\left(c\left(x\right),\lambda\left(x\right)\right)$ if and only if $\xi\left(x\right)=\frac{\partial h(c(x),\lambda(x))}{\partial\lambda(x)}=g\left(c\left(x\right),\lambda\left(x\right)\right),\forall x\in\Omega$. By (3), we have

$$\int_{\Omega}\xi\left(x\right)\lambda\left(x\right)dx=\int_{S}f\left(\sigma\right)\left(\int_{0}^{1}\xi\left(\sigma\left(t\right)\right)\left|\dot{\sigma}\left(t\right)\right|dt\right)d\sigma,$$

and thus

$$\begin{aligned}\int_{\Omega}h\left(c\left(x\right),\lambda\left(x\right)\right)dx&\geq\int_{\Omega}\xi\left(x\right)\lambda\left(x\right)dx-\int_{\Omega}\bar{h}\left(c\left(x\right),\xi\left(x\right)\right)dx\\&=\int_{S}f\left(\sigma\right)\left(\int_{0}^{1}\xi\left(\sigma\left(t\right)\right)\left|\dot{\sigma}\left(t\right)\right|dt\right)d\sigma-\int_{\Omega}\bar{h}\left(c\left(x\right),\xi\left(x\right)\right)dx\\&\geq\int_{S}f\left(\sigma\right)u_{\xi}\left(\sigma\left(0\right),\sigma\left(1\right)\right)d\sigma-\int_{\Omega}\bar{h}\left(c\left(x\right),\xi\left(x\right)\right)dx\\&=\int_{\Omega\times\Omega}u_{\xi}\left(x,y\right)\gamma\left(x,y\right)dxdy-\int_{\Omega}\bar{h}\left(c\left(x\right),\xi\left(x\right)\right)dx,\end{aligned}$$

which indicates that

$$\inf_{\boldsymbol{\lambda}\in\Lambda(\gamma)}\int_{\Omega}h\left(c\left(x\right),\lambda\left(x\right)\right)dx\geq\sup_{\boldsymbol{\xi}\in\Xi(\boldsymbol{c})}\left\{\int_{\Omega\times\Omega}u_{\xi}\left(x,y\right)\gamma\left(x,y\right)dxdy-\int_{\Omega}\bar{h}\left(c\left(x\right),\xi\left(x\right)\right)dx\right\}$$

Now assume that $\boldsymbol{\lambda}^{*}\in\Lambda\left(\gamma\right)$ is a user equilibrium solution, and let $\xi^{*}\left(x\right)=g\left(c\left(x\right),\lambda^{*}\left(x\right)\right)$, then by previous discussion we have

$$h\left(c\left(x\right),\lambda^{*}\left(x\right)\right)+\bar{h}\left(c\left(x\right),\xi^{*}\left(x\right)\right)=\xi^{*}\left(x\right)\lambda^{*}\left(x\right),\forall x\in\Omega.$$

The equilibrium conditions (4) and (5) implies that

$$f\left(\sigma\in S^{x,y}:\int_{0}^{1}\xi^{*}\left(\sigma\left(t\right)\right)\left|\dot{\sigma}\left(t\right)\right|dt=u_{\boldsymbol{\xi}^{*}}\left(x,y\right)\right)=\gamma\left(x,y\right),$$

and thus by (3) we have

$$\begin{aligned}\int_{\Omega}\xi^{*}\left(x\right)\lambda^{*}\left(x\right)dx&=\int_{S}f\left(\sigma\right)\left(\int_{0}^{1}\xi^{*}\left(\sigma\left(t\right)\right)\left|\dot{\sigma}\left(t\right)\right|dt\right)d\sigma\\&=\int_{\Omega\times\Omega}u_{\boldsymbol{\xi}^{*}}\left(x,y\right)\gamma\left(x,y\right)dxdy.\end{aligned}$$

So we find a $\boldsymbol{\lambda}^{*}$ and $\boldsymbol{\xi}^{*}$ such that

$$\begin{aligned}\int_{\Omega}h\left(c\left(x\right),\lambda^{*}\left(x\right)\right)dx&=\int_{\Omega}\xi^{*}\left(x\right)\lambda^{*}\left(x\right)dx-\int_{\Omega}\bar{h}\left(c\left(x\right),\xi^{*}\left(x\right)\right)dx\\&=\int_{\Omega\times\Omega}u_{\boldsymbol{\xi}^{*}}\left(x,y\right)\gamma\left(x,y\right)dxdy-\int_{\Omega}\bar{h}\left(c\left(x\right),\xi^{*}\left(x\right)\right)dx,\end{aligned}$$

which indicates that

$$\begin{aligned}\min_{\boldsymbol{\lambda}\in\Lambda(\gamma)}\int_{\Omega}h\left(c\left(x\right),\lambda\left(x\right)\right)dx&=\max_{\boldsymbol{\xi}\in\Xi(\boldsymbol{c})}\left\{\int_{\Omega\times\Omega}u_{\boldsymbol{\xi}}\left(x,y\right)\gamma\left(x,y\right)dxdy-\int_{\Omega}\bar{h}\left(c\left(x\right),\xi^{*}\left(x\right)\right)dx\right\}\\&=-\min_{\boldsymbol{\xi}\in\Xi(\boldsymbol{c})}\left\{\int_{\Omega}\bar{h}\left(c\left(x\right),\xi^{*}\left(x\right)\right)dx-\int_{\Omega\times\Omega}u_{\xi}\left(x,y\right)\gamma\left(x,y\right)dxdy\right\}\end{aligned}$$

□

Proof of Proposition 3

Proof. Since $g\left(c\left(x\right),\lambda\left(x\right)\right)$ is super-linear with related to $\lambda\left(x\right)$ and the fact that $\frac{\partial h(c(x),\lambda(x))}{\partial\lambda(x)} = g\left(c\left(x\right),\lambda\left(x\right)\right)$, by the first order condition we have

$$\arg\sup_{\lambda(x)\geq 0}\left\{\xi\left(x\right)\lambda\left(x\right)-h\left(c\left(x\right),\lambda\left(x\right)\right)\right\}=\lambda^{*}\text{s.t.}\lambda^{*}\left(x\right)=g^{-1}\left(c\left(x\right),\xi\left(x\right)\right)$$

which implies that $\bar{h}\left(c\left(x\right),\xi\left(x\right)\right)=\xi\left(x\right)g^{-1}\left(c\left(x\right),\xi\left(x\right)\right)-h\left(c\left(x\right),g^{-1}\left(c\left(x\right),\xi\left(x\right)\right)\right)$. Now consider the first order condition of (15) we have

$$g^{-1}\left(c\left(x\right),\xi\left(x\right)\right)+\xi\left(x\right)\frac{\partial g^{-1}\left(c\left(x\right),\xi\left(x\right)\right)}{\partial\xi\left(x\right)}-g\left(c\left(x\right),g^{-1}\left(c\left(x\right),\xi\left(x\right)\right)\right)\frac{\partial g^{-1}\left(c\left(x\right),\xi\left(x\right)\right)}{\partial\xi\left(x\right)}-\alpha_{\gamma}=0$$

Due to the fact that $g\left(c\left(x\right),g^{-1}\left(c\left(x\right),\xi\left(x\right)\right)\right)=\xi\left(x\right)$ and $\alpha_{\gamma}\left(x\right)\geq 0$, the first order condition implies that $g^{-1}\left(c\left(x\right),\xi^{*}\left(x\right)\right)=\alpha_{\gamma}\left(x\right)$. □

Proof of Corollary 2

Proof. We denote $\Xi\left(\rho\right)=\int_{r_0}^{\rho}\xi\left(r\right)dr$ and the key to lower level problem is to find the minimum $\boldsymbol{\xi}$ for the functional

$$\int_{0}^{R}\frac{4}{5}c\left(\rho\right)\left(\xi\left(\rho\right)-t_{0}\left(\rho\right)\right)\left(\frac{\xi\left(\rho\right)-t_{0}\left(\rho\right)}{\kappa t_{0}\left(\rho\right)}\right)^{\frac{1}{4}}2\pi\rho d\rho-\int_{r_0}^{R}\left(2\pi\rho k\left(\rho\right)\int_{r_0}^{\rho}\xi\left(r\right)dr\right)d\rho$$

$$J\left(\xi\left(\rho\right)\right)=2\pi\int_{r_0}^{R}L\left(\rho,\xi\left(\rho\right),\Xi\left(\rho\right)\right)d\rho$$

where

$$L\left(\rho,\xi\left(\rho\right),\Xi\left(\rho\right)\right)=\frac{4}{5}\rho c\left(\rho\right)\left(\xi\left(\rho\right)-t_{0}\left(\rho\right)\right)\left(\frac{\xi\left(\rho\right)-t_{0}\left(\rho\right)}{\kappa t_{0}\left(\rho\right)}\right)^{\frac{1}{4}}-\rho k\left(\rho\right)\Xi\left(\rho\right).$$

By Euler-Lagrange equation, a necessary condition for a minimum is $\frac{\partial L(\rho,\xi(\rho),\Xi(\rho))}{\partial\Xi(\rho)}-\frac{d}{d\rho}\left(\frac{\partial L(\rho,\xi(\rho),\Xi(\rho))}{\partial\xi(\rho)}\right)=0$, i.e.,

$$\begin{aligned}0=&-\rho k\left(\rho\right)-\frac{d}{d\rho}\left(\rho c\left(\rho\right)\left(\frac{\xi\left(\rho\right)-t_{0}\left(\rho\right)}{\kappa t_{0}\left(\rho\right)}\right)^{\frac{1}{4}}\right)\\=&-\rho k\left(\rho\right)-c\left(\rho\right)\left(\frac{\xi\left(\rho\right)-t_{0}\left(\rho\right)}{\kappa t_{0}\left(\rho\right)}\right)^{\frac{1}{4}}-\rho c'\left(\rho\right)\left(\frac{\xi\left(\rho\right)-t_{0}\left(\rho\right)}{\kappa t_{0}\left(\rho\right)}\right)^{\frac{1}{4}}-\frac{1}{4}\rho c\left(\rho\right)\left(\frac{\xi\left(\rho\right)-t_{0}\left(\rho\right)}{\kappa t_{0}\left(\rho\right)}\right)^{-\frac{3}{4}}\left(\frac{\xi\left(\rho\right)-t_{0}\left(\rho\right)}{\kappa t_{0}\left(\rho\right)}\right)'\end{aligned}$$

$$\Rightarrow 4\rho k\left(\rho\right)\left(\frac{\xi\left(\rho\right)-t_{0}\left(\rho\right)}{\kappa t_{0}\left(\rho\right)}\right)^{\frac{3}{4}}+4c\left(\rho\right)\left(\frac{\xi\left(\rho\right)-t_{0}\left(\rho\right)}{\kappa t_{0}\left(\rho\right)}\right)+4\rho c'\left(\rho\right)\left(\frac{\xi\left(\rho\right)-t_{0}\left(\rho\right)}{\kappa t_{0}\left(\rho\right)}\right)+\rho c\left(\rho\right)\left(\frac{\xi\left(\rho\right)-t_{0}\left(\rho\right)}{\kappa t_{0}\left(\rho\right)}\right)'=0$$

Let $f(\rho) = (\rho c(\rho))^4 \left(\frac{\xi(\rho)-t_0(\rho)}{\kappa t_0(\rho)}\right)$, then the above condition is equal to the following ordinary differential equations

$$4\rho k(\rho) f^{\frac{3}{4}}(\rho) + (f(\rho))' = 0,$$

with the general solution: $f(\rho) = \frac{\left(C - \int_{r_0}^{\rho} 4rk(r)dr\right)^4}{256}$ where C can be any constant. The boundary condition can be obtained by noticing that there should be no flow over the boundary, and thus $\xi(R) = t_0(R)$, which indicates that $f(R) = 0$. So the final solution for $f(\rho)$ is

$$f(\rho) = \left(\int_{\rho}^{R} rk(r)\,dr\right)^4,$$

which indicates that

$$\xi^*(\rho) = t_0(\rho)\left(1 + \kappa\left(\frac{\int_{\rho}^{R} rk(r)\,dr}{\rho c(\rho)}\right)^4\right),$$

and from then on it is easy to get

$$c^*(\rho) = \max\left\{a(\rho), \left(\frac{\int_{\rho}^{R} rk(r)\,dr}{\rho}\right)\left(\frac{4\kappa t_0(\rho)}{A(\rho)}\right)^{\frac{1}{5}}\right\}$$

□

Proof of Proposition 5

Proof. Notice that the lower level objective function of the trust-region sub-problem (23) is convex, and hence the conclusion follows by applying the first-order conditions for a constrained minimum. □