Heterogenous quadratic regularization in optimal transport in Peru

Marcelo Gallardo [*]
marcelo.gallardo@pucp.edu.pe

Manuel Loaiza [†]
manuel.loaiza@autodesk.com

Jorge Chávez [‡]
jrchavez@pucp.edu.pe

March 30, 2025

Abstract

In this paper, we extend the optimal transport model with quadratic regularization by incorporating heterogeneous congestion costs, particularly in the context of matching within the healthcare and education sectors in countries where both physical and bureaucratic congestion are significant. We analyze the mathematical properties of the model under specific cases and explore its key structural characteristics. Additionally, we present numerical examples demonstrating how this formulation more accurately captures real-world congestion effects compared to the classical optimal transport model, with a particular focus on the Peruvian context. Our main results hold under mild assumptions and establish the existence and structure of optimal solutions in the integer setting.

Keywords: optimal transport, congestion costs, quadratic regularization, matching.
JEL classifications: C61, C62, C78, D04.

[*] Department of Mathematics, Pontificia Universidad Católica del Perú (PUCP). Acknowledges insightful discussions with Professors Federico Echenique.

[†] Autodesk, Inc.

[‡] Department of Mathematics, Pontificia Universidad Católica del Perú (PUCP). Acknowledges support from Pontificia Universidad Católica del Perú DAP.

1 Introduction

Matching theory in economics studies how agents are paired based on their preferences and market constraints. The seminal work of Gale and Shapley (1962) introduced the concept of stable matching, where no two agents prefer each other over their assigned partners. This theory was later extended by Hylland and Zeckhauser (1979) in the context of house allocation and by Kelso and Crawford (1982), who incorporated transfers in two-sided matching. Alvin Roth significantly advanced these ideas, applying them to real-world scenarios such as school admissions and organ transplants, proving the effectiveness of matching mechanisms (Roth (1982), Roth and Sotomayor (1990)). He also proved that no stable matching mechanism ensures truthfulness as a dominant strategy (Roth (1982)). In school choice, Abdulkadiroğlu and Sönmez (2003) developed mechanisms that balance stability and individual preferences. Further contributions include Hatfield and Milgrom (2005), Echenique and Yenmez (2015), and the recent book by Echenique et al. (2023), which provides a comprehensive overview of these subjects.

In recent years, matching theory has been enhanced by Optimal Transport (OT) methods, a mathematical framework introduced by Monge in the 18th century and rigorously formalized by Kantorovich in the 20th century. OT addresses optimal assignment problems by minimizing transportation costs over distributions, and has been widely applied in matching markets, including student-school, patient-hospital, and worker-firm pairings. The foundational book by Cédric Villani (Villani (2009)), together with Ekeland (2010) and Ambrosio et al. (2021), provide an exhaustive mathemmatical treatment of OT. In economics, Alfred Galichon bridged OT with discrete matching problems in markets in Galichon (2016) and Galichon (2021).

The main advantage of OT is that it allows to model agents as continuous distributions rather than discrete entities. For instance, a recent study, Echenique et al. (2024), explores stability in matching through an OT framework applicable to both discrete and continuous settings. In the discrete setting, Galichon (2021) employs computational methods (SISTA algorithm) akin to Merigot and Thibert (2020) and Nenna (2020), to recover matching costs in migration problems. Further applications include models in marriage and labor markets (Dupuy and Galichon, 2014; Dupuy et al., 2019).

In this work, we develop a variant of the OT model with quadratic heterogenous regularization for the discrete setting. The model captures congestion costs and provides new insights relative to the existing literature. In particular, after addressing the problem from a theoretical perspective using basic elements of convex nonlinear programming, we present examples that illustrate our approach. The remainder of the paper is organized as follows. In Section 2, we introduce the notation and review existing models from the literature, providing the necessary background for our analysis. In Section 3, we present our new model and conduct a thorough examination of its mathematical properties. In particular, we focus on the characterization of both interior and corner solutions, which play a crucial role in microeconomic theory. We derive key theoretical results that are essential for understanding the implications of our framework. Finally, in Section 4, we apply our model to concrete economic settings, specifically analyzing inefficiencies in educational and healthcare matching markets in Peru.

Peru is one of the most traffic-congested countries in the world, leading to significant economic losses due to inefficient transportation policies and inadequate infrastructure (Martinez, 2024). Additionally, the country faces a fragile and underfunded healthcare system, as evidenced by the devastating impact of COVID-19, making Peru the most affected country globally in terms of mortality rates (Médicos Sin Fronteras, 2021). The education sector also reflects deep structural issues, with many lacking access to schooling, and even those who do often receive substandard education, as Peru consistently ranks among the lowest in international assessments such as PISA (Organisation for Economic Co-operation and Development (OECD), 2024). Overcrowding, system saturation, and congestion potentially explain this. The model we propose takes this into account. Therefore, our study is highly relevant as it provides new insights into this critical scenario, shedding light on key issues and potential policy solutions.

2 Notation and preliminaries

In this work, we denote by $X = \{x_1, \cdots, x_n\}$ and $Y = \{y_1, \cdots, y_m\}$ two sets to be matched: students and schools, patients and hospitals, workers and firms, etc. We denote by $\mathbb{Z}_+^N$ the set of positive integers in the N-dimensional real vector space $\mathbb{R}^N$. The notation $\mathcal{M}_{m\times n}$ represents the set of matrices with m rows and n columns. Each x_i may represent a group containing one or more individuals, such as groups of students. We denote by μ_i the number of individuals in these groups, referred to as mass. Similarly, ν_j denotes the capacity of y_j. For instance, it may represent the number of available spots in a school, hospital beds, among others. We also denote $I = \{1, \cdots, n\}$ and $J = \{1, \cdots, m\}$.

The classical discrete transport model assumes that the marginal cost of matching an individual from x_i to y_j is constant and equal to c_{ij}. This parameter depends on group preferences, distances, and other factors. Therefore, from the perspective of a central planner, the goal is to solve:

$$\mathcal{P}_O : \min_{\pi\in\Pi(\mu,\nu)} \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}\pi_{ij}, \tag{1}$$

where

$$\Pi(\mu,\nu) = \left\{ \pi_{ij} \geq 0 : \sum_{j=1}^{m} \pi_{ij} = \mu_i \quad \forall\, i \in I, \quad \sum_{i=1}^{n} \pi_{ij} = \nu_j \quad \forall\, j \in J \right\}. \tag{2}$$

Note that π_{ij} represents the number of individuals matched from i to j and that constraints (2) ensure that all individuals (students, patients, etc.) are assigned, and that all entities (schools, hospitals, etc.) fill their available capacity[1]. A solution to (1) is known as optimal matching or optimal transport plan. It will be denoted by π^*. To solve $\mathcal{P}_O$, linear programming techniques such as the simplex method are typically employed.

The problem (1) has been extensively studied and extended. Among these extensions is the

[1] It may not seem entirely accurate in the context of a country like Peru. However, in certain spaces or problems, the assumption is reasonable. Moreover, μ_i and ν_j can be interpreted as lower bounds. Indeed, if the central planner seeks to ensure that $\sum_j \pi_{ij} \in [\mu_i^L, \mu_i^H]$ and $\sum_i \pi_{ij} \in [\nu_j^L, \nu_j^H]$, this is equivalent to considering μ_i^L, ν_j^L in $\mathcal{P}_O$.

entropic regularization model (Carlier et al., 2020; Peyré and Cuturi, 2019)

$$\mathcal{P}_E : \min_{\pi \in \Pi(\mu,\nu)} \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}\pi_{ij} + \alpha \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{m} \pi_{ij}\log_e(\pi_{ij})}_{\mathcal{E}(\pi)},$$

with $\alpha > 0$. $\mathcal{E}(\pi)$ is continuously extended at $\pi_{ij} = 0$ using that $\lim_{x\downarrow 0} x \ln x$. Another more recent extension is the quadratic regularization model (Nutz, 2024; Lorenz et al., 2019a):

$$\mathcal{P}_Q : \min_{\pi \in \Pi(\mu,\nu)} \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}\pi_{ij} + \frac{\varepsilon}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} \pi_{ij}^2,$$

with $\varepsilon > 0$. These formulations allow for a more uniform distribution of transport mass, ensure the uniqueness of a solution, and are computationally more efficient (Merigot and Thibert, 2020).

Before introducing our model, it is important to discuss the existence of solutions to $\mathcal{P}_O, \mathcal{P}_E$ and $\mathcal{P}_Q$. The first key observation is that, given the economic context, solutions are expected to belong to $\mathbb{Z}_+^{nm}$. However, as stated, the optimization problems above do not inherently enforce that the solution lies in $\mathbb{Z}_+^{nm}$. If the problem is solved in $\mathbb{Z}_+^{nm}$, a combinatorial argument ensures the existence of a solution: Proposition 2.1 guarantees that there exists a finite number of matchings, and thus, at least one optimal matching must exist.

Proposition 2.1. In an integer setting, the number of matchings is at most $m^{\sum_{i=1}^{n}\mu_i}$.

Proof. The number of ways to assign all μ_i individuals from group i to entities is given by solutions to:

$$\pi_{i1} + \cdots + \pi_{im} = \mu_i, \quad 0 \le \pi_{ij} \le \nu_j \quad \forall\, j = 1, \cdots, m. \tag{3}$$

Disregarding the upper bounds ν_j, this reduces to a stars and bars problem (Levin, 2015). The upper bound for the number of solutions to (3) is $\binom{\mu_i+m-1}{m-1}$. Applying the multiplication principle, the total number of matchings satisfies:

$$\prod_{i=1}^{n}\binom{\mu_i+m-1}{m-1} = \prod_{i=1}^{n}\prod_{j=1}^{\mu_i}\frac{j+m-1}{j} \le \prod_{i=1}^{n}\prod_{j=1}^{\mu_i} m = m^{\sum_{i=1}^{n}\mu_i}. \qquad \blacksquare$$

The issue, as indicated, is that a priori there is no guarantee that feasible matchings belong to $\mathbb{Z}_+^{nm}$. It turns out that in the discrete linear case, we have $\pi_{ij} \in \mathbb{Z}_+$. However, in the case of optimization problems with regularization, this is no longer necessarily true (see for instance (13)). Nevertheless, the existence of a solution follows quickly from Weierstrass' Theorem (Proposition 2.2).

Proposition 2.2. Given $\mu = (\mu_1, ..., \mu_n)^T \in \mathbb{R}_{++}^n$ and $\nu = (\nu_1, ..., \nu_m)^T \in \mathbb{R}_{++}^m$, $\mathcal{P}_O, \mathcal{P}_E$ and $\mathcal{P}_Q$, always have a solution $\pi^* \in \mathbb{R}_+^{nm}$.

Proof. In each case, the objective function is continuous as it is linear. The constraint set $\Pi(\mu,\nu)$ is compact in $\mathbb{R}^{nm}$ since it is the intersection of closed sets and bounded within $[0, \sum_{i=1}^{n}\mu_i]^{nm}$. $\blacksquare$

The issue with the solution lying in $\mathbb{R}_+^{nm}$ instead of the integers is similar to the problem encountered in utility maximization: it lacks economic meaning to consume, for instance, 1.5 cars or $\sqrt{2}$ phones. However, as we will discuss in detail later, the convex and quadratic structure allows us to obtain good approximations via optimization in the real domain.

The basic linear model, as well as the entropic and quadratic regularization problems, have been extensively studied in the literature (Dupuy and Galichon, 2014; Carlier et al., 2020; Lorenz et al., 2019b; González-Sanz and Nutz, 2024; Wiesel and Xu, 2024; Nutz, 2024). We now move on to our heterogeneous quadratic costs model, which, to the best of our knowledge, along with our results, are novel contributions to the literature.

3 The model and structural properties

Traffic congestion and institutional overload are crucial factors affecting the allocation of individuals to entities such as schools and hospitals. When too many individuals are matched to the same entity, congestion costs escalate, leading to inefficiencies in both physical and bureaucratic dimensions. This phenomenon is observed in various settings:

- **Traffic congestion:** The simultaneous assignment of many students to the same school in urban areas can increase travel times, overload public transport, and generate bottlenecks in key traffic zones. The same happens with patients and hospitals Alba-Vivar (2025).
- **Medical centers overload:** Large patient inflows can overwhelm hospital resources, creating long waiting times, administrative bottlenecks, and inefficient service delivery (EsSalud, 2025a,b).
- **Bureaucratic congestion:** Excess demand for certain institutions may slow down processing times, affecting school admissions, hospital triage, and public service allocation due to outdated systems and inefficient workflows.

To model this phenomenon, we consider a strictly convex cost function with respect to the number of matched individuals $C(\pi;\theta)$, where θ is a vector of parameters. The strict convexity captures the increasing marginal costs associated with congestion. We define the cost function $C(\pi;\theta)$ as a separable and continuous function:

$$C(\pi;\theta) = \sum_{i=1}^{n}\sum_{j=1}^{m} \phi_{ij}(\pi_{ij};\theta_{ij}), \tag{4}$$

where ϕ_{ij} is structurally homogeneous[2]. The central planner's problem then becomes:

$$\min_{\pi\in\Pi(\mu,\nu)} \sum_{i=1}^{n}\sum_{j=1}^{m} \phi(\pi_{ij};\theta_{ij}), \tag{5}$$

[2] The function ϕ_{ij} does not change structurally across (i,j) pairs; whether logarithmic, exponential, or polynomial, we assume $\phi_{ij} = \phi$.

where $\Pi(\mu, \nu)$ is defined as in (2). Given that congestion leads to increasing costs, ϕ should be strictly increasing and strictly convex, transforming the problem into a convex optimization problem with linear constraints. To carry out a quantitative analysis, we assume a quadratic cost function:

$$\phi(\pi_{ij}; \theta_{ij}) = d_{ij} + c_{ij}\pi_{ij} + a_{ij}\pi_{ij}^2. \tag{6}$$

Thus, the optimization problem becomes:

$$\mathcal{P}_1 : \min_{\pi \in \Pi(\mu,\nu)} \sum_{i=1}^{n} \sum_{j=1}^{m} d_{ij} + c_{ij}\pi_{ij} + a_{ij}\pi_{ij}^2. \tag{7}$$

In here, the parameters have clear economic interpretations:

- d_{ij} represents fixed costs associated with each matching (e.g., baseline administrative or physical distance).
- $c_{ij} > 0$ corresponds to constant marginal costs, capturing individual and pair characteristics.
- $a_{ij} > 0$ introduces heterogenous congestion effects, ensuring increasing marginal costs as π_{ij} grows.

Although the Linear Independence Constraint Qualification (LICQ) condition may fail for solutions where non-negativity constraints are not binding, the convexity of the objective function and the linearity of constraints allow us to apply the Karush-Kuhn-Tucker (KKT) conditions, see Boyd (2004).

The Lagrangian function associated with (5) is given by:

$$\mathscr{L} = \sum_{\substack{1\le i\le n \\ 1\le j\le m}} \phi(\pi_{ij}; \theta_{ij}) + \sum_{i=1}^{n} \xi_i \left(\mu_i - \sum_{j=1}^{m} \pi_{ij} \right) + \sum_{j=1}^{m} \lambda_j \left(\nu_j - \sum_{i=1}^{n} \pi_{ij} \right) - \sum_{\substack{1\le i\le n \\ 1\le j\le m}} \gamma_{ij}\pi_{ij}. \tag{8}$$

The KKT first-order conditions are:

$$\begin{aligned}
\frac{\partial \mathscr{L}(\pi^*, \xi^*, \lambda^*, \gamma^*; \theta)}{\partial \pi_{ij}} = \frac{\partial \phi(\pi_{ij}^*; \theta_{ij})}{\partial \pi_{ij}} - \lambda_j^* - \xi_i^* - \gamma_{ij}^* = 0,\ \forall\ (i,j) \in I \times J \\
-\pi_{ij}^* \le 0,\ \forall\ (i,j) \in I \times J \\
\sum_{j=1}^{m} \pi_{ij}^* - \mu_i = 0,\ \forall\ i \in I \\
\sum_{i=1}^{n} \pi_{ij}^* - \nu_j = 0,\ \forall\ j \in J \\
\gamma_{ij}^* \pi_{ij}^* = 0,\ \forall\ (i,j) \in I \times J.
\end{aligned}$$

Hence, for the quadratic specification (6),

$$\pi_{ij}^* = \frac{\xi_i^* + \lambda_j^* + \gamma_{ij}^* - c_{ij}}{2a_{ij}}. \tag{9}$$

In this section, we analyze the structural properties of problem 7. The first observation is that, since the objective function is strictly convex, continuous, and the constraint set is convex, there is a unique solution. Now, if we consider $\mathbb{Z}_+^{nm} \cap \Pi(\mu, \nu)$ as the opportunity set, there exists a finite number of points where the function can be evaluated, ensuring the existence of a solution. However, uniqueness is not guaranteed. For example, minimizing $(x - 3/2)^2$ over $\mathbb{R}_+$ yields the unique solution $3/2$, but in $\mathbb{Z}_+$, there are two optimal solutions, $x^* = 1$ and $x^* = 2$.

We now focus on the characterization and properties of interior solutions, i.e., where $\pi_{ij}^* > 0$ for all i and j. We start studying the problem in $\mathbb{R}_+^{nm}$ and then we move on to the integer setting.

3.1 Structural properties in $\mathbb{R}_+^{nm}$

Proposition 3.1. With respect to problem (5), with costs given by (6), whenever $\gamma_{ij}^* = 0$ for all $(i, j) \in I \times J$, where $I = \{1, \cdots, n\}$, $J = \{1, \cdots, m\}$, the linear system obtained from (9), with respect to (ξ^*, λ^*), leads to a singular $n + m$ linear system.

Proof. Since $\gamma_{ij}^* = 0$ for all $(i, j) \in I \times J$, first order conditions lead to

$$\sum_{j=1}^m \pi_{ij}^* = \sum_{j=1}^m \frac{\xi_i^*}{2a_{ij}} + \sum_{j=1}^m \frac{\lambda_j^*}{2a_{ij}} - \sum_{j=1}^m \frac{c_{ij}}{2a_{ij}} = \mu_i, \ \forall\ i \in I \tag{10}$$

$$\sum_{i=1}^n \pi_{ij}^* = \sum_{i=1}^n \frac{\xi_i^*}{2a_{ij}} + \sum_{i=1}^n \frac{\lambda_j^*}{2a_{ij}} - \sum_{i=1}^n \frac{c_{ij}}{2a_{ij}} = \nu_j, \ \forall\ j \in J. \tag{11}$$

By setting $x = \begin{bmatrix} \xi_1^* & \cdots & \xi_n^* & \lambda_1^* & \cdots & \lambda_m^* \end{bmatrix}^T \in \mathbb{R}^{n+m}$, the linear equalities (10) and (11) on ξ_i^* and λ_j^* are described by the linear system $(\Lambda + T)x = b$, where

$$\Lambda = \text{Diag}\left(\sum_{j=1}^m \frac{1}{2a_{1j}}, \ldots, \sum_{j=1}^m \frac{1}{2a_{nj}}, \sum_{i=1}^n \frac{1}{2a_{i1}}, \ldots, \sum_{i=1}^n \frac{1}{2a_{im}} \right) \in \mathbb{R}^{n+m,n+m}.$$

$$\Upsilon = \left[\frac{1}{2a_{ij}} \right]_{\substack{1 \le i \le n \\ 1 \le j \le m}} \in \mathbb{R}^{n,m} \text{ and } T = \begin{bmatrix} 0 & \Upsilon \\ \Upsilon^T & 0 \end{bmatrix} \in \mathbb{R}^{n+m,n+m},$$

$$b = \left[\mu_1 + \sum_{j=1}^m \frac{c_{1j}}{2a_{1j}}, \ldots, \mu_n + \sum_{j=1}^m \frac{c_{nj}}{2a_{nj}}, \nu_1 + \sum_{i=1}^n \frac{c_{i1}}{2a_{i1}}, \ldots, \nu_m + \sum_{i=1}^n \frac{c_{im}}{2a_{im}} \right]^T \in \mathbb{R}^{n+m}.$$

Let $R = \Lambda + T$. If R_k denotes the $k-$th row of R, we note that $R_1 = \sum_{k=n+1}^{n+m} R_k - \sum_{k=2}^n R_k$. Hence, $\text{Det}(R) = 0$, and the claim follows. ■

Proposition 3.1 is crucial as it highlights that, even in the case of interior solutions, there is no systematic method for obtaining an analytical solution through the direct resolution of the linear system.

As usual in economics, we are interested in perform monotone or smooth comparative statics. With respect to the former (see Milgrom and Shannon (1994)), it can't be performed since $S = \Pi(\mu, \nu)$ is not a sub-lattice of $X = \mathbb{R}_+^{nm}$. Indeed, given $\pi_1, \pi_2 \in S$, in general, $\pi_1 \wedge \pi_2$ and

$\pi_1 \vee \pi_2$ do not belong to S. With respect to the latter, Proposition 3.2 explains why smooth comparative statics cannot be accomplished.

Proposition 3.2. With respect to (8), considering quadratic costs[3], we have that

$$\operatorname{Det}(J_{\pi,(\xi,\lambda)}\overline{\mathscr{L}}(\pi^*,\xi^*,\lambda^*,\overline{\theta})) = 0.$$

Proof. First, let $\pi = (\pi_{11},\dots,\pi_{1m},\cdots,\pi_{n1},\dots,\pi_{nm})^T$. Then, we define

$$D = \operatorname{Diag}(a_{11},\dots,a_{1m},\cdots,a_{n1},\dots,a_{nm}) \in \mathbb{R}^{nm,nm}_{++}$$

and $B = [b_{k\ell}] \in \mathbb{R}^{n+m,n+m}$, where

$$b_{k\ell} = \begin{cases} 1 & \text{if } k \le n \text{ and } (k-1)m < \ell \le km, \\ 1 & \text{if } n < k \le n+m \text{ and } \ell \equiv k-n \pmod{m}, \\ 0 & \text{otherwise.} \end{cases}$$

Matrix B never has full rank. Indeed, $B_1 = \sum_{k=n+1}^{n+m} B_k - \sum_{k=2}^{m} B_k$, where B_k is row k of B. Thus, since

$$J_{\pi,(\xi,\lambda)}\overline{\mathscr{L}}(\pi^*,\xi^*,\lambda^*,\overline{\theta}) = \begin{bmatrix} D & -B^T \\ -B & 0 \end{bmatrix},$$

following Gentle (2017), $\operatorname{Det}(J_{\pi,(\xi,\lambda)}\overline{\mathscr{L}}(\pi^*,\xi^*,\lambda^*,\overline{\theta})) = \operatorname{Det}(D)\operatorname{Det}(0 - BD^{-1}B^T) = 0$. ■

Although we cannot apply smooth comparative statics, the conditions of the Envelope Theorem are satisfied for π^* in the interior of Π. Therefore, by defining $V = V(\pi^*) = \sum_{i=1}^{n}\sum_{j=1}^{m}\phi_{ij}(\pi^*_{ij};\overline{\theta}_{ij})$, we can conclude that $\partial V/\partial c_{ij} = \pi^*_{ij} > 0$ and $\partial V/\partial a_{ij} = \pi^{*2}_{ij} > 0$, which is expected, as the cost of the optimal transport plan only increases if the coefficients associated with preference costs and congestion costs rise.

Note that, in general, obtaining the optimal matching π^* from (9), is quite complicated. Even if we assume an interior solution, which would simplify the equations since $\gamma^*_{ij} = 0$ automatically, we still cannot solve the linear system systematically. Note also that R not being invertible does not imply that the system has no solution. It only means that, if a solution (ξ^*,λ^*) exists, it is either not unique, or there is $\gamma^*_{ij} \neq 0$. What is unique is π^* since the objective function is strictly convex. Hence, even if we have several (ξ^*,λ^*), at the end, we obtain a unique π^*. The non uniqueness of (ξ^*,λ^*) originates from the fact that the LICQ does not hold for interior solutions.

However, from a computational perspective, our model can always be solved using standard quadratic convex optimization methods. On the other hand, when $n = m$, optimizing over $\mathbb{Z}^{nm}_+$, we can obtain an explicit solution for our model under mild assumptions. The result we present in that line in the following section is quite strong, as it allows us to obtain the explicit solution in the integer case.

[3] Following de la Fuente (2000) notation. Here $\overline{\mathscr{L}} = (\nabla_\pi \mathscr{L}, \nabla_\theta \mathscr{L})$.

3.2 Structural properties in $\mathbb{Z}_+^{nm}$

In the case of the linear model, solutions are always corner solutions (Tardella, 2010). On the other hand, in the case of entropic regularization, the solution is always interior (Nenna, 2020). The following examples show that under the quadratic setting, both interior and corner solutions could exist. Note that in $\mathcal{P}_1$, the value of d_{ij} is arbitrary, as it does not affect the solution.

Example 3.3. In this example, we show a case where the solution is interior. Consider

$$a = [a_{ij}] = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}, \; c = [c_{ij}] = \begin{bmatrix} 24 & 48 \\ 16 & 24 \end{bmatrix}, \; d = [d_{ij}] \in \mathcal{M}_{2\times 2}, \; \mu = (20, 20), \text{ and } \nu = (12, 28).$$

Consequently, we obtain $\pi^* = (7, 13, 5, 15)$, an interior solution.

Example 3.4. To illustrate a case where the solution is a corner solution, consider the following values:

$$a = \begin{bmatrix} 200 & 2 \\ 2 & 200 \end{bmatrix}, \quad c = \begin{bmatrix} 200 & 2 \\ 2 & 200 \end{bmatrix}, \; d = [d_{ij}] \in \mathcal{M}_{2\times 2}, \quad \mu = (10, 10), \text{ and } \nu = (10, 10).$$

In this scenario, the optimal solution is $\pi^* = (0, 10, 10, 0)$, a corner solution.

Now, consider adding restrictions to the parameter vector and the sizes of the sets to explicitly obtain a specific corner solutions.

Assumption 1. Let M be a positive integer strictly greater than 1. Assume that $n = m = M$ and $\mu_i = \nu_j$ for all $1 \le i, j \le M$.

Assumption 1 ensures that each school or medical center reaches full capacity with individuals from the same group.

Assumption 2. For each $1 \le i \le n$, suppose there exists $1 \le \zeta_i \le m$ such that $c_{i\,\zeta_i} < c_{i\,j}$ for all $1 \le j \le m$ with $j \ne \zeta_i$. Furthermore, assume that $\zeta_i \ne \zeta_j$ for all $1 \le i, j \le m$ with $i \ne j$.

Assumption 2 imposes that each group $i \in I$ has a unique top choice $j \in J$ based on preferences, and this top choice differs across groups.

Assumption 3. Let $\widetilde{c}_i = \min\limits_{\substack{1 \le j \le m \\ j \ne \zeta_i}} \{c_{ij}\}$ satisfy $\widetilde{c}_i > c_{i\,\zeta_i} + a_{i\,\zeta_i}\mu_i^2(1 - 1/m)$ for $1 \le i \le n$.

Assumption 3 tells us that preferences must be such that *the top choice* only based on c_{ij} is at least $a_{i\,\zeta_i}\mu_i^2(1 - 1/m)$ better than the other ones. By combining Assumptions 1, 2 and 3 we show that the solution to $\mathcal{P}_1$, in the integer setting, is given by (12). The notation $a_{i\zeta_i}$ is analogous to $c_{i\zeta_i}$ from Assumption 2.

Theorem 3.5. *Under Assumptions 1, 2 and 3, the optimal matching for $\mathcal{P}_1$ in the integer setting is given by*

$$\pi^* = [\pi_{ij}^*] = \begin{cases} \mu_i & \text{if } j = \zeta_i, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Proof. Let π be an arbitrary matching different from π^*. Then,

$$\begin{aligned} C(\pi;\theta) &= \sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij} + c_{ij}\pi_{ij} + a_{ij}\pi_{ij}^2 \\ &\geq \sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij} + \sum_{i=1}^{n}\left(\sum_{j=1}^{m} c_{ij}\pi_{ij} + a_{i\,\zeta_i}\sum_{j=1}^{m}\pi_{ij}^2\right). \end{aligned}$$

Now, consider i such that $\pi_{i\,\zeta_i} < \mu_i$. Due to the integer nature of π, $\pi_{i\,\zeta_i} \leq \mu_i - 1$. Hence

$$\begin{aligned} \sum_{j=1}^{m} c_{ij}\pi_{ij} &= c_{i\,\zeta_i}\pi_{i\,\zeta_i} + \sum_{j\neq\zeta_i} c_{ij}\pi_{ij} \\ &\geq c_{i\,\zeta_i}\pi_{i\,\zeta_i} + \widetilde{c}_i(\mu_i - \pi_{i\,\zeta_i}) \\ &= \widetilde{c}_i\mu_i - \pi_{i\,\zeta_i}(\widetilde{c}_i - c_{i\,\zeta_i}) \\ &\geq \widetilde{c}_i\mu_i - (\mu_i - 1)(\widetilde{c}_i - c_{i\,\zeta_i}) \\ &= \mu_i c_{i\,\zeta_i} + \widetilde{c}_i - c_{i\,\zeta_i}. \end{aligned}$$

On the other hand, consider the function $f : \mathbb{R}^{m-1} \to \mathbb{R}$ defined by

$$f(x_1,\ldots,x_{m-1}) = x_1^2 + \cdots + x_{m-1}^2 + (\mu_i - x_1 - \cdots - x_{m-1})^2.$$

Note that the set $x_j^* = \mu_i/m$ minimizes f. As a consequence,

$$\sum_{j=1}^{m}\pi_{ij}^2 = f(\pi_{i1},\ldots,\pi_{i\ m-1}) \geq \sum_{j=1}^{m}\left(\frac{\mu_i}{m}\right)^2 = \frac{\mu_i^2}{m}.$$

Combining these results, we have

$$C(\pi;\theta) \geq \sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij} + \sum_{i=1}^{n}\mu_i c_{i\,\zeta_i} + \widetilde{c}_i - c_{i\,\zeta_i} + a_{i\,\zeta_i}\left(\frac{\mu_i^2}{L}\right) > C(\pi^*;\theta).$$ ■

Although the assumptions required to prove Theorem 3.5 may appear strong, they align with the following situation: There is an equal number of groups on each side, as in the marriage market, membership allocations, specialized schools, and centralized assignment mechanisms. Assumption 2 then states that each group has a clear affinity with another, with no overlaps. This condition is more restrictive than what typically occurs in the marriage market or in general settings, but it applies to the examples we will discuss in the Peruvian context. This framework holds when preferences are aligned (Echenique et al., 2024). Finally, Assumption 3 is the strongest and most specific, yet it is necessary to establish the result. The intuition is that, for transportation costs not to disrupt the matching equilibrium, the given relationship must hold, ensuring that the cost $c_{i\zeta_i}$ remains sufficiently low.

Example 3.6. In this example, we illustrate numerically Theorem 3.5. Results were computed using Mathematica 14.1, `LinearOptimization`. Consider $n = m = 4$, $\mu_i = \nu_j = 20$,

$$d = \begin{bmatrix} 88 & 88 & 100 & 91 \\ 19 & 42 & 37 & 69 \\ 81 & 87 & 9 & 50 \\ 66 & 18 & 77 & 91 \end{bmatrix}, \quad c = \begin{bmatrix} 989 & 24 & 975 & 941 \\ 673 & 612 & 684 & 9 \\ 20 & 352 & 387 & 380 \\ 675 & 687 & 44 & 697 \end{bmatrix}, \quad a = \begin{bmatrix} 9 & 3 & 8 & 9 \\ 6 & 8 & 3 & 2 \\ 1 & 7 & 8 & 3 \\ 9 & 5 & 2 & 6 \end{bmatrix}.$$

The optimal matching, obtained using `QuadraticOptimization`, is

$$\pi^* = \begin{bmatrix} 0 & 20 & 0 & 0 \\ 0 & 0 & 0 & 20 \\ 20 & 0 & 0 & 0 \\ 0 & 0 & 20 & 0 \end{bmatrix},$$

Hence, the result is in accordance with Theorem 3.5.

Examples 3.3 and 3.4 demonstrate that the solution to $\mathcal{P}_1$ can be either interior or a corner solution, unlike the classical linear model. However, under the assumptions of Theorem 3.5, the solution is always a corner solution, as illustrated in Example 3.6.

The discussion regarding our model optimizing over the Euclidean space rather than the lattice $\mathbb{Z}_+^{nm}$ parallels the classical optimization models in microeconomics, where goods are assumed to be infinitely divisible. However, given the structure of the objective function—comprising a sum of convex functions and a strictly convex quadratic term—we can leverage results from the literature developed in (Hochbaum and Shanthikumar, 1990). In particular, the solution in the lattice is sufficiently close to the solution in $\mathbb{R}_+^{nm}$, depending on the coefficients of the matrix $[a_{ij}]$:

$$\|\pi_{\mathbb{Z}} - \pi_{\mathbb{R}}\| \leq C(\Theta) f(\{\lambda_\imath\}_\imath),$$

where $\lambda_\imath$ are the eigenvalues of the Hessian of the objective function, Θ represents the model parameters, and $C(\Theta)$ is a constant that depends on the parameters. For the theory of integer programming and computational issues regarding it, which yields another full and extensive analysis, see for instance Park and Boyd (2017); Hladík et al. (2019); Pia (2024).

3.3 Analysis for $n = m = 2$

Having studied the specific cases where the solution is either a corner or interior solution, we now turn to the general case for $n = m = 2$, disregarding any assumption. The following calculations were obtained using Mathematica 14.1. By solving (10) and (11), we identified four parametric solution families that require $\mu_1 + \mu_2 = \nu_1 + \nu_2$. Three of these families are discarded because they correspond to degenerate cases: the first case holds when $a_{12} + a_{22} = 0$, the second case holds when $a_{11} + a_{12} + a_{21} + a_{22} = 0$ and $\mu_2 = (2a_{12}(\nu_1 + \nu_2) + 2\nu_1(a_{21} + a_{22}) - c_{11} + c_{12} + c_{21} - c_{22})/(2a_{12} + 2a_{22})$ and the third case holds when $a_{12} + a_{22} = 0$, $a_{11} + a_{21} = 0$ and $\nu_1 = (2\nu_2 a_{22} + c_{11} - c_{12} - c_{21} + c_{22})/(2a_{21})$. These unfeasible conditions leave us with one valid solution family, given by $\xi_2^* = \xi_1^* + (2(a_{11}a_{12} + a_{12}a_{21} + a_{11}a_{22} + a_{21}a_{22})\mu_2 - 2(a_{11}a_{12} + a_{11}a_{22})\nu_1 -$

$2(a_{11}a_{12} + a_{12}a_{21})\nu_2 + (a_{12} + a_{22})(c_{21} - c_{11}) + (a_{11} + a_{21})(c_{22} - c_{12}))/(a_{11} + a_{12} + a_{21} + a_{22})$, $\lambda_1^* = (-\xi_1^* a_{21} - \xi_2^*(a_{12} + a_{21} + a_{22}) + 2(a_{12}a_{21} + a_{21}a_{22})\mu_2 - 2a_{12}a_{21}\nu_2 + a_{22}c_{21} + a_{21}c_{22} - a_{21}c_{12} - a_{12}c_{21})/(a_{12} + a_{22})$ and $\lambda_2^* = (-\xi_1^* a_{22} - \xi_2^* a_{12} - 2a_{12}a_{22}\nu_2 - a_{22}c_{12} - a_{12}c_{22})/(a_{12} + a_{22})$ where ξ_1^* is free. By plugging these equalities into (9), we obtain the optimal matching when all the resulting expressions are strictly greater than zero. A detailed analysis to guarantee that $\pi_{ij}^* > 0$ was performed by reducing inequalities programmatically, but the numerous inequalities generated are omitted here. This analysis establishes a well-defined parameter space where the solution remains interior.

Given the specific cases analyzed above, it becomes evident that there is little hope of determining analytically whether solutions are interior or corner as n and m increase beyond 2. While the examples for $n = m = 2$ allowed us to identify some conditions under which solutions are either interior or corner, as the dimension of the problem grows, these conditions become increasingly complex and indeterminate.

The case $n = m$ becomes particularly relevant when considering the healthcare sector, where certain hospital networks are designated for specific types of diseases or patients. We explore this in detail in Section 4.

Although solving $\mathcal{P}_1$ analytically in a systematic way is a rather complex challenge, one can perform numerical quadratic convex optimization to approach the solution due tot he structure of the objective function.

4 Applications

The formulation in problem $\mathcal{P}_1$ is particularly relevant in contexts where congestion costs significantly affect the allocation of resources. Unlike models with linear costs, the quadratic cost structure accounts for congestion effects indirectly by making overburdened facilities increasingly costly. This feature is crucial in understanding inefficiencies in the Peruvian healthcare and education sectors, where access is heavily determined by proximity to schools and bureaucratic efficiency in medical centers.

4.1 Healthcare: The Impact of Bureaucratic and Geographic Congestion

Congestion severely affects healthcare access in Peru, manifesting in both physical and systemic dimensions. Lima's extreme traffic congestion, ranked among the worst globally, significantly delays patient travel times, limiting access to hospitals with available capacity. The World Bank estimates that traffic congestion alone costs Peru 1.8% of its GDP annually, a pattern observed in other highly congested cities such as Mumbai, São Paulo, and Jakarta (Kikuchi and Hayashi, 2020).

Beyond geographic constraints and traffic, systemic congestion due to resource limitations and administrative inefficiencies further deteriorates healthcare delivery. Overburdened medical personnel face extreme patient inflows, contributing to burnout and operational slowdowns. With only 4 doctors per 10,000 inhabitants—far below the WHO-recommended threshold of 43—Peru's medical workforce is severely overstretched (Infobae Médicos, 2024). Hospital capacity is equally insufficient, with only 1.6 beds per 1,000 people, significantly lagging behind regional standards (Banco Mundial, 2023). Inefficient patient referral processes, bureaucratic hurdles, and insurance-based care restrictions further aggravate congestion, increasing waiting times and deferral rates (Huerta-Rosario et al., 2019; EsSalud, 2025a,b).

This congestion can be effectively captured by a quadratic formulation in our model, specifically through the term $\sum_{i,j} a_{ij}\pi_{ij}^2$, which accounts for the saturation effects when too many individuals seek care at the same facility. As patient demand grows non-linearly within a given hospital or medical subsystem, service rates deteriorate, amplifying delays. This formulation reflects not only physical crowding but also bureaucratic congestion, where administrative overload further reduces system efficiency. These factors, combined with Lima's severe traffic congestion, create a feedback loop of systemic inefficiency, reinforcing barriers to timely and effective healthcare access.

At all times, we adopt the perspective of a central planner who has individuals, their preferences, cost information, and seeks the optimal assignment. We are not asserting or assuming that, in the current reality, the market adjusts to our model; rather, this is a normative economic approach rather than a positive one.

Example 4.1. In this example, we aim to represent the healthcare sector scenario, where three groups of patients are theoretically assigned to a specific type of medical center: SIS (Sistema Integral de Salud), EsSalud, or EPS (Entidades Prestadoras de Salud). The first group consists of poor and informal individuals, the second group comprises formal workers with severe diseases, and the third group consists of formal workers with standard diseases. We do not further cluster by economic sector to keep the example simple. Additionally, we exclude wealthy informal individuals (potential criminals) or millionaires with complex diseases.

The coefficients of the matrix c reflect preferences based on costs unrelated to congestion, such as bureaucratic barriers, compatibility, etc. The choice of parameters is consistent with this approach, assigning a cost of 1 for the preferred medical center and 10 for the other two. Group $i = 1$ corresponds to informal individuals, $j = 1$ to SIS, $i = 2$ corresponds to formal workers with complex diseases, $j = 2$ to EsSalud, and finally, $i = 3$ corresponds to formal workers with standard diseases, with $j = 3$ representing EPS. In particular, the parameters used, reflecting this situations, are:

$$a = \begin{bmatrix} 2 & 1 & 2 \\ 1 & 2 & 2 \\ 2 & 1 & 2 \end{bmatrix}, \; c = \begin{bmatrix} 1 & 10 & 10 \\ 10 & 1 & 10 \\ 10 & 10 & 1 \end{bmatrix}, \; d \in \mathcal{M}_{2\times 2} \text{ and } \mu = \begin{bmatrix} 20 \\ 20 \\ 20 \end{bmatrix} = \nu.$$

The matrix a has been chosen to introduce more friction due to congestion in the optimal linear match. Then, the optimal solution π^* under this parameter configuration is:

$$\pi^* = \begin{bmatrix} 6.8074 & 7.1959 & 5.9966 \\ 8.6351 & 5.6081 & 5.7567 \\ 4.5574 & 7.1959 & 8.2466 \end{bmatrix}.$$

This solution highlights the deviations from a strict one-to-one patient allocation, as the quadratic cost terms allow for cross-assignments that would not occur in a purely linear model. For comparison, when $a = 0$, meaning there are no quadratic costs, the optimal assignment is:

$$\pi^* = \begin{bmatrix} 20 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 20 \end{bmatrix}.$$

Here, patients are strictly assigned to their designated[4] medical system, as expected in the absence of congestion effects, but in contrast with the Peruvian reality where mismatching occurs, Anaya-Montes and Gravelle (2024).

Example 4.2. In this example, we analyze a scenario where the linear costs are such that all groups i would prefer to match with $j = 3$. However, due to congestion, only those in $i = 3$ actually are matched. Think of an exclusive medical center that is far from rural areas or poor districts. The parameters are as follows:

$$a = \begin{bmatrix} 1 & 1 & 20 \\ 1 & 1 & 20 \\ 1 & 1 & 1 \end{bmatrix}, \quad c = \begin{bmatrix} 1 & 1 & 5 \\ 1 & 1 & 5 \\ 1 & 1 & 5 \end{bmatrix}, \quad d = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 10 \\ 10 \\ 10 \end{bmatrix}, \quad \nu = \begin{bmatrix} 20 \\ 20 \\ 20 \end{bmatrix}.$$

The optimal solution π^* under these conditions is:

$$\begin{bmatrix} 4.680851063829787 & 4.680851063829787 & 0.6382978723404257 \\ 4.680851063829787 & 4.680851063829787 & 0.6382978723404258 \\ 0.6382978723404259 & 0.6382978723404259 & 8.72340425531915 \end{bmatrix}.$$

This result highlights the impact of congestion costs. Even though the *fair allocation* would be to match aa third of each group with $j = 3$, $\pi^*_{33} > 10 \max\{\pi^*_{13}, \pi^*_{23}\}$.

Example 4.3. In this example we compare the standard quadratic regularization model with our proposed heterogeneous congestion cost model. Both cases share the same linear costs c_{ij} and distance factors d_{ij}, as well as the same supply and demand constraints:

$$c = \begin{bmatrix} 1.0 & 5.0 & 5.0 \\ 5.0 & 1.0 & 5.0 \\ 5.0 & 5.0 & 1.0 \end{bmatrix}, \quad d = \begin{bmatrix} 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 \end{bmatrix}$$

[4] The ideal allocation in the absence of congestion is based entirely on the costs given by c. These costs correspond to preferences, characteristics related to the patients' illness, characteristics of the medical center, etc.

$$\mu = \begin{bmatrix} 20.0 \\ 20.0 \\ 20.0 \end{bmatrix}, \quad \nu = \begin{bmatrix} 20.0 \\ 20.0 \\ 20.0 \end{bmatrix}.$$

In the standard quadratic regularization model, a_{ij} is uniform:

$$a = \begin{bmatrix} 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 \end{bmatrix}$$

yielding the optimal allocation:

$$\pi^* = \begin{bmatrix} 8.0 & 6.0 & 6.0 \\ 6.0 & 8.0 & 6.0 \\ 6.0 & 6.0 & 8.0 \end{bmatrix}.$$

In contrast, our model introduces heterogeneity in congestion costs:

$$a = \begin{bmatrix} 2.0 & 1.0 & 1.0 \\ 1.0 & 2.0 & 1.0 \\ 1.0 & 1.0 & 2.0 \end{bmatrix}$$

leading to a different optimal allocation:

$$\pi^* = \begin{bmatrix} 4.8 & 7.6 & 7.6 \\ 7.6 & 4.8 & 7.6 \\ 7.6 & 7.6 & 4.8 \end{bmatrix}.$$

Unlike the quadratic regularization model, this formulation better captures congestion differences, reducing allocations where costs are higher and redistributing demand accordingly. This results in a more realistic representation of congestion-driven inefficiencies.

It is worth mentioning that the model we have introduced is highly flexible, allowing us to analyze additional cases. For instance, instead of considering the matching between three groups of patients and the three main healthcare networks in Peru, we could group patients by type of illness and medical centers by their specialization. The existence of delays and long queues reveals frictions in the matching process, further supporting the applicability of our model.

4.2 Education: Congestion Costs and School Choice Constraints

The Peruvian education system is highly complex and decentralized, unlike centralized models in countries such as China, South Korea, and France. This decentralization has resulted in significant heterogeneity in educational quality, particularly between urban and rural areas. Unlike France, where an efficient transport network helps mitigate congestion-related issues in school assignments (Eurydice - European Commission, 2024), Peru's fragmented structure and complicates geography exacerbates disparities in access to education, infrastructure, and resources.

Despite this decentralization, our model remains relevant for understanding key educational dynamics and offers valuable insights if parts of the system, or even specific subsystems such as the High-Performance Schools (COAR), become more centralized. Indeed, as highlighted by Alba-Vivar (2025) in line with Agarwal and Somaini (2019), transportation in Lima plays a crucial role in educational access. A 17% reduction in travel time (equivalent to 30 minutes per day) increased enrollment rates by 6.3%, underscoring the importance of mobility constraints in shaping educational outcomes.

Moreover, Peru is characterized by severe congestion along major thoroughfares (World Bank, 2024; IFSA-Butler, 2024). As more individuals travel along the same routes (as Javier Prado Oeste), congestion intensifies, making it essential to incorporate congestion costs into the model. This effect cannot be captured by a linear structure, particularly when individuals are clustered by geographic location.

Additionally, stronger geographic constraints, such as those in the Andes and the Amazon, create highly congested access routes, including narrow bridges over rivers and limited transportation corridors. These natural barriers further justify the introduction of a quadratic term to account for congestion effects.

The following examples illustrate the impact of congestion costs in the proposed model.

Example 4.4. This example illustrates how introducing heterogeneous quadratic costs $a_{ij}\pi_{ij}^2$ distorts student allocation compared to a purely linear preference-based model. In many developed countries, such as France or Switzerland, well-developed metro systems allow students to access top schools regardless of distance. However, in Peru, inadequate public transportation significantly affects school choice, leading to inefficient assignments. We consider three groups of students and three types of schools, where c_{ij} represents student preferences, including perceived school quality and distance constraints. Without congestion costs, students would be perfectly sorted into their most preferred schools. The parameters are as follows:

$$a = \begin{bmatrix} 4.0 & 2.0 & 3.0 \\ 4.0 & 2.0 & 6.0 \\ 3.0 & 4.0 & 3.0 \end{bmatrix}, \quad c = \begin{bmatrix} 1.0 & 5.0 & 100.0 \\ 10.0 & 1.0 & 50.0 \\ 100.0 & 50.0 & 1.0 \end{bmatrix}, \quad d = \begin{bmatrix} 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 40.0 \\ 40.0 \\ 40.0 \end{bmatrix}, \quad \nu = \begin{bmatrix} 40.0 \\ 40.0 \\ 40.0 \end{bmatrix}.$$

When congestion costs are included, the optimal assignment is:

$$\pi^* = \begin{bmatrix} 16.40 & 17.07 & 6.53 \\ 15.07 & 17.65 & 7.29 \\ 8.53 & 5.28 & 26.19 \end{bmatrix}. \tag{13}$$

Here, students are not necessarily assigned to their most preferred schools due to congestion effects. Those who would ideally attend top schools are redirected to lower-ranked institutions, as excessive demand increases quadratic congestion costs. For comparison, when congestion costs are removed ($a = 0$), the optimal assignment is:

$$\pi^* = \begin{bmatrix} 40.0 & 0.0 & 0.0 \\ 0.0 & 40.0 & 0.0 \\ 0.0 & 0.0 & 40.0 \end{bmatrix}.$$

This result aligns perfectly with the preference-based structure of c_{ij}, as all students are assigned to their most desired schools without deviation. This example highlights how transportation inefficiencies and congestion distort the school choice process. Unlike countries with high-quality metro systems, where students can attend their ideal schools regardless of distance, in Peru, traffic congestion and poor infrastructure create a situation where even high-achieving students may not access top-tier institutions. Our model captures these effects by incorporating heterogeneous quadratic costs, providing a more realistic representation of school allocation dynamics in constrained environments.

In the Peruvian context, suppose that group $i = 1$ consists of top students, with the performance decreases towards $i = 3$. On the other hand, school $j = 1$ has the top teachers, and so on. From this perspective, the optimal assignment would be to match i with $j = i$. However, when congestion is introduced, top students may live in areas with difficult access or areas affected by a major avenue that gets heavily congested (e.g., even in La Molina, students may need to pass through Javier Prado Oeste to reach Avenida Universitaria). As a result, despite being a better fit for the best university (in terms of potential research, etc.), they end up attending a closer institution where there is less research activity.

5 Conclusions

In this paper, we developed an optimal transport model with heterogeneous quadratic regularization to account for congestion effects in matching problems. Unlike classical models that assume linear transportation costs or entropy regularization, our formulation introduces increasing marginal costs, providing greater flexibility for central planners aiming to clear excess demand effectively. By incorporating congestion costs explicitly, our model offers a more realistic representation of allocation inefficiencies caused by overcrowding.

From a theoretical perspective, we demonstrated that the optimization problem retains a convex structure and that the uniqueness of the optimal assignment is guaranteed. However, analytically characterizing the solutions remains challenging, as the system of equations derived from the KKT conditions is singular. For the particular case where the number of agent types and entities matches ($m = n$), we provided conditions under which the model yields corner solutions in the integer setting, meaning that each agent type is assigned to a single entity.

For the case $n = m$, under additional mild assumptions on the parameters, we introduce a novel result useful for integer programming applications. In particular, we highlight that imposing constraints of the form $\sum_i \pi_{ij} \leq \nu_j, \quad \sum_j \pi_{ij} \leq \mu_i$ leads to a trivial null solution, whereas using structured bounds of the form $\mu_i^L \leq \sum_j \pi_{ij} \leq \mu_i^H, \quad \nu_j^L \leq \sum_i \pi_{ij} \leq \nu_j^H$, results in the same mathematical structure as merely imposing μ_i^L and ν_j^L in the linear constraints. This suggests that penalization approaches are analytically superior to constraints in this context for analyzing excess of demand.

In terms of applications, our model is particularly useful for central planners seeking optimal allocations while accounting for frictions. In education, it captures congestion effects arising when excessive numbers of students are assigned to specific institutions, leading to infrastructure constraints and quality deterioration. In healthcare, our formulation applies to the distribution of patients across hospitals in segmented healthcare systems, such as the Peruvian case with SIS, EsSalud, and EPS, where excessive demand in certain hospitals results in long waiting times and service inefficiencies. Additionally, the model can be extended to labor markets where firms face increasing costs when hiring additional workers with similar profiles, a phenomenon observed in industries with capacity constraints.

Although we have not estimated the parameters, our examples provide a first insight into the advantages of our model. Moreover, Theorem 3.5 allows us to identify situations where the optimal matching can be computed without resorting to integer convex quadratic optimization.

Future extensions of this work aim to enhance model flexibility through four key directions:

1. **Dynamic Extensions**: Integrating *Markov Jump Linear Systems* to model time-dependent congestion dynamics, (do Valle Costa et al., 2005).

2. **Penalty-Based Formulations**: Replacing KKT-type constraints with penalization terms, as explored in Gallardo et al. (2025), improving analytical tractability.

3. **Infinite Agent Types**: Generalizing the model to continuous distributions of agent characteristics (Wang and Zhang, 2025).

4. **Stochastic Matching**: Introducing randomness in assignment costs to account for uncertainty.

These extensions will allow for a more robust framework adaptable to complex, real-world allocation problems. Moreover, advanced computational techniques, such as mixed-integer quadratic programming and nonlinear constrained optimization methods, could be employed to analyze high-dimensional and intricate cases.

A Continuous setting

In the classical optimal transport model, we consider two sets, $X \subset \mathbb{R}^{N_X}$ and $Y \subset \mathbb{R}^{N_Y}$, representing distinct populations, such as women and men, workers and firms, students and schools, or patients and doctors in hospitals. From the perspective of a central planner, the objective is to minimize the cost of matching these populations. This cost depends on the characteristics of the elements $x \in X$ and $y \in Y$ and is assumed to be linear with respect to the transported mass. The masses of X and Y are described by two finite measures, μ and ν, satisfying:

$$\mu(X) = \nu(Y) < \infty.$$

The planner seeks to ensure that all mass is matched optimally. Thus, the classical optimal transport problem is formulated as:

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c(x,y)\, d\pi(x,y),$$

where[5]

$$\Pi(\mu,\nu) = \left\{ \pi \geq 0 \mid \int_Y \pi(x,y)\, dy = \frac{d\mu}{dx}, \quad \int_X \pi(x,y)\, dx = \frac{d\nu}{dy} \right\}.$$

The measure π over $X \times Y$ represents the transport plan and is thus interpreted as a matching measure (Galichon, 2021). Consequently, the optimization problem is defined over distributions. In the main body of this work, we assumed that both X and Y are finite sets:

$$X = \{x_1, \dots, x_n\}, \quad Y = \{y_1, \dots, y_m\}.$$

Under this assumption, the measures take the discrete form:

$$\mu = \sum_{i=1}^{n} \mu_i \delta_{x_i}, \quad \nu = \sum_{j=1}^{m} \nu_j \delta_{y_j},$$

where $\delta_a(B) = 1$ if $a \in B$ and 0 otherwise (Dirac's delta measure). However, our model extends to non-discrete and infinite spaces as follows:

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c(x,y)\, d\pi(x,y) + \left(\int_{X \times Y} w(x,y) \pi(x,y)^2\, dx dy \right)^2,$$

where $w(x,y) \in L^2(X \times Y) \cap C(X \times Y)$ introduces the heterogeneity.

[5] Here, $d\mu/dx$ and $d\nu/dy$ denote the Radon-Nikodym derivatives with respect to the Lebesgue measure.

References

Abdulkadiroğlu, A. and Sönmez, T. (2003). School Choice: A Mechanism Design Approach. *The American Economic Review*, 93(3):729–747.

Agarwal, N. and Somaini, P. (2019). Revealed preference analysis of school choice models. *NBER Working Paper*, (w26505).

Alba-Vivar, F. M. (2025). Opportunity bound: Transport and access to college in a megacity. *Job Market Paper*. Department of Economics - Wake Forest University.

Ambrosio, L., Brué, E., and Semola, D. (2021). *Lectures on Optimal Transport*, volume 130 of *Unitext*. Springer.

Anaya-Montes, M. and Gravelle, H. (2024). Health Insurance System Fragmentation and COVID-19 Mortality: Evidence from Peru. *PLOS ONE*, 19(8):e0309531.

Banco Mundial (2023). Camas hospitalarias (por cada 1.000 personas) - perú. Accessed on February 21, 2025.

Boyd, S. (2004). *Convex Optimization*. Cambridge University Press.

Carlier, G., Dupuy, A., Galichon, A., and Sun, Y. (2020). SISTA: Learning Optimal Transport Costs under Sparsity Constraints. *arXiv preprint arXiv:2009.08564*. Submitted on 18 Sep 2020, last revised 21 Oct 2020.

de la Fuente, A. (2000). *Mathematical Methods and Models for Economists*. Cambridge University Press. Digital publication date: 04 June 2012.

do Valle Costa, O. L., Marques, R. P., and Fragoso, M. D. (2005). *Discrete-Time Markov Jump Linear Systems*. Probability and Its Applications. Springer.

Dupuy, A. and Galichon, A. (2014). Personality Traits and the Marriage Market. *Journal of Political Economy*, 122(6):1271–1319.

Dupuy, A., Galichon, A., and Sun, Y. (2019). Estimating Matching Affinity Matrices under Low-Rank Constraints. *Information and Inference: A Journal of the IMA*, 8(4):677–689.

Echenique, F., Immorlica, N., and Vazirani, V. V. (2023). *Online and Matching-Based Market Design*. Cambridge University Press.

Echenique, F., Root, J., and Sandomirskiy, F. (2024). Stable Matching as Transportation. Preprint submitted to arXiv on 12 Feb 2024.

Echenique, F. and Yenmez, M. B. (2015). How to Control Controlled School Choice. *The American Economic Review*, 105(8):2679–2694.

Ekeland, I. (2010). Notes on Optimal Transportation. *Economic Theory*, 42(2):437–459.

EsSalud (2025a). Dashboard de indicadores fonafe y tablero estratégico. https://app.powerbi.com/view?r=eyJrIjoiMDQwMDVlOGItNGY5Zi00ZjFjLWEyZDMtYjY1Zjk0MWVjMjcxIiwidCI6IjM0ZjMyNDE5LTFjMDUtNDc1Ni (accessed 18 March 2025).

EsSalud (2025b). Tablero de diferimento de citas. https://app.powerbi.com/view?r=eyJrIjoiN2NlMTNmNWEtODA3MS00M2UyLWE3NDAtNjcyYjZjYTQ0MmJmIiwidCI6IjM0ZjMyNDE5LTFjMDUtNDc1Ni (accessed 18 March 2025).

Eurydice - European Commission (2024). France - national education system overview. Accessed on February 21, 2025.

Gale, D. and Shapley, L. S. (1962). College Admissions and the Stability of Marriage. *The American Mathematical Monthly*, 69(1):9–15.

Galichon, A. (2016). *Optimal Transport Methods in Economics*. Princeton University Press.

Galichon, A. (2021). The Unreasonable Effectiveness of Optimal Transport in Economics. Preprint submitted on 12 Jan 2023.

Gallardo, M., Loaiza, M., and Chavez, J. (2025). Congestion and penalization in optimal transport. Submitted to Mathematical Social Sciences.

Gentle, J. E. (2017). *Matrix Algebra: Theory, Computations, and Applications in Statistics*. Springer, Cham, Switzerland, 2nd edition.

González-Sanz, A. and Nutz, M. (2024). Sparsity of quadratically regularized optimal transport: Scalar case. *arXiv preprint arXiv:2410.03353*.

Hatfield, J. W. and Milgrom, P. R. (2005). Matching with Contracts. *The American Economic Review*, 95(4):913–935.

Hladík, M., Černý, M., and Rada, M. (2019). A new polynomially solvable class of quadratic optimization problems with box constraints. *arXiv preprint*, arXiv:1911.10877.

Hochbaum, D. S. and Shanthikumar, J. G. (1990). Convex separable optimization is not much harder than linear optimization. *Journal of the ACM*, 37(4):843–862.

Huerta-Rosario, A., Huerta-Rosario, J. A., and Huerta-Rosario, J. J. (2019). Barriers to effective healthcare access in peru: An analysis of patient referrals. *Revista Peruana de Medicina Experimental y Salud Pública*, 36(2):304–311. Accessed on February 21, 2025.

Hylland, A. and Zeckhauser, R. (1979). The Efficient Allocation of Individuals to Positions. *The Journal of Political Economy*, 87(2):293–314.

IFSA-Butler (2024). Navigating Public Transportation in Peru. Accessed on February 21, 2025.

Infobae Médicos (2024). Solo hay 4 médicos por cada 10 mil habitantes en Perú: ¿cuántos son necesarios para atender a toda la población? Accessed on February 21, 2025.

Kelso, A. S. and Crawford, V. P. (1982). Job Matching, Coalition Formation, and Gross Substitutes. *Econometrica*, 50(6):1483.

Kikuchi, T. and Hayashi, S. (2020). Traffic congestion in jakarta and the japanese experience of transit-oriented development. *S. Rajaratnam School of International Studies.*

Levin, O. (2015). *Discrete Mathematics: An Open Introduction.* Taylor & Francis, fourth edition.

Lorenz, D. A., Manns, P., and Meyer, C. (2019a). Quadratically regularized optimal transport. *arXiv preprint arXiv:1903.01112.*

Lorenz, D. A., Manns, P., and Meyer, C. (2019b). Quadratically regularized optimal transport. *Applied Mathematics & Optimization.*

Martinez, M. J. (2024). Critical evaluation of transit policies in lima, peru; resilience of rail rapid transit (metro) in a developing country. *Green Energy and Intelligent Transportation Systems*, 100:100172.

Merigot, Q. and Thibert, B. (2020). Optimal Transport: Discretization and Algorithms. Preprint submitted on 2 Mar 2020.

Milgrom, P. and Shannon, C. (1994). Monotone comparative statics. *Econometrica*, 62(1):157–180.

Médicos Sin Fronteras (2021). Perú: Oficialmente el país del mundo más afectado por la covid-19. Accessed: 2025-03-13.

Nenna, L. (2020). Lecture 4 entropic optimal transport and numerics.

Nutz, M. (2024). Quadratically regularized optimal transport: Existence and multiplicity of potentials. Preprint submitted to arXiv on 10 Feb 2024.

Organisation for Economic Co-operation and Development (OECD) (2024). Pisa 2022 results (volume iv) - country notes: Peru. Technical report, OECD Publishing. Accessed: 2025-03-13.

Park, J. and Boyd, S. (2017). A semidefinite programming method for integer convex quadratic minimization. *Optimization Letters.*

Peyré, G. and Cuturi, M. (2019). Computational Optimal Transport: With Applications to Data Science. Preprint submitted on 4 June 2019.

Pia, A. D. (2024). Convex quadratic sets and the complexity of mixed integer convex quadratic programming. *arXiv preprint*, arXiv:2311.00099.

Roth, A. E. (1982). The Economics of Matching: Stability and Incentives. *Mathematics of Operations Research*, 7(4):617–628.

Roth, A. E. and Sotomayor, M. A. O. (1990). *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, volume 18 of *Econometric Society Monographs.* Cambridge University Press.

Tardella, F. (2010). The fundamental theorem of linear programming: extensions and applications. *Optimization*, 59(3):283–301.

Villani, C. (2009). *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer.

Wang, R. and Zhang, Z. (2025). Quadratic-form optimal transport. *arXiv preprint*, 2501.04658. 42 pages, 5 figures.

Wiesel, J. and Xu, X. (2024). Sparsity of quadratically regularized optimal transport: Bounds on concentration and bias. *arXiv preprint arXiv:2410.03425*.

World Bank (2024). Modernizing traffic management in lima with world bank support. Accessed on February 21, 2025.