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 19, 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 (UC Berkeley). This work is based on the preprint Congestion and Penalization in Optimal Transport; it shares the same motivations and context. However, in this work we focus on the quadratic model, while the last version of Congestion and Penalization is fully oriented towards the penalization and weighting model.

†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, demonstrating 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 Echenique et al. (2023), which provides a comprehensive overview of these subjects.

In recent years, matching theory has been enriched 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. A recent study, Echenique et al. (2024), explores stability in matching through an OT framework applicable to both discrete and continuous settings. They show that utility transformations parameterized by α yield ε-stable, welfare-maximizing, or ε-egalitarian solutions, depending on whether the transformation is convex or concave. This approach aligns with Niederle and Yariv (2009) and Ferdowsian et al. (2023).

In a 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 define the notation and present existing models from the literature. Next, in Section 3, we introduce our model. We then derive some new results and conclude with examples and extensions in Section 4. These examples account for the economic insights brought by our model, particulary explaining inefficiencies in the educational and health matching markets.

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 (MSF), 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). 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 the classical optimal transport model, there exist two sets, $X \subset \mathbb{R}^{N_X}$ and $Y \subset \mathbb{R}^{N_Y}$, representing two 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 linear with respect to the mass transported. It is assumed that the masses of X and Y, quantified by two measures μ and ν, are finite and so that: $\mu(X) = \nu(Y) < \infty$. The central planner seeks to ensure that all mass is appropriately matched. Consequently, the problem is formulated as follows:

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

where[1]

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

The distribution π is a measure over $X \times Y$, and is therefore interpreted as a matching Galichon (2021).

For the remainder of our work, we focus exclusively on the case where both X and Y are finite sets, that is, $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$. Consequently, 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. The parameters μ_i represent the number of individuals at location x_i, while ν_j similarly denotes the number of individuals at y_j. For instance, if y_j represents a hospital or a school, then ν_j corresponds to the institution's capacity. In this particular case, the optimization problem solved by the central planner is given by:

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

[1] Note that $d\mu/dx$ and $d\nu/dy$ are Radon-Nykodim derivatives w.r.t. Lebesgue measure.

where

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

and π_{ij} represents the number of individuals transported from i to j. Note that constraints (2) ensure that all individuals (students, patients, etc.) are assigned, and that all entities (schools, hospitals, etc.) fill their available capacity. Moreover, costs are linear and separable. This implies that the marginal cost of assigning an additional individual from i to j is constant, regardless of the existing assignments.

A solution to (1) is referred to as an optimal matching or optimal (transport) plan, denoted by π^*. To solve $\mathcal{P}_O$, linear programming techniques such as the simplex method are typically used.

Inspired by the entropic an quadratic regularization approaches (Carlier et al., 2020; Peyré and Cuturi, 2019; Nutz, 2024), we introduce in Section 3 our model, which includes heterogeneity in the cost structure. Let us recall that the entropic regularization problem is given by

$$\min_{\pi\in\Pi(\mu,\nu)} \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}\pi_{ij} + \sigma\pi_{ij}\ln(\pi_{ij}),$$

with $\sigma > 0$, while the quadratic regularization (Wiesel and Xu, 2024; Nutz, 2024) changes the entropy term by $(\varepsilon/2)||\pi||_2^2$. The entropy in zero takes the value zero so the function is continue.

Before introducing our model, it is important to discuss the existence of solutions to the optimization problems previously introduced. The first key observation is that, given the economic context, solutions are expected to belong to $\mathbb{Z}_+^{nm}$. However, as formulated, the optimization problems above do not inherently enforce that the solution lies in $\mathbb{Z}_+^{nm}$.

On the other hand, in the more general framework where optimization is performed over distributions, it is crucial to impose certain structure on the spaces X and Y, as well as specific properties on the function c, in order to guarantee the existence of a solution.

Furthermore, if the problem is instead 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:

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

Ignoring the upper bounds ν_j, this reduces to a stars and bars problem. 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:

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

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. Nevertheless, the existence of a solution follows from Weierstrass' Theorem (Proposition 2.2).

Proposition 2.2. Given $\mu = (\mu_1, ..., \mu_n)^T \in \mathbb{R}_{++}^n$ and $\nu = (\nu_1, ..., \nu_m)^T \in \mathbb{R}_{++}^m$, $\mathcal{P}_O$ and its variants, 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 $[0, \sum_{i=1}^n \mu_i]^{nm}$. ■

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 model has been extensively studied, along with entropic (Dupuy and Galichon, 2014; Carlier et al., 2020) and quadratic (Lorenz et al., 2019; González-Sanz and Nutz, 2024; Wiesel and Xu, 2024; Nutz, 2024) regularization, as well as in the continuous framework (Dupuy and Galichon, 2014; Echenique et al., 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

We now introduce a new variant of the optimal transport problem in the discrete setting that explicitly accounts for congestion effects. 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.
- **Medical centers overload:** Large patient inflows can overwhelm hospital resources, creating long waiting times, administrative bottlenecks, and inefficient service delivery.
- **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, capturing the increasing marginal costs associated with congestion.

3.1 Mathematical Formulation

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:

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

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.

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).

3.2 Lagrangian Formulation and KKT Conditions

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

$$\begin{aligned}\mathscr{L}(\pi,\lambda,\xi,\gamma;\theta) = &\sum_{i=1}^{n}\sum_{j=1}^{m}\phi(\pi_{ij};\theta_{ij}) + \sum_{i=1}^{n}\xi_i\left(\mu_i - \sum_{j=1}^{m}\pi_{ij}\right) + \sum_{j=1}^{m}\lambda_j\left(\nu_j - \sum_{i=1}^{n}\pi_{ij}\right)\\ &- \sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}\pi_{ij}.\end{aligned} \tag{6}$$

The KKT first-order conditions are:

$$\begin{aligned}\frac{\partial\mathscr{L}(\pi^*,\xi^*,\lambda^*,\gamma^*;\theta)}{\partial\pi_{ij}} = \frac{\partial\phi(\pi^*_{ij};\theta_{ij})}{\partial\pi_{ij}} - \lambda^*_j - \xi^*_i - \gamma^*_{ij} = 0,\ \forall\ i = 1,\cdots,n,\ j = 1,\cdots,m\\ -\pi^*_{ij} \le 0,\ \forall\ i = 1,\cdots,n,\ j = 1,\cdots,m\\ \sum_{j=1}^{m}\pi^*_{ij} - \mu_i = 0,\ \forall\ i = 1,\cdots,n\\ \sum_{i=1}^{n}\pi^*_{ij} - \nu_j = 0,\ \forall\ j = 1,\cdots,m\\ \gamma^*_{ij}\pi^*_{ij} = 0,\ \forall\ i = 1,\cdots,n,\ j = 1,\cdots,m.\end{aligned}$$

Since the objective function is strictly convex, continuous and the constraint set is convex, there is a unique solution.

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

3.3 Quadratic Cost Function and Optimal Matching

To carry out a quantitative analysis, we assume a quadratic cost function:

$$\phi(\pi_{ij};\theta_{ij}) = d_{ij} + c_{ij}\pi_{ij} + a_{ij}\pi_{ij}^2. \tag{7}$$

Thus, the optimization problem becomes:

$$\mathcal{P}_1 : \min_{\pi\in\Pi(\mu,\nu)} \left\{ \sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij} + c_{ij}\pi_{ij} + a_{ij}\pi_{ij}^2 \right\}. \tag{8}$$

The parameters have clear economic interpretations:

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

The most recent problem addressed in the literature is (8), considering a_{ij} equal. This is because the context and motivation are different, so the quadratic term acts as a regularization and does not aim to model congestion costs.

Applying KKT first order conditions, we obtain:

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

3.4 Structural Properties of the Solution

A fundamental issue in this model is determining whether solutions are interior ($\pi_{ij}^* > 0$ for all (i,j)) or a corner solutions ($\gamma_{ij}^* > 0$ for some (i,j)). The following result characterizes a structural property of the solution:

Proposition 3.1. With respect to the problem (5), with costs given by (7), 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

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

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

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

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, $\mathrm{Det}(R) = 0$, and the claim follows. ∎

Proposition 3.1 is important as it reveals that even for interior solutions, we do not have a systematic method for analytical resolution.

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 (6), considering quadratic costs[3]

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

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

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

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

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

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), $\mathrm{Det}(J_{\pi,(\xi,\lambda)}\overline{\mathscr{L}}(\pi^*, \xi^*, \lambda^*, \overline{\theta})) = \mathrm{Det}(D)\mathrm{Det}(0 - BD^{-1}B^T) = 0$. ∎

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

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 $n = m$, 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.

3.5 Integer setting

An important issue in our model is to determine whether the solution will be a corner solution ($\gamma^*_{ij} > 0$ for some $(i,j) \in I \times J$) or not. The following examples show that under the quadratic setting, both interior and corner solutions can exist.

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

$$a = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}, \; c = \begin{bmatrix} 24 & 48 \\ 16 & 24 \end{bmatrix}, \; \mu = (20,20), \;\; \nu = (12,28), \text{ so that } n = m = 2.$$

Then, we have $\pi^* = (7,13,5,15)$.

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

$$a = \begin{bmatrix} 200 & 2 \\ 2 & 200 \end{bmatrix}, \quad c = \begin{bmatrix} 200 & 2 \\ 2 & 200 \end{bmatrix}, \quad \mu = (10,10), \quad \nu = (10,10), \quad \text{so that } n = m = 2.$$

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

Note that in Example 3.3, the solution no longer satisfies that $\pi^*_{ij} = 0$ for some $(i,j) \in I \times J$, as always happens in the linear case.

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 $1 \leq i,j \leq M$.

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

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

Assumption 2 imposes that each individual is optimally matched with their top choice school, ensuring a distinct best fit for each individual. Note that Assumptions 1 and 2 imply immediately that the solution to the linear model is:

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

In fact, for any other matching $\pi \in \Pi(\mu, \nu)$,

$$C(\pi, \theta) = \sum_{i=1}^{n} \sum_{j=1}^{m} d_{ij} + c_{ij}\pi_{ij} > \sum_{i=1}^{n} \sum_{j=1}^{m} d_{ij} + \sum_{i=1}^{n} c_{i\,\zeta_i} \sum_{j=1}^{m} \pi_{ij} = C(\pi^*, \theta).$$

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

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).

Theorem 3.5. *Under Assumptions 1, 2 and 3, the optimal matching for the quadratic model in the integer setting is* (12).

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

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

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,

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

Combining these results, we have

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

Example 3.6. The following examples were computed using Mathematica 14.1. For each case, we present the parameter matrices d, c, and a (where applicable), along with the optimal matching matrix π^*, obtained using the appropriate optimization method.

For the linear model, with $n = m = 4$ and $\mu_i = \nu_j = 50$, the optimal matching was computed using `LinearOptimization`:

$$d = \begin{bmatrix} 32 & 83 & 82 & 37 \\ 47 & 75 & 56 & 45 \\ 87 & 74 & 79 & 4 \\ 40 & 55 & 94 & 14 \end{bmatrix}, \quad c = \begin{bmatrix} 76 & 77 & 83 & 6 \\ 74 & 98 & 7 & 41 \\ 6 & 86 & 8 & 70 \\ 88 & 17 & 40 & 96 \end{bmatrix}, \quad \pi^* = \begin{bmatrix} 0 & 0 & 0 & 50 \\ 0 & 0 & 50 & 0 \\ 50 & 0 & 0 & 0 \\ 0 & 50 & 0 & 0 \end{bmatrix}.$$

For the quadratic model, with $n = m = 4$ and $\mu_i = \nu_j = 20$,

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

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 $\mathcal{P}_Q$ 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.

When preferences and transportation costs are such that the minimum cost associated with individual characteristics, rather than congestion, is sufficiently lower than the others, the solution will maintain the structure of the linear model (Assumption 3). On the other hand, Assumption 1 seems somewhat restrictive. Nevertheless, its applicability remains significant. As we will see, it fits well in circumstances that we will describe in the Peruvian context of health and education.

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 $[a_{ij}]$:

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

where $\lambda_\imath$ are the eigenvalues of the Hessian of the objective function, Θ represents the model parameters, and $C(p, \Theta)$ is a constant that depends on both the parameters and the chosen norm. 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).

3.6 Analysis for $n = m = 2$

Having explored 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, as we have already explored, solving $\mathcal{P}_Q$ analytically in a systematic way is a rather complex challenge. However, fortunately, given the structure of the objective function, solving $\mathcal{P}_Q$ can be accomplished through numerical quadratic convex optimization.

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

In the Peruvian healthcare system, individuals are theoretically assigned to facilities based on their insurance type—whether EsSalud (public insurance for formal workers), SIS (universal public insurance for low-income individuals), or private insurance (EPS), and, more specifically, they are assigned based on their illness. However, the system suffers from severe inefficiencies and inadequate infrastructure, with 76% of facilities lacking sufficient capacity to provide proper care (Defensoría Pueblo, 2020), which leads, among other factors, to a poor allocation.

These inefficiencies are further illustrated by alarming statistics (Velásquez, 2020). Peru recorded the highest per capita COVID-19 mortality rate globally, highlighting the system's fragility. The shortage of medical personnel remains critical, with only 4 doctors per 10,000 inhabitants, far below the World Health Organization recommended threshold of 43 (Infobae Médicos, 2024). The deficit extends to hospital infrastructure, with only 1.6 hospital beds per 1,000 people, well below regional averages (Banco Mundial, 2023). Inefficient patient referral systems and high deferral rates further increase waiting times and restrict access to essential services (Huerta-Rosario et al., 2019; EsSalud, 2025a,b).

Geographic constraints further reinforce these inefficiencies. Traffic congestion in Lima[4] and other major cities significantly increases travel costs, deterring patients from seeking care at facilities that may have better capacity but require longer commutes. This behavior is well captured in our model by the term $\sum_{i,j} a_{ij}\pi_{ij}^2$. It is important to note that congestion is not limited to geographic factors. Bureaucratic hurdles often prevent patients from receiving specialized care at appropriate institutions (EsSalud, 2025a,b). Thus, it is necessary to consider a quadratic structure to model the fact that if many individuals of the same type seek care at the same hospital, being processed by the same subsystem within the medical center leads to saturation and congestion. We illustrate this in the following examples.

[4]The World Bank estimates that traffic congestion alone costs Peru 1.8% of its GDP annually. Similar economic losses have been reported in major cities such as Mumbai, São Paulo, and Jakarta (Kikuchi and Hayashi, 2020) due to excessive congestion.

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, EsSalud, or EPS.

The coefficients of the matrix c reflect the preferences based on costs unrelated to congestion, i.e., bureaucratic barriers, compatibility, etc. The choice of parameters is consistent with this, with a cost of 1 for the preferred medical center and 10 for the other two. The congestion costs are given by a. In particular, the parameters used are:

$$a = \begin{bmatrix} 2.0 & 1.0 & 2.0 \\ 1.0 & 2.0 & 2.0 \\ 2.0 & 1.0 & 2.0 \end{bmatrix}, \ c = \begin{bmatrix} 1.0 & 10.0 & 10.0 \\ 10.0 & 1.0 & 10.0 \\ 10.0 & 10.0 & 1.0 \end{bmatrix}, \ d = \mathbf{1}_{3\times 3} \text{ and } \mu = \begin{bmatrix} 20.0 \\ 20.0 \\ 20.0 \end{bmatrix} = \nu.$$

The optimal solution π^* under these conditions 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:

$$\pi^* = \begin{bmatrix} 20.0 & 0 & 0 \\ 0.0 & 20.0 & 0.0 \\ 0.0 & 0 & 20.0 \end{bmatrix}.$$

Here, patients are strictly assigned to their designated[5] medical system, as expected in the absence of congestion effects, but in contrast with the Peruvian reality.

Example 4.2. In this example, we analyze a scenario where the linear costs c_{ij} are identical across all assignments, meaning there are no inherent preferences between different allocation routes. However, the quadratic congestion costs a_{ij} vary, influencing the final distribution of assignments. Despite the presence of optimal routes under linear costs, congestion effects lead to deviations in the allocation. The parameters are as follows:

$$a = \begin{bmatrix} 1.0 & 2.0 & 2.0 \\ 2.0 & 1.0 & 2.0 \\ 2.0 & 2.0 & 1.0 \end{bmatrix}, \quad c = \begin{bmatrix} 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 \end{bmatrix}, \quad d = \begin{bmatrix} 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 \end{bmatrix}$$

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

[5]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.

The optimal solution π^* under these conditions is:

$$\pi^* = \begin{bmatrix} 10.0 & 5.0 & 5.0 \\ 5.0 & 10.0 & 5.0 \\ 5.0 & 5.0 & 10.0 \end{bmatrix}.$$

This result highlights the impact of congestion costs. Even though all routes have the same linear cost, the quadratic term introduces distortions in the allocation, preventing a strict adherence to any single preferred matching pattern. Instead, the system distributes assignments to mitigate excessive congestion, leading to deviations from what would be optimal under purely linear costs.

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:

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

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

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 uniform 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 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.

Note also that classical quadratic regularization is introduced from a mathematical point of view. Introducing a_{ij}, i.e., the heterogeneous structure, provides significant flexibility to model different congestion costs across pairs $(i, j) \in I \times J$.

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:

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

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.

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, meaning that each agent type is assigned to a single entity. This result is particularly relevant in sectors where specialization and supply segmentation are crucial, such as education and healthcare.

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.

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.
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.
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.

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.

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.

Defensoría Pueblo (2020). Centros de salud de lima registran graves problemas de infraestructura y falta de personal médico. Accessed on February 21, 2025.

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=eyJrIjoiMDQwMDVlOGItNGY5ZiOOZjFjLWEyZDMtYjY1ZjkOMWVjMjcxIiwidCI6IjMOZjMyNDE5LTFjMDUtNDc1Ni` (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.

Ferdowsian, A., Niederle, M., and Yariv, L. (2023). Strategic Decentralized Matching: The Effects of Information Frictions. This version: November 13, 2023.

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.*

Lorenz, D. A., Manns, P., and Meyer, C. (2019). 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 (MSF) (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.

Niederle, M. and Yariv, L. (2009). Decentralized Matching with Aligned Preferences. NBER Working Paper 14840, National Bureau of Economic Research.

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.

Velásquez, A. (2020). *Ethical Considerations of Universal Health Insurance in Peru*. Antonio Ruiz de Montoya University. `https://www.researchgate.net/publication/384054392`.

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

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.