Congestion and Penalization in Optimal Transport

Marcelo Gallardo* marcelo.gallardo@pucp.edu.pe
Manuel Loaiza† manuel.loaiza@autodesk.com
Jorge Chávez* jrchavez@pucp.edu.pe

March 31, 2025

Abstract

We introduce a novel model based on the discrete optimal transport problem that incorporates congestion costs and replaces traditional constraints with weighted penalization terms. This approach better captures real-world scenarios characterized by demand-supply imbalances and heterogeneous congestion costs. We develop an analytical method for computing interior solutions, which proves particularly useful under specific conditions. Additionally, we propose an $O((N+L)N^2L^2)$ algorithm to compute the optimal interior solution. For certain cases, we derive a closed-form solution and conduct a comparative statics analysis. Finally, we present examples demonstrating how our model yields solutions distinct from classical approaches, leading to more accurate outcomes in specific contexts, such as Peru's health and education sectors.

Keywords: Optimal transport, Congestion costs, Quadratic regularization, Matching, Penalization, Neumann series, Health economics.
JEL classifications: C61, C62, C78, D04, R41.

We gratefully acknowledge insightful discussions with Professors Federico Echenique and Amílcar Vélez, whose valuable feedback significantly improved this work, as well as former Minister of Health of Peru, Aníbal Velásquez. Dr. Velásquez provided key information regarding the Peruvian health system. We also appreciate the support from the Academic Directorate for Professors (DAP) at Pontificia Universidad Católica del Perú (PUCP).

*Department of Mathematics, Pontificia Universidad Católica del Perú (PUCP).
†Autodesk, Inc.

1 Introduction

Optimal Transport (OT) (Villani, 2009; Galichon, 2016) is a mathematical technique that, in recent years, has been integrated into economic theory, particularly in the study of matching markets (Chiappori et al., 2010; Galichon, 2021; Dupuy et al., 2019; Carlier et al., 2023; Echenique, Federico, Joseph Root and Feddor Sandomirskiy, 2024). Unlike classical matching models (Gale and Shapley, 1962; Hylland and Zeckhauser, 1979; Kelso and Crawford, 1982; Roth and Sotomayor, 1990; Abdulkadiroğlu and Sönmez, 2003; Hatfield and Milgrom, 2005; Echenique, Federico and M. Bumin, Yenmez, 2015), OT optimizes over distributions. Starting from the classical model, in which matching costs are represented by a linear function, various extensions have incorporated a regularization term in the objective function to obtain solutions with desirable properties such as sparsity. Notable examples include entropic regularization (Dupuy and Galichon, 2014; Dupuy et al., 2019; Merigot and Thibert, 2020; Galichon, 2021) and quadratic regularization (Lorenz et al., 2019; González-Sanz and Nutz, 2024; Wiesel and Xu, 2024; Nutz, 2024). Both classical OT and its regularized variants have been widely applied in analyzing matching markets, including marriage markets (Dupuy and Galichon, 2014), migration dynamics (Carlier et al., 2023), labor markets (Dupuy and Galichon, 2022), and school choice (Echenique, Federico, Joseph Root and Feddor Sandomirskiy, 2024).

The quadratic regularization model, which is the most recent, allows incorporating a congestion effect. This element is crucial as it enables the representation of scenarios where matching becomes increasingly costly. This paper introduces a new model, resulting in a convex optimization problem, built upon the quadratic regularization framework, similar to Nutz (2024), but adopting the approach of Izmailov and Solodov (2023) by replacing equality constraints with weighted penalization terms and introducing heterogeneity in the quadratic term. These elements are essential as they allow for a better modeling of situations where matching cannot be properly achieved, meaning that persistent excess demand or supply is not captured by classical models. Moreover, quadratic heterogeneity provides a more refined representation of congestion, as it allows congestion to vary across different pairs.

These properties are essential for modeling matching in developing countries, where access to healthcare and education systems is hindered by significant frictions and severe congestion resulting from inadequate infrastructure. These structural deficiencies have contributed to high mortality rates and service shortages, as highlighted during the COVID-19 pandemic. For instance, Johns Hopkins University Coronavirus Resource Center (2023) reports that Peru recorded the highest per capita COVID-19 mortality rate globally, exceeding 6,400 deaths per million inhabitants. In countries such as Peru, India, and Brazil (Kikuchi and Hayashi, 2020), congestion is particularly severe. For instance, World Bank (2024) estimates that traffic congestion alone costs Peru 1.8% of its GDP annually. Given these conditions, accounting for congestion and excess demand is crucial when modeling these dynamics.

The model presented in this paper, formulated from a social planner's perspective, provides a framework for incorporating congestion costs while accounting for excess demand in different institutional settings. This contrasts with developed countries like France or Switzerland, where

efficient infrastructure and policies mitigate such frictions. Our approach introduces a strictly convex cost structure that remains analytically tractable and flexibility through the choice of the parameters, allowing to recover situations with congestion, low congestion or even non congestion

The remainder of this paper is structured as follows. Section 2 defines the notation and establishes the preliminary concepts. Section 3 introduces the proposed model, and examines its theoretical properties. A key aspect is that the structure of the optimization problem allows for an analytical solution, and enables the construction of an $O((N+L)N^2L^2)$ algorithm to compute the optimal interior solution. Section 4 provides illustrative examples that highlight the advantages of our model, its flexibility, and its accuracy in studying scenarios where the social planner faces congestion and is unable to ensure demand-supply equilibrium. Our empirical analysis focuses on the Peruvian health and education sectors. All proofs are provided in the Appendix.

2 Preliminaries

We consider two sets, $X = \{x_1, \dots, x_N\}$ and $Y = \{y_1, \dots, y_L\}$. Each element x_i (y_j) represents an individual or a group of individuals/entities that share certain properties and are grouped into the same cluster. For example, in the marriage market (where usually $N = L$), X is the set of men and Y is the set of women. In the case of school matching, X consists of groups of students, grouped, for instance, according to their district, and Y is the set of schools. We denote by μ_i the *mass* of x_i and by ν_j the *mass* of y_j. For instance, in the marriage market, $\mu_i = \nu_j = 1$, while in the case of schools, ν_j would represent the capacity of school j. Analogously, if X were patients and Y medical care centers, then parameters ν_j would represent the capacity of the medical care center. When referring to an element of X, instead of denoting it by x_i, we usually, to simplify the notation, refer to it by i. Analogously, the elements of Y are referred to by the index j, instead of y_j. Moreover, we denote the set of indices $\{1, \dots, N\}$ by I and the set of indices $\{1, \dots, L\}$ by J. Lastly, we denote by π_{ij} the number of individuals of type i matched with j.

The problem addressed in the classic literature (Galichon, 2016; Dupuy et al., 2019; Carlier et al., 2023), from the perspective of a central planner, is to decide how many individuals from group i should be matched with $j \in J$ and so forth for each i, minimizing the matching cost[1], which is given by means of a function $C : \mathbb{R}_+^{N,L} \times \mathbb{R}^P \to \mathbb{R}$ depending on the matching $\pi = [\pi_{ij}] \in \mathbb{R}_+^{N,L}$[2], and a vector of parameters $\theta \in \mathbb{R}^P$. Moreover, the central planner must

[1] Matching individuals incurs a cost that is not limited solely to «physical» transportation costs, which certainly accounts for both ways (round trip), but also encompasses implicit costs linked to specific characteristics of i and j such as tuition fee, entrance exam, languages, sex, age, etc. This is why we refer to them as matching costs instead of transportation costs.

[2] In this work, we will mostly assume that the number of individuals matched can take values in the real positive line and not only in the positive integers. Note that this is the same issue that arises when one solves the utility maximization problem in the classical framework assuming divisible goods. Later on, we will address again this issue and explain why considering $\pi_{ij} \in \mathbb{R}_+$ allows drawing solid conclusions from an economic perspective.

ensure that there are neither excesses of demand nor supply. Hence, the central planner solves

$$\min_{\pi \in \Pi(\mu,\nu)} C(\pi;\theta), \tag{1}$$

where

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

A solution to (1) will be from now referred to as an optimal matching or optimal (transport) plan, and will be denoted by π^*. In the standard optimal transport model, separable linear costs are assumed (Galichon, 2016). This is, $C(\pi,\theta) = \sum_{i,j} c_{ij}\pi_{ij}$. It is therefore assumed that the marginal cost of matching one more individual from i with j is always the same, regardless of how many people are already matched and independent of any other variable. Hence, the central planner seeks to solve

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

To solve $\mathcal{P}_O$, one typically employs linear programming techniques, such as the simplex method. As discussed in the classical literature, the most general form of the OT problem allows for the existence of infinite types, and in such a case, the optimization is done over continuous distributions. In this paper, however, we are not going to study continuous distributions. What we do focus on, in line with the entropic regularization problem (see, for example, Carlier et al. (2023) and Peyré and Cuturi (2019)), is working with a variation of the optimization problem in the discrete setting. In the case of entropic regularization (3), the problem addressed is

$$\min_{\pi \in \Pi(\mu,\nu)} \sum_{i=1}^{N}\sum_{j=1}^{L} c_{ij}\pi_{ij} + \sigma\pi_{ij}\ln(\pi_{ij}), \tag{3}$$

with $\sigma > 0$. Given the strict convexity of $f(x) = x\ln x$, $f(0) = 0$ and $\lim_{x\downarrow 0^+} f'(x) = -\infty$, the solution is interior, i.e. $\pi^*_{ij} > 0$. Another variation is the quadratic regularization, where the problem becomes

$$\min_{\pi \in \Pi(\mu,\nu)} \sum_{i=1}^{N}\sum_{j=1}^{L} c_{ij}\pi_{ij} + \frac{\varepsilon}{2}||\pi||_2^2. \tag{4}$$

Unlike the problem (3), in the case of (4), interior solutions cannot be guaranteed[3]. In the model we present in the following section, we build upon the problem (4), making a considerable number of modifications that allow us to adapt to specific economic contexts of countries with structural problems. Before concluding this section, let us briefly note that, by a combinatorial argument, it is possible to conclude that the number of matchings is bounded by L^M in the case where $\pi_{ij} \in \mathbb{Z}_+$. However, for the case $\pi_{ij} \in \mathbb{R}_+$, considering $\mu_i, \nu_j > 0$ for all $(i,j) \in I \times J$, the compactness of $\Pi(\mu,\nu)$ and continuity of the objective functions, ensure the existence of a solution to $\mathcal{P}_O$ and its variants by Weierstrass Theorem.

[3]This is a common feature with our model, it is not straightforward to determine if the solution is interior.

3 The model

The model we propose results in the following quadratic optimization problem

$$\mathcal{P}_{CP}: \min_{\pi_{ij}\geq 0} \underbrace{\left\{ \underbrace{\alpha \sum_{i=1}^{N}\sum_{j=1}^{L} \varphi(\pi_{ij};\theta_{ij})}_{\text{Matching direct cost.}} + \underbrace{(1-\alpha)\left[\sum_{i=1}^{N} \epsilon_i \left(\sum_{j=1}^{L} \pi_{ij} - \mu_i\right)^2 + \sum_{j=1}^{L} \delta_j \left(\sum_{i=1}^{N} \pi_{ij} - \nu_j\right)^2\right]}_{\text{Costs of social objectives.}} \right\}}_{F(\pi;\theta,\alpha,\epsilon,\delta,\mu,\nu).} \tag{5}$$

where $\epsilon_1, ..., \epsilon_N$, $\delta_1, ..., \delta_L$ and $\mu_1, \cdots, \mu_N$, $\nu_1, \cdots, \nu_L$ are all non negative, and

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

The objective function in (5) represents a trade-off between the direct costs of matching, incorporating the heterogeneous congestion effect given by $\sum_{i=1}^{N}\sum_{j=1}^{L} a_{ij}\pi_{ij}^2$, and the central planner's objectives, which are defined by the targets $\mu = (\mu_1, \ldots, \mu_N)$ and $\nu = (\nu_1, \ldots, \nu_L)$.

Unlike classical models, our approach accounts for congestion and allows for excess supply or demand. Additionally, it introduces weight parameters, increasing flexibility.

Regarding the quadratic costs, they model a saturation effect in which matching more individuals from $i \in I$ with the same $j \in J$ becomes increasingly costly. For example, from the perspective of physical transportation costs, in countries with high vehicular congestion, the impact of increasing from x cars to $x+1$ on a given avenue is lower or equal to increasing from $x+n$ to $x+n+1$ with $n \geq 1$. Therefore, clustering individuals based on geographic location implies that matching many individuals from the same group to a single j congests the access route (which remains the same). The coefficient a_{ij} captures heterogeneity[4], while the quadratic term represents the previously described phenomenon[5]. Note that quadratic costs are not limited to physical transportation costs but can also represent bureaucratic costs. A hospital receives patients of the same type, and as more patients of this type arrive, the system must process an increasing number of cases. Since they share similar characteristics, the same computer or system is assumed to handle their processing. Given the precarious conditions in developing countries, increasing from x to $x+1$ patients may not significantly affect the system, but increasing from $x+n$ to $x+n+1$ with $n > 1$ might (e.g., leading to system freezes, delays, etc.).

On the other hand, the targets and weighted penalties model the fact that the central planner has specific objectives: educating (or providing healthcare to) μ_i individuals of type

[4]In some situations, the coefficient might be large, but in others—such as cases with few schools or hospitals, low traffic congestion, efficient traffic lights, etc.—the coefficient is small. Moreover, one could question whether adding a car still marginally increases costs when a route is already saturated. However, this effect only arises when the number of travelers is excessively high relative to the route's capacity. For simplicity, we omit this case, as modeling a function that is initially quadratic and later constant would unnecessarily complicate the analysis when applying FOCs.

[5]Instead of using π_{ij}^2, we could consider a general strictly increasing and convex function ψ, such as $\psi(\pi_{ij}) = e^{\pi_{ij}}$ or π_{ij}^3. However, the quadratic structure facilitates quantitative analysis and preserves the consistency of the results and modeling.

i, while ensuring that schools (or medical centers) accommodate a student (or patient) level close to ν_j. Additionally, the central planner can decide which target has greater importance through the parameters $\epsilon_1, \ldots, \epsilon_N$ and $\delta_1, \ldots, \delta_L$. The constraint $\sum_{i=1}^{N} \pi_{ij} = \nu_j$ is replaced by the penalty term $\delta_j \left[\sum_{i=1}^{N} \pi_{ij} - \nu_j\right]^2$, $\delta_j > 0$, and the constraint $\sum_{j=1}^{L} \pi_{ij} = \mu_i$ is replaced by $\epsilon_i \left[\sum_{j=1}^{L} \pi_{ij} - \mu_i\right]^2$, $\epsilon_i > 0$. The parameters ϵ_i, δ_j serve as weights. Note that we could use any $p \geq 1$ norm for the penalty. However, the quadratic structure simplifies the mathematical analysis and fulfills the intended role. By allowing deviations, as we will see in the examples, we better approximate the reality of developing countries that cannot fully ensure that demand perfectly matches supply.

Allowing for the possibility of excess supply or demand, is reasonable in some contexts, as we will see. Indeed, underdeveloped countries may not be able to ensure full coverage in education and health, making it more realistic for them to face a trade-off. However, it is natural for the central planner to seek to minimize these excesses: ensuring that children attend school, that schools or hospitals do not become overcrowded, etc.

Finally, we impose the constraint $\pi_{ij} \geq 0$ for all $(i, j) \in I \times J$. However, we do not impose upper bounds since we consider a population or universe that is arbitrarily large (a subpopulation of a sufficiently large country)[6]. Thus, the optimization is performed over the entire space $\mathbb{R}_+^{NL}$. This phenomenon also justifies the penalty terms: we no longer assume a fixed number of individuals of type i, and μ_i now represents a target that the central planner aims to achieve (how many individuals of type i should ideally be matched). Similarly, the parameters ν_j are also targets of the central planner.

In (6), despite its practical relevance, the term d_{ij}, representing fixed costs, does not influence the resolution of the problem. For this reason, when considering the parameter vector $\theta_{ij} \in \mathbb{R}^2$, we think of it as (c_{ij}, a_{ij}). Unlike more recent models in the quadratic regularization literature, we allow heterogeneity in the quadratic structure.

Having now established the model, which, to the best of our knowledge, is new in the literature[7], we focus in this section on the following theoretical problems: (i) ensuring the existence of a solution, (ii) analyzing uniqueness, (iii) addressing why optimization in $\mathbb{R}_+^{NL}$ is reasonable and why we do not resort to integer optimization, (iv) studying how to compute interior solutions, and (v) analyzing particular cases both from the analytical and numerical perspective. In the next section, we compare our model with previous ones from the literature and highlight its advantages and the new insights it provides.

Existence and uniqueness: Regarding the existence of a solution to $\mathcal{P}_{CP}$, in order to apply Weierstrass theorem to overcome the potential issue that the optimization is carried over an

[6] This significantly simplifies our analysis and does not affect the model's logic.

[7] Quadratic regularization does not involve penalization terms and assumes $a_{ij} = \varepsilon$ for all $(i, j) \in I \times J$. With respect to the classical optimal transport problem, linear costs are considered. On the other hand, entropic regularization involves Inada's conditions, which do not appear in our model. Finally, in Izmailov and Solodov (2023), only general results concerning penalization are given and this particular problem is not studied at all.

unbounded set, we can actually restrict the optimization to $\mathbb{R}_+^{NL} \cap \Omega$, where

$$\Omega = [0, R]^{NL}, \text{ with } R = N \max_{1 \leq i \leq N} \{\mu_i\} + L \max_{1 \leq j \leq L} \{\nu_j\}.$$

In fact, it is clear from the cost function F that it is strictly lower in the interior of Ω or in the axes than when evaluated in $\partial\Omega$ (without considering the axes) or outside Ω. This is a consequence of the coercivity of the objective function (Rockafellar, 1970). With respect to uniqueness, it is a consequence of the strict convexity of the objective function. Indeed, the objective function is the sum of a strictly convex function, $\sum_{i,j} \varphi(\pi_{ij}, \theta_{ij})$, with $N + L$ convex functions of the form $\varrho \left(\sum_{m=1}^{M} \eta_m - \Theta \right)^2$, with $\varrho, \Theta, \eta_m \in \mathbb{R}_+$.

Optimization carried over $\mathbb{R}_+^{NL}$: As we mentioned previously, similarly to the case of the classical demand theory, we are assuming that $\pi_{ij} \in \mathbb{R}_+$. However, just as it does not make sense to consume $\sqrt{2}$ cars, it can be also unreasonable to consider that π_{ij} is not restricted to taking values in $\mathbb{Z}_+$, since it ultimately represents the number of individuals. However, given the structure of the optimization problem—a convex quadratic optimization problem—following the classical literature on rounding methods (Beck and Fiala, 1981) and, in particular, the discrepancy between integer (Park and Boyd, 2018; Pia and Ma, 2022) and continuous solutions in the case of separable quadratic functions with linear constraints (Hochbaum and Shanthikumar, 1990), it is possible to establish bounds on the deviation of the optimal solution when transitioning from the continuous domain $\mathbb{R}_+^{NL}$ to the integer lattice $\mathbb{Z}_+^{NL}$, and ensure that it is sufficiently close. The bound depends on the eigenvalues of the Hessian matrix of the objective function[8]. Solving the problem in $\mathbb{R}_+^{NL}$ allows the use of nonlinear convex optimization techniques, yielding not only computational advantages but also analytical insights. In this work, we do not delve deeply into this aspect, but we emphasize that by adjusting the parameters, it is possible to control the bound on the norm of the difference between the solutions in the lattice and the Euclidean space.

Interior solutions: For the sake of simplicity, we take $\alpha = 1/2$. KKT first order conditions applied to (5) yield

$$\frac{\partial F}{\partial \pi_{ij}} = \frac{1}{2} \left(\varphi'(\pi_{ij}^*; \theta_{ij}) + 2\epsilon_i \left(\sum_{\ell=1}^{L} \pi_{i\ell}^* - \mu_i \right) + 2\delta_j \left(\sum_{k=1}^{N} \pi_{kj}^* - \nu_j \right) - \gamma_{ij}^* \right) = 0, \ \forall \ (i,j) \in I \times J. \quad (7)$$

Here, γ_{ij} is the associated multiplier to the inequality constraint $\pi_{ij} \geq 0$. Determining whether or not the solution is interior, is not trivial. For corner solutions, we have to iterate all possible combinations of γ_{ij}^* equal or not to zero. Formally, 2^{NL} possibilities. In general, the problem can numerically be solved. In what follows, unless the contrary is stated, we will address the case where the solution is interior. In this case, from KKT, we know that $\gamma_{ij}^* = 0$ for all $(i,j) \in I \times J$. Hence, from (7), we have $\nabla F(\pi^*) = 0$. This set of equations can be written in the compact form

[8] Specifically, the deviation is bounded by $||\pi_{\text{int}} - \pi^*||_\infty \leq O(\vartheta(H))$, where $\vartheta(H) = \lambda_{\max}(H)/\lambda_{\min}(H)$ is the condition number.

$A \begin{bmatrix} \pi_{11}^* & \pi_{12}^* & \cdots & \pi_{NL}^* \end{bmatrix}^T = b$, where

$$A = \underbrace{\text{Diag}(a_{11}, a_{12}, \ldots, a_{NL})}_{D} + \underbrace{\text{Diag}(\epsilon_1, \ldots, \epsilon_N) \otimes \mathbf{1}_{L\times L}}_{E} + \underbrace{\mathbf{1}_{N\times N} \otimes \text{Diag}(\delta_1, \ldots, \delta_L)}_{F}, \tag{8}$$

and $b = [\epsilon_1\mu_1 + \delta_1\nu_1 - c_{11}/2, \epsilon_1\mu_1 + \delta_2\nu_2 - c_{12}/2, \cdots, \epsilon_N\mu_N + \delta_L\nu_L - c_{NL}/2]^T$. The following lemma states that A is an invertible matrix.

Lemma 3.1. *The determinant of A is strictly positive, whenever all parameters are strictly positive.*

Therefore, the linear system $A\pi = b$ has a unique solution. What we still don't know is whether or not this solution belongs to $\mathbb{R}_{++}^{NL}$. If so, given the strict convexity of F, we would have determined, through an ex-post analysis, the unique solution to $\mathcal{P}_{CP}$. However, it may not always be the case that $A^{-1}b \in \mathbb{R}_{++}^{NL}$, and it is not a trivial matter to determine. Under specific cases, we will be able to do this. We propose both an analytical and a computational method to solve $A\pi = b$. The analytical method allows us, in special cases, to derive important theoretical conclusions, such as closed-form solutions, bounds, and perform comparative statics. From a computational perspective, we compare our algorithm, which exploits the structure of the matrix A, with others for solving linear systems.

3.1 Neumann's series approach

Assumption 1. Let $a_{ij} > 0$ for all $(i, j) \in I \times J$. Assume that

$$\max_{1\le i\le N}\{\epsilon_i\} \cdot L + \max_{1\le j\le L}\{\delta_j\} \cdot N < \min_{(i,j)\in I\times J}\{a_{ij}\}.$$

Assumption 1 implies that convex transport costs are large. Moreover, the fact that ϵ_i, δ_j are small follows from their interpretation as normalized weights, i.e., $\epsilon_i, \delta_j \in [0, 1]$.

Lemma 3.2. *Under Assumption 1, the following holds*

$$A^{-1} = \left(\sum_{k=0}^{\infty}(-1)^k(D^{-1}X)^k\right) D^{-1}.$$

Theorem 3.3. *Under Assumption 1,* $\lim_{n\to\infty} \pi_n = \pi^* = A^{-1}b$*, where*

$$\pi_n = S_n D^{-1}b = \left(\sum_{k=0}^{n}(-1)^k(D^{-1}X)^k\right) D^{-1}b.$$

3.2 Special cases

For the aim to explicitly compute A^{-1}, we need to impose some additional mild assumptions.

3.2.1 No interest in overcrowding or no quotas.

Assumption 2. Assume that $\delta_j = 0$ for all $j \in J$ and $D = \beta I$ for some $\beta > 0$.

Assumption 2 illustrates the case where the central planner does not care if in over or underfilling schools or hospitals ($F = 0$), and convex costs are the same across the pairs (i, j): $a_{ij} = \beta$. For instance, the latter applies when distances, routes, or bureaucratic systems are almost the same for all $(i, j) \in I \times J$.

Assumption 3. Assume that $L\epsilon_i < \min\{1, \beta\}$ for all $1 \leq i \leq N$.

In line with Assumption 1, Assumption 3 applies when convex transport costs are large.

Theorem 3.4. *Under Assumptions 2 and 3, A^{-1} is given as follows*

$$A^{-1} = \frac{I}{\beta} + \frac{1}{\beta} Diag\left(-\frac{\epsilon_1}{\beta + L\epsilon_1}, \ldots, -\frac{\epsilon_N}{\beta + L\epsilon_N}\right) \otimes \mathbf{1}_{L\times L}. \tag{9}$$

A similar result can be obtained by setting $E = 0$, i.e., when the central planner is only concerned with overcrowding or underutilization of facilities and does not care about population quotas.

Corollary 3.5. *Under Assumptions 2 and 3, the solution of $\mathcal{P}_{CP}$ is given by*

$$\pi^*_{ij} = \frac{b_{ij}}{\beta} - \sum_{\ell=1}^{L} \frac{b_{i\ell}\epsilon_i}{\beta^2 + L\epsilon_i\beta}, \tag{10}$$

provided that the right-hand side of (10) is positive.

Proof. This result follows directly from the computation of $A^{-1}b$ by using (9). ■

3.2.2 Equal weighting and identical convex costs.

Assumption 4. Let ρ and ζ be real numbers such that $\rho > 2NL\zeta > 0$, with $a_{ij} = \rho$ and $\epsilon_i = \delta_j = \zeta$ for all $(i, j) \in I \times J$.

Assumption 4 implies that the central planner assigns equal weight to each social objective and where congestion and bureaucratic costs are the same for each pair. Under this assumption, we have $D = \rho I$ and $X = \zeta Y$, where the entries of Y are given by

$$Y_{ij} = \begin{cases} 2 & i = j, \\ 1 & i \neq j \wedge (\lceil i/N \rceil = \lceil j/N \rceil \vee i \equiv j \pmod N), \\ 0 & \text{otherwise.} \end{cases}$$

This allows us to write

$$A^{-1} = \frac{1}{\rho}\left(\sum_{k=0}^{\infty}\left(-\frac{\zeta}{\rho}\right)^k Y^k\right).$$

Under Assumption 4, we will be able to establish bounds on the optimal matching, i.e., to bound the number of individuals matched across the pairs (i, j). Lemmas 3.6, 3.7 and 3.8 are used to establish Theorem 3.9.

Lemma 3.6. *Let $k \geq 1$ be a positive integer. Then*

$$\max_{1\leq i,\ j\leq NL}\left\{\left(Y^k\right)_{ij}\right\} \leq \frac{(2NL)^k}{NL}.$$

Lemma 3.7. *Let $k \geq 2$ be a positive integer.Then*

$$\frac{(NL)^{\lfloor k/2\rfloor}}{NL} \leq \min_{1\leq i,\ j\leq NL}\left\{\left(Y^k\right)_{ij}\right\}.$$

Lemma 3.8. *Under Assumptions 1 and 4, the lower and the upper bounds of $\left(A^{-1}\right)_{ij}$ can be expressed in terms of N, L, ζ and ρ,*

$$C_1(N, L, \zeta, \rho) \leq (A^{-1})_{ij} \leq C_2(N, L, \zeta, \rho), \tag{11}$$

where

$$C_1 = \frac{\zeta\left(4\zeta N^3L^3\left(2\zeta^3 - 2\zeta\rho^2 - \rho^3\right) + 8N^2L^2\rho^2\left(\rho^2 - \zeta^2\right) + \zeta NL\rho^2(2\zeta + \rho) - 2\rho^4\right)}{\rho^4\left(\zeta^2NL - \rho^2\right)(2NL - 1)(2NL + 1)}$$
$$C_2 = \frac{\zeta^2NL\rho(4NL - 1)}{\left(\rho^2 - \zeta^2NL\right)(\rho - 2NL\zeta)(\rho + 2NL\zeta)}.$$

Theorem 3.9. *Under Assumptions 1 and 4, it follows that $\pi^*_{ij} \leq NL\tilde{C}$, for all $(i, j) \in I \times J$, where*

$$\tilde{C} = \max\{|C_1|, C_2\} \cdot \max_{\substack{1\leq i\leq N\\ 1\leq j\leq L}}\left\{\left|(\epsilon_i\mu_i + \delta_j\nu_j) - \frac{c_{ij}}{2}\right|\right\}.$$

Theorem 3.9 is of particular interest as it allows us to determine, without computing the inverse of A, the maximum number of individuals that would be matched between two points i, j. In practice, this enables, for example, the establishment of capacity constraints on routes or spaces.

3.3 Algorithm for computing π^*

We now provide an efficient algorithm to compute $\pi^* \in \mathbb{R}^{NL}_{++}$. This is established in Theorem 3.10. First, let us rewrite matrix A given in (8) as follows:

$$A = \mathrm{Diag}(a_{11}, \ldots, a_{NL}) + \sum_{i=1}^{N}\left(\epsilon_i^{1/2}\mathbf{e}_i \otimes \mathbf{1}_{L\times 1}\right)\left(\epsilon_i^{1/2}\mathbf{e}_i^T \otimes \mathbf{1}_{1\times L}\right) + \sum_{j=1}^{L}\left(\delta_j^{1/2}\mathbf{e}_j \otimes \mathbf{1}_{N\times 1}\right)\left(\delta_j^{1/2}\mathbf{e}_j^T \otimes \mathbf{1}_{1\times N}\right).$$

Theorem 3.10. *For interior solutions π^*, Algorithm 1 computes π^* in $O((N + L)N^2L^2)$ time.*

Algorithm 1 OPTIMIZE $(a, b, \epsilon_1, \ldots, \epsilon_N, \delta_1, \ldots, \delta_L)$

Input: Matrices $a \in \mathbb{R}^{NL}_{++}$, $b \in \mathbb{R}^{NL}$ and parameters $\epsilon_1, \ldots, \epsilon_N, \delta_1, \ldots, \delta_L \in \mathbb{R}_{++}$
Output: $\pi^* \in \mathbb{R}^{NL}$
Initialize $A^{-1} \leftarrow \text{Diag}(1/a_{11}, \ldots, 1/a_{NL}) \in \mathbb{R}^{NL,NL}$
for $i \leftarrow 1, \ldots, N$ **do**
    Define $u^{(i)} \in \mathbb{R}^{NL}$ by $u^{(i)} \coloneqq \epsilon_i^{1/2} \mathbf{e}_i \otimes \mathbf{1}_{L\times 1}$
    $A^{-1} \leftarrow A^{-1} - \dfrac{A^{-1}u^{(i)}u^{(i)T}A^{-1}}{1 + u^{(i)T}A^{-1}u^{(i)}}$ via Sherman-Morrison formula
end for
for $j \leftarrow 1, \ldots, L$ **do**
    Define $v^{(j)} \in \mathbb{R}^{NL}$ by $v^{(j)} \coloneqq \delta_j^{1/2} \mathbf{e}_j \otimes \mathbf{1}_{N\times 1}$
    $A^{-1} \leftarrow A^{-1} - \dfrac{A^{-1}v^{(j)}v^{(j)T}A^{-1}}{1 + v^{(j)T}A^{-1}v^{(j)}}$ via Sherman-Morrison formula
end for
return $A^{-1}b$

Time	Sparse A	Galactic	Authors
$O(N^3L^3)$	No	No	Gaussian Elimination
$O((NL)^{2.81})$	No	No	Strassen (1969)
$O((NL)^{2.331645})$	Yes	Yes	Peng and Vempala (2024)
$O((NL)^{2.371339})$	No	Yes	Alman et al. (2025)
$O((N+L)N^2L^2)$	No	No	**This paper**

Table 1: Algorithms for solving our linear system. Assume A is sparse if it has $\widetilde{O}(NL)$ nonzero entries. "Galactic" refers to an algorithm wonderful in its asymptotic behavior, but is never used to actual compute anything (Lipton (2010)).

It was observed by Vassilevska (2015) that inversion can be reduced to multiplication with an equivalent runtime for Strassen (1969) and Alman et al. (2025). Even though Peng and Vempala (2024) and Alman et al. (2025) provide the best bounds, they are impractical due to large constants, leaving us with the remaining three algorithms for practical purposes. Among these, when $L = \Theta(N)$, our algorithm has the tightest upper bound compared to classical Gaussian elimination and an inversion derived from Strassen multiplication.

3.4 Comparative statics

Although we know how to compute π^* through Neumann's series or Algorithm 1, obtaining a closed-form expression for π^*_{ij} using these techniques is not straightforward. Therefore, to facilitate comparative statics, one possible approach is to approximate the matrix A^{-1} using Neumann's series. First, assume that $A^{-1} \simeq D^{-1}$. This simplification allows us to derive a closed-form expression for π^*_{ij}, providing initial insights. Under the assumption $A^{-1} \simeq D^{-1}$, we obtain:

$$\pi^*_{ij} \simeq \frac{2(\epsilon_i\mu_i + \delta_j\nu_j) - c_{ij}}{2a_{ij}}.$$

From this expression, it follows that $\partial\pi^*_{ij}/\partial a_{ij}, \partial\pi^*_{ij}/\partial c_{ij} < 0$ and $\partial\pi^*_{ij}/\partial\epsilon_i, \partial\pi^*_{ij}/\partial\delta_j$, $\partial\pi^*_{ij}/\partial\mu_i, \partial\pi^*_{ij}/\partial\nu_j > 0$. These results align with standard economic intuition. However,

under this rough approximation, we obtain $\partial\pi^*_{ij}/\partial\theta_{k\ell} = 0$ for $(k,\ell) \neq (i,j)$, which is unrealistic since we expect a substitution effect. To improve upon this, consider a refined approximation:

$$A^{-1} \sim D^{-1} - D^{-1}XD^{-1} = D^{-1} - (D^{-1})^2X.$$

From smooth comparative statics, if $\pi^* \in \mathbb{R}^{NL}_{++}$ is an interior solution to $\mathcal{P}_{CP}$ associated with the parameter vector $(\overline{\theta}, \epsilon, \delta, \mu, \nu) \in \mathbb{R}^{2NL}_{++} \times \mathbb{R}^{N}_{++} \times \mathbb{R}^{L}_{++} \times \mathbb{R}^{N}_{++} \times \mathbb{R}^{L}_{++}$, then:

$$\left[\frac{\partial\pi^*_{ij}}{\partial\theta_{k\ell}}\right] = -A^{-1}_{(\overline{\theta},\epsilon,\delta,\mu,\nu)}[I_{NL\times NL} \,|\, 2\text{Diag}(\pi^*_{11}, \cdots, \pi^*_{NL})]. \tag{12}$$

Thus, under the approximation $A^{-1} \sim D^{-1} - (D^{-1})^2X$, we obtain:

$$\left[\frac{\partial\pi^*_{ij}}{\partial\theta_{k\ell}}\right] = \left[\frac{\partial\pi^*_{ij}}{\partial c_{k\ell}} \,\middle|\, \frac{\partial\pi^*_{ij}}{\partial a_{k\ell}}\right] \simeq -\left[D^{-1} - (D^{-1})^2X \,|\, A^{-1}_{\Pi,2}\right], \tag{13}$$

where $A^{-1}_{\Pi,2}$ consists of multiplying column ij of $D^{-1} - (D^{-1})^2X$ by π^*_{ij}. From (13), if $\max_{i,j}\{\epsilon_i + \delta_j\} < 1$, then: $\partial\pi^*_{ij}/\partial\theta_{ij} < 0$ for all $(i,j) \in I \times J$, $\partial\pi^*_{ij}/\partial\theta_{k\ell} > 0$ for $i \neq k$ and $j = \ell$ or $i = k$ and $j \neq \ell$, $\partial\pi^*_{ij}/\partial\theta_{k\ell} = 0$ if $i \neq k$ and $j \neq \ell$. Then, we conclude from (13) that:

$$\partial\pi^*_{ij}/\partial c_{ij} = -(1 - (\epsilon_i + \delta_j))/a^2_{ij} < 0,$$

$$\partial\pi^*_{ij}/\partial c_{i\ell} = \epsilon_i/a^2_{ij} > 0, \quad \partial\pi^*_{ij}/\partial c_{kj} = \delta_j/a^2_{ij} > 0, \quad \partial\pi^*_{ij}/\partial c_{k\ell} = 0 \text{ if } i \neq k, j \neq \ell.$$

$$\partial\pi^*_{ij}/\partial a_{ij} = -2\pi^*_{ij}(1 - (\epsilon_i + \delta_j))/a^2_{ij} < 0, \quad \partial\pi^*_{ij}/\partial a_{i\ell} = 2\pi^*_{i\ell}\epsilon_i/a^2_{ij} > 0,$$

$$\partial\pi^*_{ij}/\partial a_{kj} = 2\pi^*_{kj}\delta_j/a^2_{ij} > 0, \quad \partial\pi^*_{ij}/\partial a_{k\ell} = 0 \text{ if } i \neq k, j \neq \ell.$$

These results are much closer to what we would expect. Indeed, we now observe a *substitution effect*: if the cost of matching individuals of type i with j increases ceteris-paribus, then the number of individuals of type i matched with ℓ (where $\ell \neq j$) increases. However, it is important to note that these results are obtained under a truncated Neumann series approximation, and should be interpreted accordingly—as an approximation. Nevertheless, note that under Assumptions 1, 2, and 3, it is possible to compute the effects of the parameters directly using (10). In such case, similar conclusions can be derived.

3.5 Case $N = L$

The case $N = L > 1$ is particularly important in the classical literature on the marriage market (Roth and Sotomayor, 1990). Similarly, as we will see in Section 4, it is of particular interest when analyzing the healthcare sector in Peru. If the solution in our model is interior, the problem reduces to solving a system of linear equations, and the condition $N = L$ improves the upper bound on the number of operations required by Algorithm 1 compared to folklore linear system solvers. On the other hand, classical transportation problems and their variants require approximation algorithms for solving convex optimization problems in finite dimensions (Merigot and Thibert (2020)).

4 Examples and applications

4.1 Health care

The Peruvian healthcare system is characterized by being a fragmented system with three main types of medical care centers: SIS (Seguro Integral de Salud), EsSalud, and EPS (Entidades Prestadoras de Salud) (Anaya-Montes and Gravelle, 2024). EPS corresponds to private health insurance offered by companies such as Rimac, Mapfre, Pacífico, La Positiva, among others. These insurances are aimed at formal workers seeking additional coverage beyond mandatory insurance. On the other hand, EsSalud is the public health insurance financed by contributions from formal workers and employers, both from the private and public sectors. Finally, SIS is a universal public insurance targeting people in poverty, informals, or without the ability to pay EPS. For the year of the pandemic (2020), SIS and EsSalud together covered more than 80% of the population, while less than 10% was covered by EPS, see Table 2.

Insurance	**Covered people**
EPS	8%
EsSalud	30%
SIS	53%

Table 2: Percentage of enrollees in Peru's healthcare system by type of medical care center in 2020, before COVID-19. At that time, Peru's population was 32,838,579 (Data Commons, 2025).

Under normal circumstances, an individual insured by SIS cannot be simultaneously enrolled in EsSalud or an EPS, and vice versa. The only permitted association is between EsSalud and EPS, where private insurance acts as a complementary coverage to the public system (Anaya-Montes and Gravelle, 2024; Velásquez, 2020). Ideally, an optimal allocation would ensure that informal workers are covered by SIS, while formal workers are appropriately distributed between EsSalud and EPS. However, in practice, overlapping affiliations occur, and individuals often seek medical care outside their designated system. Furthermore, a similar issue arises when categorizing healthcare utilization by type of illness: specialized medical centers create unintended overlaps in patient distribution across insurance networks. Additional issues related to congestion and deficiencies are detailed in Table 3.

Given Table 3, it is evident that Peru's healthcare system faces significant issues, including service inefficiencies, congestion costs, and saturation. Our model effectively captures these elements, unlike traditional matching models. Our approach can help identify critical areas for improvement, optimizing healthcare demand coverage and reducing congestion costs by analyzing the effect of parameters over π^*. It allows for the prioritization of interventions to address the most severe inefficiencies. To achieve this, estimating parameters is essential. This aligns with empirical research such as Doval et al. (2024) and the methodologies outlined in Agarwal, Nikhil and Somaini, Paulo (2023), which provide a structured framework to evaluate these inefficiencies.

Identified Problem	Quantifiable Indicator	Source
Shortage of medical personnel in primary healthcare.	12 doctors per 10,000 inhabitants, far from the WHO-recommended standard of 43.	Bendezu-Quispe et al. (2020).
Lack of hospital beds in Peru's healthcare system.	1.6 beds per 1,000 inhabitants, below the regional average.	World Bank (2020).
Congestion in neonatal intensive care units in public hospitals	50% of units experience inefficiency due to patient overcrowding.	Arrieta and Guillén (2017).
Inefficiencies in patient referral system.	High percentage of patients treated in facilities not equipped for their conditions.[9]	Soto (2019).
Coverage noncompliance, high waiting times, and some values of medical performance per hour out of range.	Coverage of up to 86% for certain complex treatments.	EsSalud (2025a).
Deferrals in certain cities are very high.	More than 23% of appointments were postponed (Jan-Mar 2025).	EsSalud (2025b).

Table 3: Issues in patient allocation within Peru's healthcare system.

In Example 5.1, we simulate three groups of patients in three healthcare networks (SIS, EsSalud, EPS). Group 3 consists of individuals who can afford an EPS for high-complexity care. High-complexity care refers to a set of less frequent and more complex health interventions, such as advanced surgical procedures and oncological treatments. Group 2 consists of formal workers who can only use EsSalud for high-complexity care. Note that they are not excluded from affording an EPS, but if they have one, it will be used exclusively for low-complexity care. Group 1 consists of the remaining individuals, including informal workers.

A particular edge case in Group 1 includes wealthy individuals engaged in illegal activities (e.g., drug traffickers or businessman avoiding taxes). These individuals are informal workers but may still afford an EPS. The central planner reasonably operates under the assumption that such cases do not exist. Moreover, it operates assuming no overlaps.[10]

Groups 1 and 3 exhibit significant differences in characteristics, such as socioeconomic status, which increases the cost of mismatching between them. The cost is even higher when there are bureaucratic or legal frictions, as seen in the case of groups 1 and 2, where an EsSalud insured individual cannot be covered simultaneously by SIS, and vice versa (Anaya-Montes and Gravelle,

[9] In 2016, the MINSA (Ministry of Health) reported a shortage of over 47,000 healthcare professionals. Additionally, 36% of medium and high-complexity facilities lacked sufficient personnel, 44% did not have adequate equipment, and 25% had infrastructure deficiencies.

[10] It is important to emphasize that our model is designed to be executed at a specific point in time. Thus, the planner does not seek overlaps, and therefore, they are not enabled in the model.

2024). Our model accounts for this heterogeneity in costs, recognizing that legal constraints impose significantly higher penalties than other sources of mismatching. For instance, while receiving treatment for a simple illness at a high-complexity facility incurs some inefficiency, the cost associated with legal barriers preventing access to appropriate healthcare is substantially greater. Moreover, incorporating penalties and weighted constraints allows the model to capture excess demand effectively. Unlike the solutions in traditional models (see Example 5.2), our model (Example 5.1) assigns almot zero or one to the match between groups 1 and 2.

Example 5.3 highlights the flexibility of our model by introducing $\varepsilon_1, \ldots, \varepsilon_N$ and $\delta_1, \ldots, \delta_L$. In the Peruvian context, the government may prioritize patients from EsSalud due to its connection to formal employment, resulting in higher weights assigned to the constraint related to μ_2. On the other hand, the goal is to prevent SIS from becoming overcrowded while maximizing facilities utilization. This objective is achieved, as the example shows that row 2 and column 1 bear the highest load without exceeding μ_i or δ_j, with respect to the other rows and columns (proportionally to the target mass).

In Example 5.4, we set $\sum_{i=1}^{N} \mu_i > \sum_{j=1}^{L} \nu_j$, which is crucial for an appropriate representation of excess demand, but additionally. Quadratic costs exacerbate the excess demand. The observed effect, due to the intentionally chosen parameters, reflects that almost no one from group 2 is matched. The parameters can certainly be adjusted to obtain more realistic values. The example illustrates how our model effectively captures excess demand, a present phenomenon in the Peruvian reality, see Table 3.

4.2 Education

The education system in Peru is highly complex due to its high degree of decentralization at both the primary and higher education levels. While this decentralization aims to improve educational management, it has generated significant disparities between urban and rural regions (Laveriano, 2010). Only a few subsystems, such as the High-Performance Schools (COAR), maintain a centralized management model, ensuring homogeneous standards (Alcázar and Balarin, 2021). However, despite not being a centralized system - which would make our model better suited - the level of congestion in Lima and its impact on education justify the introduction of a strictly convex structure. Moreover, since not everyone enrolls in school, partly due to geographic and access limitations, the penalties are well-founded.

Specifically, in Peru, infrastructure disparities and access constraints have affected educational equity (Alcázar and Balarin, 2021). Geographic barriers, particularly the Andes and the Amazon rainforest, exacerbate these inequalities by severely limiting accessibility. These mobility constraints directly impact school attendance, contributing to persistent enrollment gaps, especially in secondary education (Alba-Vivar, 2025). Tables 4 and 5 illustrate the evolution of enrollment rates in primary and secondary education, showing gradual improvement but persistent urban-rural disparities.

Area	2021	2022	2023	2024	Variation 2024/2023
National	87.1	91.3	91.3	96.0	4.7%
Urban	87.1	91.2	91.7	96.7	5%
Rural	87.1	91.7	89.8	93.6	3.8%

Table 4: Net enrollment rate in primary education in Peru (2021-2024) (INEI, 2024).

Area	2021	2022	2023	2024	Variation 2024/2023
National	80.1	81.5	86.0	88.7	2.7%
Urban	80.7	81.4	86.7	88.2	1.5%
Rural	78.1	81.8	83.6	90.0	6.4%

Table 5: Net enrollment rate in secondary education in Peru (2021-2024) (INEI, 2024).

A comprehensive study on the impact of congestion on enrollment is provided by Alba-Vivar (2025)[11], highlighting its significance, in line with the findings of Agarwal and Somaini (2019), thus, justifying the relevance of our model. Indeed, congestion is a major issue in Peru's education system, particularly in urban areas. According to World Bank (2024), Lima is one of the most congested cities in Latin America. It suffers from severe traffic bottlenecks that disproportionately affect students from lower-income districts (Alba-Vivar, 2025). When large numbers of students travel from the same location to the same school, the primary roads connecting them become saturated, increasing commuting times.

Thus, the Peruvian education system is characterized by lack of access, excessive demand, and limited supply, combined with sensitivity to physical traffic congestion, in contrast to certain education systems, such as the French one (Eurydice - European Commission, 2024; Ministère de l'Éducation Nationale et de la Jeunesse, 2024), which ensures universal education, and benefits from a much more modern transportation system. Therefore, the model we propose is well-suited to represent this situation (other cities with congestion such as Mumbai, Jakarta or São Paulo (Kikuchi and Hayashi, 2020) could also be studied).

Given these characteristics, our model better aligns with the needs of a central planner in an economic context characterized by traffic congestion and the inability to guarantee education for all. Traditional OT models, by imposing the conditions in (2), do not apply as effectively. Our model is predictive and designed to better fit reality. While there is no social planner in the Peruvian case, in the hypothetical scenario where changes are made to centralize education at different levels, the flexibility of our model becomes an advantage, allowing the social planner to better adapt to real-world conditions.

Example 5.5 is key to understand how our model performs this. We consider four student groups ($N = 4$) and three schools ($L = 3$). The groups represent: wealthy high-achieving students ($i = 1$), poor high-achieving students ($i = 2$), wealthy low-achieving students ($i = 3$), and poor low-achieving students ($i = 4$). School $j = 1$ is top-ranked and expensive, $j = 2$ has an average ranking and a mid-range price, and $j = 3$ is lower-ranked but more affordable. Transportation costs reflect the greater commuting difficulties faced by poor students, who usually use public

[11] Alba found that the 17% reduction in travel time (equivalent to 30 minutes per day) increased the enrollment rate by 6.3%.

transportation that runs along the most congested main avenues (Alba-Vivar, 2025), while linear costs capture preferences, ensuring that better students prefer better schools while weaker students do not, controlling also by monetary cost. The solutions highlight key differences: $\mathcal{P}_{CP}$ introduces quadratic penalties, leading to assignments where students with fewer resources, for whom matching is more costly due to their location and the assigned mode of transportation (as transportation in their area is precarious), are not matched. In contrast, those who have better facilities (positive correlation between socioeconomic status and the quality of transportation) are matched more easily. Moreover, high-achieving wealthy students are never matched with low-cost, low-quality institutions, and low-achieving poor students are never matched with the top, expensive school. Hence, our model predicts the complications arising from transportation costs and the unfortunate reality that education cannot be guaranteed for everyone. For example, Peru's geography excludes certain populations in the highlands and jungle, making it very costly for the central planner to complete the match. In Example 5.5, 70% of the top wealthy students are matched, but only almost 3 out of 10 of the poorer, less top-performing students are matched. In this case, both the linear and quadratic models capture the fact that preferences result in 0 individuals from group $i = 1$ being matched to $j = 3$. However, once again, they do not provide the flexibility for $\sum_j \pi^*_{ij} \neq \mu_i$, required in some contexts: for countries like Peru or others in the region in Latin America, ensuring the equilibrium is not feasible given the constraints.

5 Conclusions

This paper introduces a novel framework for analyzing mismatching, congestion effects, and supply-demand imbalances in developing economies matching markets. Our model extends the classical optimal transport framework by incorporating heterogeneous quadratic regularization and penalty terms for deviations from target allocations. Unlike traditional approaches that impose strict equality constraints, our formulation allows for more realistic depictions of inefficiencies, capturing excess demand, underutilization, and the role of heterogenous congestion costs. We have also analyzed the resulting optimization problem in detail, establishing conditions for the existence and uniqueness of solutions. Furthermore, we propose both analytical and computational methods to effectively compute interior solutions. Our approach provides not only theoretical insights but also practical tools for addressing real-world mismatching and congestion issues.

In summary, our model provides considerable flexibility, allowing for heterogeneity in congestion costs, i.e., some a_{ij} could be very small. Removing restrictions enables a better approximation of the reality in developing countries, where equilibrium equations $\Pi(\mu, \nu)$ do not hold uniformly.

Applying our model to Peru's healthcare sector highlights its ability to explain observed inefficiencies, and provide more flexibility to the central planner when they cannot ensure matching the entire population adequately, which is common in developing or poor countries. The fragmented nature of the public insurance system exacerbates mismatching, leading to suboptimal patient distribution and increased congestion in specific medical centers. Our framework captures these distortions by introducing quadratic congestion costs and penalizing deviations from optimal allocations. Although we have focused on the Peruvian case due to the

aforementioned data availability constraints, the model can be applied to centralized matching situations with heterogeneous congestion costs and excess supply and demand.

Future research could extend this framework to dynamic settings, stochastic environments where parameters evolve over time (e.g., Markov Jump Linear Systems, since at different times of the day, traffic is less sensitive to new cars), and empirical validation using real-world matching data. Determining whether the solution is interior in terms of the parameters is not a trivial matter and remains to be explored. Furthermore, exploring policy implications, such as optimal subsidy structures or decentralized decision-making mechanisms, could provide valuable information to address inefficiencies in public service delivery.

Our model aims to provide central planners with a mathematically flexible tool to approximate allocation problems (without restricting solutions to the integer domain), while allowing for imbalances between supply and demand and incorporating congestion costs. This is particularly relevant in contexts where congestion costs are significant and where, unlike in highly developed countries, ensuring universal access to healthcare and education, as well as preventing the saturation of these services, remains a major challenge.

Appendix A. Proofs

Proof of Lemma 3.1. First, $\det(D) = \prod_{(i,j)\in I\times J} a_{ij} > 0$, $\det(E) = \det(F) = 0$. On the other hand, the eigenvalues of E are non-negative since the eigenvalues of $\mathrm{Diag}(\epsilon_1, ..., \epsilon_N)$ are $\epsilon_i > 0$ and the eigenvalues of $\mathbf{1}_{L\times L}$ belong to $\{0, L\}$. Hence, the products of eigenvalues $\epsilon_i \cdot 0$ and $\epsilon_i \cdot L$ are non-negative, and so, E is positive semi-definite. Similarly, F is positive semi-definite. Thus, A is the sum of a diagonal and positive definite matrix and two other symmetric and semi-positive definite matrices. According to Zhan (2005)[12]

$$\det(A) = \det(D + E + F) \geq \det(D + E) + \det(F) \geq \det(D) + \det(E) + \det(F) > 0. \qquad \blacksquare$$

Proof of Lemma 3.2. Let $A = D + X$, where $X = E + F$. Then,

$$A^{-1} = (D + X)^{-1} = (I - (-1)D^{-1}X)^{-1}D^{-1}.$$

Then, for all $\lambda \in \sigma(D^{-1}X)$, $\lambda \leq \max_{i,j}\{1/a_{ij}\} \cdot (\lambda^E_{\max} + \lambda^F_{\max})$, where $\lambda^E_{\max} = \max_i\{\epsilon_i\} \cdot L$ and $\lambda^F_{\max} = \max_j\{\delta_j\} \cdot N$. Thus, $\left\|D^{-1}X\right\|_\sigma < 1$ [13],

$$(I - (-1)D^{-1}X)^{-1} = \sum_{k=0}^{\infty}(-1)^k(D^{-1}X)^k.$$

Then, by multiplying the series on the right hand side by D^{-1}, the claim follows. $\blacksquare$

Proof of Theorem 3.3. Define

$$\mathcal{E}_n = A^{-1} - S_n = \left(\sum_{k=n+1}^{\infty}(-1)^k(D^{-1}X)^k\right)D^{-1}.$$

On one hand $\|\pi_n - \pi^*\|_\infty = \|\mathcal{E}_n b\|_\infty \leq \|\mathcal{E}_n b\|_2$. On the other hand,

$$\|\mathcal{E}_n b\|_2 \leq \sqrt{NL}\left\|\sum_{k=n+1}^{\infty}(-1)^k(D^{-1}X)^k\right\|_\sigma \left\|D^{-1}b\right\|_\infty \leq \frac{\sqrt{NL}\left\|D^{-1}X\right\|_\sigma^{n+1}\left\|D^{-1}b\right\|_\infty}{1 - \left\|D^{-1}X\right\|_\sigma}.$$

Given $\varepsilon > 0$, let

$$N_\varepsilon = \max\left\{1, \left\lceil\left|\log_{\|D^{-1}X\|_\sigma}\left(\frac{\varepsilon\left(1 - \left\|D^{-1}X\right\|_\sigma\right)}{\sqrt{NL}\left\|D^{-1}b\right\|_\infty}\right)\right|\right\rceil\right\}.$$

For $n \geq N_\varepsilon$, we have $\|\pi_n - \pi^*\|_\infty < \epsilon$. $\blacksquare$

[12] For Minkowski's determinant inequality and its generalizations, see Marcus and Gordon (1970), Artstein-Avidan, Shiri and Giannopoulos, Apostolos and Milman, Vitali D. (2015).

[13] $\|\cdot\|_\sigma$ denotes the spectral norm.

Proof of Theorem 3.4. By using classical properties of Kronecker product, we have

$$\begin{aligned}
A^{-1} &= \frac{I}{\beta} + \left[\sum_{k=1}^{\infty}(-1)^k \left(\frac{1}{\beta}\right)^k (\mathrm{Diag}(\epsilon_1,\dots,\epsilon_N)\otimes \mathbf{1}_{L\times L})^k\right] D^{-1} \\
&= \frac{I}{\beta} + \frac{1}{\beta L}\sum_{k=1}^{\infty}(-1)^k \left(\frac{L}{\beta}\right)^k (\mathrm{Diag}(\epsilon_1^k,\dots,\epsilon_N^k)\otimes \mathbf{1}_{L\times L}) \\
&= \frac{I}{\beta} + \frac{1}{\beta L}\mathrm{Diag}\left(\sum_{k=1}^{\infty}(-1)^k \left(\frac{L\epsilon_1}{\beta}\right)^k,\dots,\sum_{k=1}^{\infty}(-1)^k \left(\frac{L\epsilon_N}{\beta}\right)^k\right)\otimes \mathbf{1}_{L\times L} \\
&= \frac{I}{\beta} + \frac{1}{\beta}\mathrm{Diag}\left(-\frac{\epsilon_1}{\beta+L\epsilon_1},\dots,-\frac{\epsilon_N}{\beta+L\epsilon_N}\right)\otimes \mathbf{1}_{L\times L}.
\end{aligned}$$

■

Proof of Lemma 3.6. The claim certainly holds for $k=1$. Now, assuming it holds for $k\geq 1$, it follows by induction that

$$\max_{1\leq i,j\leq NL}\left\{\left(Y^{k+1}\right)_{ij}\right\} = \max_{1\leq i,j\leq NL}\left\{\sum_{\ell=1}^{NL}\left(Y^k\right)_{i\ell}Y_{\ell j}\right\} \leq \sum_{\ell=1}^{NL}\frac{(2NL)^k}{NL}\cdot 2 = \frac{(2NL)^{k+1}}{NL}.$$

■

Proof Lemma 3.7. We have two distinct possibilities.

Case $k=2m$ with $m\geq 1$ **.** We now proceed by induction. We will manually verify that each $(Y^2)_{ij} = \sum_{\ell=1}^{NL} Y_{i\ell}\cdot Y_{\ell j}$ satisfies the inequality. On the diagonal we have

$$\left(Y^2\right)_{ii} = \sum_{\substack{\ell=1\\ \ell\neq i}}^{NL} Y_{i\ell}\cdot Y_{\ell i} + Y_{ii}\cdot Y_{ii} \geq 4.$$

For $i\neq j$, set

$$\ell_0 = N\left(\left\lceil\frac{j}{N}\right\rceil - \left\lfloor\frac{i-1}{N}\right\rfloor - 1\right) + i.$$

Then $\ell_0 \equiv i \pmod N$ and so $Y_{i\ell_0}\geq 1$. On the other hand,

$$\ell_0 \in \left[N\left(\left\lceil\frac{j}{N}\right\rceil - 1\right)+1, N\left\lceil\frac{j}{N}\right\rceil\right]$$

implies $\lceil \ell_0/N\rceil = \lceil j/N\rceil$. So, $Y_{\ell_0 j}\geq 1$. It follows that

$$\left(Y^2\right)_{ij} = \sum_{\substack{\ell=1\\ \ell\neq \ell_0}}^{NL} Y_{i\ell}\cdot Y_{\ell j} + Y_{i\ell_0}\cdot Y_{\ell_0 j} \geq 1.$$

Assuming $\min_{1\leq i,j\leq NL}\left\{\left(Y^{2m}\right)_{ij}\right\} \geq (NL)^m/NL$ holds for $m\geq 1$, we obtain

$$\min_{1\leq i,j\leq NL}\left\{\left(Y^{2m+2}\right)_{ij}\right\} = \min_{1\leq i,j\leq NL}\left\{\sum_{\ell=1}^{NL}\left(Y^{2m}\right)_{i\ell}\cdot\left(Y^2\right)_{\ell j}\right\} \geq \sum_{\ell=1}^{NL}\frac{(NL)^m}{NL} = \frac{(NL)^{m+1}}{NL}.$$

Case $k = 2m+1$ with $m \geq 1$ **.** We prove this by induction on m starting with the base case Y^3:

$$\left(Y^3\right)_{ij} = \sum_{\ell=1}^{NL} \left(Y^2\right)_{i\ell} \cdot Y_{\ell j} = \sum_{\substack{\ell=1 \\ \ell \neq j}}^{NL} \left(Y^2\right)_{i\ell} \cdot Y_{\ell j} + \left(Y^2\right)_{ij} \cdot Y_{jj} \geq 2.$$

Assume the statement holds for $m \geq 1$, then

$$\min_{1 \leq i,j \leq NL} \left\{ \left(Y^{2m+3}\right)_{ij} \right\} = \min_{1 \leq i,j \leq NL} \left\{ \sum_{\ell=1}^{NL} \left(Y^{2m+1}\right)_{i\ell} \cdot \left(Y^2\right)_{\ell j} \right\} \geq \sum_{\ell=1}^{NL} \frac{(NL)^m}{NL} = \frac{(NL)^{m+1}}{NL}.$$

This completes the proof. ∎

Proof of Lemma 3.8. We write A^{-1} in terms of Y

$$A^{-1} = \frac{1}{\rho} \left(I - \left(\frac{\zeta}{\rho}\right) Y + \sum_{m \geq 1} \left(\frac{\zeta}{\rho}\right)^{2m} Y^{2m} - \sum_{m \geq 1} \left(\frac{\zeta}{\rho}\right)^{2m+1} Y^{2m+1} \right)$$

and apply Lemmas 3.6 and 3.7 to bound the series as follows,

$$\frac{\zeta^2 NL}{\rho^2 - \zeta^2 NL} \leq \sum_{m \geq 1} \left(\frac{\zeta}{\rho}\right)^{2m} \left(Y^{2m}\right)_{ij} \leq \frac{4\zeta^2 N^2 L^2}{\rho^2 - 4\zeta^2 N^2 L^2}$$

$$\frac{\rho^3}{\rho(\rho^2 - \zeta^2 NL)} \leq \sum_{m \geq 1} \left(\frac{\zeta}{\rho}\right)^{2m+1} \left(Y^{2m+1}\right)_{ij} \leq \frac{8\zeta^3 N^2 L^2}{\rho(\rho^2 - 4\rho^2 N^2 L^2)}.$$

Therefore, $(A_{ij})^{-1}$ is bounded from above by

$$\frac{1}{\rho} \left(1 + \frac{4\zeta^2 N^2 L^2}{\rho^2 - 4\zeta^2 N^2 L^2} - \frac{\rho^3}{\rho(\rho^2 - \zeta^2 NL)} \right),$$

and from below by

$$\frac{1}{\rho} \left(-2\left(\frac{\zeta}{\rho}\right) + \frac{\zeta^2 NL}{\rho^2 - \zeta^2 NL} - \frac{8\zeta^3 N^2 L^2}{\rho(\rho^2 - 4\rho^2 N^2 L^2)} \right).$$

From here, (11) follows. ∎

Proof of Theorem 3.9. By triangle inequality,

$$\begin{aligned}
\pi^*_{ij} &\leq ||\pi^*||_\infty \\
&= \max_{\substack{1 \leq i \leq N \\ 1 \leq j \leq L}} \left\{ \left| \sum_{k=1}^{NL} \left(A^{-1}\right)_{(i-1)L+j \quad k} \cdot b_{\lceil k/L \rceil \quad k - L\lfloor (k-1)/L \rfloor} \right| \right\} \\
&\leq \sum_{k=1}^{NL} \max_{\substack{1 \leq i \leq N \\ 1 \leq j \leq L}} \left| \left(A^{-1}\right)_{ij} \right| \cdot \max_{\substack{1 \leq i \leq N \\ 1 \leq j \leq L}} |b_{ij}| \\
&= NL\tilde{C}.
\end{aligned}$$

∎

Proof of Theorem 3.10. Consider Algorithm 1. It is easy to see that each prefix sum of A is invertible. Hence, we can iteratively apply the Sherman-Morrison formula with a rank-1 update at each step. Then, it is clear that Lines 3 and 12 take $O\left(N^2L^2\right)$. First, the number of iterations for the for-loops on Lines 4-7 and 8-11 is $N + L$. We then show that each time we enter any for-loop, the time spent is $O\left(N^2L^2\right)$. Computing $1 + w^T A^{-1} w$ takes $O\left(N^2L^2\right)$, so the only possible optimization is finding the optimal parenthesization for the product $A^{-1}ww^TA^{-1}$. Since there are only five possible ways to parenthesize the expression, we determine by brute force that computing $(A^{-1}w)(w^TA^{-1})$ also takes $O\left(N^2L^2\right)$. This implies the desired time complexity of $O\left((N + L)N^2L^2\right)$. ■

Appendix B. Numerical examples

We define $\mathcal{P}_Q$ as the following optimization problem:

$$\mathcal{P}_Q : \min_{\pi \in \Pi(\mu,\nu)} \sum_{i=1}^{N} \sum_{j=1}^{L} \varphi(\pi_{ij}, \theta_{ij}),$$

where φ is as in (6). $\mathcal{P}_Q$ is a generalization of the quadratic regularization problem in the discrete setting.

Example 5.1. The parameters used for solving $\mathcal{P}_{CP}$ with $d = 5I_{3\times3}$ and $\alpha = 0.5$ are

$$c = \begin{bmatrix} 1 & 50 & 20 \\ 50 & 1 & 20 \\ 20 & 10 & 1 \end{bmatrix}, \; a = \begin{bmatrix} 1 & 5 & 10 \\ 5 & 1 & 2 \\ 10 & 5 & 1 \end{bmatrix}, \; \epsilon = \delta = \begin{bmatrix} 0.3 \\ 0.3 \\ 0.3 \end{bmatrix}, \; \mu = \begin{bmatrix} 100 \\ 50 \\ 20 \end{bmatrix} \text{ and } \nu = \begin{bmatrix} 90 \\ 40 \\ 40 \end{bmatrix}.$$

The optimal solution π^* obtained using Algorithm 1 in Mathematica 14.1 [14] is

$$\pi^* = \begin{bmatrix} 34.7802 & 0.19412 & 1.65935 \\ 0.10148 & 15.6978 & 3.41038 \\ 0.883807 & 0.905689 & 9.65139 \end{bmatrix}.$$

Example 5.2. Using the same parameters as in $\mathcal{P}_{CP}$ but enforcing the marginal constraints $\Pi(\mu, \nu)$ and removing penalization, the optimal solutions to $\mathcal{P}_Q$ and $\mathcal{P}_O$ are

$$\pi^*_{\mathcal{P}_Q} = \begin{bmatrix} 84.275 & 8.84062 & 6.88442 \\ 4.2985 & 30.4206 & 15.2809 \\ 1.42655 & 0.73873 & 17.8347 \end{bmatrix}, \; \pi^*_{\mathcal{P}_O} = \begin{bmatrix} 90 & 0 & 10 \\ 0 & 40 & 10 \\ 0 & 0 & 20 \end{bmatrix}.$$

Example 5.3. Using the same parameters as in $\mathcal{P}_{CP}$ but changing weighting to $\epsilon =$

[14] We also ran `QuadraticOptimization` and verified that the optimal plans coincide.

$\begin{bmatrix}0.4 & 1 & 0.2\end{bmatrix}^T$ and $\delta = \begin{bmatrix}1 & 0.5 & 0.4\end{bmatrix}^T$ leads to

$$\pi^* = \begin{bmatrix}50.7142 & 0.360177 & 1.75142\\ 4.56352 & 22.9044 & 7.05884\\ 2.37786 & 0.873057 & 9.57857\end{bmatrix}.$$

Example 5.4. Modifying the parameters with respect to Example 5.3 as follows

$$a = \begin{bmatrix}1 & 20 & 2\\ 20 & 5 & 2\\ 5 & 2 & 0.5\end{bmatrix},\ \mu = \begin{bmatrix}200\\ 50\\ 10\end{bmatrix} \text{ and } \nu = \begin{bmatrix}100\\ 20\\ 50\end{bmatrix}$$

yields

$$\pi^* = \begin{bmatrix}69.4335 & 1.23953 & 19.2527\\ 1.52132 & 6.95671 & 11.9992\\ 3.14146 & 0.282174 & 7.55862\end{bmatrix}.$$

Example 5.5. Consider the following parameters for $\mathcal{P}_{CP}$ with $d = \mathbf{1}_{4\times 3}$ and $\alpha = 0.5$:

$$c = \begin{bmatrix}0.1 & 1 & 6\\ 0.2 & 1 & 4\\ 4 & 1 & 0.2\\ 8 & 1 & 0.1\end{bmatrix},\ a = \begin{bmatrix}0.5 & 0.5 & 0.5\\ 2 & 2 & 1\\ 0.5 & 0.5 & 0.5\\ 2 & 2 & 1\end{bmatrix},\ \epsilon = \begin{bmatrix}0.2\\ 0.2\\ 0.2\\ 0.2\end{bmatrix},\ \delta = \begin{bmatrix}0.2\\ 0.2\\ 0.2\end{bmatrix},\ \mu = \begin{bmatrix}10\\ 10\\ 10\\ 10\end{bmatrix},\ \nu = \begin{bmatrix}10\\ 20\\ 10\end{bmatrix}.$$

The solution to the optimization problems are[15]

$$\pi^*_{\mathcal{P}_{CP}} = \begin{bmatrix}3.25505 & 3.89254 & 0\\ 1.20974 & 1.39412 & 0.333926\\ 0 & 3.99723 & 2.88862\\ 0 & 1.33717 & 2.17004\end{bmatrix},\ \pi^*_{\mathcal{P}_Q} = \begin{bmatrix}4.18 & 5.82 & 0\\ 3.25571 & 3.69071 & 3.05357\\ 1.25857 & 6.79857 & 1.94286\\ 1.30571 & 3.69071 & 5.00357\end{bmatrix}$$

and

$$\pi^*_{\mathcal{P}_O} = \begin{bmatrix}10 & 0 & 0\\ 0 & 10 & 0\\ 0 & 10 & 0\\ 0 & 0 & 10\end{bmatrix}.$$

[15] In this example, $\pi^*_{\mathcal{P}_{CP}}$ is not an interior solution. Therefore, it is not possible to use Algorithm 1 to solve the problem. Instead, we use `QuadraticOptimization`.

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*.

Agarwal, Nikhil and Somaini, Paulo (2023). Empirical Models of Non-Transferable Utility Matching. In Echenique, F., Immorlica, N., and Vazirani, V. V., editors, *Online and Matching-Based Market Design*, pages 530–551. Cambridge University Press.

Alba-Vivar, F. M. (2025). Opportunity Bound: Transport and Access to College in a Megacity. Accessed on March 16, 2025. Available at `https://drive.google.com/file/d/1-zQu__07sloiK2z7CAvQJ8cp3olDAU60/view?usp=drive_link`.

Alcázar, L. and Balarin, M. (2021). *Evaluación del diseño e implementación de los colegios de alto rendimiento – COAR*. MINEDU and GRADE, Lima.

Alman, J., Duan, R., Vassilevska Williams, V., Xu, Y., Xu, Z., and Zhou, R. (2025). More asymmetry yields faster matrix multiplication. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2005–2039. Society for Industrial and Applied Mathematics.

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

Arrieta, A. and Guillén, J. (2017). Output congestion leads to compromised care in Peruvian public hospital neonatal units. *Health Care Management Science*, 20(2):209–221. Accessed on March 15, 2025. Available at `https://pubmed.ncbi.nlm.nih.gov/26452716/`.

Artstein-Avidan, Shiri and Giannopoulos, Apostolos and Milman, Vitali D. (2015). *Asymptotic Geometric Analysis, Part I*, volume 202 of *Mathematical Surveys and Monographs*. American Mathematical Society.

Beck, J. and Fiala, T. (1981). Integer-making theorems. *Discrete Applied Mathematics*, 3(1):1–8.

Bendezu-Quispe, G., Mari-Huarache, L. F., Álvaro Taype-Rondan, Mejia, C. R., and Inga-Berrospi, F. (2020). Effect of Rural and Marginal Urban Health Service on the Physicians' Perception of Primary Health Care in Peru. *Revista Peruana de Medicina Experimental y Salud Pública*, 37(4):636–644.

Carlier, G., Dupuy, A., Galichon, A., and Sun, Y. (2023). SISTA: Learning Optimal Transport Costs under Sparsity Constraints. *Communications on Pure and Applied Mathematics*, 76(9):1659–1677.

Chiappori, P.-A., McCann, R. J., and Nesheim, L. P. (2010). Hedonic Price Equilibria, Stable Matching, and Optimal Transport: Equivalence, Topology, and Uniqueness. *Economic Theory*, 42(2):317–354.

Data Commons (2025). Population statistics for peru. https://datacommons.org/place/country/PER?utm_medium=explore&mprop=count&popt=Person&hl=es (accessed 18 March 2025).

Doval, L., Echenique, F., Huang, W., and Xin, Y. (2024). Social Learning in Lung Transplant Decision. Accessed on February 21, 2025. Available at arXiv:2411.10584.

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

Dupuy, A. and Galichon, A. (2022). A Note on the Estimation of Job Amenities and Labor Productivity. *Quantitative Economics*, 13:153–177.

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

Echenique, Federico, Joseph Root and Feddor Sandomirskiy (2024). Stable Matching as Transportation. Accessed on February 21, 2025. Available at arXiv:2402.13378.

EsSalud (2025a). Dashboard de indicadores fonafe y tablero estratégico. https://app.powerbi.com/view?r=eyJrIjoiMDQwMDVlOGItNGY5Zi00ZjFjLWEyZDMtYjY1ZjkOMWVjMjcxIiwidCI6IjM0ZjMyNDE5LTFjMDUtNDc1Ni04OTZlLTQ1ZDYzMzcyNjU5YiIsImMiOjR9 (accessed 18 March 2025).

EsSalud (2025b). Tablero de diferimento de citas. https://app.powerbi.com/view?r=eyJrIjoiN2NlMTNmNWEtODA3MS00M2UyLWE3NDAtNjcyYjZjYTQ0MmJmIiwidCI6IjM0ZjMyNDE5LTFjMDUtNDc1Ni04OTZlLTQ1ZDYzMzcyNjU5YiIsImMiOjR9 (accessed 18 March 2025).

Eurydice - European Commission (2024). National education systems: France overview. https://eurydice.eacea.ec.europa.eu/national-education-systems/france/overview (accessed 18 March 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. Accessed on February 21, 2025. Available at arXiv:2107.04700.

González-Sanz, A. and Nutz, M. (2024). Sparsity of Quadratically Regularized Optimal Transport: Scalar Case. Accessed on February 21, 2025. Available at arXiv:2410.03353.

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

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.

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

INEI (2024). Condiciones de Vida en el Perú - Informe Técnico 2024.

Izmailov, A. F. and Solodov, M. V. (2023). Convergence rate estimates for penalty methods revisited. *Computational Optimization and Applications*, 85(3):973–992.

Johns Hopkins University Coronavirus Resource Center (2023). Covid-19 mortality data. `https://coronavirus.jhu.edu/data/mortality` (accessed 18 March 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*.

Laveriano, N. A. (2010). The Decentralization of Education in Peru. *Educación: PUCP*, 19(37):7–26.

Lipton, R. J. (2010). Galactic Algorithms. Gödel's Lost Letter and P=NP, Blog post. Available at `https://rjlipton.com/2010/10/23/galactic-algorithms`.

Lorenz, D. A., Manns, P., and Meyer, C. (2019). Quadratically Regularized Optimal Transport. *Applied Mathematics & Optimization*.

Marcus, M. and Gordon, W. R. (1970). An extension of the Minkowski Determinant Theorem. *Cambridge University Press*.

Merigot, Q. and Thibert, B. (2020). Optimal transport: discretization and algorithms. Accessed on February 21, 2025. Available at arXiv:2003.00855.

Ministère de l'Éducation Nationale et de la Jeunesse (2024). Les chiffres clés du système éducatif. `https://www.education.gouv.fr/les-chiffres-cles-du-systeme-educatif-6515` (accessed 18 March 2025).

Nutz, M. (2024). Quadratically Regularized Optimal Transport: Existence and Multiplicity of Potentials. Accessed on February 21, 2025. Available at arXiv:2404.06847.

Park, J. and Boyd, S. (2018). A semidefinite programming method for integer convex quadratic minimization. *Optimization Letters*, 12:449–518.

Peng, R. and Vempala, S. S. (2024). Solving Sparse Linear Systems Faster than Matrix Multiplication. *Commun. ACM*, 67(7):79–86.

Peyré, G. and Cuturi, M. (2019). Computational Optimal Transport: With Applications to Data Science. *New Foundations and Trends*, 11(5-6):355–607.

Pia, A. D. and Ma, M. (2022). Proximity in Concave Integer Quadratic Programming. *Mathematical Programming*, 194:871–900.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, NJ.

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.

Soto, A. (2019). Barreras para una atención eficaz en los hospitales de referencia del Ministerio de Salud del Perú: atendiendo pacientes en el siglo XXI con recursos del siglo XX. *Revista Peruana de Medicina Experimental y Salud Pública*, 36(2):304.

Strassen, V. (1969). Gaussian elimination is not optimal. *Numerische Mathematik*, 13(4):354–356.

Vassilevska, V. (2015). CS367 Algebraic Graph Algorithms - Lectures 1 and 2 on Matrix Multiplication and Matrix Inversion. Scribed by Jessica Su. Available at `https://theory.stanford.edu/~virgi/cs367/lecture1.pdf`.

Velásquez, A. (2020). *Consideraciones éticas del aseguramiento universal de salud en el Peru*. Antonio Ruiz de Montoya University.

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. Accessed on February 21, 2025. Available at arXiv:2410.03425 .

World Bank (2020). Health at a glance: Latin america and the caribbean 2020. `https://documents1.worldbank.org/curated/en/383471608633276440/pdf/Health-at-a-Glance-Latin-America-and-the-Caribbean-2020.pdf` (accessed 18 March 2025).

World Bank (2024). Modernizing traffic management in lima with world bank support. `https://www.bancomundial.org/es/news/press-release/2024/10/15/modernizing-traffic-management-in-lima-with-world-bank-support` (accessed 18 March 2025).

Zhan, S. (2005). On the determinantal inequalities. *Journal of Inequalities in Pure and Applied Mathematics*, 6(4).