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

August 12, 2025

Abstract

We propose a new model that transforms the classical discrete optimal transport framework by incorporating heterogeneous congestion costs and replacing traditional equality constraints with weighted penalization terms. The resulting formulation is a strictly convex optimization problem that better captures demand–supply imbalances in economic matching contexts and the congestion phenomenon. We first introduce the model and establish existence and uniqueness of the optimal transport plan under general conditions. For interior solutions, we present two analytical methods—based on the Neumann series expansion and the Sherman–Morrison formula—and develop a practical $O((N+L)N^2L^2)$ algorithm for computing the optimum. We then address the case of infinitely many types, corresponding to optimal transport on measure spaces, absolutely continuous with respect to Lebesgue, and prove existence and uniqueness under reasonable assumptions via infinite-dimensional optimization methods. Finally, we illustrate the applicability of our framework with examples from Peru's health and education sectors, showing how it yields allocation patterns that differ from classical approaches and provide more accurate predictions.

Keywords: Optimal transport, Matching models, Quadratic regularization, Convex optimization, Infinite-dimensional optimization, Sherman–Morrison formula.
JEL classifications: C61, C62, C78, D04.

We gratefully acknowledge insightful discussions with Professors Federico Echenique, Amílcar Vélez and César Martinelli, and also with Carlos Cosentino, whose valuable feedback significantly improved this work and provided key insights, 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, density, uniqueness of the solution or computational advantages (Peyré and Cuturi, 2019). Notable examples include entropic regularization (Lorenz et al., 2021; Dupuy et al., 2019; Merigot and Thibert, 2020; Galichon, 2021) and quadratic regularization (Lorenz et al., 2019; González-Sanz and Nutz, 2025; Wiesel and Xu, 2024; Nutz, 2025). 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 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 (2025), 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 crucial because congestion is a real and significant phenomenon; heterogeneity allows costs to vary across pairs; marginal constraints are often not satisfied in practice; and the model we introduce provides greater flexibility for the social planner. In developing countries such as Peru, India, and Brazil, the health and education sectors face severe frictions and congestion due to insufficient infrastructure, leading to excess demand, service shortages, and high economic costs. For instance, in Peru, these problems have been reflected in the world's highest per capita COVID-19 mortality rate and in annual losses equivalent to 1.8% of GDP due to traffic congestion.

The model presented in this paper, formulated from a social planner's perspective, incorporates these elements and results in the formulation of a convex optimization problem, both in the finite-dimensional discrete case and in the infinite-dimensional setting. This framework allows for the inclusion of congestion costs while explicitly accounting for persistent excess demand in different institutional contexts. This contrasts with developed countries such as France or Switzerland, where efficient infrastructure and policies largely mitigate such frictions. Our approach introduces a strictly convex cost structure that remains analytically tractable under mild assumptions.

The paper is organized as follows. We first introduce the notation and preliminary concepts

for the discrete case. Then, in Section 2, we present the model and analyzes its theoretical properties: we establish existence and uniqueness of a solution, study the case of interior solutions, and employ the Neumann series approximation to derive closed-form expressions for the optimal solution under specific assumptions, which also facilitates comparative statics analysis. We then develop an algorithm based on the Sherman–Morrison formula to compute interior solutions in $O((N+L)N^2L^2)$ time, improving upon the upper bounds of both standard and more recent methods in the literature. Section 3 extends the analysis to the case of infinitely many types, proving existence and uniqueness in this setting. Finally, Section 4 provides illustrative examples showing how the discrete formulation can be applied in practice, with an empirical focus on the Peruvian health and education sectors.

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, which is given by means of a function $C : \mathbb{R}_+^{N,L} \times \mathbb{R}^P \to \mathbb{R}$ depending on the matching $\pi = [\pi_{ij}] \in \mathbb{R}_+^{N,L}$[1], and a vector of parameters $\theta \in \mathbb{R}^P$. Moreover, the central planner must 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, $C(\pi, \theta) = \sum_{i,j} c_{ij}\pi_{ij}$. In such model, the marginal cost of

[1] 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. This issue will be addressed again later.

matching one more individual from i with j is always the same, regardless of how many people are already matched. 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 (Ambrosio et al., 2024):

$$\mathcal{P}_\infty : \inf \left\{ \pi \in \{\text{Couplings between } \mu \text{ and } \nu\} : \int_{X \times Y} c(x,y) d\pi(x,y) \right\}.$$

The set of couplings between μ and ν—probability measures on X and Y, respectively—is

$$\begin{aligned} \hat{\Pi}(\mu,\nu) = \{\pi \in \Delta(X \times Y) : \pi(A \times Y) = \mu(A), \\ \pi(X \times B) = \nu(B), \quad \forall A \subset X,\ B \subset Y \text{ Borel}\}. \end{aligned} \tag{3}$$

In (3), Δ denotes probability distribution. It is standard to assume a cost $c : X \times Y \to \mathbb{R}_+ \cup \{\infty\}$ that is lower semicontinuous. The problem $\mathcal{P}_\infty$ has been extensively studied in the literature, in particular regarding existence of optimal couplings under these assumptions. In Section 3 we extend our framework to the case of distributions that are absolutely continuous with respect to Lebesgue measure.

Regularization problems in optimal transport have been studied for both the entropic and quadratic cases. The entropic regularization problem (Carlier et al., 2023; Peyré and Cuturi, 2019) is

$$\min_{\pi \in \Pi(\mu,\nu)} \sum_{i=1}^{N} \sum_{j=1}^{L} c_{ij}\pi_{ij} + \sigma \pi_{ij} \ln(\pi_{ij}), \tag{4}$$

with $\sigma > 0$. Given the strict convexity of $f(x) = x \ln x$, and the fact that $f(0) = 0$ and $\lim_{x \downarrow 0^+} f'(x) = -\infty$, the solution is interior, i.e. $\pi^*_{ij} > 0$. With respect to the quadratic regularization problem, the program addressed is

$$\min_{\pi \in \Pi(\mu,\nu)} \sum_{i=1}^{N} \sum_{j=1}^{L} c_{ij}\pi_{ij} + \frac{\varepsilon}{2}||\pi||_2^2, \tag{5}$$

with $\varepsilon > 0$. Unlike the problem (4), in the case of (5), interior solutions cannot be guaranteed. The literature has studied various properties of these models, including existence and uniqueness of solutions, methods for computing them either analytically or numerically, as well as extensions to the continuous case, where X and Y are not discrete and finite sets, but rather infinite continuous spaces (Villani, 2009; Dupuy and Galichon, 2014; Galichon, 2016; Wiesel and Xu, 2024).

In the following section, and in the spirit of these contributions, we introduce our model in

the discrete case, which incorporates the quadratic term but replaces the marginal constraints with penalization terms, thereby providing greater flexibility to the problem and yielding an analytically tractable solution for the interior case, as we shall see in detail.

2 The model

The model we propose results in the following finite dimensional quadratic optimization problem:

$$\mathcal{P}_{CP}: \min_{\pi_{ij}\geq 0} \underbrace{\left\{ \underbrace{\alpha \sum_{i=1}^{N}\sum_{j=1}^{L} \varphi(\pi_{ij};\theta_{ij})}_{\text{Matching direct cost.}} + \underbrace{(1-\alpha)\left[\sum_{i=1}^{N} \epsilon_i \left(\sum_{j=1}^{L}\pi_{ij}-\mu_i\right)^2 + \sum_{j=1}^{L}\delta_j\left(\sum_{i=1}^{N}\pi_{ij}-\nu_j\right)^2\right]}_{\text{Costs of social objectives.}} \right\}}_{F(\pi;\theta,\alpha,\epsilon,\delta,\mu,\nu).} \tag{6}$$

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

$$\varphi(\pi_{ij};\theta_{ij}) = d_{ij} + c_{ij}\pi_{ij} + a_{ij}\pi_{ij}^2,\ \theta_{ij} = (d_{ij}, c_{ij}, a_{ij}) \in \mathbb{R}^3_{++}. \tag{7}$$

The objective function in (6) 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 ϵ_i and δ_j, 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 i to a single j congests the access route. The coefficient $a_{ij} > 0$ captures heterogeneity[2], while the quadratic term represents the previously described phenomenon[3]. 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

[2]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.

[3]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.

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 $n \geq 1$ 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 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 therefore replaced by the penalty term $\delta_j \left[\sum_{i=1}^{N} \pi_{ij} - \nu_j\right]^2, \quad \delta_j > 0$, and the constraint $\sum_{j=1}^{L} \pi_{ij} = \mu_i$ is replaced by $\epsilon_i \left[\sum_{j=1}^{L} \pi_{ij} - \mu_i\right]^2, \quad \epsilon_i > 0$. 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 and be as close as possible to its targets.

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)[4]. 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 (7), despite its practical relevance, the term d_{ij}, representing fixed costs, does not influence the resolution of the problem. For this reason, from now, 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: the parameters a_{ij} are not all the same.

Having now established the model, which, to the best of our knowledge, is new in the literature[5], 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.

Existence and uniqueness: Regarding the existence of a solution to $\mathcal{P}_{CP}$, in order to apply

[4]This significantly simplifies our analysis and does not affect the model's logic. Moreover, the behavior of the objective function ensures that it does not diverges to $-\infty$ when $\pi_{ij} \to \infty$.

[5]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.

Weierstrass theorem to overcome the potential issue that the optimization is carried over an 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[6]. 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 (6) yield

$$\frac{\partial F}{\partial \pi_{ij}} = \frac{1}{2} \left(\varphi'(\pi_{ij}^*; \theta_{ij}) + 2\epsilon_i \left(\sum_{\ell=1}^{L} \pi_{i\ell}^* - \mu_i \right) + 2\delta_j \left(\sum_{k=1}^{N} \pi_{kj}^* - \nu_j \right) - \gamma_{ij}^* \right) = 0, \ \forall \ (i,j) \in I \times J. \quad (8)$$

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 (8), we have $\nabla F(\pi^*) = 0$. This set of equations can be written in the compact form

[6] 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

$$A = \underbrace{\text{Diag}(a_{11}, a_{12}, \ldots, a_{NL})}_{D} + \underbrace{\text{Diag}(\epsilon_1, \ldots, \epsilon_N) \otimes \mathbf{1}_{L\times L}}_{E} + \underbrace{\mathbf{1}_{N\times N} \otimes \text{Diag}(\delta_1, \ldots, \delta_L)}_{F}, \tag{9}$$

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

Proof. First, $\det(D) = \prod_{(i,j)\in I\times J} a_{ij} > 0$, and $\det(E) = \det(F) = 0$. On the other hand, the eigenvalues of E are non-negative since the eigenvalues of $\text{Diag}(\epsilon_1, ..., \epsilon_N)$ 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)[7]:

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

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.

2.1 Neumann's series approach

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

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

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 2.2. *Under Assumption 1, the following holds*

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

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

Proof. 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, $\|D^{-1}X\|_\sigma < 1$ [8],

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

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

Theorem 2.3. *Under Assumption 1,* $\lim_{n\to\infty}\pi_n = \pi^* = A^{-1}b$, *where*

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

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

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

Given $\varepsilon > 0$, let

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

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

2.2 Special cases

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

2.2.1 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 $1 \leq i \leq N$.

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

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

Corollary 2.4. *Under Assumptions 2 and 3,*

$$A^{-1} = \frac{I}{\beta} + \frac{1}{\beta} Diag\left(-\frac{\epsilon_1}{\beta + L\epsilon_1}, \ldots, -\frac{\epsilon_N}{\beta + L\epsilon_N}\right) \otimes \mathbf{1}_{L\times L}. \tag{10}$$

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

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

■

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

$$\pi_{ij}^* = \frac{b_{ij}}{\beta} - \sum_{\ell=1}^{L} \frac{b_{i\ell}\epsilon_i}{\beta^2 + L\epsilon_i\beta}, \tag{11}$$

provided that the right-hand side of (11) is positive.

Proof. This result follows directly from the computation of $A^{-1}b$ by using (10). ■

2.2.2 Equal weighting and identical strictly 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 $D = \rho I$ (classical quadratic regularization) 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 2.6, 2.7 and 2.8 are used to establish Theorem 2.9.

Lemma 2.6. *Let $k \geq 1$ be a positive integer. Then*

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

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

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

Lemma 2.7. *Let $k \geq 2$ be a positive integer.Then*

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

Proof. We have two distinct possibilities.

Case $k = 2m$ with $m \geq 1$ **.** We now proceed by induction. We will manually verify that each $\left(Y^2\right)_{ij} = \sum_{\ell=1}^{NL} Y_{i\ell}\cdot Y_{\ell j}$ 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. ■

Lemma 2.8. *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*,*

$$C_1(N, L, \zeta, \rho) \leq (A^{-1})_{ij} \leq C_2(N, L, \zeta, \rho), \tag{12}$$

where

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

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

Proof. 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 2.6 and 2.7 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, (12) follows. ■

Theorem 2.9. *Under Assumptions 1 and 4, it follows that* $\pi^*_{ij} \leq NL\tilde{C}$, *for all* $(i,j) \in I \times J$, *where*

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

Proof. By triangle inequality,

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

■

Theorem 2.9 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.

2.3 Algorithm for computing π^*

We now provide an efficient algorithm to compute $\pi^* \in \mathbb{R}^{NL}_{++}$. This is established in Theorem 2.10. First, let us rewrite matrix A given in (9) as follows:

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

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 $A^{-1} \leftarrow \mathrm{Diag}(1/a_{11}, \ldots, 1/a_{NL}) \in \mathbb{R}^{NL,NL}$
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 $v^{(j)} \coloneqq \delta_j^{1/2} \mathbf{1}_{N\times 1} \otimes \mathbf{e}_j$
    $A^{-1} \leftarrow A^{-1} - \dfrac{A^{-1} v^{(j)} v^{(j)T} A^{-1}}{1 + v^{(j)T} A^{-1} v^{(j)}}$ via Sherman-Morrison formula
end for
return $A^{-1} b$

Theorem 2.10. *For interior solutions π^*, Algorithm 1 computes π^* in $O((N+L)(NL)^2)$ time.*

Proof. 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 $1+w^TA^{-1}w$ 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 $O\left((N+L)N^2L^2\right)$. ■

Table 1: Algorithms for solving our linear system.

Time	Sparse A [9]	Galactic [10]	Reference
$O((NL)^3)$	No	No	Folklore
$O((NL)^{2.81})$	No	No	Strassen (1969)
$O((N+L)(NL)^2)$	No	No	**This paper**
$O((NL)^{2.371339})$	No	Yes	Alman et al. (2025)
$O((NL)^{2.331645})$	Yes	Yes	Peng and Vempala (2024)

Solving the linear system $A\pi = b$ efficiently has seen significant progress over the past decade for specific classes of systems, such as symmetric diagonally dominant systems Koutis et al. (2012) or when A is sparse Peng and Vempala (2024). Most recent works focus on approximation algorithms that compute A^{-1} efficiently. However, these approaches do not apply to the general case, and some of the resulting algorithms are not exact.

Matrix inversion can be reduced to matrix multiplication with equivalent runtime for many algorithms.[11] Alman et al. (2025) provides the best known bounds for matrix multiplication based on the laser method Strassen (1986); Coppersmith and Winograd (1990); Davie and Stothers (2013); Vassilevska (2012); however, these bounds are currently impractical for real-world use.

Strassen-like algorithms Strassen (1969); Winograd (1971); Pan (1982); Karstadt and Schwartz (2017) have been practically implemented, benchmarked, and employed, depending on the matrix size and hidden constants in their theoretical complexity Huang (2018). When $L \in \Theta(N)$, our algorithm achieves the tightest upper bound compared to inversion methods derived from Strassen-like matrix multiplication algorithms.

[9] Assume A is sparse if it has $\widetilde{O}(NL)$ nonzero entries.

[10] "Galactic" refers to an algorithm wonderful in its asymptotic behavior, but is never used to actual compute anything. See R. J. Lipton, *Galactic Algorithms*, *Gödel's Lost Letter and P=NP* blog, October 2010. Available at `https://rjlipton.com/2010/10/23/galactic-algorithms`.

[11] This reduction is discussed in Vassilevska Williams's lecture notes: *CS367 Algebraic Graph Algorithms, Lectures 1 and 2*, Stanford University, 2015. Scribed by Jessica Su. Available at `https://theory.stanford.edu/~virgi/cs367/lecture1.pdf`.

2.4 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:

$$\left[\frac{\partial\pi^*_{ij}}{\partial\theta_{k\ell}}\right] = -A^{-1}_{(\overline{\theta},\epsilon,\delta,\mu,\nu)}[I_{NL\times NL} \,|\, 2\text{Diag}(\pi^*_{11}, \cdots, \pi^*_{NL})]. \tag{13}$$

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

$$\left[\frac{\partial\pi^*_{ij}}{\partial\theta_{k\ell}}\right] = \left[\frac{\partial\pi^*_{ij}}{\partial c_{k\ell}} \,\middle|\, \frac{\partial\pi^*_{ij}}{\partial a_{k\ell}}\right] \simeq -\left[D^{-1} - (D^{-1})^2X \,|\, A^{-1}_{\Pi,2}\right], \tag{14}$$

where $A^{-1}_{\Pi,2}$ consists of multiplying column ij of $D^{-1} - (D^{-1})^2X$ by π^*_{ij}. From (14), 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 $j \neq \ell$. Thus, we conclude from (14) 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 (11).

3 Infinite types

In this section, we extend our model to the case of infinitely many types. This scenario arises frequently in the literature—for example, in the labor market, where we match worker types characterized by a productivity level taking values in a compact subset of the real line with firms whose technology is aligned with productivity, and whose technological level also takes values in a compact subset of the real line.

Let X and Y be separable, compact metric spaces endowed with Borel probability measures μ and ν. A transport plan π is a finite positive measure on $X \times Y$ that we assume absolutely continuous with respect to a reference product measure (e.g., Lebesgue on domains of $\mathbb{R}^d$), with Radon–Nikodym density $f = \frac{d\pi}{d(dx\,dy)} \in L^2(X \times Y)$. Denote the X- and Y-marginals of f by

$$F(x) \coloneqq \int_Y f(x,y)\,dy, \qquad G(y) \coloneqq \int_X f(x,y)\,dx.$$

We consider a heterogeneous quadratic cost $a : X \times Y \to (0,\infty)$ and a base cost $c : X \times Y \to \mathbb{R}$. As we did before, in order to relax the marginal constraints, we penalize deviations from the target densities f_μ and f_ν of μ and ν. In the most general (non-separable) form, we allow position-dependent weights $\epsilon \in L^\infty(X), \delta \in L^\infty(Y)$. The continuous penalized problem reads

$$\inf_{f \in L^2_+(X\times Y)} \left\{ \begin{aligned} &\int_{X\times Y} \left(a(x,y)\,f(x,y)^2 + c(x,y)\,f(x,y)\right) dx\,dy \\ &\quad + \int_X \epsilon(x)\left(F(x) - f_\mu(x)\right)^2 dx + \int_Y \delta(y)\left(G(y) - f_\nu(y)\right)^2 dy \end{aligned} \right\}. \tag{15}$$

The special case in which $\varepsilon = \delta = 1$ is given by

$$\inf_{\pi \in M_+(X\times Y)} \Big\{ \int_{X\times Y} \Big(a(x,y) f^2(x,y) + c(x,y) f(x,y)\Big)\,dx\,dy \\ + \|F - f_\mu\|^2_{L^2(X)} + \|G - f_\nu\|^2_{L^2(Y)} \Big\}.$$

We now prove the existence and uniqueness of a solution to the problem under mild assumptions. For this purpose, we rely on the following lemma from Brezis (2010).

Lemma 3.1. *Let E be a reflexive Banach space, and let $A \subset E$ be a closed and convex subset. Suppose that $\varphi : A \to (-\infty, \infty]$ is a convex, lower semicontinuous function that is not identically $+\infty$. If either of the following holds:*

1. *A is bounded, or*
2. $\lim\limits_{\substack{x\in A \\ \|x\|\to\infty}} \varphi(x) = \infty,$

then φ attains its infimum on A; that is, there exists $x^* \in A$ such that

$$\varphi(x^*) = \inf_{x\in A} \varphi(x).$$

Assumption 5. Let X, Y be compact metric spaces, each endowed with a probability Borel measure (normalized Lebesgue on subsets of $\mathbb{R}^d$). Let

$$a(x,y) \geq a_0 > 0 \text{ a.e.}, \qquad c \in L^2_+(X \times Y),$$

and let the marginal weights satisfy

$$\epsilon \in L^\infty(X), \quad \epsilon(x) \geq 0 \text{ a.e.}, \qquad \delta \in L^\infty(Y), \quad \delta(y) \geq 0 \text{ a.e.}$$

Compact type spaces X and Y mean heterogeneity is economically bounded (e.g., skills, hospital capacity, school quality lie within feasible ranges), so total mass and aggregates are finite. Normalizing does not change optimal policies. The weights $\epsilon \in L^\infty(X)$ and $\delta \in L^\infty(Y)$ represent the planner's tolerance for unmet demand or excess supply across types; bounded weights keep the implied shadow prices of marginal imbalances finite and prevent penalties from dominating the congestion trade-off. The curvature $a(x,y) \geq a_0 > 0$ encodes that adding traffic to any match is *always* costly at the margin (congestion is pervasive), which discourages over-concentration and yields well-behaved, strictly convex objectives. Finally, $c \geq 0$ since it represents a cost.

Remark 3.2. The sub-space $L^2_+(X \times Y)$ is convex and closed in $L^2(X \times Y)$ with respect to the topology induced by the L^2 norm. Moreover, L^2 is a reflexive Banach space (since it is a Hilbert space).

Lemma 3.3. *Under Assumption 5, the functional* $\Phi : L^2_+(X \times Y) \to (-\infty, \infty]$ *defined by*

$$\Phi : f \to \mathbb{E}[af^2 + cf] + ||\sqrt{\epsilon}(F - f_\mu)||^2_{L^2(X)} + ||\sqrt{\delta}(G - f_\nu)||^2_{L^2(Y)} \tag{16}$$

is convex. In particular, strict convexity holds whenever $f > 0$ *on a set of full measure.*

Proof. Convexity. Write $\Phi = \Phi_1 + \Phi_2 + \Phi_3$ with

$$\Phi_1(f) := \int_{X\times Y} (af^2 + cf), \;\; \Phi_2(f) := \int_X \epsilon(x) \left(F(x) - f_\mu(x)\right)^2 dx, \;\; \Phi_3(f) := \int_Y \delta(y) \left(G(y) - f_\nu(y)\right)^2 dy.$$

(i) The map $f \mapsto \int af^2$ is strictly convex because, for $0 < \theta < 1$ and $f \neq g$, under the hypothesis,

$$\begin{aligned} \int a(\theta f + (1-\theta)g)^2 &= \theta \int af^2 + (1-\theta) \int ag^2 - \theta(1-\theta) \int a(f-g)^2 \\ &< \theta \int af^2 + (1-\theta) \int ag^2. \end{aligned}$$

The map $f \mapsto \int cf$ is linear. Hence Φ_1 is strictly convex.

(ii) Let $I : L^2(X \times Y) \to L^2(X)$ and $J : L^2(X \times Y) \to L^2(Y)$ be the averaging operators $I(f) = F$, $J(f) = G$. Both I and J are bounded linear operators with $\|I\| \leq 1$ and $\|J\| \leq 1$ on

probability measure spaces:

$$\|I(f)\|_{L^2(X)}^2 = \int_X \Big| \int_Y f(x,y)\, d\mathscr{L}(y) \Big|^2 \, d\mathscr{L}(x) \tag{17}$$

$$\leq \mathscr{L}(Y) \int_X \int_Y |f(x,y)|^2 \underbrace{d\mathscr{L}(y)\, d\mathscr{L}(x)}_{\doteq dxdy} \tag{18}$$

$$= \|f\|_{L^2(X\times Y)}^2. \tag{19}$$

Restricting to $L^2_+(X \times Y)$ doesn't change this result.

Then, since $u \mapsto \|u - h\|_{L^2}^2$ is convex for fixed h, and $\epsilon, \delta \geq 0$ with $\epsilon \in L^\infty(X)$, $\delta \in L^\infty(Y)$,

$$\Phi_2(f) = \big\| \epsilon^{1/2} (\, I(f) - f_\mu \,) \big\|_{L^2(X)}^2, \qquad \Phi_3(f) = \big\| \delta^{1/2} (\, J(f) - f_\nu \,) \big\|_{L^2(Y)}^2$$

are convex as compositions of convex maps with bounded linear operators. Therefore Φ is convex; moreover, it is strictly convex due to Φ_1. ■

Lemma 3.4. *Under Assumption 5, the functional $\Phi : L^2_+(X \times Y) \to (-\infty, \infty]$ defined in (16) is lower-semicontinuous.*

Proof. Let $f_n \to f$ in $L^2(X \times Y)$.

Step 1. Let $M_g : L^2 \to L^2$ denote multiplication by g: $M_g(u) = gu$. If $g \in L^\infty$, then M_g is a bounded linear operator with $\|M_g u\|_{L^2} \leq \|g\|_{L^\infty} \|u\|_{L^2}$; hence $\|M_g\|_{L^2 \to L^2} = \|g\|_{L^\infty}$. Since $\epsilon, \delta \geq 0$ and $\epsilon, \delta \in L^\infty$, we also have $\epsilon^{1/2}, \delta^{1/2} \in L^\infty$ with $\|\epsilon^{1/2}\|_\infty = \|\epsilon\|_\infty^{1/2}$ and $\|\delta^{1/2}\|_\infty = \|\delta\|_\infty^{1/2}$. Let $I(f) = \int_Y f(\cdot, y)\, dy$ and $J(f) = \int_X f(x, \cdot)\, dx$. On probability spaces, $\|I\|_{L^2 \to L^2} \leq 1$ and $\|J\|_{L^2 \to L^2} \leq 1$; more generally $\|I\| \leq \mathscr{L}(Y)^{1/2}$ and $\|J\| \leq \mathscr{L}(X)^{1/2}$ by Cauchy–Schwarz/Fubini. Thus $I(f_n) \to I(f)$ in $L^2(X)$ and $J(f_n) \to J(f)$ in $L^2(Y)$ whenever $f_n \to f$ in $L^2(X \times Y)$. Subtraction by a fixed function is continuous: since $f_\mu \in L^2(X)$ and $f_\nu \in L^2(Y)$ are fixed,

$$\|[I(f_n) - f_\mu] - [I(f) - f_\mu]\|_{L^2(X)} = \|I(f_n) - I(f)\|_{L^2(X)} \to 0,$$

$$\|[J(f_n) - f_\nu] - [J(f) - f_\nu]\|_{L^2(Y)} = \|J(f_n) - J(f)\|_{L^2(Y)} \to 0.$$

Combining with the multiplier bound gives

$$\|\epsilon^{1/2}(I(f_n) - f_\mu) - \epsilon^{1/2}(I(f) - f_\mu)\|_{L^2(X)} \leq \|\epsilon^{1/2}\|_\infty \, \|I(f_n) - I(f)\|_{L^2(X)} \to 0,$$

$$\|\delta^{1/2}(J(f_n) - f_\nu) - \delta^{1/2}(J(f) - f_\nu)\|_{L^2(Y)} \leq \|\delta^{1/2}\|_\infty \, \|J(f_n) - J(f)\|_{L^2(Y)} \to 0.$$

Therefore $\Phi_2(f_n) \to \Phi_2(f)$ and $\Phi_3(f_n) \to \Phi_3(f)$, i.e., Φ_2 and Φ_3 are continuous (hence l.s.c.).

Step 2 (lower semicontinuity of Φ_1). The linear part is continuous:

$$\Big| \int c(f_n - f) \Big| \leq \|c\|_{L^2(X\times Y)} \, \|f_n - f\|_{L^2(X\times Y)} \to 0.$$

For the quadratic part, pass to a subsequence (not relabeled) with $f_n(x,y) \to f(x,y)$ a.e.; since

$a \geq 0$,

$$\int af^2 \;\leq\; \liminf_{n\to\infty} \int af_n^2$$

by Fatou's lemma. Therefore Φ_1 is lower semicontinuous. Summing the parts yields

$$\Phi(f) \;\leq\; \liminf_{n\to\infty} \Phi(f_n),$$

so Φ is lower semicontinuous on $L^2(X\times Y)$. As $L^2_+(X\times Y)$ is closed in $L^2(X\times Y)$, the restriction of Φ to L^2_+ is also lower semicontinuous. ■

Theorem 3.5. *Under Assumption 5, Problem (15) admits a unique minimizer in $L^2_+(X\times Y)$.*

Proof. By Lemmas 3.3 and 3.4, the objective Φ is strictly convex and lower semicontinuous on $L^2_+(X\times Y)$. Since $a(x,y)\geq a_0>0$ a.e. and the marginal penalty terms are nonnegative,

$$\Phi(f) \;\geq\; a_0\,\|f\|_{L^2}^2.$$

Existence of a minimizer then follows from Lemma 3.1 (case 2), and uniqueness follows from strict convexity. ■

Remark 3.6. If we remove the assumption $a(x,y)\geq a_0>0$ but restrict the feasible set to

$$\mathcal{A}_M \;=\; \{\, f\in L^2_+(X\times Y) \;:\; 0\leq f\leq M \text{ a.e.}\,\},$$

then $\mathcal{A}_M$ is nonempty, closed, convex, and bounded in the reflexive space L^2, hence weakly compact. Since Φ is convex and weakly lower semicontinuous (quadratic/linear terms and the marginal penalties composed with the bounded linear maps I, J), Φ attains a minimum on $\mathcal{A}_M$ by Lemma 3.1 (bounded-set case). *Uniqueness* is generally not guaranteed without $a_0>0$; it holds if additional conditions ensure strict convexity of Φ (e.g., restoring a uniform lower bound on a).

Remark 3.7. The first–order condition (variational argument $d/dt(\Phi(f+tg)|_{t=0}=0$) on the interior yields

$$\begin{aligned} B[f](x,y) &= -\frac{1}{2a(x,y)}\Big(c(x,y)+2\,\epsilon(x)\big(F(x)-f_\mu(x)\big)+2\,\delta(y)\big(G(y)-f_\nu(y)\big)\Big)\\ &= \alpha(x,y) \;-\; K[f](x,y) = f. \end{aligned}$$

where

$$\alpha(x,y) := \frac{\epsilon(x)f_\mu(x)+\delta(y)f_\nu(y)-\frac{1}{2}c(x,y)}{a(x,y)}, \qquad K[f](x,y) := \frac{\epsilon(x)F(x)+\delta(y)G(y)}{a(x,y)}.$$

On probability spaces, $I: f\mapsto F$ and $J: f\mapsto G$ satisfy $\|I\|,\|J\|\leq 1$. Hence,

$$\|B(f)-B(g)\|_{L^2} \leq \left\|\tfrac{\epsilon}{a}\right\|_{L^\infty}\|I(f-g)\|_{L^2} + \left\|\tfrac{\delta}{a}\right\|_{L^\infty}\|J(f-g)\|_{L^2} \leq \Lambda\,\|f-g\|_{L^2},$$

where $\Lambda := \left\|\epsilon/a\right\|_\infty + \left\|\delta/a\right\|_\infty$. Thus B is a strict contraction whenever $\Lambda < 1$, yielding a unique fixed point by Banach's Fixed Point Theorem. Similarly to the discrete case, we can write, under the assumption that $||K|| \leq \Lambda < 1$,

$$f^\star = (I + K)^{-1}\alpha = \sum_{n=0}^{\infty}(-1)^n K^n \alpha,$$

which is a candidate optimum provided that $f^\star$ is ex-post non–negative.

In what follows, we return to the discrete model in order to explore applications that arise once both the structural parameters and the optimal matching π^* are estimated.

4 Examples and applications

In this section, we illustrate practical applications of our model for the discrete setting (Section 2). Given the available data on Peru's health and education sectors, our examples focus on these areas. We argue that the model provides the flexibility needed by a social planner, especially in developing countries like Peru, where structural problems, limited infrastructure, and high access costs to public services are significant. We emphasize that our model is normative rather than descriptive.

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.

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).

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

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 clear that Peru's healthcare system faces serious challenges, including service inefficiencies, congestion costs, and saturation. Our model captures these features more effectively than traditional matching models and can be used to identify critical areas for improvement. In particular, it enables the optimization of healthcare demand coverage and the reduction of congestion costs by analyzing the impact of parameters on π^*. Achieving this requires reliable parameter estimation, in line with empirical work such as Doval et al. (2024) and the methodologies described in Agarwal, Nikhil and Somaini, Paulo (2023), which offer a structured framework for assessing these inefficiencies.

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

Identified Problem	**Quantifiable Indicator**
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 Soto (2019). [12]
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).

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.

[12] 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.

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.[13]

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, 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),

[13] 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.

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.

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

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 5: Net enrollment rate in secondary 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%

A comprehensive study on the impact of congestion on enrollment is provided by Alba-Vivar (2025)[14], 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

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

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 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

We presented a convex matching framework with heterogeneous quadratic congestion and penalized marginals that captures persistent imbalances (excess demand and underutilization) typical of developing economies. In the discrete formulation, we studied the existence and uniqueness of solutions, as well as approximate and exact methods to compute an interior solution. In particular, we proposed an $O((N + L)N^2L^2)$ algorithm exploiting the problem's structure and the Sherman–Morrison decomposition.

We also analyzed a general infinite–types setting, establishing existence and uniqueness under mild assumptions and providing an interior fixed–point characterization. An application to Peru's healthcare system illustrates how the model rationalizes observed bottlenecks and offers the planner controlled slack when exact feasibility is unattainable.

Future work includes several concrete directions. First, structural estimation and empirical validation using administrative matching data: estimate (c_{ij}, a_{ij}) from observed matches π^{obs} by solving, for instance,

$$\inf_{\{c_{ij},a_{ij}\}} \sum_{i,j} \left(\pi_{ij}^{\text{obs}} - \pi_{ij}^{\star}(c,a)\right)^2,$$

with parametric links such as $c_{ij} = x_{ij}'\beta + \eta_{ij}$ and $a_{ij} = z_{ij}'\gamma + \xi_{ij}$ to interpret covariates (distance, capacity, quality) and identify congestion heterogeneity.

Second, dynamic and stochastic extensions: for a panel $\{\pi_t\}$, incorporate temporal frictions via a quadratic smoothing penalty $\lambda\|\pi_t - \pi_{t-1}\|_F^2$ with $\lambda > 0$, together with time–varying parameters summarized by

$$\theta_t = \rho_\theta\, \theta_{t-1} + B\, u_t + \varepsilon_t$$

and the equilibrium relation $\pi_t = \pi^\star(\theta_t)$, where θ_t stacks (c^t, a^t, μ^t, ν^t) and u_t (e.g., infrastructure investment, staffing reallocation, subsidy intensity) is chosen to maximize discounted welfare

$$\sup_{\{u_t\}} \mathbb{E}\left[\sum_{t\geq 1} \delta^t \big(W(\pi_t) - \gamma\|u_t\|_F^2\big)\right],$$

with $\delta \in (0,1)$.

Finally, characterize interior versus boundary regimes by deriving explicit parameter regions (e.g., contraction constants based on a, ϵ, δ) and positivity conditions for the fixed–point operator.

Appendix

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

Example 5.1. The parameters used for solving $\mathcal{P}_{CP}$ with $d = 5I_{3\times 3}$ 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 [15] 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 $\epsilon = \begin{bmatrix} 0.4 & 1 & 0.2 \end{bmatrix}^T$ 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 $\alpha = 0.5$:

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

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

The solution to the optimization problems are[16]

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

and

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

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.

Ambrosio, L., Brué, E., and Semola, D. (2024). *Lectures on Optimal Transport*, volume 169 of *UNITEXT*. Springer, Cham, 2 edition.

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.

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.

[16] 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`.

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.

Brezis, H. (2010). *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, New York.

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.

Coppersmith, D. and Winograd, S. (1990). Matrix multiplication via arithmetic progressions. *Journal of Symbolic Computation*, 9(3):251–280.

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).

Davie, A. M. and Stothers, A. J. (2013). Improved bound for complexity of matrix multiplication. *Proceedings of the Royal Society of Edinburgh: Section A Mathematics*, 143(2):351–369.

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=eyJrIjoiMDQwMDVlOGItNGY5ZiOOZjFjLWEyZDMtYjY1ZjkOMWVjMjcxIiwidCI6IjMOZjMyNDE5LTFjMDUtNDc1NiO4OTZlLTQ1ZDYzMzcyNjU5YiIsImMiOjR9` (accessed 18 March 2025).

EsSalud (2025b). Tablero de diferimento de citas. `https://app.powerbi.com/view?r=eyJrIjoiN2NlMTNmNWEtODA3MSOOM2UyLWE3NDAtNjcyYjZjYTQOMmJmIiwidCI6IjMOZjMyNDE5LTFjMDUtNDc1NiO4OTZlLTQ1ZDYzMzcyNjU5YiIsImMiOjR9` (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. (2025). Sparsity of Quadratically Regularized Optimal Transport: Scalar Case. Accessed on April 10, 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.

Huang, J. (2018). *Practical fast matrix multiplication algorithms*. Ph.D. dissertation, The University of Texas at Austin, Austin, TX.

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.

Karstadt, E. and Schwartz, O. (2017). Matrix multiplication, a little faster. In *Proceedings of the 29th ACM Symposium on Parallelism in Algorithms and Architectures*, SPAA '17, page 101–110. ACM.

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*.

Koutis, I., Miller, G. L., and Peng, R. (2012). A fast solver for a class of linear systems. *Communications of the ACM*, 55(10):99–107.

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

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

Lorenz, D. A., Manns, P., and Meyer, C. (2021). Quadratically regularized optimal transport. *Applied Mathematics and Optimization*, 83:1919–1949.

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

Pan, V. (1982). Trilinear aggregating with implicit canceling for a new acceleration of matrix multiplication. *Computers and Mathematics with Applications*, 8(1):23–34.

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.

Strassen, V. (1986). The asymptotic spectrum of tensors and the exponent of matrix multiplication. In *27th Annual Symposium on Foundations of Computer Science (sfcs 1986)*, page 49–54. IEEE.

Vassilevska, V. (2012). Multiplying matrices faster than coppersmith-winograd. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, STOC'12, page 887–898. ACM.

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 .

Winograd, S. (1971). On multiplication of 2 × 2 matrices. *Linear Algebra and its Applications*, 4(4):381–388.

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).