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 11, 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 $O((N+L)(NL)^2)$ algorithm to compute the optimal interior solution, offering a tighter upper bound compared to classical numerical methods. For certain cases, we derive a closed-form solution and conduct a comparative statics analysis. Finally, we present illustrative examples demonstrating how our model produces distinct solutions from classical approaches, leading to more accurate outcomes in specific contexts. Our numerical analysis is based on statistics from 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.

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

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., 2020; Echenique et al., 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), OT optimizes over distributions, providing a more flexible and general framework. 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; Nenna, 2020; 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., 2020), labor markets (Dupuy and Galichon, 2022), and school choice (Echenique et al., 2024).

This paper introduces a new model built upon the quadratic regularization framework, similar to Nutz (2024), but adopting the approach of Izmailov and Solodov (2023) while introducing heterogeneity in the quadratic term. Our model captures elements absent in classical formulations and better aligns with real-world scenarios. Specifically, by replacing equality constraints with weighted penalization terms, the solution accommodates supply and demand imbalances, a feature particularly relevant in developing countries when modeling matching in education and healthcare markets.

Countries with developing economies often experience significant inefficiencies in education and healthcare due to excess demand, insufficient supply, mismatching, and systemic congestion. These structural issues have contributed to high mortality rates and service deficiencies, as demonstrated during the COVID-19 pandemic. For instance, Peru recorded the highest per capita COVID-19 mortality rate globally, exceeding 6,400 deaths per million inhabitants (John Hopkins University, 2023). Similar inefficiencies have been observed across Latin America, where restricted healthcare access exacerbates disparities.

A key factor behind these inefficiencies is that individuals are not properly matched due to physical barriers, bureaucratic issues, and congestion (Anaya-Montes and Gravelle, 2024; Velásquez, 2020), compounded by excess demand. In countries such as Peru, India, and Brazil (Kikuchi and Hayashi, 2020), congestion is particularly severe. For instance, the World Bank estimates that traffic congestion alone costs Peru 1.8% of its GDP annually (Banco Mundial, 2024). Given these conditions, accounting for congestion and excess demand is crucial when modeling these dynamics.

The model presented in this paper provides a framework for congestion costs while also capturing excess of demand across different institutional contexts. As such, it reflects the realities of many developing countries, contrasting with developed nations such as France or Switzerland,

where robust transportation infrastructure, efficient bureaucratic systems, and policies ensure universal access to education and healthcare.

The remainder of this paper is structured as follows. Section 2 defines the fundamental concepts and notation. Section 3 introduces the proposed model and examines its theoretical properties. Section 4 presents illustrative examples that demonstrate the advantages of our approach. Due to data availability constraints, 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. In the marriage market case, $\mu_i = \nu_j = 1$, 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 going to (matched with) j.

The problem addressed in the classic literature, 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 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)

[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.

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. Therefore, the central planner seeks to solve

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

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 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. (2020) 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 $f(x) = x\ln x$ and that the analogous of Inada's conditions are satisfied ($\lim_{x\downarrow 0} f'(x) = -\infty$), the solution is interior, i.e. $\pi^*_{ij} > 0$ (see a detailed argument in Nenna (2020)). 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 $\Pi(\mu,\nu)$ ensures the existence of a solution to $\mathcal{P}_O$ and its variants by Weierstrass Theorem.

3 The model

In this section, we present the model that we propose, inspired by the optimal transport problem with quadratic regularization, but following the approach of Izmailov and Solodov (2023). The model is derived from the very characteristics of the observed reality in certain locations. This will be explored in more detail in Section 4.

First, we need to allow the number of individuals of X who belong to i and are matched with

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

$j = 1, ..., L$, to not necessarily be μ_i[4]. Similarly, it may be the case that not all those matched with j sum to ν_j[5]. Therefore, there may be excess supply or demand. 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.

Mathematically, we model this by replacing the equality constraints defined by $\Pi(\mu, \nu)$ with penalties in the objective function. Moreover, we introduce weights for each penalty. That is, the constraint $\sum_{i=1}^{N} \pi_{ij} = \nu_j$ is replaced by the penalty term $\delta_j \left[\sum_{i=1}^{N} \pi_{ij} - \nu_j\right]^2$, with $\delta_j > 0$, and the constraint $\sum_{j=1}^{L} \pi_{ij} = \mu_i$ is replaced by $\epsilon_i \left[\sum_{j=1}^{L} \pi_{ij} - \mu_i\right]^2$, with $\epsilon_i > 0$. The parameters ϵ_i, δ_j are weights. 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.

Secondly, as is natural in some environments (see the next section), congestion costs are present. These costs reflect the fact that matching more individuals from i with the same j becomes increasingly costly. For example, from the perspective of physical transportation costs, in countries with high vehicular traffic congestion, the effect of increasing from x cars to $x+1$ passing through a certain avenue is less or equal to increasing from $x+n$ to $x+n+1$ with $n \geq 1$. Therefore, clustering groups based on geographic location means that matching many individuals from the same group to a single j congests the access route (which is the same), hence the π_{ij}^2. Therefore, we introduce the term $\sum_{i,j} a_{ij}\pi_{ij}^2$ in the cost structure. The coefficient a_{ij} captures heterogeneity[6], while the quadratic term represents the previously described phenomenon. 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. As more patients of this type arrive, the system must process a greater number of cases. Since they share the same characteristics, it is assumed that the same computer or system will 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.).

Finally, we certainly have $\pi_{ij} \geq 0$, 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)[7]. Hence, the optimization is carried out over the entire space $\mathbb{R}_+^{NL}$. This phenomenon also explains the penalties: we no longer assume a fixed number of individuals of type i, and μ_i represents now 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.

Therefore, following the described scenario, the central planner seeks to minimize costs while taking into account the objective of reaching the targets μ_i and ν_j. According to what has been specified, the problem is the following:

[4]We anticipate that μ_i will no longer be the mass of individuals of group i but rather a targeted quota for individuals of group i.

[5]Once again, ν_j is a target but not a constraint.

[6]or some situations, the coefficient might be large, but in others—such as cases where there are few schools or hospitals, lightly congested streets, good traffic lights, etc.—the coefficient is small.

[7]This considerably simplify our analysis and does not affect the logic of the model.

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

where $\epsilon_1, ..., \epsilon_N$, $\delta_1, ..., \delta_L$ and $\mu_1, \cdots, \mu_N$, $\nu_1, \cdots, \nu_L$ are all non negative, $\alpha \in [0,1]$, and

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

In Equation 6, despite its practical relevance, the term d_{ij}, representing fixed costs, does not influence the resolution of the problem at all. 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[8], 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 unbounded set, we can actually restrict the optimization to $\mathbb{R}_+^{NL} \cap \Omega$, where

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

Indeed, it is clear from the cost function F that it is strictly lower over the interior of Ω or in the axes, than when evaluated on $\partial\Omega$ (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 a 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, similar to the case of the classical demand theory, we are assuming that $\pi_{ij} \in \mathbb{R}_+$. However, just as one might argue

[8] 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 involve analogous Inada's conditions, which don't appear in our model. Finally, in Izmailov and Solodov (2023), only general results concerning penalization are given and this particular problem is not at all studied.

that it does not make sense to consume $\sqrt{2}$ cars, it is also unreasonable to consider that π_{ij} is not restricted to taking values in $\mathbb{Z}_+$, since it ultimately represents a 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 the integer and continuous solutions in the case of separable convex or quadratic functions with linear constraints (Hochbaum and Shanthikumar, 1990; Planiden and Wang, 2014; Park and Boyd, 2017; Hladík et al., 2019; Pia and Ma, 2021; Pia, 2024), 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[9]. Solving the problem in $\mathbb{R}_+^{NL}$ allows the use of nonlinear convex optimization techniques, yielding not only computational advantages but also analytical insights.

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

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

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. The proof is in the Appendix.

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

[9]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.

conclusions, such as closed-form solutions, bounds, and perform comparative statics. From a computational perspective, we compare our algorithm with traditional methods for solving linear systems. The key aspect is that we exploit the structure of the matrix A, decomposed using the Kronecker product (see Equation 8).

3.1 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., $\epsilon_i, \delta_j \in [0,1]$ and $\sum_{i=1}^{N}\epsilon_i = \sum_{j=1}^{L}\delta_j = 1$.

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 under filling 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), Theorem 3.9. Lemmas 3.6-3.8 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 re-write 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* $O((N+L)(NL)^2)$ *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 $A^{-1} \leftarrow \mathrm{Diag}(1/a_{11}, \ldots, 1/a_{NL}) \in \mathbb{R}^{NL,NL}$
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 $v^{(j)} \coloneqq \delta_j^{1/2}\mathbf{e}_j \otimes \mathbf{1}_{N\times 1}$
    $A^{-1} \leftarrow A^{-1} - \dfrac{A^{-1}v^{(j)}v^{(j)T}A^{-1}}{1 + v^{(j)T}A^{-1}v^{(j)}}$ via Sherman-Morrison formula
end for
return $A^{-1}b$

When L is similar to N (i.e. $L \in \Theta(N)$), we can compute A^{-1} in $O(N^5)$ which is significantly faster than the naive approach, see Table 1.[10]

[10] In Table 1, SPD stands for Symmetric Positive Definite, referring to matrices with efficient factorizations

Algorithm	**Time**	**Struct. Opt.**	**Rank-1 Upd.**	$N = L$
Naïve Golub and Van Loan (2013)	$O((NL)^3)$	No	No	$O(N^6)$
QR Trefethen and Bau (1997)	$O((NL)^3)$	No	No	$O(N^6)$
SVD Strang (2006)	$O((NL)^3)$	No	No	$O(N^6)$
Cholesky Higham (2002)	$O((NL)^3)$	Yes (SPD)	No	$O(N^6)$
Our Alg.	$O((N+L)(NL)^2)$	Yes	Yes	$O(N^5)$

Table 1: Computational complexity comparison of matrix inversion methods.

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:

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

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:

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

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

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

where $A^{-1}_{\Pi,2}$ consists of multiplying column ij of $D^{-1} - (D^{-1})^2 X$ 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:

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

like Cholesky $O(n^3/3)$, see Table 1. Naïve Inversion denotes direct matrix inversion via Gaussian elimination or adjugates, requiring $O(n^3)$ FLOPs, similar to QR decomposition. Our algorithm, leveraging the Sherman-Morrison formula, reduces complexity to $O((N+L)(NL)^2)$, significantly improving over traditional $O((NL)^3)$ approaches.

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

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

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

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. However, 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 when studying the marriage market Roth and Sotomayor (1990). Likewise, as we will see in Section 4, it is of particular interest when analyzing the healthcare sector in Peru. From a purely theoretical perspective, the case $N = L$ leads to a scenario where Algorithm 1 exhibits an upper-bound performance superior to classical methods used to solve the corresponding linear system. Another important result related to the case $N = L$ is presented in Appendix C.

In Section 4, we provide practical justification for our model and concrete examples of how it brings new insights. From a mathematical and theoretical perspective, if the solution is interior in our model, the problem reduces to inverting a matrix, whereas in classical transportation problems and its variants, numerical methods for solving convex optimization problems in finite dimension are required.

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. EsSalud, on the other hand, is the public health insurance financed by contributions from formal workers and employers. Finally, SIS is a universal public insurance targeting people in poverty or without the ability to pay. For the year of the pandemic (2020), SIS and EsSalud together covered more than 80% of the population, while 8% was covered by EPS, see Table 2.

Insurance	Covered people (%)	Number of affiliated	Source
EPS	8%	2,627,086	Anaya-Montes and Gravelle (2024)
EsSalud	30%	9,851,574	Anaya-Montes and Gravelle (2024)
SIS	53%	17,404,451	Anaya-Montes and Gravelle (2024)
Población Total	100%	32,838,579	Data Commons (2020)

Table 2: Number of enrollees in Peru's healthcare system in 2020 (before COVID 19).

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.

Identified Problem	Quantifiable Indicator	Source
Inadequate infrastructure in primary healthcare facilities in Lima.	76% of facilities have inadequate installed capacity.	Defensoría Pueblo (2020)
Shortage of medical personnel in primary healthcare.	4 doctors per 10,000 inhabitants, far from the WHO-recommended standard of 43.	Infobae Médicos (2024)
Lack of hospital beds in Peru's healthcare system	1.6 beds per 1,000 inhabitants, below the regional average	Banco Mundial (2023).
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	Huerta-Rosario et al. (2019)
Deferrals in certain cities are very high.	Increased waiting times.	Power BI Healthcare Data (2024a)
Coverage noncompliance, high waiting times, and some values of medical performance per hour out of range.	No access to services and inefficiencies.	Power BI Healthcare Data (2024b)

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

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 methodologies outlined in Agarwal and Somaini (2023), which provide a structured framework for assessing these inefficiencies.

In Example 5.1, we simulate three patient groups across three healthcare networks (SIS, EsSalud, EPS). Two sectors (1 and 3) are distant, creating significant matching frictions, particularly we think of informal workers accessing EsSalud and EsSalud going to SIS (Anaya-Montes and Gravelle, 2024). Our model reflects bureaucratic barriers by assigning zero matches between sectors 1 and 2 and illustrates the importance of quadratic costs. Note also that models with homogeneous quadratic regularization (Appendix C) fails to capture the heterogeneity. Example 5.3 demonstrates how incorporating penalties and weighted constraints aligns with a central planner's goals. In Peru, the government could for instance prioritize EsSalud, as it signifies formal employment, making it logical to assign higher weights to constraints associated with EsSalud. Finally, Example 5.4 highlights how strict cost convexity limits access to healthcare services generating excess of demand. Traditional transport models, even with quadratic regularization, do not fully capture these barriers, a key advantage of our model.

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 (Álvarez 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 a better fit—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 (Harvard Kennedy School Student Review, 2024; World Pulse, 2024; UNESCO, 2024). 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	5.0
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. Lima, one of the most congested cities in Latin America, suffers from severe traffic bottlenecks that disproportionately affect students from lower-income districts (Alba-Vivar, 2025; Harvard Ash Center for Democratic Governance and Innovation, 2024; IFSA-Butler, 2024; World Bank, 2024; Agence Française de Développement (AFD), 2024). When large numbers of students travel from the same location to the same school, the primary roads connecting them become saturated, increasing commuting times (América TV, 2024).

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), which is centralized, homogeneous, 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). Traditional OT models, by imposing the condition $\sum_i \mu_i = \sum_j \nu_j$, do not apply as effectively. Our model is crucial because, in countries or cities with constraints, allowing for supply or demand imbalances—i.e., schools not reaching full capacity or not all students being enrolled—is a more realistic assumption.

Example 5.5 is key to understanding the education case. 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, $j = 2$ is average, and $j = 3$ is lower-ranked. 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. The solutions highlight key differences: $\mathcal{P}_{CP}$

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

introduces quadratic penalties, leading to smoother allocations where poorer students face greater commuting constraints. $\mathcal{P}_Q$ applies quadratic regularization but matches everyone, distributing students more evenly and generates no excess of demand. $\mathcal{P}_O$ enforces too full allocation, forcing students into schools regardless of suitability, making it the least flexible model. This analysis demonstrates how different formulations capture educational constraints in distinct ways.

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. 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. Additionally, exploring policy implications—such as optimal subsidy structures or decentralized decision-making mechanisms—could provide valuable insights for addressing inefficiencies in public service delivery.

CRediT authorship contribution statement

Marcelo Gallardo: Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review and editing.
Manuel Loaiza: Formal analysis, Investigation, Software, Validation, Writing – original draft.
Jorge Chávez: Funding acquisition, Project administration, Supervision, Review and editing.

Declaration of competing interest

None.

Data availability

The data used in this study are sourced from the National Institute of Statistics and Informatics (INEI), the National Superintendence of Health (SUSALUD), and the Social Health Insurance of Peru (EsSalud).

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT-4o in order to assist with grammar correction and to make paragraphs more concise. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Acknowledgments

Chávez acknowledges financial support from the Pontificia Universidad Católica del Perú. Gallardo acknowledges insightful discussions with Professor Federico Echenique (UC Berkeley), and former Minister of Health Aníbal Velásquez. Dr. Velásquez provided key information with respect to the Peruvian health system.

Appendix A. Proofs

Proof of Lemma 3.1

Proof. 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 $\text{Diag}(\epsilon_1, ..., \epsilon_N)$ 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)[12]

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

Proof of Lemma 3.2.

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

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

[12]For Minkowski's determinant inequality and its generalizations, see Marcus and Gordon (1970), Artstein-Avidan et al. (2015).

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, $\|D^{-1}X\|_\sigma < 1$ [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. ■

Proof of Theorem 3.3.

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

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

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 $\|\pi_n - \pi^*\|_\infty < \epsilon$. ■

Proof of Theorem 3.4.

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

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

■

Proof of Lemma 3.6.

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

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

■

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

Proof Lemma 3.7.

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. ∎

Proof of Lemma 3.8.

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 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.

Proof. By triangle inequality,

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

■

Proof of Theorem 3.10.

Proof. Consider Algorithm 1. First, 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 $O((NL)^2)$. 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 $O((NL)^2)$. Computing $1 + w^T A^{-1} w$ takes $O((NL)^2)$, so the only possible optimization is finding the optimal parenthesization for the product $A^{-1} w w^T A^{-1}$. Since there are only five possible ways to parenthesize the expression, we determine by brute force that computing $(A^{-1}w)(w^T A^{-1})$ also takes $O((NL)^2)$. This implies the desired time complexity of $O((N + L)(NL)^2)$. ■

Appendix B. Numerical simulations

Next, we consider

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

Example 5.1. The parameters used for solving $\mathcal{P}_{CP}$ are:

$$d = 5I_{3\times 3}, \quad c = \begin{bmatrix} 1.0 & 50.0 & 20.0 \\ 50.0 & 1.0 & 20.0 \\ 20.0 & 10.0 & 1.0 \end{bmatrix}, \quad a = \begin{bmatrix} 1.0 & 5.0 & 10.0 \\ 5.0 & 1.0 & 2.0 \\ 10.0 & 5.0 & 1.0 \end{bmatrix},$$

$$\epsilon = \begin{bmatrix} 0.3 \\ 0.3 \\ 0.3 \end{bmatrix}, \quad \delta = \begin{bmatrix} 0.3 \\ 0.3 \\ 0.3 \end{bmatrix}, \quad \mu = \begin{bmatrix} 100.0 \\ 50.0 \\ 20.0 \end{bmatrix}, \quad \nu = \begin{bmatrix} 90.0 \\ 40.0 \\ 40.0 \end{bmatrix}, \quad \text{and } \alpha = 0.5.$$

The optimal solution π^* obtained using `cvxpy` with the ECOS solver for convex optimization in Python is

$$\pi^* = \begin{bmatrix} 33.3255 & 0.0 & 1.5258 \\ 0.0 & 14.3615 & 2.7847 \\ 0.7380 & 0.6208 & 8.3120 \end{bmatrix}.$$

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 solution to the associated $\mathcal{P}_Q$ ($\mathcal{P}_O$) is:

$$\pi^* = \begin{bmatrix} 84.27496 & 8.84062 & 6.88442 \\ 4.29850 & 30.42065 & 15.28086 \\ 1.42655 & 0.73873 & 17.83472 \end{bmatrix}, \ \pi^* = \begin{bmatrix} 90.0 & 0.0 & 10 \\ 0.0 & 40.0 & 10 \\ 0.0 & 0.0 & 20.0 \end{bmatrix}.$$

Example 5.3. Using the same parameters as in $\mathcal{P}_{CP}$ but changing weighting to $\epsilon = \begin{bmatrix} 1.0 & 0.2 & 0.2 \end{bmatrix}^T$ and $\delta = \begin{bmatrix} 1.0 & 0.2 & 0.2 \end{bmatrix}^T$ leads to

$$\pi^* = \begin{bmatrix} 59.4326 & 2.6078 & 2.8752 \\ 1.4843 & 10.0281 & 0.6203 \\ 1.7349 & 0.0911 & 5.6683 \end{bmatrix}$$

Example 5.4. Modifying the parameters with respect to Example 5.1 as follows:

$$a = \begin{bmatrix} 1.0 & 20.0 & 2.0 \\ 20.0 & 5.0 & 2.0 \\ 5.0 & 2.0 & 0.5 \end{bmatrix}, \quad \mu = \begin{bmatrix} 200.0 \\ 50.0 \\ 10.0 \end{bmatrix} \quad \text{and} \quad \nu = \begin{bmatrix} 100.0 \\ 20.0 \\ 50.0 \end{bmatrix}$$

yields

$$\pi^* = \begin{bmatrix} 50.9856 & 0.8394 & 16.7308 \\ 0.0048 & 2.9796 & 3.5361 \\ 0.5020 & 0.0000 & 7.9721 \end{bmatrix}.$$

Example 5.5. Consider the following parameters for $\mathcal{P}_{CP}$ with $\alpha = 0.5$:

$$d = \begin{bmatrix} 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 1.0 \end{bmatrix}, \quad c = \begin{bmatrix} 1.0 & 5.0 & 10.0 \\ 1.0 & 5.0 & 10.0 \\ 10.0 & 5.0 & 1.0 \\ 10.0 & 5.0 & 1.0 \end{bmatrix}, \quad a = \begin{bmatrix} 1.0 & 1.0 & 1.0 \\ 2.0 & 2.0 & 1.0 \\ 1.0 & 1.0 & 1.0 \\ 2.0 & 2.0 & 1.0 \end{bmatrix}$$

$$\epsilon = \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \end{bmatrix}, \quad \delta = \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \end{bmatrix}, \quad \mu = \begin{bmatrix} 10.0 \\ 10.0 \\ 10.0 \\ 10.0 \end{bmatrix}, \quad \nu = \begin{bmatrix} 10.0 \\ 20.0 \\ 10.0 \end{bmatrix}$$

The solution to $\mathcal{P}_{CP}$ ($\mathcal{P}_Q$ and $\mathcal{P}_O$ respectively) is:

$$\pi^* = \begin{bmatrix} 1.7896 & 1.4435 & 0.0 \\ 1.0295 & 0.8565 & 0.0 \\ 0.0 & 1.4621 & 1.6782 \\ 0.0 & 0.7873 & 1.7906 \end{bmatrix}, \ \pi^* = \begin{bmatrix} 4.4583 & 5.5417 & 0.0 \\ 3.9104 & 4.4521 & 1.6375 \\ 0.5333 & 6.1167 & 3.35 \\ 1.0979 & 3.8896 & 5.0125 \end{bmatrix}, \ \pi^* = \begin{bmatrix} 5.0 & 5.0 & 0.0 \\ 5.0 & 5.0 & 0.0 \\ 0.0 & 5.0 & 5.0 \\ 0.0 & 5.0 & 5.0 \end{bmatrix}.$$

Appendix C. Quadratic heterogeneous congestion costs

The problem $\mathcal{P}_Q$ introduced in the previous annex, as far as we know, is new in the literature. Classic quadratic regularization assumes $a_{ij} = \varepsilon$ for all $(i, j) \in I \times J$, while $\mathcal{P}_Q$ incorporates heterogeneity in the quadratic term. This heterogeneity allows us to obtain some insightful results, which we present below.

Assumption 5. Let K be a positive integer strictly greater than 1. Assume that $N = L = K$ and $\mu_i = \nu_j$ for all $1 \leq i, j \leq K$.

Assumption 5 ensures that each healthcare center or school reaches full capacity with individuals from the same group. This is, theoretically, the desired situation in the healthcare system (see Section 4) when dividing by healthcare centers attending only certain diseases or a certain type of diseases in terms of the complexity involved.

Assumption 6. For each $1 \leq i \leq N$, suppose there exists $1 \leq t_i \leq L$ such that $c_{i\,t_i} < c_{i\,j}$ for all $1 \leq j \leq L$ with $j \neq t_i$. Furthermore, assume that $t_i \neq t_j$ for all $1 \leq i, j \leq L$ with $i \neq j$.

Assumption 6 imposes that each individual is optimally matched with their top choice alternative, ensuring a distinct best fit for each individual. Note that Assumptions 5 and 6 imply immediately that the solution to the linear model is:

$$\pi^* = [\pi^*_{ij}] = \begin{cases} \mu_i & \text{if } j = t_i, \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

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

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

Assumption 7. Let $\widetilde{c}_i = \min\limits_{\substack{1\le j\le L\\ j\neq t_i}} \{c_{ij}\}$ satisfy $\widetilde{c}_i > c_{i\,t_i} + a_{i\,t_i}\mu_i^2(1 - 1/L)$ for $1 \le i \le N$.

Assumption 7 tells us that preferences between i types and j types (classes) must be such that the top choice, only based on preferences c_{ij}, and individual characteristics is at least $a_{i\,t_i}\mu_i^2(1-1/L)$ better than the other ones. We now show by combining Assumptions 5, 6 and 7 that the solution to the quadratic regularization problem with heterogeneity, in the integer setting, is given by (14).

Theorem 5.6. *Under Assumptions 5, 6 and 7, the optimal matching for the quadratic model in the integer setting is* (14).

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

$$C(\pi;\theta) = \sum_{i=1}^{N}\sum_{j=1}^{L} d_{ij} + c_{ij}\pi_{ij} + a_{ij}\pi_{ij}^2 \ge \sum_{i=1}^{N}\sum_{j=1}^{L} d_{ij} + \sum_{i=1}^{N}\left(\sum_{j=1}^{L} c_{ij}\pi_{ij} + a_{i\,t_i}\sum_{j=1}^{L}\pi_{ij}^2\right).$$

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

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

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

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

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

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

Combining these results, we have

$$C(\pi;\theta) \ge \sum_{i=1}^{N}\sum_{j=1}^{L} d_{ij} + \sum_{i=1}^{N}\mu_i c_{i\,t_i} + \widetilde{c}_i - c_{i\,t_i} + a_{i\,t_i}\frac{\mu_i^2}{L} \; > C(\pi^*;\theta). \qquad \blacksquare$$

Theorem 5.6 has economic implications, as it provides conditions on the parameters under

which the optimal matching in the transport problem with heterogeneous quadratic regularization coincides with the structure of the optimal matching in the linear model. Typically, in the quadratic model, the solution is sparse (González-Sanz and Nutz, 2024), which no longer holds when the solution is given by (14). However, introducing heterogeneous quadratic terms eliminates the corner solution property characterizing the linear model (Tardella, 2010). See Examples 5.7 and 5.8.

Example 5.7. In this example, we show a case where the solution is interior for $\mathcal{P}_Q$. Consider

$$a = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \; c = \begin{bmatrix} 12 & 24 \\ 8 & 12 \end{bmatrix}, \; \mu = \begin{bmatrix} 10 \\ 10 \end{bmatrix} \text{ and } \nu = \begin{bmatrix} 6 \\ 14 \end{bmatrix}.$$

Then, we have $\pi^* = \begin{bmatrix} 4 & 6 \\ 2 & 8 \end{bmatrix}$.

Example 5.8. To illustrate a case where the solution to $\mathcal{P}_Q$ is a corner solution, consider the following values:

$$a = \begin{bmatrix} 100 & 1 \\ 1 & 100 \end{bmatrix}, \quad c = \begin{bmatrix} 100 & 1 \\ 1 & 100 \end{bmatrix}, \text{ and } \mu = \begin{bmatrix} 5 \\ 5 \end{bmatrix} = \nu.$$

In this scenario, the optimal solution is $\pi^* = \begin{bmatrix} 0 & 5 \\ 5 & 0 \end{bmatrix}$, a corner solution.

Example 5.9. The following examples were computed using Mathematica 14.1. The parameters are taken so Theorem 5.6 can be verified. For the linear model, with $N = L = 4$ and $\mu_i = \nu_j = 50$, the optimal matching was computed using `LinearOptimization`:

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

For the quadratic model $\mathcal{P}_Q$, with $N = L = 4$ and $\mu_i = \nu_j = 20$,

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

the optimal matching, obtained using `QuadraticOptimization`, is

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

Hence, the result is in accordance with Theorem 5.6.

Finally, $\mathcal{P}_Q$ has valuable applications and interpretations, as demonstrated in the following example. Note that, up to this appendix, we have considered the case $\mathbb{Z}$. However, the next example addresses the case in which $\pi \in \Pi(\mu, \nu)$. In practice, solving $\mathcal{P}_Q$ typically requires numerical methods, as analytical solutions via Karush-Kuhn-Tucker conditions are generally infeasible for obtaining a closed-form expression systematically. Both smooth and monotone comparative statics cannot be performed. This follows from the fact that $\Pi(\mu, \nu)$ is neither a surface nor a lattice (de la Fuente, 2000; Milgrom and Shannon, 1994).

Example 5.10. Consider the following case with the given parameters:

$$d = 5I_{3\times 3}, \quad c = \begin{bmatrix} 1.0 & 5.0 & 100.0 \\ 10.0 & 1.0 & 50.0 \\ 100.0 & 50.0 & 1.0 \end{bmatrix} \text{ and } a = \begin{bmatrix} 2.0 & 1.0 & 1.5 \\ 2.0 & 1.0 & 3.0 \\ 1.5 & 2.0 & 1.5 \end{bmatrix}$$

and $\mu_i = \nu_j = 30$. The optimal solution to the problem $\mathcal{P}_Q$ ($\mathcal{P}_O$) is:

$$\pi^* = \begin{bmatrix} 26.7466 & 3.2534 & 0.0000 \\ 0.7329 & 25.2260 & 4.0411 \\ 2.5205 & 1.5205 & 25.9589 \end{bmatrix}, \ \pi^* = \begin{bmatrix} 30.0 & 0.0 & 0.0 \\ 0.0 & 30.0 & 0.0 \\ 0.0 & 0.0 & 30.0 \end{bmatrix}.$$

This demonstrates that the problem with quadratic heterogeneity allows for modeling situations where, for example, individuals, despite having a higher affinity for a particular institution, may choose not to attend if reaching it involves traversing a highly congested route. In Peru, this is exemplified by Avenida Javier Prado, one of the most congested roads in Lima, where drivers can take up to two hours to travel from Magdalena to La Molina, a distance of approximately 17 km (TomTom Traffic Index, 2024; Infobae Javier Prado, 2024). In contrast, in France, this distance can be covered by metro in about 30 minutes (Seat61 - The Man in Seat 61, 2024).

References

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

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

Agarwal, N. and Somaini, P. (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.

Agence Française de Développement (AFD) (2024). Developing an efficient and sustainable public transport system in lima. Accessed on February 21, 2025.

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

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

América TV (2024). Traffic congestion near schools in lima affects morning commutes. Accessed on February 21, 2025.

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 February 21, 2025.

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

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

Banco Mundial (2024). Modernizando la gestión del tráfico en lima con apoyo del banco mundial. Accessed: February 21, 2025.

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

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

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 (2020). Población total de perú (2020). Accessed on February 21, 2025.

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

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

Doval, L., Echenique, F., Huang, W., and Xin, Y. (2024). Social learning in lung transplant decision. *arXiv preprint*. https://arxiv.org/abs/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, F., Root, J., and Sandomirskiy, F. (2024). Stable Matching as Transportation. Preprint submitted to arXiv on 12 Feb 2024.

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

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

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

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

Golub, G. H. and Van Loan, C. F. (2013). *Matrix Computations*. Johns Hopkins University Press, 4th edition.

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

Harvard Ash Center for Democratic Governance and Innovation (2024). Lima's public transport system: An analysis of traffic congestion and accessibility. Accessed on February 21, 2025.

Harvard Kennedy School Student Review (2024). Connecting schools to reduce student dropout: A peruvian case. Accessed on February 21, 2025.

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

Higham, N. J. (2002). *Accuracy and Stability of Numerical Algorithms*. SIAM, 2nd edition.

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

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

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

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

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

INEI (2024). Living conditions in peru: Technical report 2024. Accessed on February 21, 2025.

Infobae Javier Prado (2024). Avenida Javier Prado congested: Drivers take up to two hours from Magdalena to La Molina. Accessed on February 21, 2025.

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

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

John Hopkins University (2023). Mortality Analysis by JHU Coronavirus Resource Center. Accessed: February 21, 2025.

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

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

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

Marcus, M. and Gordon, W. R. (1970). An extension of the minkowski determinant theorem. *Cambridge University Press*. Received 21st September 1970.

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

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

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

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

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

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

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

Pia, A. D. and Ma, M. (2021). Proximity in concave integer quadratic programming. *arXiv preprint*, arXiv:2006.01718.

Planiden, C. and Wang, X. (2014). Most convex functions have unique minimizers. *arXiv preprint*, arXiv:1410.1078.

Power BI Healthcare Data (2024a). Statistical Analysis of Healthcare System Performance in Peru - Report 1. Accessed on February 21, 2025.

Power BI Healthcare Data (2024b). Statistical Analysis of Healthcare System Performance in Peru - Report 2. Accessed on February 21, 2025.

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.

Seat61 - The Man in Seat 61 (2024). Train travel in france: Metro and regional transport. Accessed on February 21, 2025.

Strang, G. (2006). *Linear Algebra and Its Applications*. Cengage Learning, 4th edition.

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

TomTom Traffic Index (2024). Lima Traffic Report: Most Congested Roads in Peru. Accessed on February 21, 2025.

Trefethen, L. N. and Bau, D. (1997). *Numerical Linear Algebra*. SIAM.

UNESCO (2024). Study on the situation and proposals for education in the amazon region. Accessed on February 21, 2025.

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

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

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

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

World Pulse (2024). Peru: Education Barriers in the Andes. Accessed on February 21, 2025.

Zhan, S. (2005). On the determinantal inequalities. *Journal of Inequalities in Pure and Applied Mathematics*, 6(4):Article 105.

Álvarez Laveriano, N. (2010). The decentralization of education in peru. *Educación: PUCP*, 19(37):7–26. Accessed on February 21, 2025.