Mean Field Game models for maritime traffic and applications to real data:

how to anticipate the unpredictable?

Charles-Albert Lehalle
CMAP, Ecole Polytechnique, IP Paris
(Louis Bachelier Fellow, Paris, France)
Co-director with Pr. V. Perchet of the
"**AI for Markets and Quantitative Investment**" MScT Program by Polytechnique-ENSAE

Goldman Sachs Quant Finance Seminar
Paris – Jan. 23, 2026

Based on a joined work with Giulia Livieri (LSE)

Positioning: What Are We Taking About?

Alternative data (satellite images, geolocation, etc) give (partial) information on the current state of the economy [Capponi and Lehalle, 2023]. In principle, they **unlock two drawbacks** of financial data (market prices and flows, fundamentals, etc):

- real-world phenomena are **more stationary** that economic/financial ones (less exposed to economic conditions and business cycles);
- they can be checked against a ground truth.

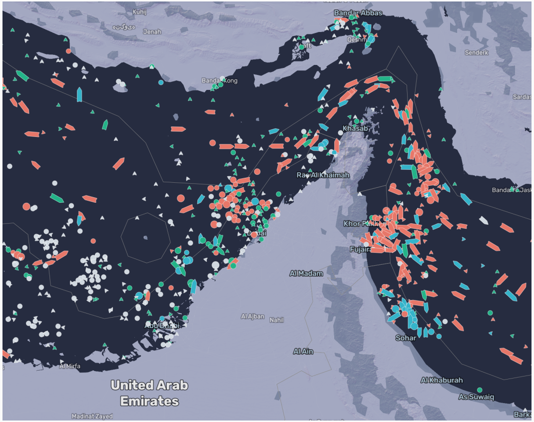
And obviously financial prices are highly influenced by **nowcasting** (that contains information on forecasts since the real world has inertia) that can be built using such datasets.

Data on (almost real-time) **maritime traffic** are available; what kind of nowcasting indexes can be built using that?

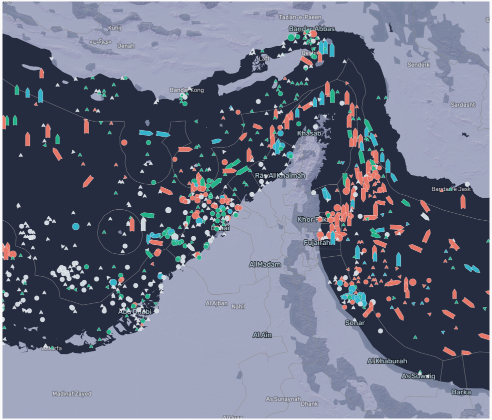
- You can rely on standard econometric techniques, typical seasonality + trend modelling, telling you what flows of goods are expected there and there. You rely on **stationarity of the stochastic processes** you are considering.
- But: **What if a port is closed?** because of a war (Ukraine), pandemic (China), wildfire (California), etc. All **these stationarity assumption fall**...

What Is Different Between These Two Pictures?

Maritime traffic (2025-06-24) [US bombings the 22nd]



Maritime traffic (2026-01-18)



We used the Shipfix data (sourced from a decentralized email exchange among ship owners); the past 10 years, billions of data points have been collected, anonymized, and aggregated across 243 cargo types.

Our (Mean Field Games) Approach To Solve This Problem

The rules of the games played by the ship owners is not changing, can we **infer the parameters of these rules** given some data points?

What are these parameters

- **expected gains** to ship a unit of good from port i to port j
- **transportation costs** from i to j (adjusted by tariffs)
- **congestion costs** at j.

If we model this game and can express observables (i.e. goods effectively routed from i to j this week) as formulas of the parameters of the game: we have a chance to **invert the process** and build **estimators of the parameters of the game**.

This is the story of this paper: we started a mix of empirical and theoretical explorations three years ago, using different datasets and during internships and Master projects.
☞ Thanks to Evan Chow, Michele Bergami, Borel Martial Domgue Defo, Simone Moawad, and Barath Raaj Suria Narayanan.

The Journey Of This Talk

A Mean Field Game Of Shipping: Setting The Scene

A Mean Field Game Of Shipping: Solving A Simplified Version

A Statistical Procedure To Invert The Problem

A First Application Of Our Statistical Procedure (And Its Limitations)

A Mean Field Game Of Shipping: Setting The Scene

A Mean Field Game Of Shipping: Existing Work

Existing papers are **exploring limited aspects of the problem** (f.i. [Anderson et al., 2008] for two ports: Busan and Shanghai; [Kaselimi et al., 2011, Do et al., 2015, Yip et al., 2014, Ignatius et al., 2018] for a two-steps game for the transportation between two ports, or when ports are competing with each others). They are very useful to **understand** the standard shapes for **the different sources of costs and gains**.

Their main limitation comes from the standard combinatorial difficulty in games with a lot of players, it is where **Mean Field Games** (MFG) are supposed to help.

On the MFG side, [Baillon and Cominetti, 2008] provide a framework for Markovian traffic equilibrium, and [Guéant, 2015] one for congestion effects on graphs. [Tanaka et al., 2020] Is probably the closest work to ours; a big difference is that the players of their games are the ships, whereas **for us the players are the owners of the shipping companies**, or (under standard MFG assumptions of interchangeability) the ports.
We find this more convenient and it allows us to obtain results that we can use for our end goal: the estimations of the parameters of the game.

All these papers (plus a long discussion with Jean-Michel Lasry 6 years ago) have been very useful for our work.

Variable Of Interest For Our Mean Field Game

$\Phi^n_{i,j}(t^+)$ is the flow of ships from seaport i to seaport j, carrying good n; once the fleet's capacity at i is known. We consider the total capacity of the fleet F^n for a category of good is fixed during the observation phase; **our control will be the transition matrix** $Q^n_{i,j}$ in:

$$\Phi^{n,Q}_{i,j}(t^+) = F^n P^n_i(t)\, Q^n_{i,j}(t^+), \quad \varphi^n_i := F^n P^n_i \text{ is the waiting capacity at } i.$$

- The **expected margin**:

$$\mathrm{M}^{n,Q}_i = \sum_{j=1}^{K} \Phi^{n,Q}_{i,j}(t^+)[v^n_j - v^n_i] := \sum_{j=1}^{K} \Phi^{n,Q}_{i,j} M^n_{i,j}.$$

- The expected **transportation cost** at the decision time:

$$C^{n,Q}_{i,\cdot} = c_n \sum_{\substack{j=1\\ j\neq i}}^{K} \left|\Phi^{n,Q}_{i,j}(t^+)\right|^{\gamma} g(T_{i,j}).$$

- The **cost of congestion**, or crowding:

$$R^Q_i(t^+) = \sum_{j=1}^{K} r_j \left| \sum_{n=1}^{N} \sum_{\substack{\ell=1\\ \ell\neq i}}^{K} \Phi^{n,Q}_{\ell,j}(t^+ + T_{i,j} - T_{\ell,j}) + \sum_{n=1}^{N} \Phi^{n,Q}_{i,j}(t^+) \right|^{\rho}.$$

Overall Objective Function Of Our Mean Field Game

$$\mathcal{J}^N_{i,n}(Q^n_{i,\cdot}, Q_{-i,\cdot}) = \underbrace{\sum_{j=1}^{K} \Phi^{n,Q}_{i,j} M^n_{i,j}}_{\text{expected gain}} - \underbrace{\sum_{j=1}^{K} r_j \left| \sum_{\substack{m=1\\ m\neq n}}^{N} \sum_{\ell=1}^{K} \Phi^{m,Q}_{\ell,j} + \sum_{\substack{\ell=1\\ \ell\neq i}}^{K} \Phi^{n,Q}_{\ell,j} + \Phi^{n,Q}_{i,j} \right|^{\rho}}_{\text{congestion cost}} - \underbrace{c_n \sum_{\substack{j=1\\ j\neq i}}^{K} \left| \Phi^{n,Q}_{i,j} \right|^{\gamma} g(T_{i,j})}_{\text{transportation cost}}.$$

We will need to consider the expectation of this function, a very important point to notice is that at the limit, the game should stationaries, and the flows φ^n_i waiting at port i to transport good n is the eigenvector of Q^n, thus the stationarized maximization to solve function (written here for a quadratic version) becomes

$$\max_{\forall i,n\; Q^n_{i,\cdot}} \left\{ \sum_{j=1}^{K} \varphi^n_i \cdot Q^n_{i,j} M^n_{i,j} - \sum_{j=1}^{K} r_j \left[\sum_{\substack{m=1\\ m\neq n}}^{N} \varphi^m_j + \varphi^n_j \cdot Q^n_{\ell,j} + \varphi^n_i \cdot Q^n_{i,j} \right]^2 - c_n \sum_{\substack{j=1\\ j\neq i}}^{K} (\varphi^n_i \cdot Q^n_{i,j})^2 g(T_{i,j}) \right\}.$$

An important simplification is due to the negligible effect of one decision with respect to the mean field of shippings (*Proposition 1* in the paper); formally $\int_{\ell\neq i} \varphi^n_\ell Q^n_{\ell,j}\, dm(\ell) = \varphi^n_j$.

A Mean Field Game Of Shipping: Solving A Simplified Version

Theorem

Consider the constrained maximization problem of *the previous equation*. For every origin seaport $i \in \{1,\dots,N\}$, every destination port $j \in \{1,\dots,N\}\setminus\{i\}$ and for every good $n \in \{1,\dots,N\}$, let

$$\widetilde{M}^n_{i,j} := \frac{M^n_{i,j}}{2}, \quad w^n_{i,j} = \frac{1}{r_j + c_n\, g(T_{i,j})1_{\{j\neq i\}}}, \quad \underline{w}^n_{i,j} := \frac{w^n_{i,j}}{\sum_{\ell=1}^{K} w^n_{i,\ell}}, \quad \varphi^\bullet_j = \sum_{m=1}^{N} \varphi^m_j$$

where $1_{\{j\neq i\}}$ indicates the case in which the seaport i is different from the seaport j. Then the solution is given when it exists by

$$Q^n_{i,j} = \frac{w^n_{i,j}}{\varphi^n_i}\left\{\widetilde{M}^n_{i,j} - \sum_{\ell=1}^{K} \underline{w}^n_{i,\ell}\,\widetilde{M}^n_{i,\ell} - \left[r_j\varphi^\bullet_j - \sum_{\ell=1}^{K}\underline{w}^n_{i,\ell}\, r_\ell\varphi^\bullet_\ell\right]\right\} + \underline{w}^n_{i,j}.$$

The terms are clear: w is invert proportional to the (transportation and congestion) costs that normalize to $\underline{w}$, and the operation $x \mapsto (x - \underline{w}^\top x)$, applied to margins and congestion flows, maps a variable to its relative value, cost-wise.

Where Does It Comes From?

You need Lagrange multipliers to keep each rows of the stochastic matrix Q^n in the simplex

$$\frac{\lambda}{\varphi_i^n} =: \lambda_i^n = M_{i,j}^n - 2r_j(\varphi_j^\bullet + \varphi_i^n q_j) - 2c_n g(T_{i,j})\varphi_i^n q_j \mathbb{1}_{\{j\neq i\}}.$$

It gives you to optimal controls and an implicit fixed point (between Q and its stationary solution φ):

(1) $$Q_{i,j}^n = \frac{w_{i,j}^n}{2\varphi_i^n}\left(M_{i,j}^n - 2r_j\varphi_j^\bullet - \lambda_i^n\right), \quad \text{keep in mind } \varphi_j^n = \sum_{i=1}^K \varphi_i Q_{i,j}^n,$$

where the Lagrange multipliers are:

$$\lambda_i^n = \frac{1}{\sum_{j=1}^K w_{i,\ell}^n}\left(\sum_{\ell=1}^K w_{i,\ell}^n (M_{i,\ell}^n - 2r_\ell \varphi_\ell^\bullet)\right) - 2\frac{1}{\sum_{\ell=1}^K w_{i,\ell}^n}\varphi_i^n.$$

Moreover, this allows to obtain an equation in φ^n only:

Existence Condition

$$\left[1 + r_j \sum_{i=1}^K w_{i,j}^n\right]\cdot \varphi_j^n = \left(\sum_{i=1}^K w_{i,j}\right)\overline{M}_j^n + \sum_{\ell=1}^K \left(\sum_{i=1}^K w_{i,j}\underline{w}_{i,\ell}^n\right) r_\ell\, \varphi_\ell^\bullet + \sum_{\ell=1}^K \varphi_\ell^n\, \underline{w}_{\ell,j}.$$

A Statistical Procedure To Invert The Problem

A Statistical Procedure To Invert The Problem

Our goal is clear now: go from data to $\underline{w}$, then from $\underline{w}$ to the cost terms r_j, c_n, $g(T_{i,j})$ and to the relative values of the goods $\tilde{v}$ that can be deduced from the relative margins: $\widetilde{M}_{i,j}^{(n)} = v_j^{(n)} - v_i^{(n)}$.

The subtle point we have is due to the stationarity of the equation and the non deterministic observed flow. We propose to rely on this formulation of the theoretical flows:

$$
\begin{aligned}
\left\lfloor \varphi_i^n Q_{i,j}^n \right\rfloor_t &= w_{i,j}^n \left\lfloor \widetilde{M}_{i,j}^n - \sum_{\ell=1}^{K} \underline{w}_{i,\ell}^n \, \widetilde{M}_{i,\ell}^n \right\rfloor_\infty + \sum_{\ell=1}^{K} w_{i,j}^n (\underline{w}_{i,\ell}^n - 1_{\{\ell=j\}}) \, r_\ell^n \left\lfloor \sum_{m=1}^{N} \varphi_\ell^m \right\rfloor_{t:} + \underline{w}_{i,j}^n \left\lfloor \varphi_i^n \right\rfloor_t \\
Y_{i,j}^n(t) &= A_{i,j}^n + \sum_{\ell=1}^{K} B_{i,j,\ell}^n \, X_\ell^{\cdot}(t:) + C_i^n \, X_i^n(t) + \epsilon_{i,j}^n(t),
\end{aligned}
$$

meaning that if one builds the database of $\{Y_{i,j}^n(t), X_\ell^{\cdot}(t:), X_i^n(t)\}$ and performs $n \times i \times j$ linear regression, one gets good orders of magnitude of

$$
A_{i,j}^n \simeq w_{i,j}^n \left\lfloor \widetilde{M}_{i,j}^n - \sum_{\ell=1}^{K} \underline{w}_{i,\ell}^n \, \widetilde{M}_{i,\ell}^n \right\rfloor_\infty, \quad B_{i,j,\ell}^n \simeq w_{i,j}^n \sum_{\ell=1}^{K} (\underline{w}_{i,\ell}^n - 1_{\{\ell=j\}}) \, r_\ell^n, \quad C_{i,j}^n \simeq \underline{w}_{i,j}^n.
$$

You could have in mind a global maximum likelihood approach, but here he opt for the two stage regression, hence the second stage is, to solve simultaneously for all n, i, j

$$
A_{i,j}^n \simeq \frac{\tilde{v}_j^n - \tilde{v}_i^n}{r_j + c_n g(T_{i,j}) 1_{\{\ell=j\}}} \rightsquigarrow \min_{r_j, c_n, \tilde{v}_i^n, \tilde{v}_j^n} \sum_{\substack{i,j=1 \\ i \neq j}, n}^{K} \left\| A_{i,j}^n \cdot (r_j + c_n g(T_{i,j}) 1_{\{\ell=j\}}) - (\tilde{v}_j^n - \tilde{v}_i^n) \right\|^2.
$$

A First Application Of Our Statistical Procedure (And Its Limitations)

A First Application Of Our Statistical Procedure: Good and Country Selection

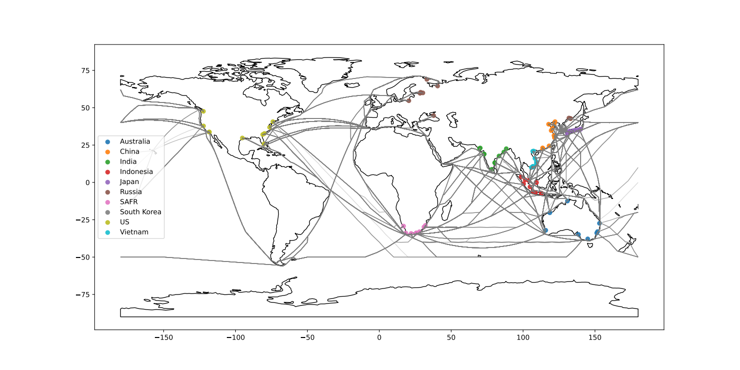
We use ShipFix data, **focus on one good only (Coal)** that is the most present in the dataset, and keep **5 main countries** that import and 5 main countries that export.

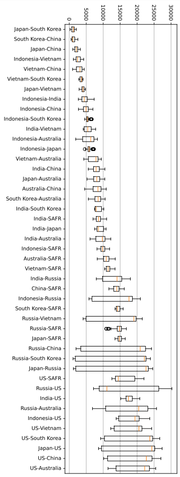
This is an approximation: for now, **we sacrifice stability of the numerical procedure to granularity**.

Imports	India	China	Vietnam	Japan	S.K.
Indonesia	165,802	164,098	26,333	11,316	15,152
Australia	82,115	31,210	19,276	19,470	8,412
SAFR	52,190	2,297	4,518	0,909	0,635
US	25,635	5,629	–	2,962	0,531
Other	17,443	17,727	2,521	2,841	7,082
India	15,275	0,768	0,148	–	–
Mozambique	13,292	–	2,059	–	–
Russia	11,707	22,043	3,711	3,985	10,611
China	7,034	3,712	2,149	2,985	2,664
Colombia	2,805	–	–	0,823	–
Singapore	2,726	2,389	0,420	–	–
Philippines	–	6,757	–	–	–
Canada	–	2,935	0,469	2,134	1,716

Exports	Indonesia	Australia	SAFR	US	Russia
India	165,982	81,776	52,145	25,588	11,623
China	164,523	31,176	2,283	5,636	22,045
Other	39,718	29,586	22,092	28,025	20,836
Thailand	26,347	–	–	–	–
Vietnam	26,197	19,053	4,500	–	3,676
Philippines	21,150	–	–	–	–
South Korea	15,100	8,410	–	–	10,617
Bangladesh	12,736	–	1,605	–	–
Japan	11,334	19,410	–	2,934	3,987
Taiwan	10,499	11,129	–	–	3,232
Malaysia	7,486	3,372	–	–	–
Australia	–	5,819	–	–	–
Brazil	–	6,715	1,214	11,155	2,175

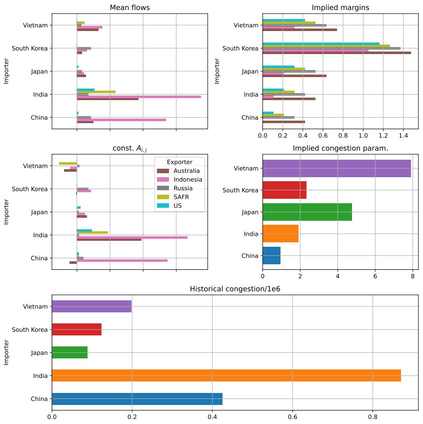
Extract of Table 2. The table shows average yearly imports and exports in the dataset, for the period from January 2, 2015, to May 17, 2025. SAFR stands for South Africa.



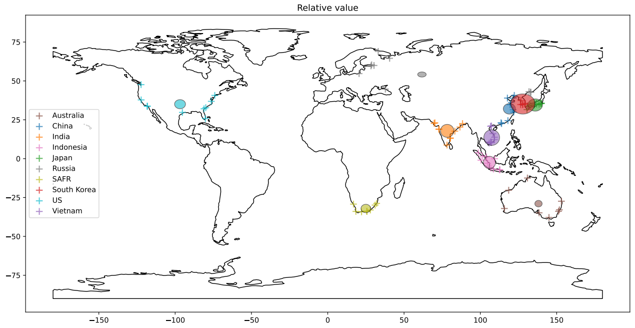


Ten ports selected by country and all our computed routes (darker gray means shared segments).

On the right: the level of approximation from 10 ports to one country is bad when the location of ports is diverse →



← mean exports (top left), the obtained coefficients of $(A_{i,j})$ of the linear regression (middle left), and the mean expected congestion on destination countries j (bottom panel). Implied congestion coefficient $\hat{r}$ (middle right) and expected margin, i.e. $\tilde{v}_j - \tilde{v}_i$ (top right).



Examples: for the model, Indian flows are explained by medium expected margins but a low congestion cost coefficient whereas the Vietnam flows are explained by attractive margins and an opposite congestion cost coefficient (do not forget the transportation costs: SAFR to Vietnam ≫ Indonesia to Vietnam).

Time To Conclude: Anticipate The Unpredictable

A Mean Field Game Of Shipping: Setting The Scene

A Mean Field Game Of Shipping: Solving A Simplified Version

A Statistical Procedure To Invert The Problem

A First Application Of Our Statistical Procedure (And Its Limitations)

A Good Baseline Model For A ML World Model?

Back to the quadratic version, the maximisation taking place globally is:

$$\max_{\forall i,n\ Q^n_{i,\cdot}} \left\{ \sum_{j=1}^{K} \varphi_i^n \cdot Q^n_{i,j} M^n_{i,j} - \sum_{j=1}^{K} r_j \left[\sum_{\substack{m=1\\ m\neq n}}^{N} \varphi_j^m + \varphi_j^n \cdot Q^n_{\ell,j} + \varphi_i^n \cdot Q^n_{i,j} \right]^2 - c_n \sum_{\substack{j=1\\ j\neq i}}^{K} (\varphi_i^n \cdot Q^n_{i,j})^2 g(T_{i,j}) \right\}.$$

If one fit embeddings $(e_i)_i$ for ports and $(g_n)_n$ for goods and write the vector of optimal control $Q^n_{i,\cdot}$ at port i for good n as $Q^n_{i,j} = \Phi_\theta(e_i, e_j, g_n)$, you get a global optimisation

$$\max_{e_\cdot, g_\cdot, \theta} \left\{ \sum_{j=1}^{K} \varphi_i^n \cdot \Phi_\theta\begin{bmatrix} e_i \\ e_j \\ g_j \end{bmatrix} M^n_{i,j} - \sum_{j=1}^{K} r_j \left[\sum_{\substack{m=1\\ m\neq n}}^{N} \varphi_j^m + \varphi_j^n \cdot \Phi_\theta\begin{bmatrix} e_\ell \\ e_j \\ g_j \end{bmatrix} + \varphi_i^n \cdot \Phi_\theta\begin{bmatrix} e_i \\ e_j \\ g_j \end{bmatrix} \right]^2 - c_n \sum_{\substack{j=1\\ j\neq i}}^{K} \left(\varphi_i^n \cdot \Phi_\theta\begin{bmatrix} e_i \\ e_j \\ g_j \end{bmatrix}\right)^2 g(T_{i,j}) \right\}.$$

Take-Aways And Next Steps

- To go beyond the natural stationarity of a problem, you can **lift it** at the level of **the rules followed by the participants of the game** to fuel **what if scenarios**.
- Mean field games is a possible direction: if you model the game such a way that the parameters of the game reflect in some observables, you may invert them and get the parameters.

☞ Of course it is a first step, and they are a lot of approximations:

- (for the theory) The stochasticity of the problem should **introduce some convexity** in the equations (preventing closed form formula) + an ϵ-Nash equilibrium theorem should be provided;
- (for the statistical approach) A **global maximum likelihood** should replace our layered approach;
- (for the empirical part) It should be tested on **more granular datasets** (port-level and not country-level) such that conclusions can be confronted in detail with real congestion shocks.

Thank you for your attention: A Throw of the Dice Will Never Abolish Chance

charles-albert.lehalle@polytechnique.edu

References i

Anderson, C. M., Park, Y.-A., Chang, Y.-T., Yang, C.-H., Lee, T.-W., and Luo, M. (2008).
A game-theoretic analysis of competition among container port hubs: the case of busan and shanghai.
Maritime Policy & Management, 35(1):5–26.

Baillon, J.-B. and Cominetti, R. (2008).
Markovian traffic equilibrium.
Mathematical Programming, 111(1):33–56.

Capponi, A. and Lehalle, C.-A. (2023).
Machine Learning and Data Sciences for Financial Markets: A Guide to Contemporary Practices.
Cambridge University Press.

Do, T. M. H., Park, G.-K., Choi, K., Kang, K., and Baik, O. (2015).
Application of game theory and uncertainty theory in port competition between hong kong port and shenzhen port.
International Journal of e-Navigation and Maritime Economy, 2:12–23.

References ii

Guéant, O. (2015).
Existence and uniqueness result for mean field games with congestion effect on graphs.
Applied Mathematics & Optimization, 72(2):291–303.

Ignatius, J., Tan, T. S., Dhamotharan, L., and Goh, M. (2018).
To cooperate or to compete: A game theoretic analysis on ports in malaysia and singapore.
Technological and Economic Development of Economy, 24(4):1776–1800.

Kaselimi, E. N., Notteboom, T. E., and De Borger, B. (2011).
A game theoretical approach to competition between multi-user terminals: the impact of dedicated terminals.
Maritime Policy & Management, 38(4):395–414.

Tanaka, T., Nekouei, E., Pedram, A. R., and Johansson, K. H. (2020).
Linearly solvable mean-field traffic routing games.
IEEE Transactions on Automatic Control, 66(2):880–887.

Yip, T. L., Liu, J. J., Fu, X., and Feng, J. (2014).
Modeling the effects of competition on seaport terminal awarding.
Transport Policy, 35:341–349.

References iii