Technology with unequal gains: Steamship and globalization

Shotaro Beppu*, Cesar Ducruet, Kenmei Tsubota, Ryuichi Shibasaki

December 2023

Abstract

Transportation technologies allow more market integration. The benefit could be unequal due to the adoption of this technology. This paper studies the rapid adoption of steamships in the late 19th century and its effect on the First Globalization and the Great Divergence. For this, this paper uses a novel deep-learning method to digitize historical shipping data to analyze the transition from sailing to steamships and the change in worldwide shipping patterns. Using the change in duration, this paper finds the advent of steamships increased trade and growth on average. However, colonized countries experienced fewer gains. To understand the mechanism, this paper incorporates differences in shipping technology to trade with heterogeneous firms. The estimate points out that the advantage of steamships might not have been captured due to the high cost of adoption for trade links involving colonies. This provides further insights into how shipping technologies affect economic activity through trade and how the fixed cost of adoption matters in gains from such technology.

*E&E lab, The University of Chicago (email: shotarob@uchicago.edu). I am grateful for the helpful comments from Professor Akihiko Matsui, Professor Yasuyuki Sawada, Professor Ryuichi Shibasaki, Professor Kenmei Tsubota, Dr. Cesar Ducruet, and participants in the Development Economics Study Group at UTokyo. The author is grateful to Dr. Cesar Ducruet (CNRS) for sharing shipping data. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n.[313847] "World Seastems". All errors and omissions are my own.

1 Steamship and unequal growth

How can technology and integration lead to growth? This paper uses the rapid adoption of steamships to illustrate how market integration induced by technology leads to unequal gains and how the fixed cost of adoption matters. The faster and bigger steamships replaced sailing ships within a couple of decades. During this period (1880-1900), we observed an explosion of trade (the First Era of Globalization) as well as divergent trends in countries' development (the Great Divergence). Due to limitations in data, we lacked an understanding of how steamships affected this global trend. At a higher level, steamships-induced changes give insights into how technology that increases integration leads to growth. The availability of technology is not often enough for gains to materialize. Growth disproportionately comes to those wanting to take up the technology, potentially exacerbating inequality. The steamship provides an ideal case study of how different countries decided to adopt one such technology and whether this helps explain global inequality.

To investigate these questions, this paper uses historical shipping data, Lloyd's Shipping Index. Although the data provides us with comprehensive information on trips of ships, either sailing or steamships, from 1880 onward, the use was limited due to data availability. This paper digitized this index en masse using deep learning. The methodology uses recent advances in optical character recognition, identifying texts in images, and table structure recognition, identifying the table structure in images. The lack of either makes using large historical data in table formats difficult. While the machine learning community made substantial advances in character recognition for historical texts, table structure recognition remains a problem even in contemporary documents. As such, this paper extended recent work in structure recognition using graph neural networks to digitize those tables. The methodology's success provides more than 30,000 trip data in 1880, 1890, and 1900 with information on the type and size of each ship, the origin and destination port of the trip and the duration.

The changes brought forth by steamships were substantial in maritime shipping and economic growth. First, steamship usage increased dramatically. In 1880, only 20% of the total tonnage was steamships. In 1900, the share of steamships increased to about 80%. The rapid adoption could be explained by how fast and big those steamships were compared to sailing ships. Steamships drove down the duration between ports by more than half on average. In addition, the size of steamships in tonnage was about three times that of sailing ships. Using this change in duration and gravity equation models in international trade, this paper finds steamships led to a significant increase in countries' GDP per capita. The effect was driven by being closer to larger markets (market access) rather than being closer to larger suppliers (supplier access). This average gain, however, masked an important heterogeneous effect. In particular, while high-income countries benefited more from market and supplier access, colonized countries made less gains from both channels. An important fact that might explain this is the lower adoption of steamships by ports in colonized countries. Steamships induced a large shipping pattern change, and yet ports in colonized ports were, on average, 15 percentage points less in adoption, and a significant share of ports were still exclusively using sailing ships. This paper thus looks into when a reduction in duration leads to gains from trade using port-to-port level shipping patterns.

This paper combines the heterogeneous impact and the steamship adoption by incorporating the adoption of shipping technology in (M. Melitz, 2003). Specifically, firms with different productivity do not export, export using sailing ships, or export using steamships. This is explained by the two differences between sailing and steamships: the duration and the entry cost. If it is relatively harder to export using steamships, only the most productive firms export using steamships as they can compensate for the higher fixed cost. If the fixed cost is low enough, on the other hand, all the firms previously using sailing ships use steamships. In addition, firms

that could not export using sailing ships can now export due to better technology. Thus, the fixed cost of using the steamship can explain the large differences in trade gains. Using the share of tonnage carried by steamships and the differences in trade volumes in 1880 and 1900, the model estimates the fixed cost between countries. This paper finds that the estimated fixed cost is higher for colonized countries, implying that these countries could not fully benefit from steamships. The cause of the differences in the fixed cost, such as port investment, competition, and extractive colonial relationships, which can be incorporated into the general equilibrium, is a fruitful area for further studies.

The remainder of the paper is as follows. Section 2 provides historical background of the late 19th century and literature in shipping, trade, and development that helps understand the transition from sail to steamship and the resulting gains. Section 3 details the digitization process of the shipping data as well as the data used both at the shipping and country level. Section 4 documents steamship adoption and its effect on country development. Section 5 provides a theoretical framework to understand how a reduction in duration and fixed cost of adoption affects trade and welfare. Section 6 estimates the fixed cost implied by the model using data. Section 7 concludes.

2 Background and literature

2.1 Period of Change

The period from 1880 to 1914 is called the First Era of Globalization. With the lowest tariff in history, trade between countries increased exponentially (Pascali, 2017). Nations in South America, for example, reversed its protectionism policy and reduced tariffs greatly (Britannica, 2023). Great Britain dominated the sea as it expanded all over the world. At the same time, countries such as the United States and Germany were rapidly expanding, challenging the dominance of Great Britain. In parallel, the period also coincides with what we call the Great Divergence where the Western civilization took off while other regions fell behind. This period was coined "Belle Époque" in French, representing optimism, peace and economic growth. On the other hand, the European powers moved inland in Africa, claiming most of the land during this period (New Imperialism). India was under the rule of Great Britain and lost its place to the U.S. in terms of GDP (Bolt & van Zanden, 2014). The decline of China and other Southeast Asian nations also occurred in this period. Related to this paper, the First Sino-Japanese War, which took place in 1894-1895, was the first warfare where modern steam-propelled ironclad warships were used in battle. The demise and the division of China accelerated thereafter.

In this global setting, numerous innovations came out. Telephones, transatlantic cables, light bulbs and automobiles began to be used from 1870 to 1890. Steam engine and electricity was the key invention. This propelled the Second Industrial Revolution and drastically changed transportation. Railroads and automobiles expanded across the world and changed transportation on land. Steamships combined with triple-expansion engines began to be used in the 1880s, allowing long-distance travel to be feasible for the first time. Until the advent of the diesel engine after World War II, steamships became the dominant mode of transportation at sea. During this era, along with inventions in refrigerators and metal, steamships allowed safer and faster transportation. The main advantage of steamships compared to sailing ships was also due to less concern for wind at sea. As (Pascali, 2017) uses in its analysis, steamships travelled a shorter route where sailing ships had to take a detour due to the wind. In addition, the use of steamships further accelerated due to the opening of the Suez Canal in 1869, as the wind conditions in the Mediterranean and the Red Sea meant steamships were much more feasible than sailing (Fletcher, 1958).

This paper, then, investigates how this innovation in the sea affected economic growth and, in particular, the Great Divergence.

2.2 Transportation, trade, and development

This paper uses three strands of literature in economics. First is empirical analyses of gains from trade. Second is how transportation matters in trade. Third is how technologies can affect the global pattern of trade. In addition, due to the methodologies used, this paper provides a novel method to digitize historical documents, which will be explained in the data section.

First, this paper builds on empirical work investigating trade gains. The ideal regression researchers would run is welfare on openness, often proxied by trade volumes. This, however, has omitted variable bias as higher GDP per capita likely leads to more trade for variety and more trade is associated with better institutions to promote industries that promote growth. Numerous research employed instrumental variable approaches to tackle this issue (see (Donaldson, 2015) for further review). The first of such works was (Frankel & Romer, 1999), which uses distance to other countries as an instrument. As distance is likely correlated with other variables that affect welfare (proximity to other large developed countries for instance), subsequent works used transportation investments as a natural experiment. (Feyrer, 2021) uses the switch from ocean shipping to aeroplanes as an exogenous change in distance, (Feyrer, 2019) used the closure and reopening of the Suez Canal, and (Donaldson & Hornbeck, 2016) used the expansion of railroad networks to see gains within the U.S. These provide positive estimates from more openness. However, some research documented the negative consequences of better proximity to other regions. (Faber, 2014) finds that regions in China that had better access to the national highway experienced a decrease in economic activity. (Campante & Yanagizawa-Drott, 2018) uses the discontinuous adoption of large aeroplanes in passenger travel to show that although the overall effect of aeroplanes is positive, the poorest places did not benefit. On the other hand, (Okoye et al., 2019) shows that the colonial railway benefited the least accessible places. The most related result is from (Pascali, 2017). This used the change in the distance travelled by sailing to steamships and found that the gains from trade only accrued to countries with strong institutions. This paper investigates why we observe those negative effects for select regions. The idea explored in this paper that technology such as steamships does not lead to immediate take-up is applicable to many situations where the benefit of those technologies is related to past economic conditions and does not lead to convergence.

Second, this paper considers port and shipping an important aspect of economic growth. In addition to the papers above using sea distance, numerous papers used changes in shipping and ports to answer this question. (Heiland et al., 2022), for example, used the 2016 Panama Canal expansion as a quasi-natural experiment and showed that expansion not only affected trade flows directly exposed to the canal but also indirect effects on world trade due to the shipping network. In terms of economic activity, (Galiani et al., 2023) used the opening of the Panama Canal on economic activity in the U.S. (Coşar & Fajgelbaum, 2016) looked at the interplay of regions that are nearby ports and ports themselves and how more trade induces specialization across cities. An important aspect of these analyses are ports. Ports are a source of efficiency gains by investment. Inefficient ports, on the other hand, are as if they impose additional tariffs on goods arriving. (Blum et al., 2019) considers inventory management and shows that improvement, especially in poor countries, can increase trade. (Feenstra & Ma, 2014) showed that port efficiency is important in the extensive margin, i.e. trade facilitation. Other papers used the advent of containerization to see how port improvements affect various outcomes. A number of papers looked into the increase in trade (Bernhofen et al., 2016) (Coşar & Demir, 2018). Others looked into local effects around ports that integrated into the world market (Brooks et al., 2021). Related to the paper's time period, (P. Fajgelbaum & Redding, 2022) uses Argentina dur-

ing the late 19th century to show how closeness to ports during this time induced more trade and structural transformation. With digitized data on shipping data, this paper investigates how steamships changed the port structure.

Third, this paper uses theoretical frameworks in trade. First, the empirical estimate largely follows the market access approach (S. Redding & Venables, 2004) (Donaldson & Hornbeck, 2016) (S. J. Redding & Turner, 2015). (S. Redding & Venables, 2004), in particular, expands the (Krugman, 1980) model of economies of scale and monopolistic competition to show that the closeness to other countries increases welfare. Papers stemming from (M. Melitz, 2003) used (Krugman, 1980) to incorporate heterogeneous firms. As (Helpman et al., 2008) estimated the trade relationship, this model features a fixed cost of entry and makes only the more productive firms export. This additional extensive margin for trade is shown to dominate compared to the intensive margins of trade when using a Pareto distribution for firm productivity (Chaney, 2008). (Arkolakis et al., 2012) shows that the gains from trade can be captured by the initial trade share and trade elasticity. To account for gains from transportation specifically, papers such as (Allen & Arkolakis, 2022) incorporated transport networks and more structure to the local economic activity to derive changes in wages. Due to the data this paper has, this paper uses an extended version of the canonical trade model to see the gains from steamships. Steamships can be considered a technology that reduces variable trade costs but may introduce higher fixed costs upfront. From this perspective, this paper is related to models incorporating market access (Arkolakis et al., 2021), reduction in duration as a quality upgrade (D. L. Hummels & Schaur, 2013), binary technology adoption (Bustos, 2011). This paper is further related to understanding the unequal gains from trade (M. J. Melitz & Redding, 2015) (P. D. Fajgelbaum & Khandelwal, 2016).

3 Digitized shipping and trade data

This paper uses two types of data. One is shipping data at the port level and another is trade and economic development at the country level. For port-level data, this paper digitizes shipping log data. For country-level data, this paper uses data created by (Pascali, 2017) and TRADHIST (Fouquin & Hugot, 2016). This paper combines these data for the years 1880, 1890, and 1900.

3.1 Port level data

3.1.1 Overview

For shipping data, this paper uses a traditional shipping index, "Lloyd's Shipping Index." Since the late 16 century, the merchant company Lloyd's List has recorded maritime flows of merchant vessels. In 1880, the company started publishing Lloyd's Shipping Index, reporting weekly the flows of ships globally for both sailing and steamships. Importantly, these publications detail the movements of most of the world's fleet, including vessels insured by other companies, making it highly representative of global maritime trade. Although this provides us with the most comprehensive movements of ships at this time, it was traditionally used only for exploratory purposes (Ducruet & Itoh, 2022). Figure 1 shows an example of the first few rows of the tables. In this case, the index records, from the left, the association the ship belongs to, the ship name, the master of the ship, the tonnage in brackets, the nationality, the type of ship, the origin and when it left the destination, and the latest whereabouts often accompanied by indicators of status (e.g., "Ar" means Arrived). With this information across multiple weeks, it is possible to describe how a particular ship moved to different places in detail based on characteristics such as nationality, ship type, and tonnages. Furthermore, it is possible to observe how the shipping

Figure 1: Example of 1900's Index covering sailing ships

Reg.	Ship	Master	Ton.	Flag	Rig	From	For	Latest Reports
R V	E A O'Brien	*Pratt*	(1038)	Br	bq	Manilla Apr 4	Boston	ArSept10—ForBuenosAyres
R V	E B Sutton	*Carter*	(1639)	Am	s	Honolulu Oct 13	New York	
* R	E C Mowatt	*Hersey*	(1026)	Am	bq	Philadelphia Sept6	Table Bay	Pd Marcus Hook Sept 6
*	E J Spence	*Stronach*	(519)	Br	bq	Singapore July 26	Mauritius	Ar Sept 12
V	E J Spicer	*Cochran*	(1268)	Br	s	Table Bay July 21	Ncstle(NSW)	Ar Spt2—ForWSCAmerica
R	E K Wood	*Hansen*	(452)	Am	sc	Tacoma Aug 25	Haiphong	
V	E S Hocken	*Willcock*	(249)	Br	bq	Clyde Aug 4	Table Bay	Ar Oct 25
*	Eagle Crag	*Shimmin*	(1347)	Br	bq	Cardiff Sept 15	CaletaBuena	Pd Barry Island Sept 15
*	Earl Cadogan	*Williams*	(1334)	Br	bq	Ncstle(NSW) Aug24	Antofagasta	Ar Oct 15
*	Earl Derby	*Mackintosh*	(961)	Brbq		Brisbane Sept 8	Ncstle(NSW)	ArSpt15--ForWSCAmerica

network has evolved over the years. Above all, this data provides the total tonnage from one port to another, allowing port-level analysis combined with country-level data. This research uses three years of data (1880, 1890, and 1900) from October to December due to limitations in data. Each year, this research obtains shipping movements for both sailing and steamships. An important limitation of this data is that it does not contain any information on cargo and where those were unloaded. Thus, even if one ship travelled from one port to another, it may well be possible that that cargo went to another destination. This paper shows later, however, that the total tonnage is highly correlated with the total trade at the country level, suggesting that the tonnage is a good proxy for trade between countries.

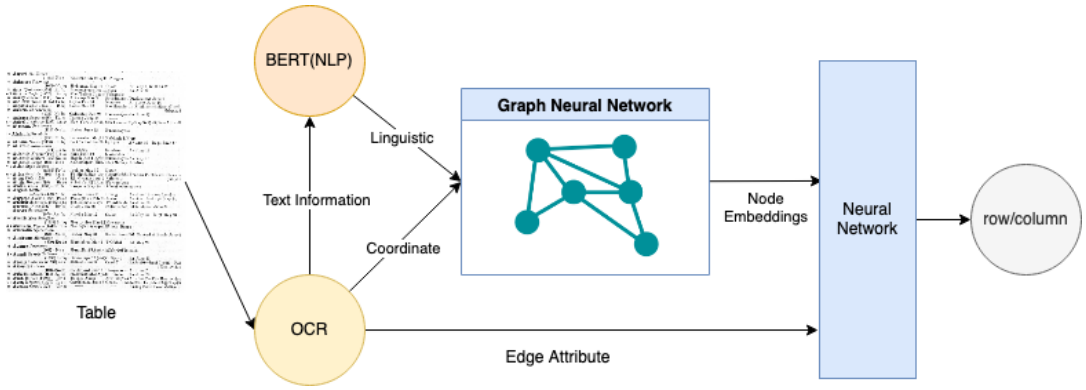
This paper further extends the database of ports created by (Ducruet et al., 2018). Here, for each port, the data includes the longitude and latitude of the port and the nearest city. This paper further categorizes each city by its location to the port: coastal, downstream, inland, and upstream. This category is useful to obtain as the difficulty of ports to be used for steamships depended on the geographical characteristics as this paper later shows in the data. For example, a number of upstream ports were unsuited for steamships as it was hard to expand (Ducruet & Itoh, 2022).

The issue we have when using duration is that there is only data when there is a ship travelling. As duration itself is a product of geographical characteristics (wind and ocean conditions), for the unobserved duration, this paper computed the distance similar to (Pascali, 2017) and used predicted sailing/steamship time using observed observations. This will be used in understanding the intensive and extensive margin.

3.1.2 Digitization

The data were only available in images. This paper digitized these images extending explored methods in deep learning. Recently, digitizing historical documents using deep learning has started to emerge in the field of economics. LayoutParser (Shen et al., 2021), for example, created an easy interface combining recent advances in optical character recognition to be used to obtain texts from documents. However, these tools are not sufficient to create structured datasets such as tables. Current research still requires intensive post-processing of these texts to associate which text belongs to which row. This is indispensable as, in the case of the shipping data, we need to know which ship went to which port, which is hard to obtain due to distortions in the data. Matching these variables to one ship is difficult using simple heuristics due to irregularities such as spanning rows which are prevalent in this dataset. Thus, this paper extends the methodologies in table structure recognition (Gatos et al., 2005) (Kasar et al., 2013) (Hao et al., 2016) (Rashid et al., 2017) (Schreiber et al., 2017) (Dai et al., 2017). This field, in addition to optical character recognition to recognize texts in documents, uses further machine learning

Figure 2: Overview of Digitization Architecture



to create tables. This paper specifically explores graph neural networks due to recent success (Qasim et al., 2019). As past work only explored these methodologies in tables created by Latex or HTML files, this paper is the first to explore this usage in historical documents. In particular, this paper provides a case where a graph neural network is fairly successful in obtaining data from historical tables where there are irregularities in texts as well as structures such as spanning rows.

The structure of the digitization architecture is illustrated in figure 13. This architecture has three main parts: construction of graphs (from image to graph using OCR), learning text representations (BERT and Graph Neural Network), and prediction of relationships (The final proponent using a neural network). This study uses the OCR tool, Google Cloud Vision, to construct graphs to extract texts and their location. By extracting locations, this study obtains the closeness and the direction between pairs of texts, which the later stage uses as edge attributes. The extracted texts are encoded via the BERT model, a prominent natural language processing model representing text well in many instances (Devlin et al., 2019). During training, each text is matched to hand-labelled data, creating a graph for each table. For learning text representations, this study uses a graph neural network to learn the graph's structure and outputs a numerical representation for each text extracted. For the final stage, predicting relationships between texts, this study uses a standard neural network given the text's representations and edge attributes. The output is the probability of whether the pair of texts belong to 1) the same row, 2) the same column, and 3) the same cell. The latter two stages use the PyTorch geometric library (Fey & Lenssen, 2019) built on PyTorch (Paszke et al., 2019) in order to test out multiple models.

This paper tests the feasibility of this digitization process by comparing hand-coded tables for each year in a different month. The detailed methodology, result and validation using historical facts are devoted in the appendix. In the end, this methodology was able to capture 60% of the total rows of shipping movements. Whether this is a good result is conditional on future progress on table structure recognition for historical documents. Yet, the general conclusion is that the data obtained is accurate enough to observe yearly port-to-port level tonnage as well as average duration. Since Lloyd's Index is separated by sailing and steamships, the extent to which we see the adoption of steamships is accurate.

3.2 Country level data

The second data is trade. For this, this research uses data compiled from (Pascali, 2017) and (Fouquin & Hugot, 2016). The former is said to be the most comprehensive trade data for this period (1880 to 1900); it is a database for bilateral trade covering the second half of the nine-

Table 1: Variables for economic development

Name	Description	Source
GDP	Country wealth	*
Population	Country population	*, ***
Per Capita Income	Individual wealth	*, ***
Urbanization +50	% population in cities with more than 50,000	*
Urbanization +100	% population in cities with more than 100,000	*
Colony	Colonized by a country or not	*, **
Tariffs	Tariff rate for each trading partner or overall	**
Constraints on the executive	Institutionalized constraints on power by POLITY IV	*

* indicates (Pascali, 2017), **(Fouquin & Hugot, 2016), and *** (Bolt & van Zanden, 2014).

teenth century constructed from primary sources and, overall, the data consist of almost 24,000 bilateral trade observations for nearly 1,000 distinct country pairs (Pascali, 2017). On the other hand, "TRADHIST" (Fouquin & Hugot, 2016) has an advantage in length due to the purpose of exploring the two modern waves of globalization: the First Globalization of the nineteenth century and the post-World War II Second Globalization (Fouquin & Hugot, 2016). Overall, they provided about 1.9 million bilateral trade observations from 1827 to 2014. After matching countries in the shipping data described in section 3.1, the former had more coverage; hence, this research primarily uses the data from (Pascali, 2017). This research uses the latter for analyzing the long-term and robustness checks.

In addition to trade volumes, this research also uses 344 entries on total exports and 154 entries on the share of exports in non-agricultural products (37 countries every 5 years from 1845 to 1905, with gaps) compiled by (Pascali, 2017). (Pascali, 2017) writes, "approximately half a million entries on exports by product were collected from primary sources; a SIC (rev1) code was then assigned to each of these products; and, finally, the share of exports that did not belong to the SIC categories 0 (food and live animals), 1 (beverages and tobacco), 4 (animal and vegetable oils and fat) was computed."

Lastly, this research uses various economic indicators to gain insights into how steamships affect development. This paper uses data compiled by (Pascali, 2017) and (Fouquin & Hugot, 2016). Summary statistics of the variables collected are described in the table 1. In addition to these variables, this research also used geographic information such as whether a pair of countries is contiguous and the latitude/longitude of countries (Pascali, 2017).

4 Changes in shipping, trade, and development

Using both port-level and country-level datasets, this paper first documents 1) how steamships were adopted, 2) the effect on countries' welfare, and 3) the change in port-level trade that leads to the theoretical framework.

4.1 Changes in geography

4.1.1 Adoption of steamships

The Lloyd's Shipping Index provides us with the travel of ships categorized into either sailing or steamships. We can thus therefore see which travels (port to port) were most used by steamships. Note that the advent of steamships in international trade was in the 1860s. As the time period we have is from 1880, we still observe a number of steamships at this time. Figure 3 shows the total tonnage of ships travelled in the data set. The sailing ship was still the major mode of

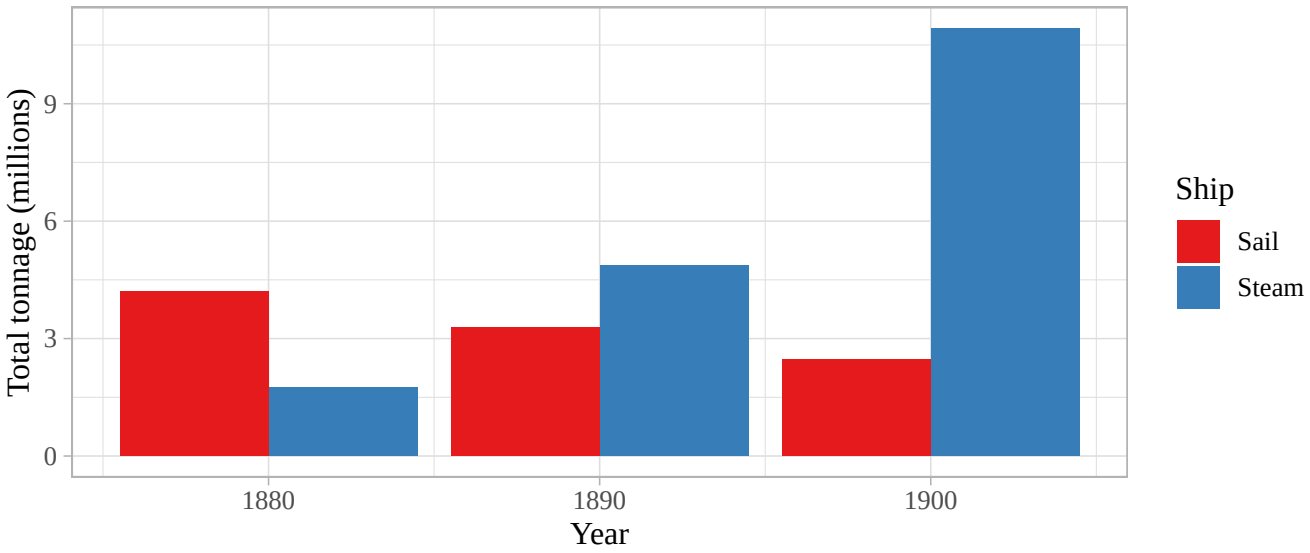


Figure 3: Total tonnage by year

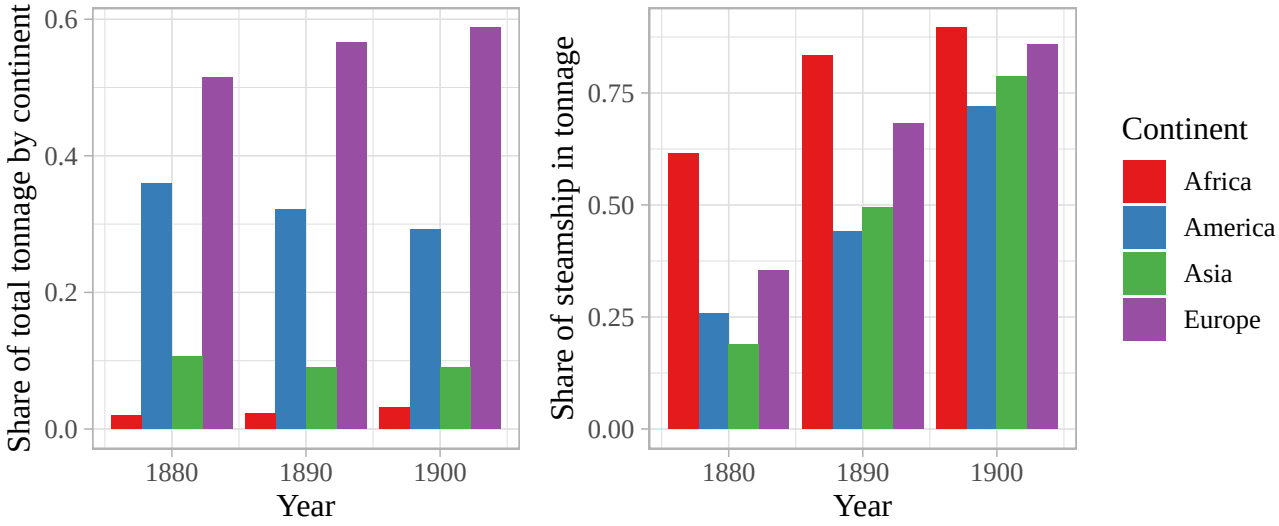
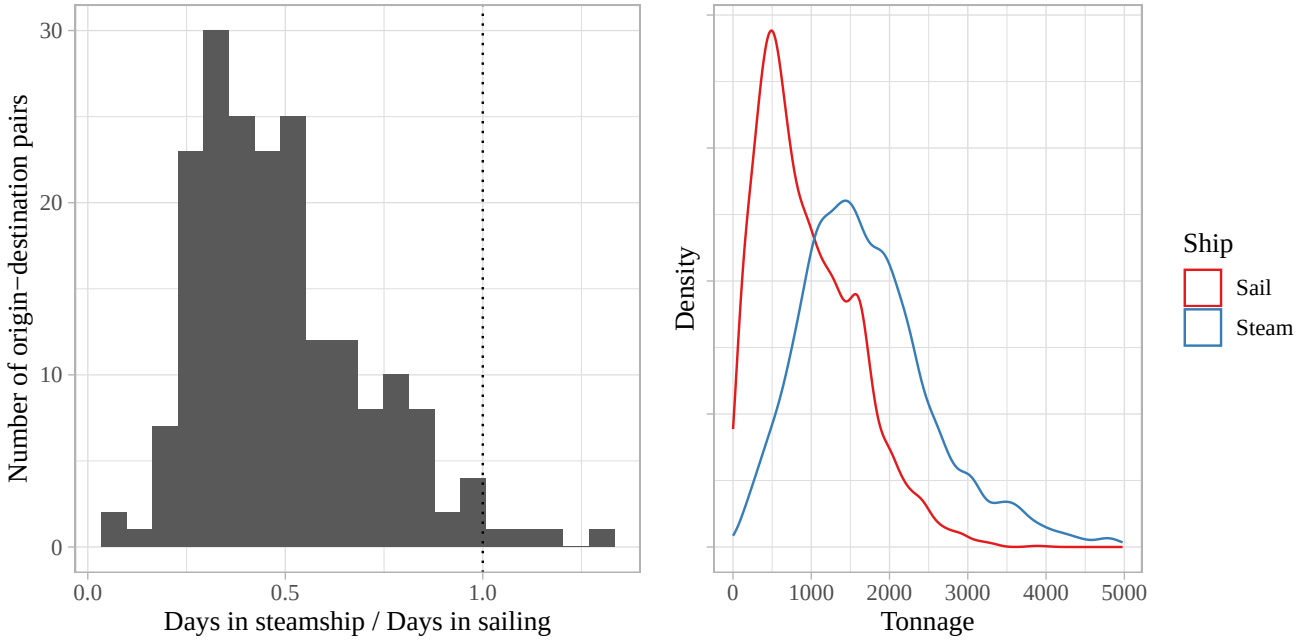


Figure 4: Tonnage and steamship share by continent

shipping in 1880 but steamship took over in 1890 and 1900. Steamship use exploded in 1900 surpassing 10 million tonnage travelled. Sailing ships, on the other hand, declined.

Figure 4 shows how this adoption differs by continent. First, the left figure shows that ships are predominantly involved in Europe or America (in particular North America). The share of tonnage going through European ports is more than half for all years and combined with America's ports, the share goes up to about 80%. The adoption of steamships, on the other hand, is fairly similar across continents. This shows the share of the tonnage of steamships compared to the total tonnage that passed through ports in each continent. Across all continents, the share of steamships is trending upward. One noteworthy point is that steamship adoption was faster in Africa and Europe. The other continents, Asia and America, caught up in 1900 to a similar level. We see that the adoption of steamships in 1900 was above 75% from about 25% in 1880.



(a) Duration

(b) Ship size

Figure 5: Differences between sailing and steamship

4.1.2 Steamships were faster and bigger

What drove the use of steamships? Figure 5 illustrates that steamships were both faster and bigger. Figure 5a shows the relative average duration of a steamship compared to the average duration of a sailing ship for the same origin-destination pair. Thus a value below 1 (the vertical dotted line) means that steamships were faster than sailing ships for a particular route. We can see that, in all but a few of the routes, steamships are faster to reach. In fact, the median duration change is 50%. In addition, figure 5b shows the density of tonnage of ships by ship type. We can see steamships were much bigger compared to sailing ships.

4.2 Changes in trade and development

How did this change in geography affect the trade and growth of countries? For this, this paper uses the exogenous change in duration induced by steamships. Specifically, this paper employs the market access approach (S. Redding & Venables, 2004) (Donaldson & Hornbeck, 2016).

First, this paper tests the assumption of exogeneity in duration change. The duration change came from various geographical characteristics, such as the wind and ocean currents. Yet, it is possible that, by chance, this duration is correlated with observable that affect trade and growth. Running a univariate regression of the relative change in duration for steamships on various port-level and country-level shows that the duration change can be considered good-as-random.

To estimate the impact of the steamship, this paper first runs the following regression

$$y_{c,t} = \alpha_0 + \alpha_1 \Delta duration_{c,1800} + \varepsilon_{c,t} \tag{1}$$

where $y_{c,t}$ is the outcome (GDP per capita) for country c at year t and $\Delta duration_{c,t}$ is the duration change for country c defined by

$$\Delta duration_{c,1800} = \sum_j (\hat{duration}^{sail}_{cj,1880} - \hat{duration}^{steam}_{cj,1880})\, population_{j,1880} \tag{2}$$

where $\hat{duration}_{cj,1880}$ is the predicted duration from country c and j in year 1880 using either sailing or steamships. This is akin to (Pascali, 2017) but using the actual duration as a measure of closeness between country c and j. Column 1 of table 2 shows that being closer to the rest of the world led to positive growth in GDP per capita. This differs from the null result obtained by (Pascali, 2017). There are two reasons for this result. First, the measurement error using predicted distance from distance likely underestimated the estimates. Second, (Pascali, 2017) instruments this change in duration for openness to trade. This instrument, however, likely violates exclusion restriction as steamships likely induced not just trade volumes but also immigration. This paper, therefore, asks about the effect of steamships, not the effect of trade, including all the possible benefits of being closer to other countries. This paper then defines another $\Delta duration_{c,1880}$ where the duration is now from country j to c. Column 2 reports that the effect of being closer to large supplying countries is positive but smaller than being closer to exporting markets. Column 3 shows that conditional on being close to exporting markets, closer to supplying countries has a null effect on per capita GDP. Note that this result is consistent with the micro-founded gravity model (Eaton & Kortum, 2002)(M. Melitz, 2003). First, being closer to larger markets to exports allows domestic firms to export more, raising wages. On the other hand, being closer to supplying markets has both positive and negative effects on welfare. First, as the country can import goods cheaply, this lowers the price index of the country, increasing real wages. However, the firms in the country will now face increased competition from foreign firms lowering the wages. Thus, while the decrease in duration change for exporting is predicted to be positive, the effect of being closer to supplying markets is ambiguous. The estimation shows this is indeed the case. For changes in duration for suppliers, these two effects cancel out.

Estimation using these gravity models (S. Redding & Venables, 2004) (Donaldson & Hornbeck, 2016) (Faber, 2014) used the market access approach to capture those effects precisely. Specifically, this paper defines market access and supplier access as

$$\begin{aligned} MarketAcess_{c,t} &= \sum_j duration_{cj,t}^{1-\sigma} m_j \\ SupplierAcess_{c,t} &= \sum_j duration_{jc,t}^{1-\sigma} s_j \end{aligned} \tag{3}$$

and run the following regression.

$$\Delta y_{c,t} = \beta_0 + \beta_1 \Delta MarketAcess_{c,t} + \beta_2 \Delta SupplierAcess_{c,t} + \nu_{c,t} \tag{4}$$

where $\Delta z_{c,t} = z_{c,t} - z_{c,1880}$. This follows from the gravity equation (S. Redding & Venables, 2004) and here m_j and s_i are proxied by the country's GDP. To run this estimation, we require σ the elasticity of trade with respect to trade cost. Past papers such as (Eaton & Kortum, 2002), (Donaldson, 2018), and (P. Fajgelbaum & Redding, 2022) use price information to estimate the elasticity. However, this was not possible in this paper due to data limitations. As such, this paper runs the conventional gravity equation with origin, exporter, and time-fixed effect with duration to estimate σ. This assumes that the trade cost is proportional to the changes in duration between sailing and steamships. Using PPML due to many trade flows with 0, the estimated σ is 1.08. This relatively small value is consistent with (Jacks & Pendakur, 2010) that documents low sensitivity of trade volumes with respect to freight rates and recent estimates using panel data of trade volumes and distance (Head & Mayer, 2014). Columns 4 to 6 in the table 2 present the result using market/supplier access clustering standard errors at the country level due to the treatment (being closer to certain markets) are assigned at the country-level. Consistent with the

Table 2: Effect of change in duration

Dependent Variables:	ln(GDP per capita)			$\Delta GDPpercapita$		
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
ln(Duration Change Exp)	0.705***		0.733**			
	(0.252)		(0.327)			
ln(Duration Change Imp)		0.381	-0.047			
		(0.261)	(0.305)			
$\Delta MarketAccess$				0.061***		0.102***
				(0.012)		(0.024)
$\Delta SupplierAccess$					0.065***	-0.065*
					(0.023)	(0.035)
Fixed-effects						
Year	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
R^2	0.183	0.057	0.184	0.465	0.237	0.447
Observations	63	63	63	59	59	58

Clustered (country,year)) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

theory, increasing market access (4) and supplier access (5) leads to a statistically significant positive effect on GDP per capita. Column 6, however, shows that, controlling for market access, supplier access negatively affects GDP per capita. This suggests a negative effect of being closer to larger supplying markets not correlated with market access.

Where does this negative effect come from? This paper shows that the increase in market and supply access decreases gains for colonized countries. Table 3 runs regression 4 with the interaction of whether the country is colonized or not and the institution level from Polity IV. Columns 1 to 3 interact with market and supply access with colonized indicator. This shows that while both access has a positive effect on the changes in GDP per capita, the colonized country reduced their welfare with an increase in those accesses. The magnitude suggests that colonized countries had null and negative effects from market and supplier access, respectively. This cannot be explained by low institution levels in colonized countries. Columns 4 to 6 show interacts access with whether the country had a below median executive constraints or not. The results show that both access was beneficial for all institution levels, suggesting that there are other mechanisms that led to a loss in welfare due to more integration. The next reduced analysis looks into port-level shipping data and gives insight into the mechanism.

4.3 Changes in shipping pattern

What could drive the heterogeneous effect of market and supplier access? First, this paper shows the substantial impact of steamships on shipping patterns by simply analyzing whether shipping patterns were destroyed or created due to the introduction of steamships. For this, table 4 shows the following regression results.

$$I_{ij} = \alpha_0 + \alpha_1 \Delta_{ij} + \alpha_2 \Delta^{steam-sail}_{ij,1900} + \alpha_3 \Delta^{steam-sail}_{ij,1880} + \alpha_i + \alpha_j + \varepsilon_{ij,t} \tag{5}$$

Here, Δ_{ij} is the difference in duration from port i to j between 1900 and 1880. $\Delta^{steam-sail}_{ij,t}$

Table 3: Heterogeneous effect of change in market and supply access

Dependent Variable:	$\Delta GDPpercapita$			
Model:	(1)	(2)	(3)	(4)
Variables				
$\Delta MarketAccess$	0.062***		0.031	
	(0.015)		(0.019)	
Colony	-0.122***	-0.097***		
	(0.037)	(0.028)		
$\Delta MarketAccess\times$ Colony	-0.087***			
	(0.030)			
$\Delta SupplierAccess$		0.091***		0.002
		(0.028)		(0.023)
$\Delta SupplierAccess\times$ Colony		-0.123***		
		(0.036)		
Instituion below median			-0.069**	-0.084***
			(0.029)	(0.029)
$\Delta MarketAccess\times$ Instituion below median			0.028	
			(0.018)	
$\Delta SupplierAccess\times$ Instituion below median				0.065*
				(0.038)
Fixed-effects				
Year	Yes	Yes	Yes	Yes
Fit statistics				
R^2	0.528	0.404	0.509	0.307
Observations	59	59	59	59

Clustered (country,year)) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 4: Steamships impact on link destruction/creation

Dependent Variables:	Ship link destroyed	Ship link created
Model:	(1)	(2)
Variables		
Duration change between 1900 and 1880	-1.22***	1.65***
	(0.251)	(0.325)
Steamship advantage in 1900	-1.03***	0.782***
	(0.135)	(0.112)
Steamship advantage in 1880	-1.13***	0.342
	(0.319)	(0.362)
Fixed-effects		
Port origin	Yes	Yes
Port destination	Yes	Yes
Fit statistics		
Observations	6,314	6,314
R^2	0.5092	0.5475
Within R^2	0.0634	0.0395

Clustered (Port origin) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

is the difference in duration by steamships minus that of sailing ships in year t (either 1880 or 1900). The outcome variable I_{ij} is either a dummy variable whether 1) there were no ships in 1900 while there were in 1880 (destroyed) and 2) ships in 1900 while there were no ships in 1880 (created). First, we can observe that links that did not benefit from a shorter duration time are likely to be destroyed (column (1)), while those that benefited from a shorter duration were created. Furthermore, links in which steamships were quicker than sailing ships in 1880 or 1900 were less likely to be destroyed and more likely to be created (column (2)).

Given this change, however, the adoption was not equal for all countries. Specifically, steamship adoption was slower for colonized countries. Table 5 describes the rate of adoption. Column 1 reports the share of ports that only have sailing, column 2 is for ports with steamships only, and column 3 is for the share of steamships in total tonnage. We can see that, compared to 1880, there are about 40 percentage points fewer ports only with sailing ships, about 30 percentage points more ports with steamships only, and an increase of 40 percentage points in terms of share of steamships. However, focusing on ports in colonized countries, there were 17 percentage points more likely to have sailing ships, eight percentage points fewer ports with steamships only, and 13 percentage points less steamship share in 1900.

The port-level analysis suggests that 1) the introduction of steamships caused a change in maritime trade structure but 2) the colonized countries experienced relatively less of those changes. If conventional theory that trade, or openness to trade, provides positive benefits, this implies that the colonized countries were left out of the large technological change. The following theoretical section makes this argument more precise.

5 Steamship adoption: Theory

To further understand the differences in gains from steamships and the adoption rate, this paper uses the canonical (M. Melitz, 2003) with two different shipping technologies. Specifically, the

Table 5: Changes in port and ship composition

Dependent Variables:	Only sailing	Only steam	Share steam
Model:	(1)	(2)	(3)
Variables			
Constant	0.649***	0.067***	0.177***
	(0.025)	(0.021)	(0.021)
Year 1900	-0.379***	0.279***	0.424***
	(0.032)	(0.028)	(0.027)
Year 1880 × Colony	0.050	0.010	-0.030
	(0.052)	(0.045)	(0.043)
Year 1900 × Colony	0.173***	-0.082**	-0.133***
	(0.044)	(0.038)	(0.037)
Fit statistics			
R^2	0.135	0.097	0.217
Observations	1,096	1,096	1,096

IID standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

marginal cost of using steamships and sailing ships is determined by the duration, but the fixed cost of using it is also different.

5.1 Trade with heterogeneous firms

This paper first illustrates how geography and fixed cost matter in the trade pattern observed. This follows (M. Melitz, 2003), (Chaney, 2008), and lecture notes by Treb Allen. This paper then builds upon this canonical model to analyze how steamships affects trade patterns and welfare.

5.1.1 Set up

World Denote the set of countries as S and denote $i \in S$ as an origin country and j as a destination country. In each country $i \in S$, there is an exogenous measure L_i of workers who supply a unit of labour inelastically. Let the wage of a worker in i w_i. Further, assume that labour is the only factor of production.

Demand A representative consumer in j gets utility U_j from the consumption of goods shipped by all other firms in all other countries with CES preferences over varieties.

$$U_j = \left(\sum_{i \in S} \int_{\Omega_{ij}} (q_{ij}(\omega))^{\frac{\sigma}{\sigma-1}} d\omega \right)^{\frac{\sigma-1}{\sigma}} \tag{6}$$

Here, $q_{ij}(\omega)$ is the quantity of ω shipped from country j and σ is the elasticity of substitution.

Supply For the supply side, following (Krugman, 1980), there is a continuum Ω of possible varieties the world can produce, and suppose that every firm in the world produces a distinct variety $\omega \in \Omega$. A firm uses $\frac{1}{\varphi}$ unit of labour to produce a unit of its variety. φ is drawn from a cumulative distribution function $G_i(\varphi)$ for firms in i. Let the set of varieties produced by firms

in country i be denoted $\Omega_i \subset \Omega$. There is a mass M_i of firms from country i and that firm must incur a fixed cost $f_{ij} > 0$ to export to destination $j \in S$. In addition to the fixed cost, firms are subject to iceberg trade costs $\{\tau_{ij}\}_{i,j\in S}$. This means that destination i must ship $\tau_{ij} \geq 1$ for a good to arrive in j. Equality holds when the good is shipped to its own country $i = j$.

5.1.2 Obtaining the gravity equation

Demand The properties of the CES function imply that a consumer in country $j \in S$ demands good $\omega \in \Omega$ following

$$q_{ij}(\omega) = p_{ij}(\omega)^{-\sigma} Y_j P_j^{\sigma-1} \tag{7}$$

where Y_j is the income of country j and

$$P_j = \left(\sum_{i\in S} \int_{\Omega_i} p_{ij}(\omega)^{1-\sigma} d\omega \right)^{\frac{1}{1-\sigma}} \tag{8}$$

is the Dixit-Stiglitz price. To obtain the total trade flows, we aggregate across all firms in country i and since the amount spent on a variety ω from j is $p_{ij}q_{ij}(\omega)$, we get

$$X_{ij} = Y_j P_j^{\sigma-1} \int_{\Omega_i} p_{ij}(\omega)^{1-\sigma} d\omega \tag{9}$$

Supply A firm with productivity φ maximizes profits by solving

$$\max_{q_{ij}(\varphi)_{j\in S}} \sum_{j\in S} \left(p_{ij}(\varphi) q_{ij}(\varphi) - \frac{w_i}{\varphi} \tau_{ij} q_{ij}(\varphi) - f_{ij} \right) \quad \text{s.t. } q_{ij}(\varphi) = p_{ij}(\varphi)^{-\sigma} Y_j P_j^{\sigma-1}. \tag{10}$$

From the first order condition, the firm will set the price as follows conditional on selling to destination j

$$p_{ij}(\varphi) = \frac{\sigma}{\sigma - 1} \frac{w_i}{\varphi} \tau_{ij}. \tag{11}$$

The total revenue becomes

$$x_{ij}(\varphi) = \left(\frac{\sigma}{\sigma - 1} \frac{w_i}{\varphi} \tau_{ij} \right)^{1-\sigma} Y_j P_j^{\sigma-1} \tag{12}$$

and the profit is

$$\begin{aligned} \pi_{ij}(\varphi) &= \left(p_{ij}(\varphi) - \frac{w_i}{\varphi} \tau_{ij} \right) q_{ij}(\varphi) \\ &= \frac{1}{\sigma} x_{ij}(\varphi) \end{aligned} \tag{13}$$

Trade flow Denote $\mu_{ij}(\varphi)$ and M_{ij} to be the equilibrium probability density function of the productivities of and the number of firms from country i selling to j. Then the average price can be written as

$$\begin{aligned}\int_{\Omega_i} p_{ij}(\omega)^{1-\sigma} d\omega &= \int_0^{\infty} M_{ij}\left(\frac{\sigma}{\sigma-1}\frac{w_i}{\varphi}\tau_{ij}\right)^{1-\sigma} \mu_{ij}(\varphi)d\varphi \\ &= \left(\frac{\sigma}{\sigma-1}w_i\tau_{ij}\right)^{1-\sigma} M_{ij}\int_0^{\infty}\varphi^{\sigma-1}\mu_{ij}(\varphi)d\varphi \\ &= \left(\frac{\sigma}{\sigma-1}w_i\tau_{ij}\right)^{1-\sigma} M_{ij}\tilde{\varphi}_{ij}^{\sigma-1}\end{aligned} \tag{14}$$

where $\tilde{\varphi}_{ij}$ captures the average productivity of products from i selling to j. From equation 9,

$$X_{ij} = \left(\frac{\sigma}{\sigma-1}\right)^{1-\sigma}(\tau_{ij}w_i)^{1-\sigma}M_{ij}\tilde{\varphi}_{ij}^{\sigma-1}Y_jP_j^{\sigma-1} \tag{15}$$

Selection in exporting A firm in i will only export to j if and only if its profit is at least the fixed cost of entering, i.e. $\pi_{ij}(\varphi) \geq f_{ij}$. From the equilibrium revenue and profit, the threshold productivity of firms can be written as

$$\begin{aligned}&\frac{1}{\sigma}\left(\frac{\sigma}{\sigma-1}\frac{w_i}{\varphi}\tau_{ij}\right)^{1-\sigma} Y_jP_j^{\sigma-1} \geq f_{ij} \Leftrightarrow \\ &\qquad \varphi \geq \varphi_{ij}^* \equiv \left(\frac{\sigma f_{ij}\left(\frac{\sigma}{\sigma-1}w_i\tau_{ij}\right)^{\sigma-1}}{Y_jP_j^{\sigma-1}}\right)^{\frac{1}{\sigma-1}}.\end{aligned} \tag{16}$$

Using the distribution of productivity, the trade flow becomes

$$X_{ij} = \left(\frac{\sigma}{\sigma-1}w_i\tau_{ij}\right)^{1-\sigma} M_i\left(\int_{\varphi_{ij}^*}^{\infty}\varphi^{\sigma-1}dG_i(\varphi)\right)Y_jP_j^{\sigma-1}. \tag{17}$$

This shows that as the threshold productivity of exporting firms decreases, we will observe equal or larger trade flows than before.

The Pareto distribution (Chaney, 2008) uses the Pareto distribution to get crisp results on the extensive and the intensive margin. Suppose $\varphi \in [1,\infty]$ and

$$G_i(\varphi) = 1 - \varphi^{-\theta_i} \quad \theta_i > \sigma - 1. \tag{18}$$

Then the gravity equation becomes

$$X_{ij} = C_1(\tau_{ij}w_i)^{-\theta_i} f_{ij}^{\frac{\sigma-\theta_i-1}{\sigma-1}} M_i(Y_iP_j^{\sigma-1})^{\frac{\theta_i}{\sigma-1}} \tag{19}$$

where $C_1 = \sigma^{\frac{\sigma-\theta_i-1}{\sigma-1}}\left(\frac{\sigma}{\sigma-1}\right)^{-\theta_i}\frac{\theta_i}{\theta_i+1-\sigma}$

5.2 Gains from steamships

5.2.1 Introducing steamships

This paper now considers sailing and steamships in this canonical model. Let $s = \{sail, steam\}$ denote the type of ship used. The difference in ship type manifests in two ways. One is the

change in the iceberg trade cost τ_{ij}. As steamships can move faster, $\tau_{ij}^{steam} \leq \tau_{ij}^{sail}$ in most cases. The second is the change in the fixed cost of entry. As steamships were bigger and also required additional investments (such as storage for coal and suitable geography) the fixed cost to arrive at the port was different by ship type. Thus, $f_{ij}^{steam} \neq f_{ij}^{sail}$. Which one is lower depends on port investment and thus, the year. For example, if i invested heavily in port infrastructure at year t, $f_{ij,t}^{steam} > f_{ij,t+1}^{steam}$. In this case, if the investment is large enough $f_{ij,t+1}^{steam} < f_{ij,t+1}^{sail}$.

Given this framework, firms will decide which mode to use. The profit using s and exporting is

$$\pi_{ij}^{s}(\varphi) = \frac{1}{\sigma}\left(\frac{\sigma}{\sigma-1}\frac{w_i}{\varphi}\tau_{ij}^{s}\right)^{1-\sigma} Y_j P_j^{\sigma-1} - f_{ij}^{s} \tag{20}$$

Let $f_{ij}^{steam} = \rho_{ij} f_{ij}^{sail}$ and $\tau_{ij}^{steam} = \xi_{ij}\tau_{ij}^{sail}$. Then the thresholds for exporting using sailing ships and steamships and the productivity in which the profit using steamships and sailing ships is equal are

$$\varphi_{ij,sail}^{*} = \left(\frac{\sigma f_{ij}^{sail}\left(\frac{\sigma}{\sigma-1} w_i \tau_{ij}^{sail}\right)^{\sigma-1}}{Y_j P_j^{\sigma-1}}\right)^{\frac{1}{\sigma-1}} \tag{21}$$

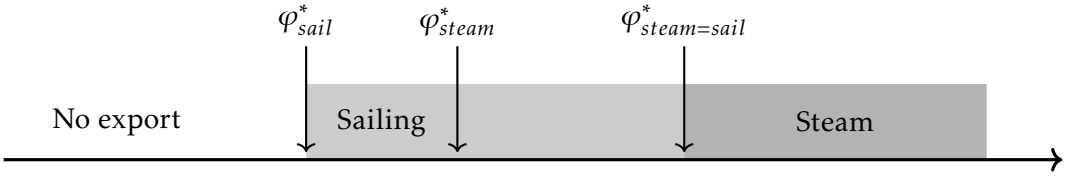
$$\varphi_{ij,steam}^{*} = \varphi_{ij,sail}^{*}\left(\rho_{ij}\xi_{ij}^{\sigma-1}\right)^{\frac{1}{\sigma-1}} \tag{22}$$

$$\varphi_{ij,steam=sail}^{*} = \varphi_{ij,sail}^{*}\left(\frac{\rho_{ij}-1}{\xi_{ij}^{1-\sigma}-1}\right)^{\frac{1}{\sigma-1}} \tag{23}$$

Given these productivity thresholds, we need to consider two cases. For simplicity, the origin destination is omitted. Importantly, this paper assumes $\rho \leq 1$, i.e., steamships trade cost is weakly lower in all origin destination links. This comes from the empirical fact that the duration is almost always less for steamships and to obtain tractability in estimating trade flow by ships.

(I) When using steamship is harder to export: $\varphi_{sail}^{*} \leq \varphi_{steam}^{*}$
In this case, it must be that $\xi \geq 1$ as $\rho \leq 1$. That is, the fixed cost is higher for steamships than sailing ships. Given this, we obtain $\left(\frac{\rho-1}{\xi^{1-\sigma}-1}\right) \geq \rho\xi^{\sigma-1}$. Therefore, the firms that exports and which ships they use will be as in figure 7. As shown, this implies that even if steamships can be used, due to lower profitability steamships are only used by firms that can satisfy the profit condition.

Figure 6: Export and ship used when cost of using steamship is high



(II) When using steamship is easier to export: $\varphi_{sail}^{*} > \varphi_{steam}^{*}$
Here, we consider two cases: $\xi < 1$ and $\xi \geq 1$. $\xi < 1$ is when the profit condition is lower than the entry condition of steamships. Thus, firms with productivity above φ_{steam}^{*} all export

using steamships. Similarly, when $\xi \geq 1$, since the entry condition implies that $\xi < \left(\frac{1}{\rho}\right)^{\frac{1}{\sigma-1}}$ this also means $\varphi^*_{steam=sail} < \varphi^*_{steam}$. Therefore, when steamship is easier to export, all firms with productivity above φ^*_{steam} exports using steamships.

Figure 7: Export and ship used when cost of using steamship is low

These cases illustrate how lowering the fixed cost for steamships have on export patterns. First, lowering the fixed cost allows firms that did not export before to export. Specifically firms with productivity $\varphi \in [\varphi^*_{steam}, \varphi^*_{sail}]$ now exports. Second, more firms can benefit from steamships. Those who could earn positive profit using steamships now fully uses steamships.

5.2.2 Gains

To compute the gains from trade under different shipping technologies, this paper now considers the decision to enter the market. Let $f_i = f_{ii}$ be the fixed cost to operate in the domestic market. The cutoff of productivity for domestic activity is therefore

$$\varphi^*_i = \left(\frac{\sigma f_i \left(\frac{\sigma}{\sigma-1} w_i\right)^{\sigma-1}}{Y_i P_i^{\sigma-1}} \right)^{\frac{1}{\sigma-1}} \tag{24}$$

To better understand the gains, this section assumes symmetric countries, i.e. $Y_j = Y, P_i = P, w_i = w, \tau_{ij} = \tau, f_{ij} = f_x \; i \neq j$. This implies that there are no terms of trade, and thus, we set $w_i = 1 \forall i$. First, consider the case where there are only type of ship. Then, the cutoff for exporting is

$$\varphi^*_s = \varphi^* \tau^s \left(\frac{f_x}{f}\right)^{\frac{1}{\sigma-1}} \tag{25}$$

As in (Krugman, 1979) expected profits must be equal to the fixed cost of entry

$$f_e = \bar{\pi}(1 - G(\varphi^*))\frac{1}{\delta} \tag{26}$$

and the expected profit is given by

$$\bar{\pi} = \tilde{\pi}_d(\tilde{\varphi}_d) + p_x \tilde{\pi}_x(\tilde{\varphi}_x) \tag{27}$$

where p_x is the probability of exporting and $\tilde{\pi}_d(\tilde{\varphi}_d), \tilde{\pi}_x(\tilde{\varphi}_x)$ are expected profit from domestic and exporting, respectively. Using the properties of the model, we obtain

$$\bar{\pi} = f\left[\left(\frac{\tilde{\varphi}(\varphi^*_x)}{\varphi^*}\right)^{\sigma-1} - 1\right] + \frac{1 - G(\varphi^*_x)}{1 - G(\varphi^*)} f_x \left[\left(\frac{\tilde{\varphi}_x(\varphi^*)}{\varphi^*_x(\varphi^*)}\right)^{\sigma-1} - 1\right] \tag{28}$$

where $\tilde{\varphi}(\varphi^*_x) = \int_0^{\varphi^*_x} \varphi^{\sigma-1} dG(\varphi)$ and $\tilde{\varphi}_x(\varphi^*) = \int_{\varphi_*}^{\infty} \varphi^{\sigma-1} dG(\varphi)$. From this, equations 26 and 28 solves φ^*. and the Price index,

$$P = \left(\frac{\sigma f}{L}\right)^{\frac{1}{\sigma-1}} \frac{\sigma}{\sigma-1}\frac{1}{\varphi^*} \tag{29}$$

This is the welfare when there are only sailing ships and only steamships as in case (II). (Arkolakis et al., 2012) shows that with a Pareto distribution, the welfare is defined by the domestic trade share and trade elasticity. The trade share is

$$\begin{aligned} \lambda_{ii} &\equiv \frac{X_{ii}}{X_i} \\ &= \frac{f^{\frac{1-\theta}{\sigma-1}}}{(|S|-1)\tau^{-\theta} f_x^{\frac{1-\theta}{\sigma-1}} + f^{\frac{1-\theta}{\sigma-1}}} \end{aligned} \tag{30}$$

and the welfare gains from shipping technology s compared to autarky is

$$\begin{aligned} W_s &\equiv 1/P \\ &= \lambda_{ii}^{\frac{1}{1-\sigma}} \end{aligned} \tag{31}$$

Using the assumption that $\theta > \sigma - 1$, we can see that the gains in case (II) are larger than in the case of only sailing ships due to lower τ and small enough f_x.

In case (I), we observe both sailing and steamships. Thus, the expected profit becomes

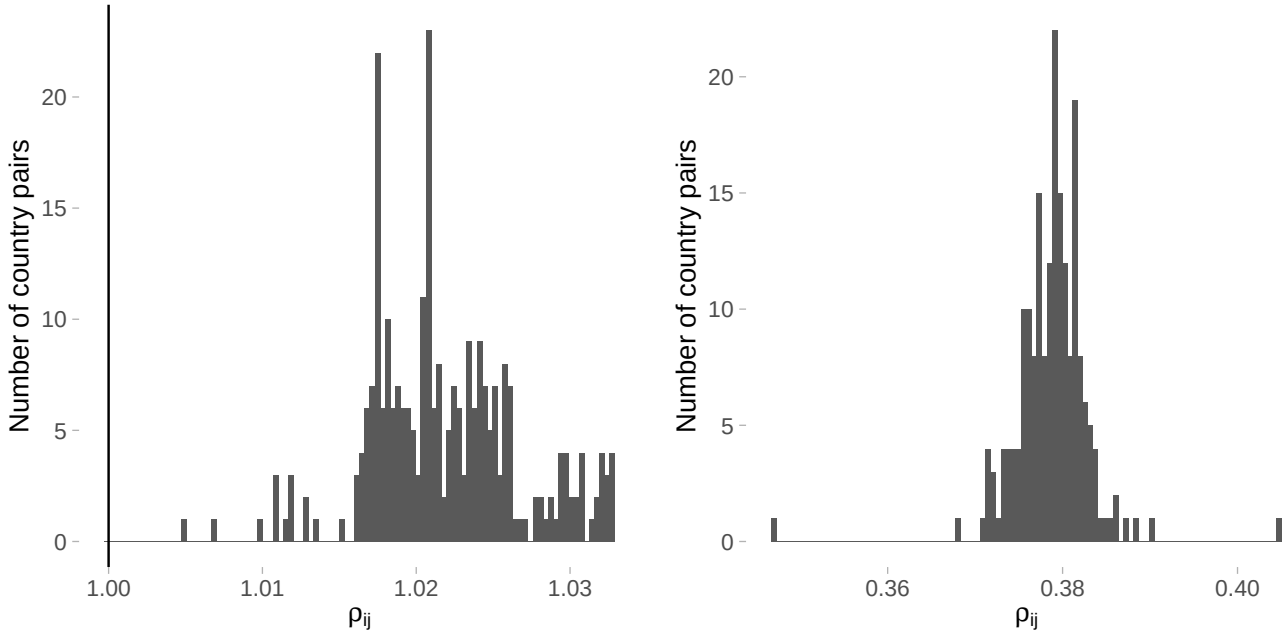
$$\begin{aligned} \bar{\pi} = & f\left[\left(\frac{\tilde{\varphi}(\varphi^*_{sail})}{\varphi^*}\right)^{\sigma-1} - 1\right] \\ & + \frac{G(\varphi^*_{steam=sail}) - G(\varphi^*_{sail})}{1 - G(\varphi^*)} f_{sail}\left[\left(\frac{\tilde{\varphi}_x(\varphi^*_{sail}) - \tilde{\varphi}_x(\varphi^*_{steam=sail})}{\varphi^*_{sail}}\right)^{\sigma-1} - 1\right] \\ & + \frac{1 - G(\varphi^*_{steam=sail})}{1 - G(\varphi^*)} f_{steam}\left[\left(\frac{\tilde{\varphi}_x(\varphi^*_{steam=sail})}{\varphi^*_{steam}}\right)^{\sigma-1} - 1\right] \end{aligned} \tag{32}$$

Together with equation 26, this solves φ^* and therefore the price index (equation 29). For example, consider the case when $\delta = 0.95$, $\sigma = 2.5$, $\theta = 4.25$, $f_e = 1$, $f_{domestic} = 3$ $f_{sail} = 6$, $\tau_{sail} = 1.4$, $\xi = 0.8$, $L = 1$. Given this, the welfare in case I, steamships and sailing ships coexisting, is 0.098, while in case II, steamships only, is 0.123. Thus, the difference in fixed cost, in this case, leads to differences in welfare gains.

6 Role of fixed cost in adoption: Estimation

6.1 Estimating the fixed cost of adoption

Here, this paper estimates the fixed cost of entry f_{ij}. This paper considers case (I) from the theory. To use the panel structure of the data, recall that the share of steamships in terms of quantity in period $t+1$ is given by



(a) Case (I)

(b) Case (II)

Figure 8: Distribution of ρ_{ij}

$$\begin{aligned}\mu_{ij} &= \frac{\int_{\varphi^*_{steam=sail}}^{\infty} q_{ij}(\varphi) dG(\varphi)}{\int_{\varphi^*_{sail}}^{\infty} q_{ij}(\varphi) dG(\varphi)} \\ &= \frac{\xi_{ij}}{1 - \left(\frac{\rho_{ij}-1}{\xi_{ij}^{1-\sigma}-1}\right)^{\frac{1}{1-\sigma}}}\end{aligned} \tag{33}$$

In the data, we observe both μ_{ij}, the share of steamships in terms of tonnage, and ξ_{ij}, the relative change in trade cost under a number of assumptions. First, the tonnage of ships is sufficient to know the quantity sold from i to j. This is a restrictive assumption. It is possible that the ship contains a fraction of goods that will be transported to another place, and ships may not be used in full capacity proportional to their ship size. Another assumption is trade cost is proportional to duration. This is also a restrictive assumption as the marginal costs of sailing ships and steamships differ. For example, the iceberg cost itself is subject to change depending on the cost of coal. With additional considerations, such as endogenous freight rates, more work is needed to understand the full effect of steamships. This paper considers subsequent estimates as a baseline for future work. Given those assumptions, we can compute the relative fixed cost between sailing and steamships by solving equation 33. The resulting distribution of ρ_{ij} and the relative productivity cutoff for steamship and sailing ship is shown in figure 8a. The estimate shows that the difference in fixed cost is about 2% more for steamships.

Second, consider case (II). For this, this paper looks not between modes but between years and looks at trade links that switched from all sailing ships in 1880 to all steamships in 1900. The share of such change is about 50% of all trade links. The change in trade volume implies

$$\frac{x_{ij,t+1}}{x_{ij,t}} = \Delta w_i^{-\theta_i} \Delta Y_j^{\frac{\theta_i - 1}{\sigma - 1}} \Delta P_j^{\theta_i - 1} \rho_{ij}^{\frac{\sigma - \theta_i}{\sigma - 1}} \xi_{ij}^{-\theta_i} \tag{34}$$

Table 6: Correlation between fixed cost and country characteristics

Dependent Variable:	log(rho)					
Year	Full sample	1880	1900	Full sample	1880	1900
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Constant	-1.063***	-0.507*	-1.323***	-1.081***	-0.212	-1.299***
	(0.169)	(0.245)	(0.199)	(0.216)	(0.196)	(0.240)
Colony	0.442	0.297	0.589*			
	(0.271)	(0.448)	(0.306)			
above top 25pc GDP				-0.135	-1.264**	0.600
				(0.529)	(0.311)	(0.865)
Fit statistics						
R^2	0.072	0.052	0.134	0.004	0.847	0.042
Observations	36	10	26	18	5	13

IID standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

where $\Delta z = \frac{z_{t+1}}{z_t}$. Taking the log gives us the following specification, which we can run using origin and destination fixed effects and setting the shape parameter, θ_i, of the Pareto distribution as 4.25 as in (M. J. Melitz & Redding, 2015).

$$\log x_{ij,t+1} - \log x_{ij,t} = \alpha_i + \alpha_j - \theta_i \log \xi_{ij} + \frac{\sigma - \theta_i}{\sigma - 1} \log \rho_{ij} \tag{35}$$

Finally, this paper considers when links switch from using both sailing and steamship to only steamships. Similarly using the difference in quantity traded, we obtain

$$\log x_{ij,t+1} - \log x_{ij,t} = \alpha_i + \alpha_j + \log \frac{\rho_{ij}^{\frac{\sigma-\theta_i}{\sigma-1}} \xi_{ij}^{-\theta_i}}{(1+\xi_{ij}^{-\sigma})\left(\frac{\rho_{ij}-1}{\xi_{ij}^{1-\sigma}-1}\right)^{\frac{\sigma-\theta_i}{\sigma-1}} - 1} \tag{36}$$

The resulting distribution of fixed entry costs relative to sailing ships for links fully transitioned to steamships are shown in figure 8b. We can see that even though the full transition is still possible with a higher ρ due to a decrease in variable cost (duration), the fixed cost is much lower than sailing ships. Specifically, the fixed cost is centred around 0.38, implying those ports are substantially easier for steamships to enter.

With these estimates, we can now test whether it is indeed the case that colonized countries have higher costs for steamships to enter. First, table 6 shows the correlation between colonized and countries in the top quartile of GDP in 1880 and the fixed cost of steamship relative to sailing. We can see that the fixed cost is generally higher for colonized countries while lower for high-income countries. This suggests the importance of the fixed cost in explaining the heterogeneous impact of opening up to trade. As the theoretical framework illustrated, even if variable cost falls, the country will not gain from increased openness if adoption is prevented by high fixed entry.

6.2 Discussion

This section explored whether the differences in fixed cost affected steamship adoption and whether this could explain the differences in gains from changes in duration across countries. The key idea is that even if variable costs decrease, allowing countries to export and import more easily, some countries could not access this technology fully due to the high barrier to use. This paper discusses some hypotheses as to why this might be the case.

First is port investments. Similar to container shipping (Ducruet et al., 2020), ports for steamships required modernization. The decision to modernize and improve ports is endogenous. One would consider the effect on the local economy and allocation of land to other sectors. It also requires coordination among countries to adopt steamships. Including in the model whether to invest in ports and thus lower f_{ij}^{steam} will be able to close the model and provide a valid counterfactual analysis.

The second is markups in the transportation sector. A number of papers in recent years have shown (Djankov et al., 2010) (D. Hummels et al., 2009) (D. Hummels, 2007) (Asturias, 2020) (Brancaccio et al., 2020) the importance of price setting by the shipping sector on trade volumes. Although this has been largely ignored in the past with the "iceberg" cost, the price setting is especially important in the case of new shipping technology. As historical data shows (Jacks et al., 2008), the adoption of steamships did not immediately lead to a decline in freight rates. Market power by the steamships likely influenced this result. If, for example, the entry of steamships were high for certain links, steamships could command a higher markup in this case. Thus, the endogenous price setting is affected by the fixed cost of entry.

Last is search friction. Even if variable trade costs decreased, colonized countries did not export more to other countries (shown in the appendix). In fact, for countries colonized by Great Britain, more ships travelled to Great Britain and ships owned by British companies were used more often. This suggests that past relationships before the adoption partially decided the pattern of trade afterwards. If the colonizers could enter the market disproportionately easier, this may allow them to take advantage of it.

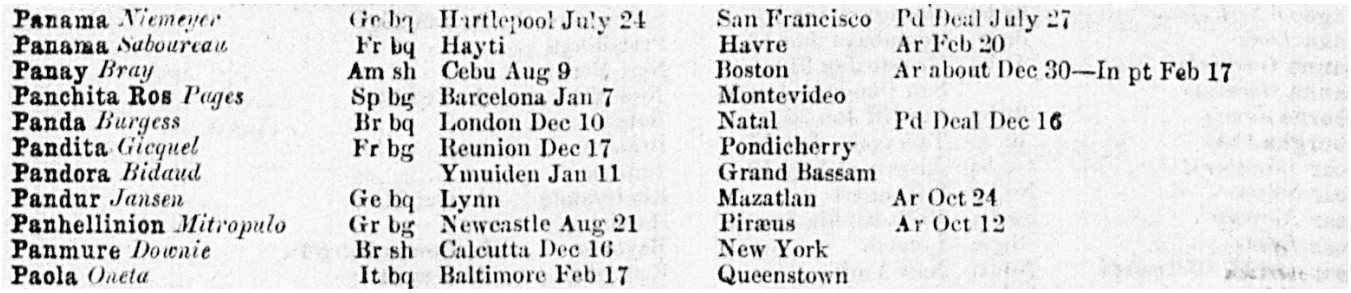
These hypotheses illustrate the possible role of shipping and ports in promoting more trade and extortion through high entry barriers. This effect could be particularly pronounced in the age of disruptive technologies such as steamships. A further investigation into how this invention changed the shipping industry and, henceforth, trade patterns and inequality is a fruitful area of research.

7 Conclusion

This paper used the adoption of steamships to investigate what could drive unequal gains from a reduction in trade cost. Using the exogenous changes in duration due to the adoption of steamships, this paper first shows that increased integration to large markets benefits countries, but there are relative welfare losses when a country becomes closer to a large supplier. This effect, however, is driven by colonized countries. The hypothesis this paper tested is whether steamship adoption was unequal, leading to such results. For this, this paper implemented a novel deep-learning method to digitize historical shipping data en masse and combined the canonical trade model with technology adoption. The estimate of the model shows that colonized countries had a higher barrier to using steamships, reducing the gains from trade. This may have led to producers in those countries losing out on the benefits of market integration. The results and the data introduced have larger implications for inequality, technology adoption, and trade. The period of steamships coincides with the Great Divergence, where the Western civilization took off while other regions stagnated. How technologies affected this result and

Reg.	Ship	Master	Ton.	Flag	Rig	From	For	Latest Reports
R V	E A O'Brien	*Pratt*	(1038)	Br	bq	Manilla Apr 4	Boston	ArSept10—ForBuenosAyres
R V	E B Sutton	*Carter*	(1639)	Am	s	Honolulu Oct 13	New York	
* R	E C Mowatt	*Hersey*	(1026)	Am	bq	Philadelphia Sept6	Table Bay	Pd Marcus Hook Sept 6
*	E J Spence	*Stronach*	(519)	Br	bq	Singapore July 26	Mauritius	Ar Sept 12
V	E J Spicer	*Cochran*	(1268)	Br	s	Table Bay July 21	Nestle(NSW)	Ar Spt2—ForWSCAmerica
R	E K Wood	*Hansen*	(452)	Am	sc	Tacoma Aug 25	Haiphong	
V	E S Hocken	*Willcock*	(249)	Br	bq	Clyde Aug 4	Table Bay	Ar Oct 25
*	Eagle Crag	*Shimmin*	(1347)	Br	bq	Cardiff Sept 15	CaletaBuena	Pd Barry Island Sept 15
*	Earl Cadogan	*Williams*	(1334)	Br	bq	Nestle(NSW)Aug24	AntofagastaAr	Oct 15
*	Earl Derby	*Mackintosh*	(961)	Br	bq	Brisbane Sept 8	Nestle(NSW)	ArSpt15--ForWSCAmerica

Figure 9: Example of 1900's LSI covering sailing ships



Panama *Niemeyer*	Ge bq	Hartlepool July 24	San Francisco	Pd Deal July 27
Panama *Sabourcau*	Fr bq	Hayti	Havre	Ar Feb 20
Panay *Bray*	Am sh	Cebu Aug 9	Boston	Ar about Dec 30—In pt Feb 17
Panchita Ros *Pages*	Sp bg	Barcelona Jan 7	Montevideo	
Panda *Burgess*	Br bq	London Dec 10	Natal	Pd Deal Dec 16
Pandita *Gicquel*	Fr bg	Reunion Dec 17	Pondicherry	
Pandora *Bidaud*		Ymuiden Jan 11	Grand Bassam	
Pandur *Jansen*	Ge bq	Lynn	Mazatlan	Ar Oct 24
Panhellinion *Mitropulo*	Gr bg	Newcastle Aug 21	Piræus	Ar Oct 12
Panmure *Downie*	Br sh	Calcutta Dec 16	New York	
Paola *Oneta*	It bq	Baltimore Feb 17	Queenstown	

Figure 10: Example of 1880's LSI covering sailing ships differing from 1900 in terms of format and available information

what would be a better way for trade benefits to be equal has important implications today.

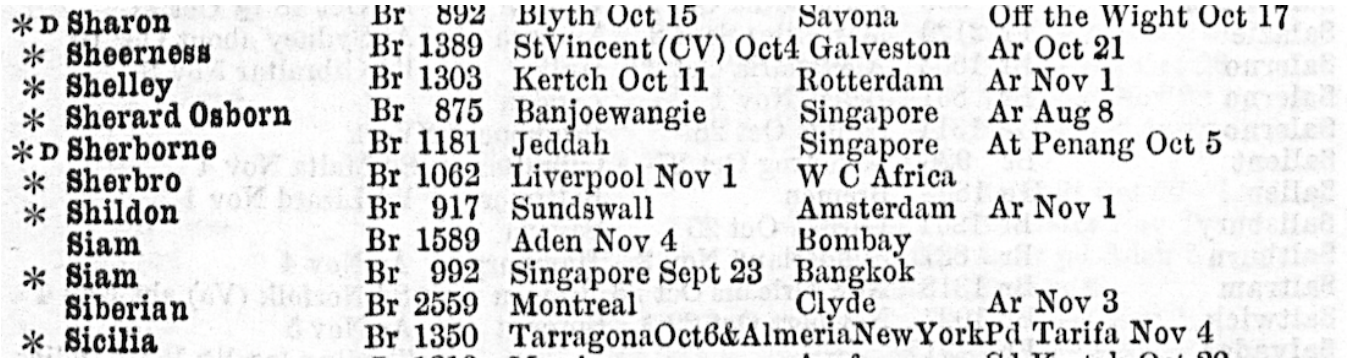
8 Appendix

A Digitization

Important for digitization, the format varies across years. Fig.9 is an excerpt from 1880. We can observe that 1) the text quality, as well as the spacing, is different, 2) there is information on association and tonnage, and 3) more empty cells. Fig. 10, in contrast, have much more white regions and also less information (no tonnage information). Fig.11 shows that the tonnage was recorded after nationality and not in brackets for steamships. In some years (as in 1900 for sailing ships), due to more information, the index recorded a ship in two rows as shown in Fig.12. With such heterogeneity, creating a rule-based method applicable to multiple years for extracting sufficient information is challenging. For all these images, there are no border lines to separate cells which poses difficulties in digitizing (Qasim et al., 2019).

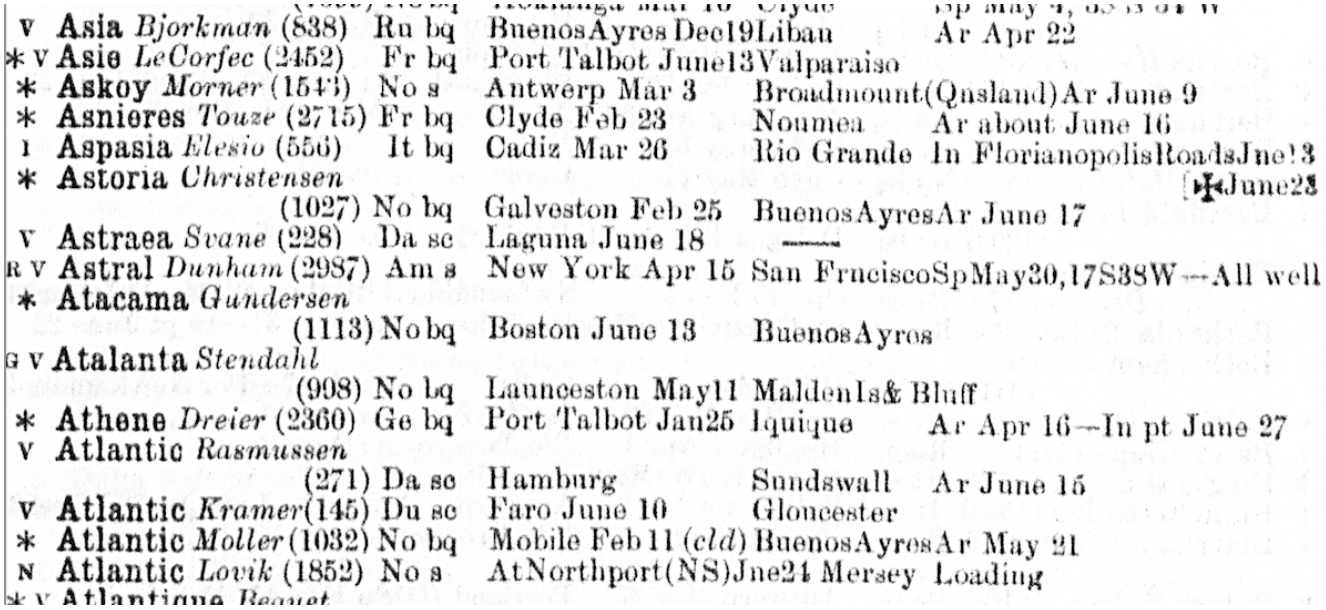
A.1 Overview

General Given these difficulties, this study uses deep learning to recognize the structure of this index. Deep learning in this setting has mainly two advantages. One is that deep learning can incorporate rules that structure this table (such as the order of the columns and how rows are organized) and sort the anomalies in the table (such as rows spanning for longer rows). Another is that, with different training data, the framework can apply to other format types with minor changes.



* D **Sharon**	Br	892	Blyth Oct 15	Savona	Off the Wight Oct 17
* **Sheerness**	Br	1389	StVincent (CV) Oct4	Galveston	Ar Oct 21
* **Shelley**	Br	1303	Kertch Oct 11	Rotterdam	Ar Nov 1
* **Sherard Osborn**	Br	875	Banjoewangie	Singapore	Ar Aug 8
* D **Sherborne**	Br	1181	Jeddah	Singapore	At Penang Oct 5
* **Sherbro**	Br	1062	Liverpool Nov 1	W C Africa	
* **Shildon**	Br	917	Sundswall	Amsterdam	Ar Nov 1
Siam	Br	1589	Aden Nov 4	Bombay	
* **Siam**	Br	992	Singapore Sept 23	Bangkok	
Siberian	Br	2559	Montreal	Clyde	Ar Nov 3
* **Sicilia**	Br	1350	TarragonaOct6&Almeria	NewYork	Pd Tarifa Nov 4

Figure 11: Example of 1890's LSI covering steamships with tonnage in a different location



V **Asia** *Bjorkman* (838)	Ru	bq	BuenosAyres Dec19	Liban	Ar Apr 22
* V **Asie** *LeCorfec* (2452)	Fr	bq	Port Talbot June13	Valparaiso	
* **Askoy** *Morner* (1543)	No	s	Antwerp Mar 3	Broadmount(Qnsland)	Ar June 9
* **Asnieres** *Touze* (2715)	Fr	bq	Clyde Feb 23	Noumea	Ar about June 16
I **Aspasia** *Elesio* (556)	It	bq	Cadiz Mar 26	Rio Grande	In FlorianopolisRoadsJne13
* **Astoria** *Christensen* (1027)	No	bq	Galveston Feb 25	BuenosAyres	✠June28 Ar June 17
V **Astraea** *Svane* (228)	Da	sc	Laguna June 18	——	
R V **Astral** *Dunham* (2987)	Am	s	New York Apr 15	San FrnciscoSpMay30,17S38W	—All well
* **Atacama** *Gundersen* (1113)	No	bq	Boston June 13	BuenosAyros	
G V **Atalanta** *Stendahl* (998)	No	bq	Launceston May11	Maldenls& Bluff	
* **Athene** *Dreier* (2360)	Ge	bq	Port Talbot Jan25	Iquique	Ar Apr 16—In pt June 27
V **Atlantic** *Rasmussen* (271)	Da	sc	Hamburg	Sundswall	Ar June 15
V **Atlantic** *Kramer* (145)	Du	sc	Faro June 10	Gloucester	
* **Atlantic** *Moller* (1032)	No	bq	Mobile Feb 11 (*cld*)	BuenosAyres	Ar May 21
N **Atlantic** *Lovik* (1852)	No	s	AtNorthport(NS)Jne24	Mersey	Loading
* V **Atlantique** *Beguet*					

Figure 12: Example of 19000's LSI with spanning rows

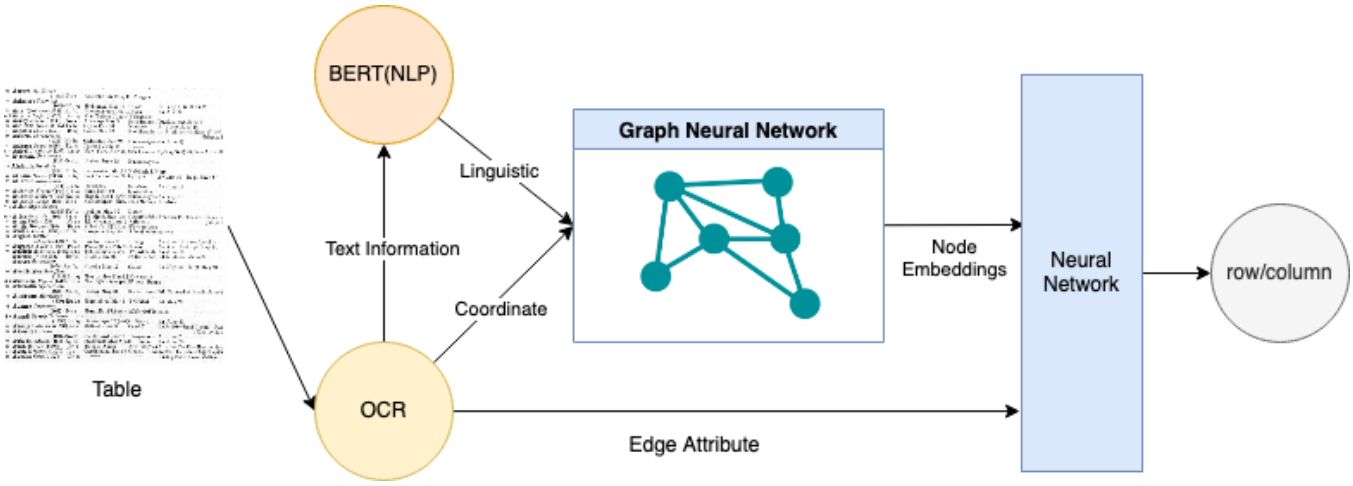


Figure 13: Overview of Digitization Architecture

This study uses a graph neural network, a type of deep learning framework for modeling each table. As described in more detail in subsequent sections, the graph neural network predicts the relationship between texts using the graph structure of the tables. This method is recently emerging in Table Structure Recognition (TSR). TSR refers to sorting data into organized tables, so that the data is in the correct relationships. Combined with Optical Character Recognition (OCR), TSR allows us to construct data frames necessary for statistical analysis(Qasim et al., 2019) (Chi et al., 2019) (Xiao et al., 2022). However, graph neural network has only been implemented when using a large number of tables extracted via HTML or TeX files. To the best of our knowledge, this study is the first to 1) apply this method to extracting historical documents and 2) construct training data only from hand-labeled data, which does not specify relationships between each text with its corresponding location.

Architecture The methodology is illustrated in Fig.13. This architecture has three main parts: construction of graphs (from image to graph using OCR), learning text representations (BERT and Graph Neural Network), and prediction of relationships (The final proponent using neural network). This study uses the OCR tool, Google Cloud Vision, to construct graphs to extract texts and their location. By extracting locations, this study obtains the closeness and the direction between pairs of texts, which the later stage uses as edge attributes. The extracted texts are encoded via the BERT model, a prominent natural language processing model representing text well in many instances(Devlin et al., 2019). During training, each text is matched to hand-labeled data, creating a graph for each table using the method illustrated in section A.2. For learning text representations, this study uses a graph neural network to learn the graph's structure and outputs a numerical representation for each text extracted. For the final stage, predicting relationships between texts, this study uses a standard neural network given the text's representations and edge attributes. The output is the probability of whether the pair of texts belong to 1) the same row, 2) the same column, and 3) the same cell. The latter two stages use the PyTorch geometric library(Fey & Lenssen, 2019) built on PyTorch(Paszke et al., 2019) in order to test out multiple models.

Problem definition Here, this study illustrates the problem the architecture tries to solve. To recognize table structure, we need to distinguish three types of relationships: whether a pair of words are in the same row, column, and cell. For example, consider a ship name "Earl Cadogan." Between "Earl" and "Cadogan," we want to know that the pair belongs to the same row, column, and cell. For two ships named "Earl Cadogan" and "Eagle Crag," we want to know that "Earl"

and "Eagle" are not in the same row and cell but are in the same column. Without loss of generality, the following explains the problem of whether a pair of words are in the same row.

Define a table as a graph G where $G = \langle N, R \rangle$. Here, N is the set of words ($n \equiv |N|$) in the table as nodes, and $R \in \{0,1\}^{n \times n}$ is the set of edges connecting any two nodes. $R_{ij} = 1$ implies that the two words i and j are in the same row and not in the same row when $R_{ij} = 0$. To predict R, it is possible to utilize the following. For each node, we can get 1) positional features ($F_{point} \in \mathbb{R}^{v \times 8}$), which are the four coordinates of the bounding box for each text (x and y coordinates of the top left, top right, bottom left, and bottom right of the bounding box), 2) linguistic features ($F_{lin} \in \mathbb{R}^{v \times l}$ where l is the number of linguistic features) which is characteristics related to the detected text itself, and 3) image features ($F_{image} \in \mathbb{R}^{v \times im}$ where im is the number of image features) which is features based on scanned images. Edge features ($F_{edge} \in \mathbb{R}^{v \times v \times e}$ where e is the number of edge features for each pair of nodes) such as the direction of the edge, as well as distance, can be incorporated as well. Then, the problem formulation is given a function g with inputs $F_{point}, F_{lin}, F_{image}, F_{edge}$ and output $R' \in [0,1]^{v \times v}$, minimize $\mathcal{L}(R, R')$ where $\mathcal{L}$ is the loss function. In other words $\hat{g} = \arg\min$For the loss function, since it is a binary variable, this research uses binary cross-entropy loss.

A.2 Creating graph

Optical character recognition Often the literature on table structure learning uses data already in the format of G, a one-to-one mapping of nodes with their respective features, and a complete relationship between other nodes(Qasim et al., 2019). This study, however, does not have such data. Instead, this study uses hand-labeled data of one week for each year/ship type (sailing or steamships) in the same format as the figures in section **??**. Namely, whether a pair of words are in the same row/column/cell can be identified, but F_{point}, F_{image}, and F_{edge} are unknown. This study used the Google Cloud Vision for Optical Character Recognition (OCR) to create those inputs following (Saito et al., 2022). This gives the positional features, F_{point}, as bounding boxes with the corresponding text.

Matching Since there are multiple occurrences of the exact text in the table (e.g., "Ar" meaning "Arrived" in the latest reports section), a simple search and find is insufficient. A different "Ar" in the hand-labeled data can be matched to a different "Ar." Then, the positional features are inconsistent with the word's correct place. The study used the following steps to match the positional features with the actual word. For illustration, consider the case of Fig. 14, where there are three rows. Suppose we want to extract the second row v_2 which is composed of the words "Br," "February," "Port A," and "Arrived". Further, suppose that due to the OCR, "Br" was not detected. In other rows there are same words (such as "Port A" and "Arrived" in all rows) so there needs to be distinguished as different rows when creating the graph. The method is the following.

1. For each words in the row (ground-truth), get the same words regardless of its position. In the example (Fig. 15), the words in purple (candidate texts) are the words to extract. There are multiple duplicates such as "Arrived."

2. From those candidate texts, get the most frequent range of the top left y-coordinate. In the example, the texts are divided into 5 ranges and the range in which there are the most top left point of the boundary box is in the 3rd range. This is indicated as the "Most Observed Range." Define the words in this range as "Within Range Words," the words in yellow.

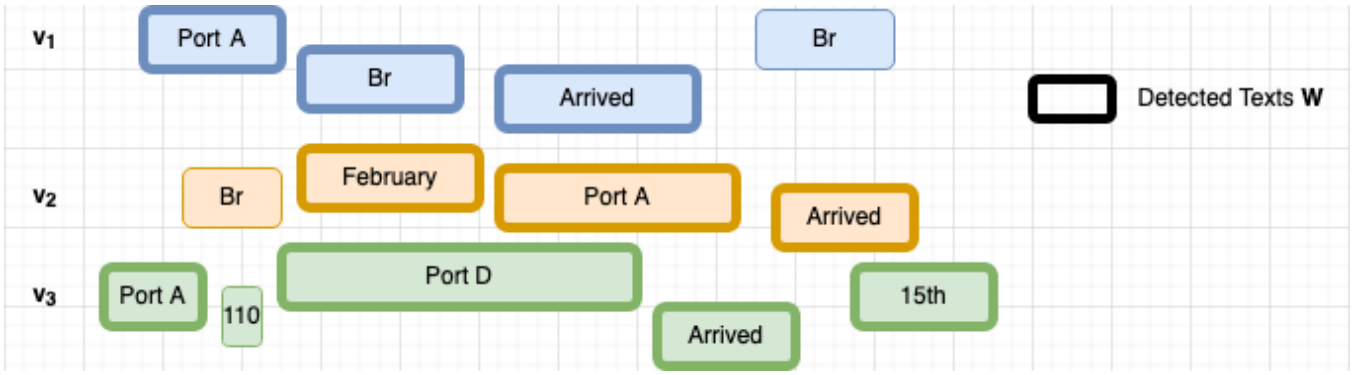


Figure 14: Example of ground-truth

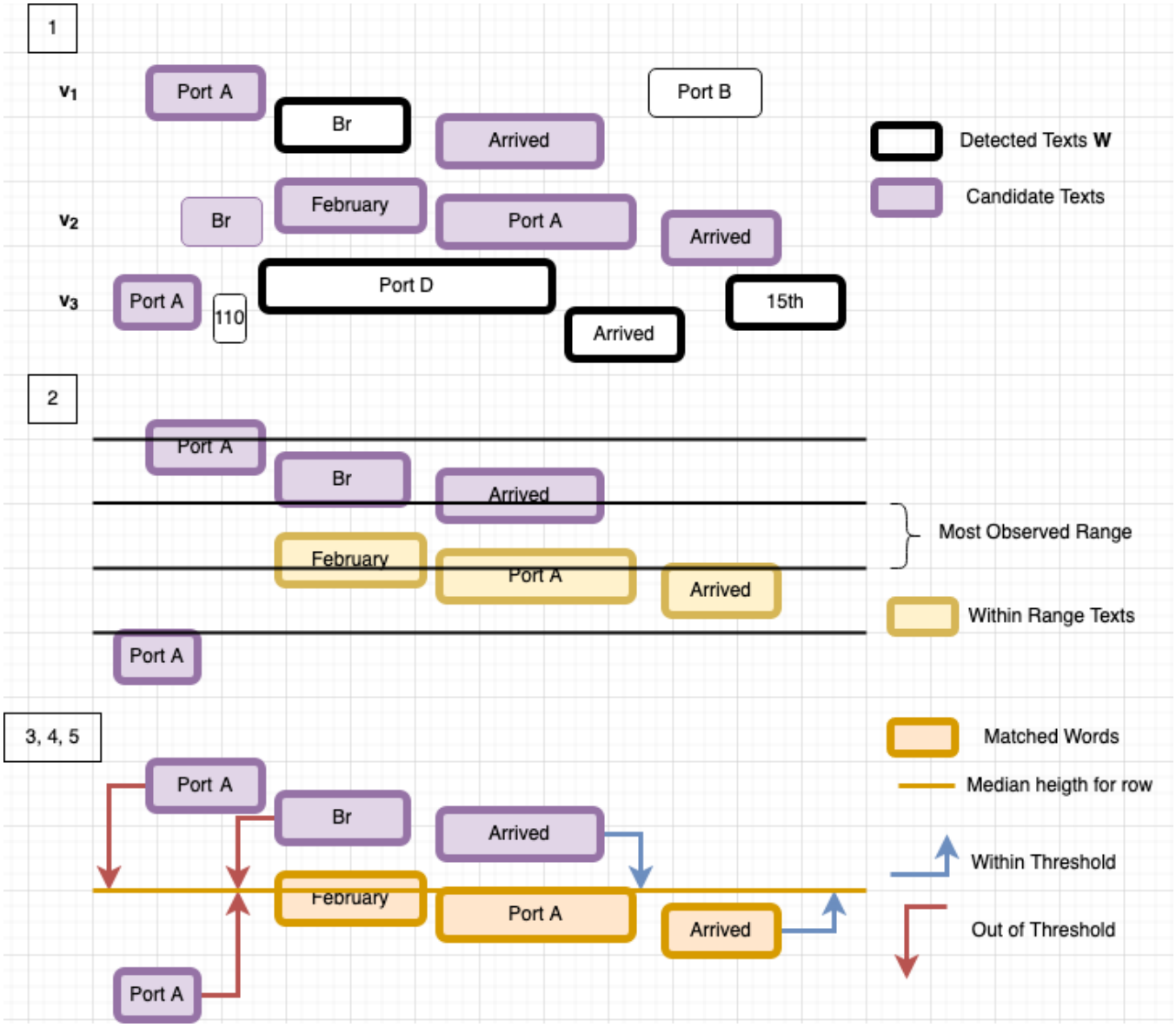


Figure 15: Illustration of steps

3. From the "Within Range Words" get the median top left coordinate. In the example, this is shown as "Median height for row," the orange line. This will be the basis of getting words in the second row.

4. From this "Median height for row," get the closest one for each word in the row. For example, since row 2 consists of "Br," "February," "Port A," and "Arrived", we select "February" and the closest of "Port A" and "Arrived".

5. For words that are not matched (in the example "Br") it is possible that due to detection in OCR, the true ones were not detected. Matching the closest one in this case would thus be a mistake (in the example, "Br" from the first row will be matched). To avoid this case, set some threshold value to avoid this. In other words, we will select words only within a certain distance from "Median height for row." With this the "Br" from row 1 is not be considered as row 2. For the digitization of LSI, the threshold was set to be about 2 rows apart in consideration of spanning rows.

In step 2, this study calculates the range most common across words as it is likely that we will have similar vertical coordinates for words in each row the most. This study employs a range as the exact vertical coordinates are different. For rows spanning, for example, vertical coordinates in the first row are about 30 pixels apart from the vertical coordinates in the second row. Step 5 is to ensure that the matched words are not too far from the median as given the range (for the example, it was 10), as identical words in neighboring rows might be included in step 4.

Training data Given that this study matches extracted texts via OCR to hand-labeled data, the final step for creating a graph is to label edges between pairs of matched texts. Without loss of generality, consider labeling whether a pair of text is in the same row. The easiest way is to label edges that do not indicate row as 0 and label edges as 1 otherwise. However, this will require $O(v^2)$ to train all edges, which is impractical. Furthermore, the overwhelming majority will be labeled 0 since only a couple of texts belong to the same row. For these problems, this study deletes the negative edges that connect vertices on average five rows apart. This uses the fact that the table has at most two spanning rows. With this procedure, the graph used for training will contain a labeled edge indicating a pair of words is in the same row or not and node features consisting of text and its location (bounding box). In the end, this study constructs, for each year of 1880, 1890, and 1900, about 60 tables each for steamships and sailing ships. Due to limited hand-labeled data, training data for each year comes from one of the weekly indexes. Using this training data, this study predicts the table structure for the other 9 weeks of index each year/ship type.

A.3 Model

The general idea is simple. The location of each word can be inferred from other words around. For example, "Ar" for Arrived in the latest reports column in 1900 will usually have a date type ("Sept 9") to its left and a specific port name to its right. Vertically, similar words (such as "Pt" for Passed) or blank cells will be found above or below. Based on this reasoning, a graph neural network model obtains the embedding for each text. For each word, representations from its surroundings get aggregated to the representation of the word itself. Then, this study uses a neural network to predict whether a sampled pair of texts is in the same row, column, and cell. In the neural network, this study incorporates edge attributes as well as the node embeddings gathered from the graph neural network.

Graph neural network This study implemented the graph neural network in PyTorch Geometric(Fey & Lenssen, 2019). In particular, this study uses GraphSAGE(Hamilton et al., 2018), which is a prominent algorithm in graph neural networks. GraphSAGE is an inductive algorithm for computing node embedding. In general terms, this algorithm finds rules iteratively to classify similar texts and thus output similar embeddings for similar texts in a similar location. This algorithm's usefulness is that it efficiently uses node attributes to generate representations on previously unseen data(Hamilton et al., 2018) efficiently. This useful property applies to this setting as we would like to describe unseen tables; tables in week 30 after training the model using index in week 35, for example. As it is an inductive algorithm, we can run GraphSAGE for multiple iterations to find new rules every time using surrounding information and thus get more intricate node embeddings for each text. To elaborate, in each iteration, GraphSAGE aggregates neighboring nodes' representations. Mathematically, suppose the representation for $v \in V$ is given by h_v^k where k is for the k^{th} iteration. Then given a function for aggregation, $f_{aggregate}$, the aggregated node representation a_v is given by

$$a_v^t = f_{aggregate}(\{h_u^t | u \in N(v)\}) \tag{37}$$

where $N(v)$ is the neighbouring vertices of v. Then the updated node representation is

$$h_v^{t+1} = f_{update}(a_v^t, h_v^t) \tag{38}$$

where f_{update} is a differentiable function. The study(Hamilton et al., 2018) shows that a single-layer neural network followed by a max-pooling operator is promising. This study, however, uses a simple mean operator to minimize training time. In addition, this study uses up to 3 layers since more layers will likely create a uniform representation for neighboring nodes due to the table's structure. The node embeddings use linguistic features created by the BERT model and the coordinates of the bounding boxes for each text. This study then passes this to a neural network to predict the labels of edges.

Neural network Given the node embeddings of two sampled nodes, this study used a standard neural network with 4 layers. However, in addition to the node embeddings, this study uses edge attributes in this stage. In other words, the inputs to the neural network are a concatenated vector of two node embeddings and the edge attributes. At the last layer, this study employs the sigmoid function to output the probability that a given edge represents a column, row, cell, or no connection. During training, this study uses the binary cross entropy loss function and backpropagates the error to the parameters of both the neural network and GraphSAGE.

Training Recall that the data this study tried to digitize spans three different years (1880, 1890, and 1900) and has different formats across years, as well as steamships and steamships. Thus, this study constructed a table structure model for each year for each ship type (steamships or sailing ships). Thus, the next step could be the possible use of pre-training the model(Hu et al., 2020) to create a universal model. For training, this study splits tables into samples; 70% of the tables are used for fitting, 10% for validation and fine-tuning, and the rest (10%) for testing their accuracy. This study randomly sampled 50 edges with positive and negative edges of equal length for each epoch after extracting node representations for the table of interest.

A.4 Result

Link prediction First, this study presents the results regarding links. In particular, this study presents precision and recall rates. This rate was averaged over all test tables. Precision is the

Table 7: Correct row prediction

Year	Ship	Precision	Recall
1880	Steam	0.84	0.71
1880	Sail	0.85	0.74
1890	Steam	0.83	0.75
1890	Sail	0.86	0.77
1900	Steam	0.81	0.75
1900	Sail	0.88	0.72

Table 8: Correct column prediction

Year	Ship	Precision	Recall
1880	Steam	0.93	0.80
1880	Sail	0.90	0.82
1890	Steam	0.91	0.81
1890	Sail	0.90	0.84
1900	Steam	0.88	0.82
1900	Sail	0.86	0.82

Table 9: Row Matching

Year	Ship	Precision
1880	Steam	0.67
1880	Sail	0.71
1890	Steam	0.67
1890	Sail	0.66
1900	Steam	0.65
1900	Sail	0.66

percentage of true positives out of total predicted positives, and recall rate is the percentage of true positives out of total true positives. We can first observe that the results are similar by year. This probably comes from the fact that this study created models based on the year and ship type of interest alone. We also observe that the precision rate is high for both types of prediction. This means that if two texts are said to be in the same row, this is likely the case. The recall rate is lower than the precision rate probably comes from missed texts by OCR. Since some words are missing due to failed OCR, words in the same row or columns could not be connected. In any case, a lower recall rate compared to the precision rate suggests that the method used in this paper misses observation rather than creating a false observation.

Row matching In addition, this study also shows a measure of perfect matching. In other words, the number of rows with identical wordings in the hand-labeled data and the observations is complete, given column-wise segmentation. This study constructed rows in the following way. First, ship names in hand-labeled data are matched. Given this, this study collects all the text connected to the ship's name. The result is in table 9.

B Calculation of Distance

Figure 16 is the grid (https://github.com/eurostat/searoute) that was used to calculate the shortest distance between ports. For steamships, the distance was calculated using the Suez Canal while, for sailing ships, the canal wasn't used due to weather conditions and historical evidence **stopford˙maritime˙2009**. With this grid, this research used Djikstra's algorithm with consideration to the actual distance on the globe.

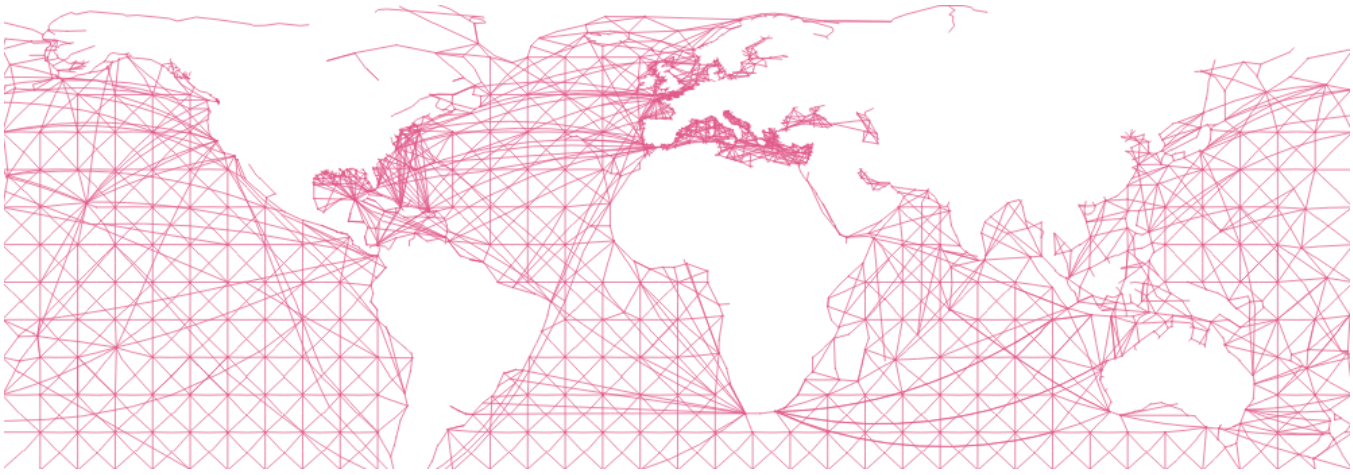


Figure 16: The grid used for calculating distance

References

Allen, T., & Arkolakis, C. (2022). The Welfare Effects of Transportation Infrastructure Improvements. *The Review of Economic Studies*, rdac001.

Arkolakis, C., Costinot, A., & Rodríguez-Clare, A. (2012). New Trade Models, Same Old Gains? *American Economic Review*, *102*(1), 94–130.

Arkolakis, C., Ganapati, S., & Muendler, M.-A. (2021). The Extensive Margin of Exporting Products: A Firm-Level Analysis. *American Economic Journal: Macroeconomics*, *13*(4), 182–245.

Asturias, J. (2020). Endogenous transportation costs. *European Economic Review*, *123*, 103366.

Bernhofen, D. M., El-Sahli, Z., & Kneller, R. (2016). Estimating the effects of the container revolution on world trade. *Journal of International Economics*, *98*, 36–50.

Blum, B. S., Claro, S., Dasgupta, K., & Horstmann, I. J. (2019). Inventory Management, Product Quality, and Cross-Country Income Differences. *American Economic Journal: Macroeconomics*, *11*(1), 338–388.

Bolt, J., & van Zanden, J. L. (2014). The Maddison Project: Collaborative research on historical national accounts. *The Economic History Review*, *67*(3), 627–651.

Brancaccio, G., Kalouptsidi, M., & Papageorgiou, T. (2020). Geography, Transportation, and Endogenous Trade Costs. *Econometrica*, *88*(2), 657–691.

Britannica. (2023). History of Latin America - Colonialism, Revolution, Independence — Britannica.

Brooks, L., Gendron-Carrier, N., & Rua, G. (2021). The local impact of containerization. *Journal of Urban Economics*, *126*, 103388.

Bustos, P. (2011). Trade Liberalization, Exports, and Technology Upgrading: Evidence on the Impact of MERCOSUR on Argentinian Firms. *American Economic Review*, *101*(1), 304–340.

Campante, F., & Yanagizawa-Drott, D. (2018). Long-Range Growth: Economic Development in the Global Network of Air Links*. *The Quarterly Journal of Economics*, *133*(3), 1395–1458.

Chaney, T. (2008). Distorted Gravity: The Intensive and Extensive Margins of International Trade. *American Economic Review*, *98*(4), 1707–1721.

Chi, Z., Huang, H., Xu, H.-D., Yu, H., Yin, W., & Mao, X.-L. (2019, August). Complicated Table Structure Recognition.

Coşar, A. K., & Demir, B. (2018). Shipping inside the box: Containerization and trade. *Journal of International Economics*, *114*, 331–345.

Coşar, A. K., & Fajgelbaum, P. D. (2016). Internal Geography, International Trade, and Regional Specialization. *American Economic Journal: Microeconomics*, *8*(1), 24–56.
Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable Convolutional Networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, 764–773.
Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
Djankov, S., Freund, C., & Pham, C. S. (2010). Trading on Time. *The Review of Economics and Statistics*, *92*(1), 166–173.
Donaldson, D. (2015). The Gains from Market Integration. *Annual Review of Economics*, *7*(1), 619–647.
Donaldson, D. (2018). Railroads of the Raj: Estimating the Impact of Transportation Infrastructure. *American Economic Review*, *108*(4-5), 899–934.
Donaldson, D., & Hornbeck, R. (2016). Railroads and American Economic Growth: A "Market Access" Approach*. *The Quarterly Journal of Economics*, *131*(2), 799–858.
Ducruet, C., Cuyala, S., & El Hosni, A. (2018). Maritime networks as systems of cities: The long-term interdependencies between global shipping flows and urban development (1890–2010). *Journal of Transport Geography*, *66*, 340–355.
Ducruet, C., & Itoh, H. (2022). The spatial determinants of innovation diffusion: Evidence from global shipping networks. *Journal of Transport Geography*, *101*, 103358.
Ducruet, C., Juhász, R., Nagy, D. K., & Steinwender, C. (2020). All Aboard: The Effects of Port Development. *National Bureau of Economic Research Working Paper Series*, *No. 28148*(Journal Article).
Eaton, J., & Kortum, S. (2002). Technology, Geography, and Trade. *Econometrica*, *70*(5), 1741–1779.
Faber, B. (2014). Trade Integration, Market Size, and Industrialization: Evidence from China's National Trunk Highway System. *The Review of Economic Studies*, *81*(3 (288)), 1046–1070.
Fajgelbaum, P., & Redding, S. J. (2022). Trade, Structural Transformation, and Development: Evidence from Argentina 1869–1914. *Journal of Political Economy*, *130*(5), 1249–1318.
Fajgelbaum, P. D., & Khandelwal, A. K. (2016). Measuring the Unequal Gains from Trade *. *The Quarterly Journal of Economics*, *131*(3), 1113–1180.
Feenstra, R. C., & Ma, H. (2014). Trade Facilitation and the Extensive Margin of Exports. *The Japanese Economic Review*, *65*(2), 158–177.
Fey, M., & Lenssen, J. E. (2019, April). Fast Graph Representation Learning with PyTorch Geometric.
Feyrer, J. (2019). Trade and Income—Exploiting Time Series in Geography. *American Economic Journal: Applied Economics*, *11*(4), 1–35.
Feyrer, J. (2021). Distance, trade, and income — The 1967 to 1975 closing of the Suez canal as a natural experiment. *Journal of Development Economics*, *153*, 102708.
Fletcher, M. E. (1958). The Suez Canal and World Shipping, 1869-1914. *The Journal of Economic History*, *18*(4), 556–573.
Fouquin, M., & Hugot, J. (2016). Two Centuries of Bilateral Trade and Gravity data: 1827-2014, 39.
Frankel, J. A., & Romer, D. H. (1999). Does Trade Cause Growth? *American Economic Review*, *89*(3), 379–399.
Galiani, S., Jaramillo, L. F., & Uribe-Castro, M. (2023, August). Market Access and Migration: Evidence from the Panama Canal Opening during the First Great Migration.
Gatos, B., Danatsas, D., Pratikakis, I., & Perantonis, S. J. (2005). Automatic Table Detection in Document Images. In S. Singh, M. Singh, C. Apte, & P. Perner (Eds.), *Pattern Recognition and Data Mining* (pp. 609–618). Springer.

Hamilton, W. L., Ying, R., & Leskovec, J. (2018, September). Inductive Representation Learning on Large Graphs.

Hao, L., Gao, L., Yi, X., & Tang, Z. (2016). A Table Detection Method for PDF Documents Based on Convolutional Neural Networks, 287–292.

Head, K., & Mayer, T. (2014, January). Chapter 3 - Gravity Equations: Workhorse,Toolkit, and Cookbook. In G. Gopinath, E. Helpman, & K. Rogoff (Eds.), *Handbook of International Economics* (pp. 131–195, Vol. 4). Elsevier.

Heiland, I., Moxnes, A., Ulltveit-Moe, K. H., & Zi, Y. (2022). Trade From Space: Shipping Networks and The Global Implications of Local Shocks *. (Generic).

Helpman, E., Melitz, M., & Rubinstein, Y. (2008). Estimating Trade Flows: Trading Partners and Trading Volumes *. *Quarterly Journal of Economics*, *123*(2), 441–487.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., & Leskovec, J. (2020, February). Strategies for Pre-training Graph Neural Networks.

Hummels, D. (2007). Transportation Costs and International Trade in the Second Era of Globalization. *Journal of Economic Perspectives*, *21*(3), 131–154.

Hummels, D., Lugovskyy, V., & Skiba, A. (2009). The trade reducing effects of market power in international shipping. *Journal of Development Economics*, *89*(1), 84–97.

Hummels, D. L., & Schaur, G. (2013). Time as a Trade Barrier. *American Economic Review*, *103*(7), 2935–2959.

Jacks, D. S., Meissner, C. M., & Novy, D. (2008). Trade Costs, 1870–2000. *American Economic Review*, *98*(2), 529–534.

Jacks, D. S., & Pendakur, K. (2010). Global Trade and the Maritime Transport Revolution. *The Review of Economics and Statistics*, *92*(4), 11.

Kasar, T., Barlas, P., Adam, S., Chatelain, C., & Paquet, T. (2013). Learning to Detect Tables in Scanned Document Images Using Line Information. *Proceedings of the 2013 12th International Conference on Document Analysis and Recognition*, 1185–1189.

Krugman, P. (1979). Increasing returns, monopolistic competition, and international trade. *Journal of International Economics*, *9*(4), 469–479.

Krugman, P. (1980). Scale Economies, Product Differentiation, and the Pattern of Trade. *The American Economic Review*, *70*(5), 950–959.

Melitz, M. (2003). The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity. *Econometrica*, *71*(6), 1695–1725.

Melitz, M. J., & Redding, S. J. (2015). New Trade Models, New Welfare Implications. *THE AMERICAN ECONOMIC REVIEW*, *105*(3).

Okoye, D., Pongou, R., & Yokossi, T. (2019). New technology, better economy? The heterogeneous impact of colonial railroads in Nigeria. *Journal of Development Economics*, *140*, 320–354.

Pascali, L. (2017). The Wind of Change: Maritime Technology, Trade, and Economic Development. *American Economic Review*, *107*(9), 2821–2854.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2019). Automatic differentiation in PyTorch.

Qasim, S. R., Mahmood, H., & Shafait, F. (2019, July). Rethinking Table Recognition using Graph Neural Networks.

Rashid, S. F., Akmal, A., Adnan, M., Aslam, A., & Dengel, A. (2017). Table Recognition in Heterogeneous Documents Using Machine Learning, 777–782.

Redding, S., & Venables, A. J. (2004). Economic geography and international inequality. *Journal of International Economics*, *62*(1), 53–82.

Redding, S. J., & Turner, M. A. (2015). Transportation Costs and the Spatial Organization of Economic Activity. In *Handbook of Regional and Urban Economics* (pp. 1339–1398, Vol. 5). Elsevier.

Saito, T., Shibasaki, R., Murakami, S., Tsubota, K., & Matsuda, T. (2022). Global Maritime Container Shipping Networks 1969–1981: Emergence of Container Shipping and Reopening of the Suez Canal. *Journal of Marine Science and Engineering*, *10*(5), 602.

Schreiber, S., Agne, S., Wolf, I., Dengel, A., & Ahmed, S. (2017). DeepDeSRT: Deep Learning for Detection and Structure Recognition of Tables in Document Images. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, *01*, 1162–1167.

Shen, Z., Zhang, R., Dell, M., Lee, B. C. G., Carlson, J., & Li, W. (2021). LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis.

Xiao, B., Simsek, M., Kantarci, B., & Alkheir, A. A. (2022, March). Table Structure Recognition with Conditional Attention.